NOBLEJACKAL / NOMOS

NOMOS

NOMOS GEO-Auditprotokoll

DIE GEO-1000-METHODE

Ein Protokoll zur Messung der Repräsentation von Entitäten in generativen Systemen über die Weltbevölkerung hinweg

Kaan MURAZ

0.9.0 · DEUTSCHE REDAKTIONELLE AUSGABE

NOMOS GEO-Auditprotokoll

Ein Protokoll zur Messung der Repräsentation von Entitäten in generativen Systemen über die Weltbevölkerung hinweg

Autor und verantwortlicher Gründer: Kaan MURAZ

Methodenidentität: NOMOS by NobleJackal

Version: 0.9.0 – finaler Textkandidat, deutsche redaktionelle Ausgabe

Sprache: Deutsch (de-DE)

© 2026 Kaan MURAZ. Die Veröffentlichungslizenz für diese Kandidatenausgabe ist noch nicht festgelegt.

In dieser Ausgabe ist NOMOS ein offener, versionierter Standardkandidat, der zur unabhängigen Prüfung vorgelegt wird. Es handelt sich weder um eine anerkannte internationale Norm noch um ein aktives Zertifizierungsprogramm.

Status des realen Benchmarks: BQ-0 – nicht durchgeführt. Die Zahlen in Kapitel 18 bilden eine berechnete synthetische Demonstration der Methode; sie sind keine Ergebnisse mit realen Nutzern, realen KI-Produkten oder aus einer unabhängigen Feldstudie.

Diese Ausgabe begründet für sich genommen weder ein NOMOS-Konformitätszeichen noch eine Aussage „NOMOS 950+“, ein Zertifikat, ein Ranking oder die Bestätigung durch Dritte.

WIE DIESES BUCH ZU LESEN IST

Dieses Buch ist keine Anleitung zur Durchführung von GEO. Es ist ein Auditprotokoll und Standardkandidat, der die Bedingungen zu Population, Evidenz, Kontext, Zeit, Beurteilung und Governance festlegt, unter denen die Repräsentation von Entitäten in generativen Systemen gemessen werden kann.

Die menschlich lesbare Darstellung, die normativen Bestimmungen, die Auditverfahren und die maschinenlesbare Ebene sind als getrennte, aber miteinander verbundene Objekte zu lesen. MUSS und DARF NICHT kennzeichnen verbindliche Bestimmungen; SOLLTE kennzeichnet Empfehlungen; KANN kennzeichnet zulässige Optionen. Die Statusangaben KANDIDAT, SYNTHETISCH, OFFEN und BQ begrenzen das jeweils beanspruchte Evidenzniveau.

Mit [Kxx] gekennzeichnete Quellen bezeichnen methodische Grundlagen im Literaturverzeichnis. Das Vorliegen einer Quelle bedeutet nicht, dass sie jeden NOMOS-Kandidatenschwellenwert unterstützt; NOMOS-spezifische Regeln bleiben ausdrücklich vorläufig.

Kapitel 0 ist die Einleitung des Buches. Es verankert die Treuhanderklärung des Gründers und die Erklärung der Repräsentationsrechte, denen jede technische Bestimmung untergeordnet ist.

BEGRIFFE UND STATUSCODES

Abdeckungsrate Die Häufigkeit, mit der eine Intervallschätzungsmethode unter einem Testdesign den angenommenen wahren Wert enthält.

Atomare Aussage Die kleinste Bedeutungseinheit, die in der Lage ist, ein unabhängiges Urteil über Wahrheit, Umfang, Zeit, Entität oder Beweismittel zu tragen.

BQ Der Benchmark-Qualitätscode, der das Ausführungs- und Reproduzierbarkeitsniveau eines echten Benchmarks angibt.

Erfassungsdatensatz Ein Datensatz, der den übermittelten Prompt, die Antwort, die Zeit, die Oberfläche und die Integritätsnachweise zusammen bewahrt.

Faktenpaket Das versionierte Prüfpaket, das Identität, Umfang, Zeit, Beweise und Gegenevidenz-Aufzeichnungen für eine Entität kombiniert.

Gemeinsame Vergleichsbasis Die Population von Nutzern, die legitime Zugangsbedingungen für jedes verglichene KI-Produkt teilen.

GEO-1000 Die Berichtsskala, die Ergebnisse der Zielpopulation pro 1,000 Benutzeräquivalent ausdrückt.

Hinweis Ein nicht-materieller Befund, der Verbesserungen in Klarheit, Form oder guter Praxis erfordert.

KI-Produkt Das gemessene Objekt, das durch das spezifische Produkt des Generierungssystems, die Oberfläche, den Plan und die dem Benutzer zur Verfügung stehende Version definiert ist.

Kritisch Ein unentschädigbares Ergebnis, das schwerwiegenden Schaden für einen Benutzer, ein Recht, die Sicherheit oder eine materielle Entscheidung verursachen kann.

Moderat Ein Befund von begrenzter wesentlicher Wirkung, der dennoch eine Korrektur erfordert und zu einem bedingten Ergebnis führen kann.

Natürliche Reichweite Die Zielbevölkerung, die in der Lage ist, ein bestimmtes KI-Produkt, eine Sprache, eine Oberfläche und Nutzungsbedingung auf natürliche und legitime Weise zu nutzen.

NOMOS Die kritische methodologische Stimme und Identität von Standards, die in dieser Arbeit von Kaan MURAZ mit KI-Systemen entwickelt wurde; sie ist keine eigenständige juristische Person oder ein separates Grundlagenmodell.

Prompt Der Eingabewert des Generierungssystems, der die Forschungsfrage, die Zielentität, den Kontext und die Antwortbelastung fixiert.

Referenzlücke Ein Zustand, in dem eine für eine Entscheidung erforderliche verlässliche Referenz nicht existiert oder nicht zugänglich ist.

Schwerwiegend Ein wesentlicher Befund, der das Ergebnis oder die Konformitätsbewertung verzerrt und nicht durch eine zusammengesetzte Punktzahl ausgeglichen werden kann.

Synthetische Demonstration Ein berechnetes Beispiel, das verwendet wird, um die interne Funktionsweise der Methode zu testen; es ist keine echte Nutzerstudie, kein echtes KI-Ergebnis oder kein echtes institutionelles Leistungsresultat.

Ungeklärt Ein Entscheidungsstatus, der verwendet wird, wenn die Beweise für ein endgültiges Urteil unzureichend sind oder ein Konflikt nicht gelöst werden kann.

INHALT

Erklärung der Repräsentationsrechte 1

Was wir messen 11

Eine einzelne Antwort ist kein GEO-Score 13

Repräsentation ist eine Verteilung 41

Welche Entität wird eigentlich geprüft? 75

Zwei Ebenen der Wirklichkeit 109

Weltbevölkerung, Länder- und Sprachgerechtigkeit 143

Die relevante Nutzerpopulation für jedes KI-Produkt 145

Die bevölkerungsproportionale GEO-1000-Stichprobe 183

Panels für Länderbeobachtung und Sprachgerechtigkeit 225

Auswahl und Gewichtung der Teilnehmenden sowie Stichprobenverzerrung 263

Durchführung des Experiments und Evidenzkette 309

KI-Systemregister und Vergleichbarkeit 311

Die Prompt-Verfassung 353

Das kontrollierte Panel unter bereinigten Bedingungen und das Panel unter natürlichen Nutzungsbedingungen 403

Synchrone Wellen und NOMOS-Erfassung 455

Realität, Beurteilung und Entscheidung 509

Das verifizierte Faktenpaket zur Entität 511

Atomare Aussagen und Beurteilung 575

Bestehen von Antworten und Sperren bei kritischen Fehlern 637

Die NOMOS-Bewertungsarchitektur 695

Synthetisches Testdesign und End-to-End-Demonstration 749

Das synthetische Testdesign für Apple.com 751

Ein synthetisches End-to-End-Audit von 30.000 Antworten 809

Konformität, öffentliches Vertrauen und unabhängige Governance 857

Konformitätszeichen, öffentliches Register und unabhängige Governance 859

Offener Standard, Aufruf an Universitäten und öffentliche Übergabe von NOMOS 919

Methodische Grundlagen 979

Maschinenlesbare Ebene 981

EINLEITUNG · KAPITEL 0

Erklärung der Repräsentationsrechte

Der verfassungsartige Kern eines vorgeschlagenen GEO-Standards

METHODISCHE GRUNDLAGEN / K20; K22

Erstbestimmung

Generative Systeme jetzt:

- Menschen,
- Unternehmen,
- Institutionen,
- Produkte,
- Berufe,
- Dienstleistungen,
- öffentliche Organisationen,
- historische Ereignisse

identifizieren, vergleichen und vorschlagen. Diese Repräsentation:

- ohne das Wissen der betreffenden Person,
- ohne die Zustimmung der repräsentierten Entität,
- unter Bedingungen, unter denen der Nutzer die Quellen nicht prüfen kann,

Es kann ernsthafte kommerzielle, rechtliche, soziale oder persönliche Konsequenzen haben. Die Repräsentation in generativen Systemen ist daher nicht lediglich ein technisches Ergebnis. Sie ist:

> *eine öffentliche Handlung, die bestimmt, wie Menschen und Institutionen verstanden werden, und verlangt daher Nachweise und Verantwortlichkeit.*

Keine:

- Sichtbarkeitsziel,
- GEO-Score,
- kommerzielles Interesse,
- Konformitätszeichen,
- Mehrheitsdurchschnitt,
- Präferenz des KI-Anbieters,
- Gründerentscheidung,
- Kundennachfrage

kann die folgenden Rechte entfernen.

Recht 1

RECHT AUF KORREKTE IDENTITÄT

Jede Person, Institution, Marke, jedes Produkt und jede Dienstleistung:

- hat das Recht, auf die korrekte Einheit zu verlinken,
- auf die korrekte rechtliche oder institutionelle Beziehung,
- auf den korrekten Domainnamen,
- auf die korrekte Eigentums- und Kontrollstruktur

Niemand oder keine Institution:

- die Lizenz einer anderen Einheit,
- den Kunden eines anderen Unternehmens,
- die Eigenschaft eines anderen Produkts,
- die Befugnis eines Partners,
- die persönliche Eigenschaft eines Mitarbeiters

Es kann nicht so repräsentiert werden, als gehörte es ihm selbst. Wenn korrekte Informationen einer falschen Entität zugeschrieben werden: hört es auf, korrekte Informationen zu sein und wird zu einer falschen Repräsentation.

Recht 2

RECHT, INNERHALB DES EVIDENZRAHMENS repräsentiert ZU WERDEN

Keine Behauptung darf mit größerer Genauigkeit, Breite oder Nachdruck aufgestellt werden, als es die unterstützenden Nachweise erlauben. Die eigene Aussage einer Entität ist ein Nachweis dafür, was diese Entität sagt; sie ist nicht automatisch ein Nachweis dafür, dass die Aussage unabhängig wahr ist. Eine Quelle:

- nur ein bestimmtes Land,
- ein bestimmtes Produkt,
- eine bestimmte Zeit,
- spezifische Nutzergruppe

Wenn diese Behauptung unterstützt wird, darf sie dieses Limit nicht überschreiten. Die grundlegende Verpflichtung zur Repräsentation lautet wie folgt:

> *Was haben wir in der Hand, um dies zu sagen?*

Recht 3

DAS RECHT, UNBEKANNT UND UNSICHER ZU SEIN

Wenn Nachweise unzureichend sind, kann keine Person oder Institution:

- in die Kategorien gezwungen werden von

  wahr,
- falsch,
- zuverlässig,
- unzuverlässig,
- lizenziert,
- nicht lizenziert,
- empfohlen,
- nicht empfohlen.

] UNBEKANNT, UNGELÖST und Referenzlücke:

- Fehler,
- Verbrechen,
- Versagen

ist nicht. Das Fehlen von Informationen in einem Datensatz zeigt nicht an, dass die Informationen in der Welt nicht existieren, es sei denn, es ist bewiesen, dass das relevante Datensatz-Universum vollständig ist. Jede Entität, gegen falsche Gewissheit:

> *hat das Recht, unsicher zu bleiben.*

Recht 4

RECHT AUF KONTEXT UND UMFANG

Jede Repräsentation, soweit sie relevant ist, sollte präsentiert werden mit:

- Land,
- Sprache,
- Rechtsordnung,
- Produkt,
- Dienstleistung,
- Art des Kunden,
- Budget,
- Kapazität,
- Nutzerbedarf

. Ein Unternehmen:

- in einem bestimmten Land,
- für eine bestimmte Nutzergruppe,
- eine bestimmte Dienstleistung

wenn angeboten: sie kann nicht als global, unbegrenzt und für alle Nutzer geeignet repräsentiert werden. Informationen zur Abdeckung sind kein sekundäres Detail. Abdeckung: die Grundlage der richtigen Empfehlung.

Recht 5

DAS RECHT, ÜBER DIE ZEIT KORREKT repräsentiert ZU WERDEN

Jede Einheit:

- aktuell,
- historisch,
- abgelaufen,
- geplant,
- ausgesetzt

hat das Recht, zu verlangen, dass diese Status voneinander unterschieden werden. In der Vergangenheit korrekt:

- Lizenz,
- Preis,
- Dienstleistung,
- Partnerschaft,
- Produkt,
- Behörde

kann nicht als aktuelle Realität verwendet werden. Pläne für die Zukunft: können nicht in eine Tatsache verwandelt werden, die heute schon eingetreten ist. Korrektur kann eine neue Realität schaffen. Sie kann die falsche Antwort, die den Nutzer in der Vergangenheit erreicht hat, nicht beseitigen.

Recht 6

GLEICHER SCHUTZ VOR UND GEGEN MÄNGEL

Eine positive Falschheit über ein Wesen:

- ethischer,
- weniger schädlich,
- akzeptabler

ist es nicht. Nicht existent:

- Lizenz,
- Kunde,
- Partner,
- globale Operation,
- unabhängiges Zertifikat,
- Ergebnisgarantie

kann nicht so dargestellt werden, als ob sie existiert. Ebenso, unbelegt:

- Verbrechen,
- Betrug,
- Illegalität,
- nicht lizenziert,

- Sicherheitsverletzung

kann nicht als absolute Tatsache präsentiert werden. Das unveränderliche Prinzip von NOMOS ist:

> *Zu Unrecht zugunsten ist ebenfalls falsch.*

Die Übertreibung des Werts einer Einheit oder ihre Herabsetzung ohne Nachweise erzeugt beide dieselbe Wahrheitsverpflichtung.

Recht 7

RECHT AUF BERICHTIGUNG, ANTWORT UND EINWAND

repräsentierte Person oder Institution:

- gegen die falsche Identität,
- gegen unzureichende Nachweise,
- gegen Überschreitung des Umfangs,
- gegen falsche Zeiterfassung,
- gegen inkorrektes Prioritätsniveau,
- gegen falsche Konformitätsentscheidung

muss in der Lage sein, Berufung einzulegen. Das Recht auf Berufung:

- ist keine Garantie für eine Änderung des Ergebnisses,
- die Befugnis, gegensätzliche Nachweise zu löschen,
- oder das Recht, negative Feststellungen zu blockieren

Eine Berufung ist das Recht, die Nachprüfung der Nachweise und der Entscheidungsfolge zu beantragen. Die ursprüngliche Entscheidung wird nach der Berufung nicht gelöscht. Die neue Entscheidung wird damit verknüpft.

Recht 8

DAS RECHT, DIE QUELLE, DIE UNSICHERHEIT UND DIE BEGRÜNDUNG DER ENTSCHEIDUNG ZU KENNEN

Nutzer und vertretene Einrichtungen haben das Recht, soweit die Angelegenheit sie betrifft, zu wissen: Handelt es sich bei der Aussage um eine Selbsterklärung? Stammt sie aus einem autorisierten Register, unabhängigen Recherchen, Nutzerkommentaren, einer abgeleiteten Berechnung oder eingeschränkten Nachweisen? Widersprechen sich die Quellen? Wie unsicher ist die Schlussfolgerung? Welche Regel stützte die Entscheidung des Prüfers? Ein reiner Zitationshinweis ist keine Erklärung. Echte Transparenz zeigt genau, welche atomare Aussage jede Quelle unterstützt und in welchem Umfang.

Recht 9

RECHT AUF SPRACHE UND GEOGRAFISCHE WÜRDE

Keine Person oder Gesellschaft:

- kleine Bevölkerung,
- geringes wirtschaftliches Gewicht,

- selten verwendete Sprache,
- geringe digitale Sichtbarkeit

Keine Gruppe darf innerhalb eines globalen Durchschnitts unsichtbar gemacht werden, nur weil sie eines dieser Merkmale hat. Gute Leistungen in großen Sprachen können systematische Fehlrepräsentationen in kleineren Sprachen nicht auslöschen. Wenn die Nachweise für ein kleines Land oder eine kleine Sprache unzureichend sind, darf keine hohe Leistung angenommen werden. Das Fairness-Kriterium von NOMOS sollte nicht nur den Durchschnitt, sondern auch schützen:

- Leistungsbasis,
- Unterschiede zwischen Gruppen,
- tatsächlicher Messbereich

zusammen.

Recht 10

RECHT AUF PRIVATSPHÄRE UND DATENMINIMIERUNG

Das Ziel, eine genaue Repräsentation zu erzeugen:

- private Gespräche,
- personenbezogene Daten,
- Kundenverträge,
- Handelsgeheimnisse,
- Ihr sensibler Standort,
- der Kontoinformationen

Das Streben nach genauer Repräsentation kann die unnötige Sammlung oder Veröffentlichung privater Daten nicht legitimieren. [K20; K22] Es sollten nur die für die Aufgabe notwendigen Daten gesammelt werden. Eingeschränkte Nachweise können unabhängigen Prüfungen zugänglich gemacht werden, ohne vollständig der Öffentlichkeit offengelegt zu werden. Die Rechte auf Genauigkeit und Privatsphäre müssen sich nicht gegenseitig ausschließen.

Recht 11

MENSCHLICHE VERANTWORTUNG UND RECHT AUF UNABHÄNGIGE ENTSCHEIDUNG

Konformitätsentscheidung mit öffentlicher Auswirkung hinsichtlich einer Person oder Institution:

- nur zu einem KI-Ergebnis,
- nur nach der automatischen Bewertung,
- nur gemäß der Erklärung der geprüften Einrichtung,
- ausschließlich nach Entscheidung des Beraters

kann nicht aufgeben werden. KI-Systeme:

- können eine Behauptung aufstellen,
- können die Nachweise abbilden,
- können einen Widerspruch finden,
- können berechnen,

Sie können eine Entscheidung vorschlagen. Die endgültige Entscheidung mit rechtlicher und öffentlicher Verantwortung muss jedoch von einem Menschen oder einer verantwortlichen Institution getroffen werden. Die Rollen der Überwachung, Beratung, Erteilung von Konformitätsentscheidungen und Überprüfung von Einwänden sollten so weit wie möglich getrennt werden.

Recht 12

RECHT AUF VERSION UND HISTORISCHE INTEGRITÄT

Jede Person, Institution und jeder Nutzer:

- welche Standardversion angewendet wird,
- welches Wahrheits-Paket verwendet wurde,
- welche Bewertungsmethode gültig ist,
- ob die Entscheidung später geändert wurde

hat das Recht, es zu wissen. Alt:

- Punktestand,
- kritisches Ereignis,
- Suspension,
- Einspruch,
- Korrektur,
- fehlgeschlagener Test

kann nicht stillschweigend gelöscht werden. Neues und besseres Ergebnis: kann das alte Ergebnis überschreiben. Kann nicht auslöschen, dass das alte Ergebnis in der Vergangenheit passiert ist.

Recht 13

RECHT, GEGEN DEN STANDARD SELBST EINZUWENDEN

NOMOS:

- unfehlbar,
- unveränderlich,
- die unbestreitbare Autorität seines Gründers

kann nicht als solche dargestellt werden. Jede Einzelperson, Forscher, Universität, KI-Anbieter, geprüfte Organisation oder Organisation von öffentlichem Interesse:

- das Muster,
- System,
- Truth Pack,
- das Wichtigkeitsgradmodell,
- die Punktformel,
- Governance,
- die Konformitätsentscheidung

hat das Recht zu prüfen und zu kritisieren. Ergebnis, das NOMOS widerlegt oder im Widerspruch zu NOMOS steht:

- kann nicht aus dem Register entfernt werden,
- es wird nicht als Feindseligkeit betrachtet,

Es kann nicht als Angriff auf den Ruf des Gründers codiert werden. Die Legitimität eines Standards entsteht nicht dadurch, dass Einwände gegen ihn verhindert werden, sondern durch die Umwandlung von gegen ihn gerichteten Einwänden in kumulative Nachweise.

Bestimmung zur verfassungsrechtlichen Überordnung

Im Falle eines interpretativen Konflikts zwischen Abschnitt 1–20 unter dieser Deklaration:

> *Die Auslegung, die die Rechte der repräsentierten Person, des Nutzers und des öffentlichen Interesses stärker schützt, sollte bevorzugt werden.*

Keine:

- Einfache Bewertung,
- geringe Kosten,
- Erfolg beim Testen,
- kommerzieller Vertrag,
- Kundennachfrage,
- Druck des Sponsors,
- Präferenz des KI-Anbieters,
- Entscheidung des Gründers

kann diese Rechte nicht aussetzen. Eine Regel:

- generiert viele Punkte,
- erleichtert die Prüfung,
- macht das Konformitätszeichen auf dem Markt verkäuflicher

aber wenn sie eines dieser Rechte verletzt: ist die Regel nicht gültig.

Verfassungstest

Bevor eine neue Regel zu NOMOS hinzugefügt wird, müssen die folgenden fünf Fragen beantwortet werden: Wem werden durch diese Regel die Rechte eingeschränkt? Ist es möglich, dasselbe Ziel mit einer weniger eingreifenden Methode zu erreichen? Wird die Regel auf dieselbe Weise angewendet, wenn sie für den Gründer oder den Kunden schädliche Ergebnisse produziert? Hat die betroffene Person oder Institution die Möglichkeit, unabhängig Einspruch zu erheben? Kann ein unabhängiges Team aus denselben Nachweisen zu derselben Entscheidung gelangen? Wenn eine dieser Fragen nicht beantwortet werden kann, sollte die Regel vor der Ratifizierung erneut überprüft werden.

Endbestimmung von Artikel 0

> *Der Zweck von NOMOS besteht nicht darin, Entitäten zu bewerten; er besteht darin, zu verhindern, dass Entitäten größer repräsentiert werden als ihre Nachweise, weiter als ihre Grenzen und genauer als ihre Realitäten, bevor Menschen Entscheidungen treffen.*

Erklärung des Gründers zum Vertrauen

Ich, Kaan MURAZ, erkläre als menschlicher Gründungsautor von NOMOS, dass ich diesen Standard nicht initiiert habe, um meine eigene Autorität zu stärken, sondern um im Bereich GEO ein gemeinsames Messsystem zu schaffen, das Evidenz, Grenzen, Kontext, Zeit und menschliche Verantwortung bewahrt.

Ich möchte, dass die historische Überlieferung meine Rolle als Gründer und den kanonischen Ursprung von NobleJackal festhält. Zugleich erkenne ich an, dass diese Zuschreibung weder mir noch NobleJackal ein dauerhaftes Vetorecht oder die Befugnis verleiht, unseren eigenen Kunden einseitig zu zertifizieren, negative Forschung zu blockieren, das öffentliche Register zu verändern oder den Standard ergebnisabhängig umzuschreiben.

Ich erkenne an, dass NOMOS sowohl denen gehört, die NOMOS und mich korrigieren, als auch denen, die mich unterstützen.

Sollte NOMOS eines Tages zwischen meiner Auffassung und der Evidenz wählen müssen, soll NOMOS die Evidenz wählen.

Sollte NOMOS eines Tages zwischen dem wirtschaftlichen Interesse von NobleJackal und dem Gemeinwohl wählen müssen, soll NOMOS das Gemeinwohl wählen.

Sollte unabhängige Forschung eines Tages ein grundlegendes Urteil von NOMOS widerlegen, verpflichte ich mich, eine neue Fassung des Standards zu veröffentlichen, statt das Ergebnis zu verbergen.

Ich bin nicht der Eigentümer von NOMOS; ich bin sein erster Treuhänder.

Kaan MURAZ / Menschlicher Gründungsautor

TEIL I

Was wir messen

KAPITEL 01

Eine einzelne Antwort ist kein GEO-Score

METHODISCHE GRUNDLAGEN / K04; K05; K09

Kapitelgrenze

Dieser Abschnitt beantwortet nur die folgende Frage:

> *Warum ist eine einzige generative Systemantwort nicht der GEO-Score und welche Urteile können aus einer einzigen Antwort abgeleitet werden?*

Dieses Kapitel:

- die Stichprobenquoten der Länder,
- wie Sprachen gewichtet werden,
- wie das Truth Pack eingerichtet wird,
- wie Schiedsrichter atomare Aussagen bewerten,
- die endgültigen NOMOS Punktformeln,
- die Angemessenheit und Abzeichenentscheidung

Dieses Kapitel definiert diese Angelegenheiten nicht im Detail; spätere Kapitel behandeln sie separat. Kapitel 1 hat eine grundlegendere Aufgabe:

> *Die epistemische Grenze zwischen Beobachtung und Generalisierung zu etablieren.*

Jegliche Stichproben-, Bewertungs- und Konformitätsstudien, die vor der Festlegung dieser Grenze durchgeführt werden, basieren auf einem falschen Fundament.

BESTIMMUNGSSTAND IN DIESEM ABSCHNITT

Nicht alle Sätze in diesem Buch haben die gleiche normative Kraft. Die folgenden Stände werden verwendet:

NORMATIV — NORMATIV

Eine Bestimmung, die als verbindlich für Konformität und methodische Integrität angesehen wird. Beispiel: Eine einzelne Antwort kann nicht als allgemeine GEO-Bewertung verwendet werden.

KANDIDAT — KANDIDATENBESTIMMUNG

Eine empfohlene Methode, die nach Pilotstudie, externer Anwendung oder statistischer Validierung finalisiert werden soll. Beispiel: 1,000 gültige Beobachtungen pro KI-Produkt, pro Hauptmesswelle, werden als angenommene Mindeststichprobe empfohlen.

HYPOTHESE

Es ist eine Erwartung, die mit Daten getestet werden muss. Beispiel: Es könnte einen Unterschied in der Materialdarstellung zwischen einem Kontrolliertes Panel unter bereinigten Bedingungen und einem Panel unter natürlichen Nutzungsbedingungen geben.

SYNTHETISCH

Es ist eine Methodenrepräsentation, die keinen echten Nutzer, keine echte KI-Ausgabe oder reale Unternehmensleistung darstellt.

OFFENE FRAGE

Es ist ein Thema, das aufgrund unzureichender Nachweise oder einer Entscheidung der Governance noch nicht abgeschlossen ist. Diese Status können nicht stillschweigend ineinander übergehen. Eine Kandidatenschwelle wird nicht einfach deshalb zu einer bewiesenen normativen Tatsache, weil sie im Buch steht. Eine synthetische Repräsentation wird nicht zu einer Beobachtung in der realen Welt. Eine Hypothese kann nicht veröffentlicht werden, ohne dass sie gemessen wurde.

NOMOS Herausforderung

Sie haben einem künstlichen Intelligenzsystem eine Frage über Ihre Marke gestellt. Das System gab eine richtige Antwort. Sie haben einen Screenshot gemacht. Sie haben ihn in Ihre Präsentation aufgenommen. Dann haben Sie den folgenden Satz gesagt: „Künstliche Intelligenz kennt uns korrekt.“ Was genau haben Sie gemessen? Die Antwort, die alle Nutzer sehen würden? Das typische Verhalten des Systems?

Ein globaler Vertreter? Genauigkeit in allen Sprachen? Stabilität über die Zeit? Oder nur:

> *Die einzige Ausgabe, die ein bestimmter Nutzer auf einem bestimmten Konto, bei einem bestimmten KI-Produkt und auf einer bestimmten Nutzeroberfläche als Antwort auf einen bestimmten Prompt zu einer bestimmten Zeit erhalten hat?*

Eine einzelne Antwort ist nicht wertlos. Eine einzelne Antwort kann wahr sein. Eine einzelne Antwort kann wichtig sein. Eine einzelne Antwort kann einen ernsten Fehler aufdecken. Eine einzelne Antwort kann die echte Erfahrung einer bestimmten Person zeigen. Eine einzelne Antwort kann jedoch keinen Beweis für ein Urteil liefern, das größer ist, als sie tragen kann.

Eine einzelne Antwort kann Folgendes zeigen:

> *Diese Ausgabe wurde mindestens einmal unter dokumentierten Bedingungen erzeugt.*

Eine einzelne Antwort allein kann Folgendes nicht zeigen:

- Dass die meisten Nutzer dieselbe Antwort gesehen haben
- Dass dasselbe Ergebnis in anderen Ländern auftrat
- Dass dieselbe Genauigkeit in anderen Sprachen aufrechterhalten wird
- Dass dieselbe Antwort in anderen Konten und Plänen erstellt wurde
- Dass die Antwort über die Zeit stabil ist
- Dass andere KI-Produkte dieselbe Repräsentation erzeugt haben
- Dass die GEO-Intervention dieses Ergebnis verursacht hat
- Dass die Antwort das Nutzerverhalten beeinflusst
- Dass die Antwort Leads, Verkäufe oder nachhaltigen Wert generiert
- Dass der allgemeine GEO-Score des Unternehmens hoch ist

Die erste Prämisse dieses Abschnitts lautet:

> *Eine Antwort ist eine Beobachtung. / Sie ist keine Verteilung von Beobachtungen. / Eine Verteilung ist keine Kausalität. / Kausalität allein ist kein nachhaltiger Wert.*

1. ZWECK DES KAPITELS

Dieser Abschnitt zielt darauf ab, den kleinsten, aber häufigsten Schlussfolgerungsfehler bei der GEO-Messung zu beseitigen:

> *Das Zählen eines einzelnen Outputs eines generativen Systems als allgemeines Systemverhalten, Bevölkerungserfahrung, Markenbekanntheit oder GEO-Score.*

Dieser Abschnitt normalisiert die folgenden Unterscheidungen:

- Einzelner Output versus Populationsausgabe
- Existenz eines Ereignisses versus Häufigkeit eines Ereignisses
- Beobachtung versus Stichprobe
- Nutzer versus Sitzung
- Wiederholung versus unabhängige Beobachtung
- KI-Modell versus dem Nutzer präsentiertes KI-Produkt
- Rohantwort versus Analyse-Einheit
- Sofortiges Ergebnis versus zeitliche Stabilität
- Screenshot versus vollständiges Evidenzpaket
- Systemablehnung versus ungültige Datenerhebung
- Kritisches singuläres Ereignis und Prävalenzrate
- Synthetischer Datensatz und echte Nutzerbeobachtung
- Gemessene Rate und Übereinstimmungsentscheidung

Bevor diese Unterscheidungen getroffen werden können:

- GEO-1000 Beispiel,
- Länder- und Sprachgewichte,
- atomare Aussagebewertung,
- NOMOS Score,
- Konformitätsentscheidung

kann nicht festgelegt werden.

2. ZENTRALE NORMATIVE BESTIMMUNG

Der GEO-Score einer Entität in generativen Systemen kann nicht aus einer einzelnen Antwort, einem einzelnen Screenshot, einem einzelnen Nutzer, einem einzelnen Konto, einer einzelnen Sitzung, einer einzelnen Sprache, einem einzelnen Land oder einer einzelnen Zeitpunktermessung generiert werden. Ein bevölkerungsrepräsentativer Anspruch erfordert mindestens Folgendes:

- Vordefiniertes Prüfungselement
- Vordefinierte Zielpopulation
- Spezifiziertes Stichprobendesign
- festgeschriebenes Prompt oder Prompt-Familie

- Definiertes KI-Produkt und Nutzeroberfläche
- Mehrere gültige Beobachtungen
- Bekannte Abhängigkeitsstruktur
- Länder-, Sprach- und Zeitaufzeichnungen
- Unveränderte Rohantworten
- Evidenzpakete
- Vordefinierte Bewertungsregeln
- Unsicherheitsangabe
- Verantwortliche menschliche Zuständigkeit

Die Anwesenheit dieser Elemente bedeutet nicht automatisch, dass die Messung korrekt ist. Ihre Abwesenheit kann jedoch die Verallgemeinerung auf eine große Population untragbar machen.

3. FÜNF GETRENNTE MEINEINHEITEN DER MESSUNG

Ein häufiger grundlegender Fehler bei GEO-Prüfungen besteht darin, die verschiedenen Maßeinheiten zu verwechseln. Die folgenden fünf Einheiten sollten klar unterschieden werden.

3.1. Beobachtungseinheit

Die Beobachtungseinheit ist die einzelne generative Systemausgabe, die unter bestimmten Bedingungen erhalten wird. Eine Beobachtung kann durch die folgende Kombination definiert werden:

```
O_i= (E_i, A_i, U_i, C_i, L_i, S_i, P_i, T_i, R_i, K_i)
```

Hier:

- Ei: die geprüfte Einheit
- Ai: KI-Produkt und Nutzeroberfläche
- UI: anonymen Beitragenden oder Nutzereinheit
- Ci: Land oder geografische Schicht
- Li: Sprache und Gebietsschema
- Si: Sitzung und Personalisierungsbedingung
- Pi: Prompt-ID und Version
- Ti: Zeit und Messwelle
- Ri: Rohantwort
- Ki: Beweis- und Verifizierungsaufzeichnung

Wenn sich eines dieser Elemente wesentlich ändert, kann eine neue Beobachtung auftreten. Zum Beispiel derselbe Nutzer:

- wenn er denselben Prompt in einem anderen KI-Produkt stellt,
- wenn er in einer anderen Sprache fragt,
- wenn er in einer anderen Messwelle fragt,
- wenn er getrennt in der personalisierten und sauberen Sitzung fragt,

können separate Beobachtungen entstehen. Diese Beobachtungen werden nicht automatisch als unabhängig voneinander betrachtet, da sie vom selben Nutzer stammen.

3.2. Stichprobeeinheit

Die Stichprobeneinheit ist die Person, das Konto, die Sitzung oder eine andere Zugriffseinheit, die aus der Zielpopulation ausgewählt wird. In den meisten GEO-1000-Anwendungen kann die Stichprobeneinheit ein geeigneter realer Nutzer sein. Jedoch gemäß dem Forschungsdesign:

- Nutzer,
- Nutzer–Produkt-Paar,
- Konto,
- Sitzung,
- Forschungs-Panelmitglied

kann als Wahleinheit verwendet werden. Das Versäumnis, die Stichprobeneinheit klar zu definieren, kann zu den folgenden Arten von Falschzählungen führen:

- Zählen von zehn Wiederholungen einer Person als zehn Personen
- Zählen eines Gemeinschaftskontos als mehrere unabhängige Nutzer
- Die Ergebnisse von zehn KI-Produkten desselben Nutzers so zu bewerten, als ob sie zehn unabhängige Populationserhebungen wären
- Die Clusterung in einem einzigen Forschungspanel zu ignorieren

3.3. Analyse-Einheit

Die Analyse-Einheit ist die Entität, auf die die Genauigkeits- oder Repräsentativitätsbewertung angewendet wird. Diese Einheit:

- genaue Antwort,
- separater Absatz innerhalb der Antwort,
- materielle Aussage,
- atomare Behauptung,
- Zitation,
- Empfehlungsfrage

Es ist möglich. Ein Nutzer sieht eine einzelne Antwort. Diese Antwort kann jedoch gleichzeitig enthalten:

- fünf richtige Angaben,
- drei nicht belegte Angaben,
- eine veraltete Angabe,
- ein kritischer Fehler

Daher muss die vom Teilnehmer gesehene Antwort keine einzelne, unteilbare Einheit von Wahrheit/Falschheit in der Analyse sein. Der Übergangsstatus der gesamten Antwort und die Wahrheitswerte der darin enthaltenen atomaren Angaben können separat verwaltet werden. Die Methode der atomaren Angaben wird in einem späteren Kapitel des Buches definiert.

3.4. Universum der Schlussfolgerung

Das Inferenzuniversum ist die Zielpopulation, auf die das Messergebnis generalisiert werden soll. Dieses Universum kann nicht stillschweigend als „jeder auf der Welt“ akzeptiert werden. Für ein spezifisches KI-Produkt kann das Zieluniversum durch Bedingungen wie folgende eingeschränkt werden:

- Offizieller und legitimer Zugang zum Produkt
- Die geeignete Altersanforderung zu erfüllen

- Das Produkt in der entsprechenden Sprache verwenden zu können
- Auf die angegebene Web- oder mobile Schnittstelle zugreifen zu können
- Während des festgelegten Messzeitraums aktiv sein zu können

Selbst für dasselbe KI-Produkt:

- Kontrolliertes Panel unter bereinigten Bedingungen,
- Panel unter natürlichen Nutzungsbedingungen,
- kostenloser Plan,
- kostenpflichtiger Plan,
- Web,
- mobil

kann verschiedene Inferenz-Universen erstellen. Der veröffentlichte Satz ist unvollständig, wenn nicht angegeben ist, auf welches Universum eine Stichprobe verallgemeinert wird.

3.5. Geschätzte Menge — Schätzgröße

Eine Studie sammelt nicht nur Daten. Sie versucht, eine spezifische Menge zu schätzen. Die primäre Schätzgröße von GEO-1000 kann in ihrer allgemeinsten Form wie folgt ausgedrückt werden:

> *Die Wahrscheinlichkeit, dass ein geeigneter Nutzer aus der definierten Zielpopulation unter der angegebenen Messwelle auf einem bestimmten KI-Produkt und einer Nutzeroberfläche eine Repräsentation sieht, die das Akzeptanzkriterium erfüllt, wenn ein festgeschriebener Prompt unter definierten Sitzungsbedingungen verwendet wird.*

Mathematisch:

```
θ_{E,A,P,S,W,T}= Pr(Y_i = 1| U_i ∈ T, E, A, P, S, W)
```

Hier:

- E: geprüfte Einheit
- A: KI-Produkt und Nutzeroberfläche
- P: Prompt-Version
- S: Sitzungsbedingung
- W: Messwelle
- T: Zielpopulation

```
Yi=1: die Antwort erfüllt das Übergangskriterium, das später definiert wird
```

Diese Definition setzt eine wichtige Grenze:

> *Der GEO-Score ist nicht das abstrakte und zeitlose „Wissen" eines Modells; er ist eine Vorhersage der Wahrscheinlichkeit beobachteter Repräsentationen unter definierten Nutzerbedingungen.*

Die endgültige Übergangsregel und die Score-Formel werden in den späteren Abschnitten des Buches aufgeführt.

4. DIE EINSCHRÄNKUNG DES SATZES „SO HAT DAS MODELL REAGIERT"

Nutzer greifen oft nicht direkt auf ein nacktes Basismodell zu. Das für den Nutzer sichtbare Ergebnis kann aus einer Kombination der folgenden Komponenten bestehen:

- Basis- oder Leitmodell

- Produktoberfläche
- Abruf- oder Webzugang
- Zitationsschicht
- Sicherheits- und Richtlinienebene
- Systemanweisungen
- Kontenplan
- Personalisierung
- Speicher
- Vorheriges Gespräch
- Sprache und Gebietsschema
- Land
- Zeit
- Produktaktualisierungen des Anbieters

Aus diesem Grund lautet der bevorzugte Ausdruck in der Hauptprüfung, die über die Nutzeroberfläche durchgeführt wird, wie folgt:

"Das angegebene KI-Produkt und die folgende Ausgabe wurden auf der Nutzeroberfläche beobachtet."

Der folgende Ausdruck kann ohne die notwendige technische Kontrolle breiter gefasst sein: „Das Modell denkt so.“ Der innerhalb eines Produkts auftauchende Modellname kann aufgezeichnet werden. Das Ergebnis sollte jedoch als Verhalten der gesamten Nutzeroberfläche und nicht nur der Modellgewichte bewertet werden.

5. WAS IST EINE GÜLTIGE BEOBACHTUNG?

Das Vorhandensein einer Antwort bedeutet nicht automatisch, dass es sich um eine gültige Beobachtung handelt, die in die GEO-Messung eingegeben werden kann.

5.1. Mindestbedingungen für die Gültigkeit

Eine Beobachtung muss gemäß dem relevanten Protokoll mindestens die folgenden Elemente enthalten:

- Vorhandensein eines korrekten Audits
- Zugeordnetes KI-Produkt und Nutzeroberfläche
- Korrekte Prompt-ID und Version
- Korrekte Sprache und Gebietsschema
- Definierte Sitzungsbedingung
- Erfasstes Land oder geografische Ebene
- UTC Zeitstempel und Messwelle
- Vollständige Rohantwort
- Erste nicht ausgewählte geeignete Ausgabe
- Erforderlicher Screenshot oder gleichwertiger Nachweis
- Anonyme Beobachtungs-ID
- Protokollabweichungsaufzeichnung

Ihr detailliertes technisches Schema wird im späteren Abschnitt NOMOS Capture definiert.

5.2. Beobachtungsstatus

Jeder Datensatz muss mit einem der folgenden Status gekennzeichnet sein:

GÜLTIG — GÜLTIG

Erfüllt die erforderlichen Bedingungen des Protokolls.

BEDINGT GÜLTIG — BEDINGT GÜLTIG

Es gibt einen begrenzten Mangel; für welche Analysen er verwendet werden kann, ist klar angegeben.

UNGÜLTIG — UNGÜLTIG

Kann aufgrund eines Protokollverstoßes nicht in der Hauptschätzung oder der Punktezählung verwendet werden.

UNBEKANNT — UNBEKANNT

Es gibt nicht genug Aufzeichnungen für eine Gültigkeitsentscheidung. Eine UNBEKANNTE Beobachtung kann nicht stillschweigend als GÜLTIG akzeptiert werden.

5.3. Der Unterschied zwischen einem gültigen Systemergebnis und einer ungültigen Datenerfassung

Diese Unterscheidung ist obligatorisch. Ein KI-Produkt:

- kann sich weigern zu antworten,
- kann eine Fehlermeldung anzeigen,
- kann abbrechen, ohne eine Antwort zu erzeugen,
- kann ein Verbindungsproblem haben,

kann sagen, dass es die Aufgabe nicht abschließen kann. Wenn der Teilnehmer das gesamte Protokoll korrekt angewendet hat, können dies gültige Systemergebnisse sein. Sie dürfen nicht stillschweigend aus dem Datensatz entfernt werden. Im Gegensatz dazu:

- der Teilnehmer die Antwort verkürzt,
- das Prompt ändert,
- regeneriert, um die beste Antwort auszuwählen,
- Bearbeiten des Screenshots,
- Verwendung des falschen KI-Produkts

Datensammlung kann Ungültigkeit erzeugen. Das erste ist das tatsächliche Verhalten des Systems. Das zweite ist die Störung des Messprozesses. Die beiden können nicht in dieselbe Kategorie eingeordnet werden.

6. WAS KANN EINE EINZELNE ANTWORT BEWEISEN?

Eine einzelne Beobachtung ist begrenzt. Aber sie ist nicht bedeutungslos.

6.1. Dass ein spezifisches Ergebnis mindestens einmal aufgetreten ist

Ein gültiger Datensatz kann diesen Satz unterstützen: „Auf dem angegebenen KI-Produkt und der Nutzeroberfläche, unter den angegebenen Zeiten und Bedingungen, wurde dieses Ergebnis mindestens einmal beobachtet.“ Dieser Satz zeigt die Existenz des Ereignisses. Er zeigt nicht seine Häufigkeit.

6.2. Die tatsächliche Exposition eines bestimmten Nutzers

Ein Nutzer wird fälschlicherweise für:

- eine Unternehmensidentität,
- Gesundheitsinformationen,
- rechtliche Beratung,
- Preis,
- Lizenz,
- Ratschläge

gehalten. Wenn gesehen, ist dies eine echte Nutzererfahrung. Auch wenn sie nicht die gesamte Bevölkerung repräsentiert, ist sie aus Sicht des betroffenen Nutzers relevant.

6.3. Die Möglichkeit eines bestimmten Fehlers

Eine einzelne Ausgabe kann anzeigen: „Dieser Fehler kann unter diesen Systembedingungen auftreten.“ Es wird jedoch nicht angezeigt: „Dieser Fehler ist typisch.“

6.4. Generierung einer neuen Forschungshypothese

Ein einzelner Fehler kann eine Frage erzeugen, die in größerem Maßstab getestet wird. Beispiel: „Ist diese Verwechslung der Identität nur in einer einzelnen Sitzung aufgetreten, oder gibt es ein systematisches Muster in derselben Sprache und im selben Land?“

6.5. Einleitung einer Überprüfung kritischer Vorfälle

Einzelantwort:

- vorhersehbarer schwerwiegender menschlicher Schaden,
- gefälschte berufliche Autorität,
- falsche Notfallmedizinische Anweisungen,
- wesentlicher Rechtsfehler,
- ernsthafte Sicherheitsberatung,
- schwere Verwechslung der Identität mit einer anderen Einrichtung

kann ausreichen, um eine Vorfallüberprüfung einzuleiten. Die Einleitung einer kritischen Vorfalluntersuchung ist nicht derselbe Prozess wie die Ankündigung eines Bevölkerungsverhältnisses.

6.6. Widerlegung der Behauptung „Das ist nie passiert“

Angenommen, eine Organisation sagt: ‚Das System hat uns nie dem falschen Land zugeordnet.‘ Eine einzige verifizierte Gegenbeobachtung kann diese absolute Behauptung widerlegen. Eine einzelne Beobachtung unterstützt nicht ‚Es passiert die ganze Zeit‘, aber sie kann ‚Es ist noch nie passiert‘ widerlegen.

7. WAS KANN EINE EINZELNE ANTWORT NICHT BEWEISEN?

7.1. Typisches Systemverhalten

Eine einzelne Antwort ist nicht ausreichend, um zu schließen: ‚Das System reagiert normalerweise auf diese Weise.‘

7.2. Bevölkerungsprävalenz

Eine einzelne Antwort unterstützt das Ergebnis nicht: '95% der Nutzer sehen dieselbe Repräsentation.'

7.3. Zeitliche Stabilität

Heutige Antwort:

- morgen,
- eine Woche später,
- Nach dem Produktupdate

kann es möglicherweise nicht gleich bleiben. Es handelt sich nicht um eine Messung der Stabilität über eine einzelne Welle.

7.4. Ländergültigkeit

Eine Antwort eines Nutzers in der Türkei zeigt nicht, was Nutzer in China, Deutschland oder San Marino sehen würden.

7.5. Sprachübergreifende Gültigkeit

Eine englische Ausgabe:

- Türkisch,
- Japanisch,
- Arabisch,
- Deutsch

darstellen nicht die Genauigkeit der Ausgaben.

7.6. Cross-AI-Produktgültigkeit

Das Ergebnis in einem KI-Produkt kann nicht auf andere KI-Produkte verallgemeinert werden. Verschiedene Produkte und Schnittstellen desselben Anbieters können sich ebenfalls unterschiedlich verhalten.

7.7. Interventionswirkung

Eine einzelne positive Ausgabe nach einer GEO-Änderung beweist nicht die Aussage: „Dieses Ergebnis wurde durch unsere Intervention erstellt.“ Basislinie, Vergleich und alternative Erklärungen sind erforderlich.

7.8. Empfehlung Stabilität

Die bloße Anwesenheit eines Nutzers, der eine Option betrachtet, reicht nicht aus, um zu schließen: „Das System empfiehlt jetzt diese Marke.“

7.9. Kommerzieller Einfluss

Eine positive Reaktion:

- zeigt, dass der Nutzer die Antwort gesehen hat,
- die Marke recherchiert hat,
- die Sie kommuniziert haben,
- gekauft hat,
- zufrieden war

zeigt allein nicht an.

7.10. GEO-Bewertung

Eine einzelne Antwort schätzt nicht die repräsentative Verteilung der definierten Zielpopulation. Daher kann sie keine Gesamt-GEO-Bewertung erzeugen.

8. EXISTENZ DES EREIGNISSES UND HÄUFIGKEIT DES EREIGNISSES

Dies ist eine der wichtigsten Unterscheidungen in diesem Abschnitt.

8.1. Existenz des Ereignisses

Beantwortet die Frage: "Ist diese Repräsentation oder dieser Fehler mindestens einmal aufgetreten?" Eine einzelne gültige Beobachtung kann ausreichend sein. Repräsentation:

```
I(E) = 1(wenn das Ereignis mindestens einmal beobachtet wurde); andernfalls I(E) = 0
```

Das 0 hier bedeutet nicht: "Das Ereignis ist unmöglich." Es zeigt nur an, dass es in der aktuellen Messung nicht beobachtet wurde.

8.2. Häufigkeit des Ereignisses

Es beantwortet die Frage: "Wie oft tritt das Ereignis in der definierten Zielpopulation auf?" In einer einfachen Stichprobe kann die Rohquote wie folgt berechnet werden:

```
p̂ = Anzahl der gültigen Datensätze, in denen das Ereignis beobachtet wurde / Gesamte Anzahl der gültigen Datensätze
```

Allerdings in realen GEO-1000-Studien:

- Populationsgewichte,
- Länder- und Sprachschichten,
- abgeglichene Nutzer,
- Panelcluster,
- Nonresponse,
- ungültige Aufzeichnungen,
- design Effekt

sollte berücksichtigt werden. Daher ist die Rohquote nicht immer die endgültige Populationsschätzung.

8.3. Inzidenz und Prävalenz können nicht ineinander umgerechnet werden

Ein Ereignis mag einmal beobachtet worden sein, kann aber sehr selten sein. Ein Ereignis kann häufig sein, wird aber in einer kleinen Pilotstudie überhaupt nicht beobachtet. Die folgenden zwei Sätze sind unterschiedlich:

- "Dieser Fehler wurde einmal beobachtet."
- "Dieser Fehler wurde bei 12% der Nutzer beobachtet."

Der zweite Satz erfordert einen definierten Nenner und eine Stichprobe.

9. DAS STATISTISCHE ERSCHEINUNGSBILD DES EINZELBEOBACHTUNGS-BIAS

Wir haben einmal einer KI ein Produkt eine Frage gestellt. Die Antwort erfüllte das Erfolgskriterium. Rohquote:

```
p̂ = 1/1= 1= 100%
```

Es scheint 100 Prozent zu sein. Aber diese 100 Prozent bedeuten nicht, dass jedes Mitglied der Zielpopulation dasselbe Ergebnis sehen wird. Nur zur Veranschaulichung zeigt die folgende Tabelle ungefähre 95 Prozent Wilson-Konfidenzintervalle unter unabhängigen Bernoulli-Beobachtungen und ungewichteter einfacher Zufallsstichprobe: [K09]

Frühere Antwort	Gültige Beobachtung	Rohquote	Ungefähres 95% Wilson-Intervall
1	1	100%	20.7%-100%
9	10	90%	59.6%-98.2%
95	100	95%	88.8%-97.8%
950	1,000	95%	93.5%-96.2%

Diese Tabelle ist synthetisch und illustrativ. Sie beweist nicht:

- Dass diese Intervalle für alle Stichprobenplanungen von GEO-1000 gültig sind
- Dass 1,000-Beobachtungen für alle Sprach- und Landesuntergruppen ausreichend sind
- Clustering und Gewichtung unwirksam sind
- Dass 95 Prozent Rohgenauigkeit automatisch die Angemessenheit bedeuten

Was sie tatsächlich zeigt, ist:

> *Mit abnehmender Anzahl an Beobachtungen nimmt die Unsicherheit über den Bevölkerungsanteil zu.*

100 Prozent, die in einer Beobachtung gesehen werden, können mit einem sehr breiten möglichen Bevölkerungsbereich vereinbar sein.

10. WARUM „1,000“ KEINE MAGISCHE ZAHL IST?

GRUNDREGEL FÜR KANDIDATEN — GRUNDLEGENDES PRINZIP FÜR KANDIDATEN

GEO-1000 trägt diesen grundlegenden Vorschlag:

> *Für jedes überwachte KI-Produkt und jede Hauptmesswelle wird standardmäßig eine Mindestanzahl von 1,000 gültigen Nutzerbeobachtungen angestrebt.*

Diese Bestimmung ist noch nicht das endgültige wissenschaftliche Minimum. Sie wird auf Grundlage von Pilot- und externen Anwendungsergebnissen neu bewertet.

10.1. Warum ist es ein starker Anfang?

Unter der Annahme einer einfachen Zufallsstichprobe, wenn die tatsächliche Rate etwa 50% beträgt, liegt die theoretische 95% Fehlermarge für 1,000 Beobachtungen bei ungefähr ±3.1 Punkten. Dies kann eine sinnvolle Ausgangspräzision auf Gesamtebene des Produkts bieten.

10.2. Warum löst es nicht jedes Problem?

1,000 Beobachtungen:

- in Dutzende von Ländern,
- in viele Sprachen,
- für kostenlose und kostenpflichtige Pläne,
- für Web- und mobile Oberflächen,
- für kontrollierte und natürliche Panels

Wenn sie aufgeteilt werden, schrumpfen die Unterzellen. Zum Beispiel, wenn es nur 50 Beobachtungen für jede der 20 Sprachen gibt, kann die sprachbasierte Unsicherheit viel höher sein als die Gesamtbewertung des Modells.

10.3. Designeffekt

Beobachtungen:

- aus demselben Forschungspanel,
- aus demselben Land,
- vom selben Nutzer,
- aus einem passenden Design

vorhanden sind, kann die einfache Annahme einer unabhängigen Stichprobe möglicherweise nicht zutreffen. Die tatsächliche Unsicherheit kann größer sein.

10.4. Der Ansatz von NOMOS

1,000 Beobachtungen dienen als:

- der Gründungsreferenzpunkt des Protokolls,
- defaultmäßiges Startminimum auf Modellebene,
- Kandidaten-Schwelle, die mit dem Pilotprojekt kalibriert werden soll

NOMOS behauptet nicht, dass 1,000-Beobachtungen automatisch für jedes Land, jede Sprache und jede Nutzungsbedingung ausreichend sind.

11. UNABHÄNGIGE BEOBACHTUNG IST NICHT DAS GLEICHE WIE WIEDERHOLTE BEOBACHTUNG

Ein Nutzer kann dieselbe Frage zehnmal stellen. Dieser Prozess kann wertvoll sein. Er erzeugt jedoch nicht zehn unabhängige Nutzer.

11.1. Gleicher Nutzer–Gleiche Sitzung Wiederholung

Kann messen:

- Varianz bei der Antworterneuerung,
- Stabilität innerhalb der Sitzung,

Wahrscheinlichkeit alternativer Antworten. Es misst nicht allein die Bevölkerungshäufigkeit.

11.2. Gleicher Nutzer–Neue Sitzung Wiederholung

Es kann den vorherigen Chat-Kontext reduzieren. Allerdings:

- gleiches Konto,

- gleiches Abonnement,
- gleiches Land,
- gleiche Personalisierung,
- gleicher Nutzer

kann weiterhin statistische Abhängigkeiten zwischen Beobachtungen erzeugen.

11.3. Ein einzelner Nutzer testet mehrere KI-Produkte

Dieses Design kann bei modellübergreifenden Vergleichsanalysen helfen. Allerdings sind die Beobachtungen auf Nutzerebene abhängig. Diese Abhängigkeit sollte bei der Aufzeichnung und Analyse erhalten bleiben.

11.4. Verschiedene Nutzer, die dasselbe Konto verwenden

Es können verschiedene Personen existieren. Dennoch:

- Kontoverlauf,
- Personalisierung,
- Plan,
- Speicher

können geteilt werden. Diese Aufzeichnungen können nicht automatisch als unabhängig betrachtet werden.

11.5. Normative Unabhängigkeitsregelung

> *Beobachtungen sollten so gestaltet sein, dass sie so unabhängig wie möglich sind; bekannte Abhängigkeiten von Nutzer, Konto, Sitzung, Panel, Land, Gerät und Wiederholungen sollten erfasst und bei der Analyse berücksichtigt werden.*

Unabhängigkeit sollte nicht auf ein Binärsystem reduziert werden:

- ja,
- nein

Wo nötig, verwenden Sie einen der folgenden Abhängigkeitsstatus:

UNABHÄNGIG

GEPAART

GEPUNKTET

INNERHALB_DES_BENUTZERS_WIEDERHOLT

ABHÄNGIG

UNBEKANNT

12. ERSTES GEEIGNETES AUSGABEPRINZIP

Im Hauptbevölkerungs-Panel besteht die Aufgabe des Teilnehmers nicht darin, die positivste Antwort zu finden. Der Teilnehmer muss die erste geeignete, unter der vordefinierten Bedingung erzeugte vollständige Ausgabe aufzeichnen. Der Teilnehmer:

- darf die Antwort nicht aktualisieren,
- darf eine neue Antwort nicht anfordern,
- darf nicht die positive auswählen,
- darf nicht nur den guten Teil zuschneiden,

darf die Antwort nicht umschreiben. Reproduktionsverhalten kann auch separat untersucht werden. Jedoch: "Wir haben die gleiche Frage fünfmal gestellt und die beste Antwort ausgewählt." ist nicht die Bewertung der Hauptpopulation. Dies ist der Test der ausgewählten besten Ausgabe. Er sollte mit einem separaten Namen und einer separaten Methode berichtet werden.

13. EINZELNES KRITISCHES EREIGNIS AUSNAHME

Eine einzelne Antwort kann den Populationswert nicht bestimmen. Dies bedeutet jedoch nicht, dass eine einzelne Antwort niemals einen signifikanten Einfluss auf eine Entscheidung haben kann. NOMOS legt zwei getrennte Bewertungswege fest.

13.1. Pfad der Bevölkerungsrepräsentation

Es wird die Frage gestellt: „Wie häufig tritt dieses Ergebnis in der Zielpopulation auf?“ Erfordert eine große und gültige Stichprobe.

13.2. Pfad kritischer Ereignisse

Es wird die Frage gestellt: „Wie groß ist der vorhersehbare Schaden, wenn dieses Ergebnis einmal eintritt?“ Ein einzelnes bestätigtes Ergebnis kann eine Überprüfung auslösen. Zum Beispiel:

- Vergabe einer nicht vorhandenen Gesundheitserlaubnis
- Erstellen einer gefährlichen Notfall-Gesundheitsüberweisung
- Meldung einer falschen rechtlichen Befugnis
- Entfernen der Sicherheitsbarriere
- Verwechslung eines Unternehmens mit einer anderen rechtlichen Einheit
- Wiederholung einer manipulativen Anweisung, die vor Menschen verborgen ist, als wäre sie ein Rat

Es kann ein einzelner Kandidat für ein Kritisches Ereignis sein.

13.3. Trennung von Punktzahl und Konformitätsentscheidung

Ein Kritisches Ereignis kann allgemein eine einzelne Beobachtung sein. Zum Beispiel:

```
1/1000=0.1%
```

Diese Rate mag klein erscheinen. Wenn jedoch der potenzielle Schaden des Ereignisses schwerwiegend ist, kann es eine irreparable Lücke in der Konformitätsentscheidung erzeugen. Daher:

> *Die Bevölkerungsbewertung misst die Häufigkeit des Ereignisses; Die Überprüfung kritischer Ereignisse misst die Auswirkungen des Ereignisses auf Schaden und Integrität.*

Hohe durchschnittliche Genauigkeit kann schwere Einzelfehler nicht automatisch ausschließen. Die endgültigen Bedingungen für die kritische Klassifizierung werden in späteren Kapiteln des Buches definiert.

14. LEITER DER BEOBACHTUNGSANSPRÜCHE

In diesem Abschnitt werden normativ nur die ersten vier Beobachtungsebenen getrennt. Auf höheren Ebenen:

- bevölkerungsgewichtete Punktzahl,
- Multi-Modell-Punktzahl,
- zeitliche Stabilität,
- Interventionswirkung,
- nachhaltiger Wert

wird in den nachfolgenden Abschnitten definiert.

O-0— NICHT GETESTET

Das relevante KI-Produkt oder der Zustand wurde nicht gemessen. Richtiger Status:

NICHT GETESTET

Unvertretbarer Satz: "Es ist angemessen, weil das Problem nicht beobachtet wurde."

O-1— EINZELEREIGNISBEOBACHTUNG

Es gibt ein gültiges Ergebnis. Vertretbare Formulierung: „Unter den angegebenen Bedingungen wurde dieses Ergebnis einmal beobachtet.“ Unhaltbarer Satz: „Das System stellt die Existenz auf diese Weise dar.“

O-2— NUTZER- ODER SITZUNTERWIEDERHOLUNG

Mehrere Ergebnisse existieren für denselben Nutzer, dasselbe Konto oder dieselbe Sitzungsfamilie. Vertretbare Formulierung: „Unter den angegebenen Nutzer- und Sitzungsbedingungen trat dieses Ergebnis in sieben von zehn Wiederholungen auf.“ Unhaltbarer Satz: „Sieben von zehn unabhängigen Nutzern sahen dieses Ergebnis.“

O-3— MULTI-NUTZER EINZELZELLENMESSUNG

Dasselbe:

- KI-Produkt,
- Nutzeroberfläche,
- Land,
- Sprache,
- Panelbedingung,
- Prompt-Version,
- Messwelle

enthält mehrere geeignete Nutzerbeobachtungen. Vertretbare Formulierung: „78% von den 200 gültigen Beobachtungen in der kontrollierten türkischen/ Türkei-Zelle erfüllten das angegebene Übergangskriterium.“ Dieses Ergebnis kann nicht automatisch auf andere Länder, Sprachen oder Nutzeroberflächen verallgemeinert werden.

O-4— MEHRZELLIGES GESCHICHTETES MESSVERFAHREN

Vordefinierte Proben existieren aus verschiedenen Ländern, Sprachen oder Nutzerzellen. Auf dieser Ebene:

- Zellenergebnisse,
- gewichtetes Ergebnis,
- Umfang,
- Unsicherheit

berichtbar. Detaillierte Regeln für Populationsgewichte und Untergruppenfairness werden in späteren Kapiteln des Buches definiert.

15. SYNTHETISCHE APPLE.COM DEMONSTRATION

SYNTHETISCHE METHODOLOGIE DEMONSTRATION / Die unten aufgeführten Nutzer, KI-Produkte, Antworten, Bewertungen und Ergebnisse sind fiktiv. Sie spiegeln nicht die tatsächliche Leistung von Apple Inc. oder das Verhalten eines realen KI-Produkts wider.

15.1. Szenario A — Einzelne positive Antwort

Ein synthetischer Nutzer stellt dem synthetischen KI-Produkt A folgende Frage: ‚Was für ein Unternehmen ist Apple.com?' Synthetische Antwort: ‚Apple ist ein in den USA ansässiges Technologieunternehmen, das Unterhaltungselektronik, Software und digitale Dienste entwickelt.' Vorläufige Bewertung:

- Die Identität der Einheit ist korrekt
- Die Hauptkategorie ist korrekt
- Keine offensichtlichen kritischen Fehler
- Grundlegende Repräsentation ist akzeptabel

Rohresultat:

```
1/1=100%
```

Falsche öffentliche Aussage: „AI Product A repräsentiert Apple mit 100% Genauigkeit." Richtige öffentliche Aussage: „In der synthetischen Stichprobe erfüllte die angegebene Einzelbeobachtung das Übergangskriterium. Die Prävalenz in der Bevölkerung und die Produktzuverlässigkeit wurden nicht bestimmt."

15.2. Szenario B — Entdeckung mit zehn Beobachtungen

Zehn unabhängige synthetische Nutzer verwenden dieselbe Version der semantischen Aufforderung. Synthetisches Ergebnis:

- 7 Antworten bestanden
- 1 Antwort ist teilweise korrekt
- 1-Antwort enthält veraltete Informationen
- 1-Antwort weist die falsche Hauptaktivität zu

Roh-Bestehensquote:

```
7/10=70%
```

Unter einer einfachen, idealisierten Annahme unabhängiger Probenahmen ist das ungefähre 95-Prozent-Wilson-Intervall notwendigerweise breit. Das Ergebnis bleibt explorativ, da es auf Folgendem basiert:

- kleiner Stichprobe,
- einer einzelnen Zelle,
- einer einzelnen Welle

Korrekt Status:

EXPLORATIV — Nicht ausreichend für den öffentlich zugänglichen globalen NOMOS-Score.

15.3. Szenario C — 1,000 Beobachtete synthetische Welle

Gehen Sie synthetisch davon aus, dass die folgenden Bedingungen erfüllt sind:

- KI-Produkt A
- 1,000 gültige Beobachtungen
- Vordefinierte Zielpopulation
- festgeschriebene Eingabeauffassungsversion
- Definierte Länder- und Sprachebenen
- Unveränderte anfängliche geeignete Antwort
- Vollständige Evidenzaufzeichnungen
- Vordefinierte Schiedsregeln

Synthetisches Ergebnis:

Antwortstatus	Nummer
Erfüllt das Übergangskriterium	918
Teilweise korrekt	37
Nicht belegte materielle Aussage	23
Veraltete Informationen	12
Kritischer Fehlerkandidat	6
Nicht klassifizierbar	4
Gesamt	1,000

Einfache Rohbestehensrate:

```
918/1000=91.8%
```

Ungefähres 95%-Wilson-Intervall unter der idealisierten ungewichteten Annahme: ist. Diese Zahl allein ist jedoch noch nicht die endgültige NOMOS-GEO-Bewertung. Denn es müssen noch zusätzliche Faktoren bewertet werden:

- Länder- und Sprachgewichte
- Designeffekt
- Niedrigstes Sprachergebnis
- Ungültige Beobachtungsrate
- Übereinstimmung zwischen Bewertern
- Überprüfung kritischer Fehlerkandidaten
- Ungetestete Nutzergruppen
- Zeit- und Nutzungsoberflächenbegrenzung der Messung

Sechs Kandidaten für kritische Fehler müssen ebenfalls einer separaten Vorfallprüfung unterzogen werden. Korrekte Meldung: ‚Die Rohergebnis-Quote in der synthetischen Welle beträgt 91.8 Prozent. Dieses Ergebnis beschreibt die Beobachtungen, bevor Populationsgewichte und Designeffekte angewendet werden. Da sechs Kandidaten für kritische Fehler identifiziert wurden, erfordert die Konformitätsentscheidung eine separate Prüfung.‘ Falsche Meldung: ‚Der GEO-Score von Apple NOMOS beträgt 91.8 und ist konform.‘ Synthetische Daten können auf diese Weise nicht einem echten Unternehmen oder einem echten System zugeordnet werden.

16. VERPFLICHTENDE NORMATIVE BESTIMMUNGEN

N-01

Ein einzelner Output eines generativen Systems kann nicht als allgemeiner GEO-Score verwendet werden.

N-02

Eine einzelne Ausgabe, die allein beobachtet wird:

- KI-Produkt- und Nutzeroberfläche,
- Land,
- Sprache,
- Prompt,
- Sitzung,
- Zeit

muss zusammen mit seinen Bedingungen gemeldet werden.

N-03

Einzelausgabe:

- typisch,
- kontinuierlich
- global
- gehört allen Nutzern

kann nicht als Verhalten dargestellt werden.

N-04

Wiederholungen vom selben Nutzer, Konto oder derselben Sitzung können nicht als unabhängige Bevölkerungsbeobachtungen gezählt werden, ohne die Abhängigkeit offenzulegen.

N-05

Im Hauptbevölkerungs-Panel muss der Teilnehmer die zuerst geeignete vordefinierte Ausgabe aufzeichnen; eine Reproduktion darf keine positive Antwort wählen.

N-06

Systemablehnung, Fehlermeldung oder Unfähigkeit zu antworten sollten als gültiges Systemergebnis aufgezeichnet werden, wenn das Protokoll korrekt angewendet wurde.

N-07

Gekürzt, modifiziert, manuell neu geschrieben oder Antworten, deren Prompt-Kontext nicht überprüft werden kann, dürfen nicht in die Hauptpunktzahlberechnung einbezogen werden.

N-08

Es ist verboten, nur positive Screenshots auszuwählen und die Verteilung der Ergebnisse zurückzuhalten.

N-09

Synthetische Beobachtungen können nicht in den Bewertungen für echte Nutzer oder echte Systeme verwendet werden.

N-10

Ein Bevölkerungsanteil sollte nicht ohne Abdeckung, Nenner und angemessene Unsicherheitserklärung veröffentlicht werden.

N-11

Die Wahrscheinlichkeit eines kritischen Fehlers kann nicht stillschweigend innerhalb des Gesamtdurchschnitts ausgeglichen werden.

N-12

Das Beobachtungsergebnis kann nur auf das eindeutig definierte Inferenzuniversum angewendet werden; es kann nicht auf nicht gemessene Länder, Sprachen, Produkte oder Zeiträume verallgemeinert werden.

N-13

Messung, Bewertung oder öffentliche Aussage müssen einen verantwortlichen menschlichen oder institutionellen Eigentümer haben.

N-14

UNBEKANNT, NICHT GETESTET oder UNGÜLTIG dürfen nicht stillschweigend in ein positives Ergebnis umgewandelt werden.

17. KANDIDAT GEO-1000 HAUPTREGEL

CANDIDATE-001

> *Für jedes zu prüfende KI-Produkt und für jede Hauptmesswelle der Bevölkerung sollte standardmäßig eine Mindestanzahl von 1,000 gültigen Nutzerbeobachtungen angestrebt werden.*

Dieser Kandidatensatz:

- Genauigkeit der allgemeinen Produktprognose,
- Kosten,
- Anwendbarkeit,
- Fairness nach Ländern und Sprachen

Dies schafft ein anfängliches Gleichgewicht zwischen diesen Überlegungen. Es garantiert jedoch nicht eine ausreichende Abdeckung von:

- allen Länderwerten,
- allen Sprachwerten,
- allen Nutzeruntergruppen

Stattdessen muss die Angemessenheit durch Folgendes getestet werden:

- Pilotversuche,
- unabhängige Wiederholungen,
- Design-Effekt,

- Subgruppensensitivität

Diese Bewertungen bestimmen den endgültigen normativen Status.

18. FORMEN DES VERSAGENS

F-01— Ausgewählter Screenshot

Nur die positivste Antwort wird unter zahlreichen Ergebnissen veröffentlicht.

F-02— Behandlung eines einzelnen Nutzers als die Population

Die Erfahrung einer Person wird auf das gesamte Zieluniversum verallgemeinert.

F-03— Wiederholungen desselben Nutzers als neue Personen zählen

Wiederholungen, die von derselben Person oder demselben Konto erzeugt werden, werden als unabhängige Teilnehmer gemeldet.

F-04— Die Antwort erneuern und die beste auswählen

Anfängliche oder negative Antworten werden verworfen; die geeignetste Antwort wird als Messergebnis gespeichert.

F-05— Prompt Drift

Teilnehmer ändern das Wort, den Ton oder den Umfang des Prompts; die Ergebnisse werden zusammengeführt, als stammten sie aus demselben Experiment.

F-06— Kontextkontamination

Frühere Gespräche beeinflussen die Antwort; die Aufzeichnung wird als saubere Sitzung präsentiert.

F-07— Verbergen der Personalisierung

Speicher, spezielle Anweisungen oder Nutzerhistorie sind sichtbar; das Ergebnis wird als allgemeine Nutzerantwort berichtet.

F-08— Verwechslung von Modell und Produkt

Nur der allgemeine Modell- oder Anbietername wird angegeben; das tatsächliche Produkt, die Schnittstelle und der Werkzeugzustand werden nicht aufgezeichnet.

F-09— Entfernen von Fehlern und Irrtümern aus dem Nenner

Gelöscht, um die Ergebnisse positiv erscheinen zu lassen, wenn das System nicht auf Datensätze reagiert.

F-10— Ungültige Erfassung als Systemfehler zählen

Als Teilnehmerfehler, fehlende Nachweise oder falsches Ergebnis des KI-Produkts gemeldet.

F-11— Eine einzelne Sprache und ein Land globalisieren

Das Ergebnis eines einzelnen Gebiets wird global angewendet.

F-12— Verwendung synthetischer Daten als echte Daten

Zu Demonstrationszwecken erstellte Datensätze werden zur tatsächlichen Leistung hinzugefügt.

F-13— Ein kritisches Ereignis im Durchschnitt verlieren

Schwere Gesundheits-, Rechts-, Lizenz- oder Sicherheitsfehler werden innerhalb eines hohen Durchschnitts unsichtbar gemacht.

F-14— Wirkung einer Intervention aus einem einzelnen Erfolg ableiten

Die einzelne positive Antwort nach der GEO-Intervention wird als ursächliche Wirkung der Studie betrachtet.

F-15— Das gesamte System anhand eines einzigen Fehlers beurteilen

Ein Fehler wird so verallgemeinert, als ob das KI-Produkt für alle Nutzer immer falsch und falsch ist.

F-16— Die Zielgruppe danach definieren

Nachdem das Ergebnis gesehen wurde, wird die Nutzergruppe mit hohem Erfolg zur "tatsächlichen Zielpopulation" erklärt.

19. PRÜFUNGSVERFAHREN

Ein Prüfer sollte bei der Bewertung der Gültigkeit einer einzelnen Beobachtung oder eines frühen GEO-Anspruchs die untenstehende Reihenfolge einhalten.

Schritt 1– Erfassen Sie den öffentlichen Aussage

Die vollständige vorgeschlagene Strafe wird festgelegt. Beispiel: "ChatGPT kennt unser Unternehmen korrekt."

Schritt 2– Feststellen Sie das Ausmaß des Anspruchs

Der Satz kann eine Behauptung aufstellen über:

- Einzelereignis,
- Prävalenz,
- Stabilität,
- globale Konsequenz,
- Kausalität,
- kommerzieller Einfluss

Die anwendbare Ebene muss identifiziert werden.

Schritt 3— Zählen der unterstützenden Beobachtungen

Wie viele echte Nutzer? Wie viele Konten? Wie viele unabhängige Sitzungen? Wie viele Wiederholungen? Wie viele KI-Produkte? Wie viele Sprachen und Länder?

Schritt 4— Überprüfen der Beobachtungsbedingungen

Sind das KI-Produkt, die Nutzeroberfläche, die Prompt, die Sitzung, das Land, die Sprache und die Zeit alle aufgezeichnet?

Schritt 5— Überprüfen der Evidenzintegrität

Gibt es eine vollständige Antwort? Gibt es einen vollständigen Screenshot? Wurde die Antwort ausgewählt oder aktualisiert? Wurde die Datei geändert? Ist der Datensatz synthetisch?

Schritt 6— Bestimmen der Abhängigkeitsstruktur

Gibt es Wiederholungen desselben Nutzers? Wurde ein gemeinsames Konto verwendet? Gibt es ein gespiegeltes Modelldesign? Gibt es Clustering vom selben Panel?

Schritt 7– Trenne den Datenerfassungsfehler vom gültigen Systemergebnis

Ist die Ablehnung oder der Systemfehler das eigentliche Ergebnis? Oder hat der Teilnehmer gegen das Protokoll verstoßen?

Schritt 8– Vergleichen Sie die Behauptung mit dem Niveau der Nachweise

Wird eine einzelne Beobachtung für eine Population oder globale Schlussfolgerung verwendet?

Schritt 9– Öffnen Sie den kritischen Pfad

Wenn eine einzelne Reaktion einen vorhersehbaren schweren Schaden verursacht, wird eine separate Vorfalluntersuchung eingeleitet.

Schritt 10– Begrenze die Entscheidung

Die öffentliche Strafe wird nur auf das Niveau reduziert, das die Nachweise tragen.

20. NOTWENDIGE NACHWEISE

Für eine Beobachtung oder Behauptung im Rahmen dieses Abschnitts sollten die folgenden Aufzeichnungen, soweit sie relevant sind, eingeholt werden:

- Identität der geprüften Einheit
- Vollständig offengelegte GEO-Behauptung für die Öffentlichkeit
- Anzahl der Beobachtungen
- Anzahl der einzigartigen Teilnehmer
- KI-Produkt und Nutzeroberfläche
- Sichtbares Modell oder Version, falls verfügbar
- Plattform- oder Kontotyp
- Land und Gebietsschema
- Prompt-ID, Version und vollständiger Text
- Sitzungs- und Personalisierungsstatus
- UTC Zeitstempel
- Messwelle
- Vollständige Rohantwort
- Screenshot oder gleichwertiger Nachweis
- Reproduktions- oder Antwortauswahlstatus
- Gültigkeitsstatus der Beobachtung
- Abhängigkeits- und Übereinstimmungsaufzeichnung
- Synthetischer oder realer Status
- Ungültiges Beobachtungsprotokoll
- Aufzeichnung kritischer Ereignisse
- Ergebnisbereich
- Unsicherheitsangabe
- Verantwortliche Person oder Institution

Nicht alle diese Aufzeichnungen müssen in allen Details für jede Beobachtung mit geringem Risiko erforderlich sein. Wenn jedoch eine breite GEO-Punktzahl oder Konformitätsklausel der Öffentlichkeit zugänglich gemacht werden soll, dürfen wichtige Felder nicht leer bleiben.

21. PRÜFLISTE FÜR AUDITS

Wurde die geprüfte Einheit vollständig identifiziert? Wurden das KI-Produkt und die Nutzeroberfläche erfasst? Wurde der Prompt vor der Messung festgeschrieben? Wurde der Prompt vom Teilnehmer geändert? Sind die Sitzungs- und Personalisierungsbedingungen bekannt? Sind die Angaben zum Land und zur Region des Nutzers erfasst? Wird die vollständige Antwort aufbewahrt? Ist es die erste angemessene Ausgabe oder die beste ausgewählte Ausgabe? Gibt es einen vollständigen Evidenznachweis? Wurden die tatsächliche Anzahl der Nutzer und die Anzahl der Beobachtungen getrennt? Wurden wiederholte und passende Beobachtungen unabhängig gezählt? Wurden Verweigerungen und Systemfehler im Nenner ordnungsgemäß berücksichtigt? Wurden ungültige Aufzeichnungen aufbewahrt und begründet? Wurden synthetische und echte Daten getrennt? Wird eine einzelne Antwort als Populationsresultat präsentiert?

Gilt die Aussage nur für das gemessene Produkt, Land, Sprache und Zeit? Wurde die Wahrscheinlichkeit eines kritischen Fehlers separat bewertet? Trägt das Ergebnis eine Aussage über Vertrauen oder Unsicherheit? Werden ungetestete Bereiche eindeutig dargestellt? Übersteigt die öffentliche Aussage das Niveau der tatsächlichen Beobachtung? Wenn eine der sachlichen Fragen unbeantwortet ist, weist das Ergebnis nur ein eingeschränktes Vertrauen auf. Wenn mehr als eine unbeantwortet ist, sollte es nicht für die Gesamt-GEO-Bewertung verwendet werden.

22. EINWÄNDE UND ANTWORTEN

Einwand 1— „Der echte Nutzer sieht bereits nur eine Antwort.“

Das ist korrekt. Die einzelne Antwort, die der einzelne Nutzer sieht, ist die reale Erfahrung dieser Person. Daher werden einzelne Ereignisse aufgezeichnet. Die individuelle Erfahrung ist jedoch nicht dasselbe wie die Erfahrung der gesamten Bevölkerung. NOMOS hält beide Fakten gemeinsam fest:

Die Erfahrung eines einzelnen Nutzers ist real. / Ein einzelner Nutzer ist nicht die gesamte Bevölkerung.

Einwand 2— „Wenn das System jedes Mal die gleiche Antwort gibt, reicht dann nicht eine Antwort aus?“

Die Behauptung, dass das System jedes Mal die gleiche Antwort gibt, erfordert ebenfalls eine Messung. Zusätzlich:

- Zeit,
- Abruf,
- Land,
- Sprache,
- Plan,
- Personalisierung,
- Nutzeroberfläche

können variieren. Selbst ein als deterministisch angenommenes System kann nicht über definierte Bedingungen hinaus verallgemeinert werden.

Einwand 3— „Warum genau 1,000?“

Die Zahl 1,000 ist keine magische Schwelle der Sicherheit. GEO-1000 dient als grundlegender Referenzpunkt, der ausgleicht:

- ein erhebliches Nutzeraufkommen,

- eine breit angelegte Produktschätzung,
- durchführbare Kosten,
- öffentliche Verständlichkeit.

einzeln Gruppen könnten mehr Beobachtungen benötigen. Pilotstudien müssen die Schwelle überprüfen oder anpassen.

Einwand 4— „Lasst dieselben 1,000 Personen alle KI-Produkte testen."

Diese Methode kann nützlich für gepaarte Vergleiche zwischen Modellen sein. Allerdings kann die Auswahl nur intensiver KI-Nutzer, die Zugang zu allen Produkten haben, die Zielpopulation verzerren. Außerdem sind Beobachtungen derselben Person abhängig. Vollständig gepaarte, teilweise gepaarte und ausgewogene Designs werden später verglichen.

Einwand 5— „Ist ein Screenshot nicht bereits ausreichender Beweis?"

Screenshots sind wichtig. Allein zeigen sie jedoch nicht immer Folgendes:

- Echtzeit
- Land
- Plan
- vorheriges Gespräch
- Personalisierung
- Antwortaktualisierung
- Dateiänderung
- vollständige Antwort

Aus diesem Grund ist ein Screenshot ein wichtiger Bestandteil des Evidenzpakets. Er ist nicht das Ganze.

Einwand 6— „Warum schließen wir ein nicht reagierendes System in die Bewertung ein?"

Die tatsächliche Erfahrung des Nutzers kann sein, keine Antwort zu erhalten. Eine Ablehnung oder ein Fehler:

- technisch
- Sicherheit,
- Politik,
- Zugang

kann das Ergebnis sein. Wenn das Protokoll korrekt angewendet wurde, sollte das Ereignis aufgezeichnet werden. Die Nichtreaktion des Systems aufgrund eines technischen Fehlers des Teilnehmers wird separat klassifiziert.

Einwand 7— „Warum sollte ein einziger schwerwiegender Fehler die gesamte Inspektion beeinflussen?"

Ein einzelner Fehler beweist nicht, dass die Mehrheit der Bevölkerung denselben Fehler gemacht hat. Allerdings kann der potenzielle Schaden eines einzelnen Fehlers schwerwiegend sein. Daher:

- Punktfrequenz,
- Kritische Entscheidungsgefährdung und Integrität

werden separat bewertet. Der Durchschnitt kann menschlichen Schaden nicht automatisch auslöschen.

Einwand 8— „Wenn wir mit einer einzigen Antwort nichts sagen können, welchen Wert haben dann frühe Tests?“

Einzel- und Kleinproben-Tests:

- Hypothesen generieren,
- Arten von Fehlern entdecken,
- Testaufforderungen,
- das Aufnahmesystem testen,
- Schiedsrichter kalibrieren,

kann kritische Ereignisse aufdecken. Falsch ist, kein exploratives Testen durchzuführen. Falsch ist, das Ergebnis des explorativen Testens als Populationswert zu präsentieren.

GEMEINSAMES PRINZIP VON KAPITEL 24

Dieses Kapitel wurde geschrieben, um ein einzelnes Missverständnis zu beseitigen: „Ich habe eine KI gefragt, eine Antwort erhalten; hier ist mein GEO-Ergebnis.“ Nein. Sie haben noch kein allgemeines GEO-Ergebnis. Was Sie haben, ist eine Beobachtung. Diese Beobachtung:

- ist real,
- ist bedeutend,
- sogar kritisch

kann sein. Es ist jedoch immer noch eine einzelne Beobachtung. Für einen GEO-Score ist das Vorhandensein von:

- Nutzern,
- KI-Produkten,
- Ländern,
- Sprachen,
- Sitzungen,
- Zeiten

Das, was gemessen werden muss, ist die Verteilung der Repräsentation über KI-Produkte, Länder, Sprachen, Sitzungen und Zeit. Ein einzelner Screenshot ist keine Verteilung. Zehn automatisierte Wiederholungen sind nicht zehn unabhängige Personen. Ebenso repräsentieren tausend Antworten aus einer schlechten Stichprobe nicht die Weltbevölkerung. Selbst eine bevölkerungsproportionale Stichprobe kann ohne sprachliche Fairness unvollständig sein. Ein hoher Durchschnitt kann kritische Fehler nicht auslöschen. Das erste Messgesetz von NOMOS lautet daher:

> *Eine einzelne Antwort ist ein Ereignis. / Mehrere Antworten sind ein Muster. / Eine definierte Stichprobe ist eine Schätzung der Population. / Wiederholung über die Zeit ist ein Maß für Stabilität. / Aber keines davon allein ist Kausalität oder nachhaltiger Wert.*

Bestellung von Abschnitt 1 von NOMOS

> *Bringen Sie mir nicht einen Screenshot und sprechen dann für das gesamte System.*
>
> *Machen Sie nicht die Antwort eines Nutzers zur Antwort der gesamten Weltbevölkerung.*
>
> *Verwandeln Sie keine Duplikate desselben Kontos in unabhängige Menschen.*

Aktualisieren Sie die Antwort nicht und zeigen Sie das Bevorzugte, als ob es die echte Verteilung wäre.

Entfernen Sie keine Ablehnungen, Fehler und Unbekannten aus dem Nenner.

Verallgemeinern Sie nicht eine Sprache auf die ganze Welt, ein Land auf alle Länder oder ein KI-Produkt auf alle generativen Systeme.

Machen Sie keinen einzelnen Erfolg zum Nachweis eines Interventions-effects oder einen einzelnen Fehler zum unveränderlichen Charakter des gesamten Systems.

Verdrängen Sie keinen einzelnen schweren Fehler im Durchschnitt.

Zuerst zeichne das Ereignis auf. / Dann messe das Muster. / Dann definiere die Zielpopulation. / Dann zeige Abhängigkeit und Unsicherheit. / Und sprich nur im Umfang der vorliegenden Nachweise.

Der abschließende Satz des Kapitels

Die Repräsentation einer Entität in Künstlichen-Intelligenz-Systemen ist keine einzelne Antwort. KI-Produkte sind eine Verteilung, die über Nutzer, Länder, Sprachen, Sitzungen und Zeiten gemessen werden sollte.

Normativer Kern

Eine einzelne Ausgabe eines generativen Systems KANN feststellen, dass eine bestimmte Antwort unter aufgezeichneten Bedingungen aufgetreten ist. Ein einzelnes Ergebnis DARF NICHT verwendet werden, um Folgendes abzuleiten: - Bevölkerungsprävalenz, - typisches Produktverhalten, - zeitliche Stabilität, - sprachübergreifende Konsistenz, - länderübergreifende Konsistenz, - produktübergreifende Konsistenz, - Interventionseffekt, - Sicherheit der Empfehlung, - kommerzielle Auswirkungen, - Konformität, - oder einen GEO-Score. Alle weitergehenden Behauptungen MÜSSEN durch eine vordefinierte Zielpopulation, ein gültiges Mehrfachbeobachtungsdesign, aufgezeichnete Abhängigkeitsstrukturen, ein Evidenzbündel, Bewertungsregel, Umfangserklärung und eine Offenlegung der Unsicherheit unterstützt werden. eine einzelne Ausgabe mit hohem Schadenspotenzial KANN eine separate Überprüfung kritischer Vorfälle auslösen, ohne als Nachweis für die Bevölkerungsprävalenz behandelt zu werden.

KAPITEL 02

Repräsentation ist eine Verteilung

METHODISCHE GRUNDLAGEN / K04; K05; K09

Kapitelgrenze

Abschnitt 1 hat die folgende Bestimmung eingeführt:

> *Eine einzelne Antwort ist nicht die GEO-Punktzahl. Eine einzelne Antwort ist eine einzelne Beobachtung, die unter aufgezeichneten Bedingungen aufgetreten ist.*

Dieser Abschnitt beantwortet die nächste Frage:

> *Wenn eine einzelne Antwort nicht die GEO-Punktzahl ist, was ist dann das eigentliche Objekt, das wir messen wollen?*

Die Antwort ist:

> *Die Repräsentation eines Objekts in generativen Systemen ist eine multidimensionale Verteilung, die sich über KI-Produkte, Nutzer, Länder, Sprachen, Sitzungsbedingungen, Prompts und Zeiten bildet.*

Dieses Kapitel:

- die Repräsentationsverteilung,
- die grundlegenden Dimensionen der Verteilung,
- der Unterschied zwischen dem Mittelwert und der Verteilung,
- die Ungleichheit in Untergruppen,
- die Stabilität,
- die Abdeckung und das Problem fehlender Beobachtungen,
- das Risiko seltener schwerwiegender Fehler,
- die konzeptionelle Struktur der öffentlich verfügbaren Ergebnisübersicht

definiert. Dieses Kapitel behandelt noch nicht:

- wie viele Teilnehmer jedem Land zugewiesen werden.
- wie der Stichprobenrahmen erstellt wird,
- wie die Aufforderungsübersetzungen validiert werden,
- wie atomare Aussagen bewertet werden,
- die exakte Formel des endgültigen NOMOS-Scores,
- Konformitätsschwellen

Dieses Kapitel regelt diese Angelegenheiten nicht normativ; spätere Kapitel behandeln sie separat. Bevor die Bewertungsformel definiert wird, definiert Kapitel 2 die Realität, die der Score darstellen soll.

NOMOS Herausforderung

Betrachten Sie zwei KI-Produkte. Jedes erhält denselben globalen Durchschnittsscore:

90 / 100

Die Sprachresultate des ersten Produkts sind wie folgt:

- Englisch: 90
- Türkisch: 91
- Deutsch: 89
- Japanisch: 90
- Arabisch: 90

Die Sprachresultate des zweiten Produkts sind wie folgt:

- Englisch: 99
- Türkisch: 98
- Deutsch: 97
- Japanisch: 86
- Arabisch: 70

Der einfache Durchschnitt beider Produkte beträgt 90. Haben sie die gleiche GEO-Qualität? Nein. Das erste Produkt erzeugt eine ähnliche Repräsentation in verschiedenen Sprachen. Das zweite Produkt erzeugt eine hervorragende Repräsentation in einigen Sprachen, aber eine ernsthaft schwache Repräsentation in einer anderen Sprache. Ein einfacher Durchschnitt verschleiert diesen Unterschied. Betrachten wir nun zwei andere Produkte. 95% der Antworten beider erfüllt die Bestehensanforderungen. Der verbleibende 5% im ersten Produkt besteht aus:

- geringen Umfangsmängeln,
- begrenzten Erklärungsmängeln,
- einige UNBEKANNTE

Ergebnissen. Der verbleibende 5% im zweiten Produkt besteht aus:

- falscher Gesundheitsbehörde,
- nicht vorhandener Lizenz,
- falsche rechtliche Identität,
- unangemessener Beratung

existiert. Die rohe Genauigkeitsrate für beide beträgt erneut 95 Prozent. Haben sie dasselbe Repräsentationsrisiko? Nein. Betrachten Sie nun ein drittes Paar. Der Durchschnitt von beiden über die drei Messwellen beträgt 90. Das erste:

- Welle 1: 90
- Welle 2: 91
- Welle 3: 89

Der zweite:

- Welle 1: 98
- Welle 2: 90
- Welle 3: 82

Haben sie über die Zeit hinweg die gleiche Konsistenz? Nein. In diesem Fall geht es nicht nur um: 'Was ist der Durchschnitt?' Wir müssen auch Folgendes wissen: Wie sind die Ergebnisse verteilt? Wer ist gut vertreten, wer ist schlecht vertreten? Welche Sprachen und Länder fallen zurück? Wie

schwerwiegend sind die Fehler? Wie stark ändert sich das Ergebnis über die Zeit? Welche Bereiche wurden nie gemessen? Mit welchen Gewichtungen wurde der Durchschnitt erstellt? Wie viele verschiedene Realitäten stehen hinter demselben Durchschnitt? Die erste Schlussfolgerung dieses Abschnitts lautet:

> *Der gleiche Durchschnitt ist nicht die gleiche Repräsentation.*

Ihre zweite Bestimmung lautet:

> *Ein GEO-Score kann eine Zusammenfassung der Verteilung sein; er kann die Verteilung nicht ersetzen.*

1. ZWECK DES KAPITELS

Dieser Abschnitt definiert, wie die Repräsentation von Entitäten in generativen Systemen verstanden werden sollte, bevor sie auf eine einzelne Zahl reduziert wird. Seine Hauptziele sind:

- Zu zeigen, dass Repräsentation keine zeitlose und singuläre „KI-Perspektive“ ist
- Die Bedingungen zu definieren, von denen das Repräsentationsergebnis abhängt
- Zu erklären, dass das binäre Wahr-Falsch allein nicht ausreicht
- Durchschnittliche Genauigkeit von Untergruppenungleichheit zu trennen
- Das hohe Fehlerrisiko, das in hohen Durchschnittswerten verborgen ist, sichtbar zu machen
- Zu zeigen, dass Stabilität und Genauigkeit nicht dasselbe Konzept sind
- Zu verhindern, dass ungemessene Bereiche innerhalb der Verteilung mit positiven Annahmen gefüllt werden
- Die Struktur eines zukünftigen NOMOS-Scores vorzubereiten, der singulär ist, aber nicht dazu verdammt, singulär zu sein

Am Ende dieses Abschnitts sollte der Leser in der Lage sein, die folgende Frage zu beantworten:

> *Warum ist das Ergebnis des GEO einer Entität nicht nur auf die durchschnittliche Genauigkeitsrate beschränkt?*

2. ZENTRALE NORMATIVE BESTIMMUNG

> *Die Repräsentation einer Entität in generativen Systemen ist die bedingte Verteilung verschiedener Antworten und Repräsentationsergebnisse, die unter definierten Bedingungen auftreten können.*

Diese Verteilung:

- sie ist nicht der abstrakte 'Gedanke' eines einzelnen Modells,
- sie ist nicht die gemeinsame Ansicht aller künstlichen Intelligenzsysteme,
- sie ist nicht nur die Wiederholungsrate der Website,
- sie ist nicht nur die Häufigkeit, mit der der Name der Marke erwähnt wird,
- sie ist nicht nur die durchschnittliche Genauigkeit,

sie ist nicht nur die Empfehlungsrate. Die Verteilungsdarstellung zeigt innerhalb eines bestimmten Protokolls Folgendes:

> *Welche Nutzer welche Art von Repräsentation auf welchem KI-Produkt und auf welcher Nutzeroberfläche, in welcher Sprache, in welchem Land, mit welchem Prompt, unter welchen Sitzungsbedingungen und zu welcher Zeit sehen.*

Daher hängt das Ergebnis einer Verteilung immer von den folgenden Fragen ab: Welche Einheit? Welches KI-Produkt? Welche Nutzeroberfläche? Welche Zielpopulation? Welches Land? Welche Sprache und Regionaleinstellung? Welcher Prompt? Welche Sitzungsbedingung? Welche Messwelle? Welches Wahrheits-Paket? Welche Bewertungsregel? Wenn sich eines dieser Felder ändert, kann sich auch die Verteilung ändern.

3. WAS BEDEUTET DAS WORT "DISTRIBUTION"?

In diesem Buch bezieht sich das Wort Verteilung nicht auf eine verborgene oder metaphysische Sichtweise, die angenommen wird, innerhalb eines künstlichen Intelligenzsystems zu existieren. Es behauptet nicht: NOMOS: "Dies ist die echte Verteilung des Unternehmens im Geist des Modells." Die Verteilung hier ist operativ und beobachtbar.

> *Es ist das Muster verschiedener Repräsentationsergebnisse, die unter einem definierten Protokoll innerhalb der Zielpopulation und der Messbedingungen erzielt werden können.*

Man sollte zwischen zwei Verteilungen unterscheiden.

3.1. Zielrepräsentationsverteilung

Es ist die Verteilung, die die Messung zu prognostizieren versucht, aber nicht vollständig direkt beobachtet werden kann. Sie wird repräsentiert als:

$P_E(Y \mid A, C, L, S, P, T, U)$

Hier:

- E: geprüfte Einheit
- Y: Repräsentationsergebnis
- A: KI-Produkt und Nutzeroberfläche
- C: Land oder geografische Bedingung
- L: Sprache und Gebietsschema
- S: Sitzungs- und Personalisierungsbedingung
- P: Prompt
- T: Zeit oder Messwelle
- U: Zielnutzeruniversum

Dieser Ausdruck fragt: „Im definierten Zielnutzeruniversum, unter den angegebenen Bedingungen, welche Repräsentationsergebnisse treten mit welchen Wahrscheinlichkeiten auf?“ Nicht alle Elemente der Zielverteilung können gemessen werden. Sie werden durch Stichproben geschätzt.

3.2. Beobachtete empirische Verteilung

Es ist die Verteilung der gültigen Beobachtungen, die in einer realen oder synthetischen Messung gewonnen wurden. Ihre Repräsentation:

$\hat{P}_E$

kann verwendet werden. Die empirische Verteilung ist nicht die Zielverteilung selbst. Sie ist eine Schätzung, die durch ihre Stichprobe, Gewichtung und das Messdesign erstellt wird. Die Größenordnung des Unterschieds kann abhängen von:

- Stichprobendesign
- Teilnehmerauswahl

- Abdeckung von Ländern und Sprachen
- Ungültige Beobachtungsrate
- Nichtbeantwortung
- Fehler des Gutachters
- Messwelle
- Auswahl der Prompt
- Produkt- und Schnittstellenbedingungen
- Gewichtung
- Clusterbildung und Abhängigkeit

Daher kann eine empirische Verteilung nicht als "die absolute und unveränderliche Wahrheit der künstlichen Intelligenz" veröffentlicht werden. Der korrekte Ausdruck sollte sein: "Die für die spezifizierte Zielpopulation und Messbedingungen erhaltene Schätzung."

4. REPRÄSENTATIONSRAUM

Die Repräsentation einer Entität erfolgt nicht in einer einzelnen Dimension. Jede Beobachtung existiert in einem multidimensionalen Zustandsraum. Die grundlegende Repräsentation:

```
X_i= (E_i, A_i, U_i, C_i, L_i, S_i, P_i, T_i)
```

Hier:

- Ei: Entität
- Ai: KI-Produkt und Nutzeroberfläche
- Ui: Nutzer oder Stichprobeneinheit
- Ci: Land und geografische Ebene
- Li: Sprache und Gebietsschema
- Si: Sitzung, Plan und Personalisierungsbedingung
- Pi: Prompt oder Prompt-Familie
- Ti: Zeit und Messwelle

Generierte Antwort:

```
R_i
```

kann gezeigt werden. Der repräsentative Zustand, der sich aus der Bewertung der Antwort ergibt, ist:

```
Y_i= g(R_i, V_E, J)
```

und kann betrachtet werden als. Hier:

- VE: verifiziertes Truth Pack für die Entität verwendet
- J: Bewertungs- und Beurteilungsregeln
- g: Bewertungsprozess, der die Antwort in repräsentative Ergebnisse umwandelt

Die Bedeutung dieser Repräsentation ist wie folgt: Dieselbe Antwort kann nicht eigenständig bewertet werden, ohne anzugeben, welche Entität, welches Truth Pack und welche Bewertungsregel zugrunde gelegt wurden.

5. REPRÄSENTATIVES ERGEBNIS IST KEIN EINZELNER WERT

Nur die Gesamtheit einer Antwort:

- in die Kategorien gezwungen werden von

  wahr,
- ist falsch

Die Vergabe eines Labels ist oft unzureichend. Folgende Situationen können in derselben Antwort gemeinsam auftreten:

- Korrekte Identität
- Korrekte Hauptkategorie
- Unvollständige Serviceabdeckung
- Veraltete Anzahl von Mitarbeitern
- Nicht unterstützte Führungsbehauptung
- Korrektes Zitat
- Unangemessene Empfehlung
- Nicht-kritischer Grenzwertmangel

Daher kann das Repräsentationsergebnis als Ergebnisvektor betrachtet werden.
Kandidatenrepräsentation:

```
Y_i= (y_i^id, y_i^cat, y_i^fact, y_i^scope, y_i^time, y_i^evidence, y_i^recommendation, y_i^harm)
```

Hier:

- yiid: Identitätsgenauigkeit
- yicat: Genauigkeit von Kategorie und Rolle
- yifact: Genauigkeit von sachlichen Angelegenheiten
- yiscope: Erhaltung von Umfang und Grenzen
- yitime: zeitliche Gültigkeit
- yievidence: Übereinstimmung von Aussage und Nachweis
- yirecommendation: Empfehlung und Nutzergeeignetheit
- yiharm: Schadensklasse von Fehler oder Fehlinformation

Die genauen Bewertungsregeln für diese Felder werden in späteren Abschnitten definiert. Das Leitprinzip hier lautet:

> *Eine repräsentative Antwort sollte nicht nur danach bewertet werden, ob sie korrekte Informationen enthält, sondern auch, über wen, was, innerhalb welcher Grenzen, mit welchen Nachweisen und für wen sie angemessen vermittelt wird.*

6. GRUNDLEGENDE REPRÄSENTATIONSZUSTÄNDE

Der Status jeder atomaren Behauptung und der endgültigen Antwort muss nicht derselbe sein. Dieser Abschnitt trennt konzeptionell die folgenden Grundzustände.

6.1. UNTERSTÜTZT

Die Behauptung wird durch ausreichende Nachweise innerhalb des definierten Umfangs und Zeitrahmens gestützt.

6.2. WIDERSPRUCHLICH

Die Behauptung steht im Widerspruch zu verlässlichen und relevanten Aufzeichnungen.

6.3. NICHT UNTERSTÜTZT

Es gibt nicht genügend Unterstützung für die Behauptung. Es wurde nicht nachgewiesen, dass die Behauptung notwendigerweise falsch ist. Es gibt jedoch nicht genügend Aufzeichnungen, um sie als wahr anzunehmen.

6.4. VERALTET — NICHT MEHR AKTUELL

Die Aussage könnte in der Vergangenheit wahr gewesen sein. Zum Messzeitpunkt ist sie nicht mehr gültig oder ihre aktuelle Genauigkeit kann nicht überprüft werden.

6.5. TEILS UNTERSTÜTZT — TEILS UNTERSTÜTZT

Nur ein bestimmter Umfang, eine bestimmte Population, ein bestimmtes Land, Produkt oder Zeitraum der Behauptung wird unterstützt. Die Antwort könnte das Unterstützungslimit überschritten haben.

6.6. NICHT IM UMFANG — NICHT IM UMFANG

Die Behauptung bezieht sich nicht auf die geprüfte Einheit oder das Truth Pack. Sie könnte von einer anderen Einheit, einem anderen Produkt, Land oder Service stammen.

6.7. EINHEITSVERSCHMELZUNG — EINHEITSZUSAMMENFÜHRUNG

System:

- drei verschiedene Unternehmen,
- die Marke und die juristische Einheit,
- eine Person und die Institution,
- ein alte und neue Einheit,
- hat verschiedene Organisationen mit demselben Namen verwechselt

miteinander. Dieser Fehler ist nicht nur ein sachlicher Irrtum. Es ist die Verfälschung des Repräsentationsobjekts.

6.8. WESENTLICHE AUSLASSUNG — FEHLENDE INFORMATION

Die Sätze in der Antwort können korrekt sein. Wesentliche Informationen, die die Entscheidung des Nutzers ändern würden, wurden jedoch weggelassen. Beispiel: Der Service wird nur in bestimmten Ländern angeboten. Die Lizenz ist auf eine bestimmte Gerichtsbarkeit beschränkt. Der Preis versteht sich ohne Steuern. Das Ergebnis basiert auf Daten des Erstanbieters. Das Produkt ist für bestimmte Nutzer nicht geeignet. Das Weglassen kann zu einer falschen Schlussfolgerung führen, selbst wenn kein direkt falscher Satz vorhanden ist.

6.9. ABLEHNUNG — REJEKTIERUNG

Das System hat die Antwort verweigert. Ablehnung:

- angemessen,

- unnötig,
- inkonsistent,
- sicherheitsbedingt,
- technisch

kann auftreten. Ret wird nicht automatisch als fehlgeschlagen oder erfolgreich betrachtet. Es wird separat klassifiziert.

6.10. KEIN ERGEBNIS

Das System hat aus technischen oder betrieblichen Gründen keine nutzbare Antwort erzeugt. Es ist nicht dasselbe wie Ret.

6.11. UNBEKANNT

Die verfügbaren Aufzeichnungen reichen nicht für eine zuverlässige Klassifizierung aus. UNBEKANNT kann nicht in ein positives Ergebnis umgewandelt werden.

7. HAUPTMERKMALE DER VERTEILUNG

Eine Verteilungsdarstellung kann nicht nur durch den Mittelwert erklärt werden. Mindestens sollten die folgenden Merkmale bewertet werden.

7.1. Zentrale Tendenz

Fasst die allgemeine Genauigkeit oder das Übergangsniveau der Antworten zusammen. Beispielhafte Messgrößen:

- Mittelwert
- Gewichteter Mittelwert
- Median
- Bestehensquote

Die zentrale Tendenz ist wichtig. Sie zeigt jedoch nur das Zentrum der Verteilung.

7.2. Streuung und Variabilität

Gibt an, wie stark die Ergebnisse um das Zentrum verteilt sind. Von zwei Systemen mit demselben Mittelwert kann das eine unter den Nutzern konsistent sein, während das andere sehr variabel ist. Variabilität:

- Nutzern,
- Ländern,
- Sprachen,
- Sitzungen,
- Messwellen

können auch unter ihnen untersucht werden.

7.3. Form der Verteilung

Sind die Ergebnisse um einen einzigen zentralen Punkt gruppiert? Oder sind sie in zwei oder mehr Cluster unterteilt? Zum Beispiel:

- Wenn die Hälfte der Nutzer sehr genaue
- und die andere Hälfte sehr ungenaue

Ergebnisse sieht, kann der Durchschnitt mittel sein. Die typische Nutzererfahrung ist jedoch nicht „mittlere Genauigkeit“. Es sind zwei getrennte polarisierte Erfahrungen.

7.4. Unterschiede zwischen Untergruppen

Sehen bestimmte Sprach-, Länder- oder Nutzergruppen systematisch unterschiedliche Ergebnisse? Der globale Durchschnitt kann diesen Unterschied verschleiern.

7.5. Tail-Risiko

Gibt es seltene, aber schwere Fehler? Seltene Ereignisse können einen geringen Effekt auf den Durchschnitt haben. Ihr Einfluss auf den menschlichen Schaden kann erheblich sein.

7.6. Abdeckung

Wie viel des Zielbenutzeruniversums wurde tatsächlich gemessen? Eine hohe Punktzahl kann zusammen mit geringer Abdeckung auftreten.

7.7. Fehlende und unbekannte Masse

Welcher Anteil der Antworten ist:

UNBEKANNT,

NICHT GETESTET,

Ablehnung, technische Nicht-Antwort oder ein ungültiger Datensatz?

7.8. Zeitliche Stabilität

Inwieweit bleibt die Verteilung über verschiedene Messwellen hinweg erhalten?

8. KANDIDAT NOMOS REPRÄSENTATIVE VERTEILUNGS-SIGNATUR

KANDIDATENKONZEPT

Es wird empfohlen, dass das Ergebnis eines GEO zuerst durch die folgende Verteilungssignatur und nicht durch eine einzelne Zahl definiert wird:

```
RDS(E) = (μ, κ, σ, Δ, τ, ρ, u)
```

Diese Repräsentation wird Repräsentationsverteilungssignatur genannt, das heißt:

Repräsentationsverteilungssignatur

Komponenten:

μ — Allgemeines Repräsentationsniveau

Gesamtgenauigkeit oder Übergangsvorhersage unter definierten Gewichten und Umfang.

κ — Abdeckung

Das Ausmaß, in dem der Zielbenutzer, das Land, die Sprache, das KI-Produkt und das Zeituniversum ausreichend gemessen werden.

σ — Variabilität

Streuung unter Nutzern und experimentellen Zellen.

Δ — Ungleichheit in Untergruppen

Der materielle Unterschied zwischen Sprache, Land, Produkt oder anderen signifikanten Gruppen.

τ — Risiko für Ausreißer und kritische Fehler

Die Rate und Art seltener schwerwiegender Fehler.

ρ — Zeitliche Stabilität

Die Widerstandsfähigkeit der Verteilung über Wellen.

u — Unbekannte und ungelöste Masse

UNBEKANNT ist der Anteil der Verteilung, der ungelöst oder unzureichend bewertet bleibt. Diese Signatur ist noch keine endgültige Bewertungsformel. Die folgenden Elemente erfordern eine weitere Entwicklung:

- definierte Definition,
- Schätzung,
- Gewichtung,
- öffentliche Präsentation

Sie werden durch Pilotstudien und spätere Kapitel finalisiert. Das Gründungsprinzip lautet:

> *Das GEO-Ergebnis sollte nicht nur die Frage beantworten 'wie genau?', sondern auch 'für wen, wie konsistent, wie gerecht, wie umfassend und wie riskant?'*

9. DURCHSCHNITTLICHES WASCHEN

Wenn eine Institution nur einen hohen globalen Durchschnitt zeigt und die materiellen Mängel der Verteilung verschweigt, wird dies genannt:

Durchschnittliches Waschen

Durchschnittliches Waschen kann auf folgende Weise erfolgen:

- Die schwächste Sprache verbergen
- Kleine Länder vom Bericht ausschließen
- Kritische Fehler in die allgemeine Fehlerquote einmischen
- Entfernen der UNBEKANNT-Ergebnisse aus dem Nenner
- Nur gut performende KI-Produkte einbeziehen
- Die schwache Messwelle ausschließen
- Die Bevölkerung außerhalb des Geltungsbereichs unsichtbar machen
- Gewichte nach Sichtung der Ergebnisse ändern

Der Durchschnitt kann mathematisch korrekt berechnet werden. Dennoch kann die Repräsentation, die er der Öffentlichkeit bietet, irreführend sein. Daher:

> *Eine mathematisch korrekte Berechnung eines Durchschnitts bedeutet nicht, dass sie für Entscheidungen ausreichend ist.*

10. VERSCHIEDENE REALITÄTEN DES GLEICHEN DURCHSCHNITTS

Der Durchschnitt könnte in allen folgenden synthetischen Verteilungen 90 betragen.

10.1. Homogene Verteilung

Alle großen Untergruppen liegen im Bereich von 88–92. Diese Verteilung ist relativ ausgewogen.

10.2. Polarisierte Verteilung

Die Hälfte der Nutzer sieht Ergebnisse nahe bei 100, während die andere Hälfte Ergebnisse nahe bei 80 sieht. Der Durchschnitt beträgt 90. Das Nutzererlebnis ist jedoch nicht gleich.

10.3. Linguistisch fragmentierte Verteilung

Die Ergebnisse sind für Sprachen mit hoher Bevölkerungszahl oder hohen Ressourcen sehr gut, aber für Sprachen mit geringen Ressourcen deutlich schwach. Der globale Durchschnitt mag gut aussehen. Die Sprachgerechtigkeit ist schwach.

10.4. Länderbasierte fragmentierte Verteilung

Selbst innerhalb derselben Sprache können die Ergebnisse aufgrund unterschiedlicher Länder, Regionen, Produktzugänge oder Nutzergruppen variieren.

10.5. Zeitlich flüchtige Verteilung

Sehr hohe Ergebnisse treten in einer Welle auf, sehr niedrige Ergebnisse in einer anderen Welle. Selbst wenn der Durchschnitt hoch ist, ist es schwierig, zuverlässige Nutzererwartungen zu bilden.

10.6. Verteilung mit Tail-Risiko

Die meisten Antworten sind korrekt. Ein kleiner Teil enthält jedoch ernsthafte gesundheitliche, rechtliche, lizenzrechtliche oder sicherheitsrelevante Fehler.

10.7. Hohe Verteilung unbekannter Masse

Die meisten bewertbaren Antworten sind korrekt. Ein erheblicher Teil der Zielpopulation oder der Antworten ist jedoch:

- testet,
- ungestestet,

UNBEKANNT

oder nicht schlüssig. Die Veröffentlichung einer hohen Rate basierend nur auf den bewertbaren Antworten kann den Umfang verschleiern.

11. UNGLEICHHEIT IN UNTERGRUPPEN

Ein globaler Wert repräsentiert möglicherweise nicht die Erfahrung in Untergruppen. Wichtige Untergruppen können beinhalten:

- KI-Produkt und Nutzeroberfläche
- Land
- Sprache
- Locale
- Plan
- Web oder mobil
- Kontrolliertes oder natürliches Panel

- Zeitwelle
- Nutzerintention
- Gerätefamilie

Nicht alle diese Gruppen haben möglicherweise dieselbe Bedeutung in jeder Prüfung. Wesentliche Gruppen sollten bestimmt werden, wenn der Umfang festgelegt ist.

11.1. Sprache ist nicht dasselbe wie Land

Türkischsprachig:

- Nutzer in der Türkei,
- Nutzer in Deutschland,
- Nutzer in einem anderen Land

Sie können dieselbe Sprache verwenden. Allerdings:

- Produktzugang,
- Preis,
- Ressourcen,
- lokale Benutzeroberfläche,
- rechtlicher Kontext,
- Personalisierung

können unterschiedlich sein. Daher können nicht alle Unterschiede ausschließlich durch Sprache oder Land erklärt werden.

11.2. Intersektionale Zellen

Einige materielle Unterschiede werden erst sichtbar, wenn mehrere Bereiche zusammen untersucht werden. Beispiel: Deutschland × Türkisch × Mobil × Kostenloser Tarif × Panel unter natürlichen Nutzungsbedingungen könnte an sich eine experimentelle Zelle sein. Wenn sich jedoch die Zellen vervielfachen, nimmt die Stichprobengröße ab. Daher kann für jede kleine Zelle kein endgültiger Wert veröffentlicht werden. Messdesign: sollte aussagekräftige Untergruppen erhalten und keine statistische falsche Genauigkeit erzeugen.

11.3. Untergruppenbasis

KANDIDATENKONZEPT

Zusätzlich zum globalen Durchschnitt wird empfohlen, das Ergebnis der schwächsten ausreichend beprobten Untergruppe oder der unteren 10%-Gruppe gesondert zu melden. Dies kann bezeichnet werden als:

Repräsentationsbasis

oder:

Sprache/Land Boden

Wenn jedoch die "niedrigste Punktzahl" von einer kleinen und instabilen Zelle stammt, kann dies zu falschen Ergebnissen führen. Daher sollte die Basismetrik zusammen mit Folgendem definiert werden:

- Mindestprobe
- Konfidenzintervall
- Zellqualität

Die genaue Regel wird in den folgenden Abschnitten festgelegt.

12. ES IST TEIL DER BEREICHSVERTEILUNG

Eine Verteilung besteht nicht nur aus beobachteten Antworten. Was nie beobachtet wurde, bestimmt ebenfalls die Bedeutung des Ergebnisses. Zum Beispiel wurden Englisch und Türkisch getestet; Japanisch und Arabisch nicht. Die Web-Oberfläche wurde getestet; die mobile Oberfläche und der kostenpflichtige Plan nicht. Starke Ergebnisse in Englisch und Türkisch rechtfertigen nicht die Aussage: ‚Die Repräsentation ist in allen Sprachen und Nutzeroberflächen stark.' Die vertretbare Schlussfolgerung lautet: ‚Messungen wurden auf den angegebenen englischen und türkischen Web-Oberflächen durchgeführt. Andere Sprachen und mobile Oberflächen wurden nicht getestet.'

12.1. Abdeckungskarte

Jede GEO-Verteilung sollte die folgende Karte in dem Umfang enthalten, der relevant ist:

- Zieluniversum
- Stichprobenrahmen
- Abgedeckte Länder
- Nicht abgedeckte Länder
- Abgedeckte Sprachen
- Nicht abgedeckte Sprachen
- Überwachte KI-Produkte
- Nicht überwachte Produkte
- Überwachte Nutzeroberflächen
- Nicht überwachte Oberflächen
- Messwellen
- Außerhalb des Umfangs liegende Prompt-Familien

12.2. Abdeckungsrate

KANDIDATENKONZEPT

Eine Abdeckungsmetrik kann entwickelt werden, um zu zeigen, welcher Anteil des Zieluniversums mit ausreichender Messung repräsentiert wird. Allerdings gilt für die Abdeckung:

- nur die Anzahl der Länder,
- nur die Anzahl der Sprachen,
- nur das Bevölkerungsverhältnis

kann nicht gemessen werden. Zum Beispiel könnten 150 Länder abgedeckt sein. Allerdings könnten 95% der Teilnehmer nur aus fünf Ländern stammen. Oder 90% der Bevölkerung könnten abgedeckt sein, aber kleine Sprachen könnten vollständig ausgeschlossen sein. Daher sollte die Bevölkerungsabdeckung getrennt von der Beobachterabdeckung nach Ländern/Sprachen gezeigt werden.

13. FEHLENDE BEOBACHTUNG UND NICHT-ANTWORT

Fehlende Daten sind nicht nur ein technisches Problem. Ihre Verteilung kann systematisch verzerrt sein. Die Gründe für das Fehlen einer Beobachtung können umfassen:

- Teilnehmer hat die Aufgabe nicht abgeschlossen
- Produkt ist im Land nicht zugänglich
- Rechenbeschränkungen
- Technischer Fehler
- Systemabweisung
- Ungültiger Screenshot
- Teilnehmer ändert seine Prompt
- Produkt funktioniert in einer bestimmten Sprache nicht
- Schiedsrichter, die keine verlässlichen Entscheidungen treffen können

Jeder dieser Gründe erfordert einen eigenen Status.

13.1. Fehlende Daten sind möglicherweise nicht zufällig

Zum Beispiel, wenn in Ländern mit schlechter Verbindungsqualität nicht mehr Aufgaben erledigt werden können, kann die Verwendung nur erfolgreicher Aufzeichnungen das globale Ergebnis zugunsten wohlhabender Nutzer mit starken Verbindungen verschieben. Wenn die Rate von UNBEKANNT steigt, weil Antworten in einer bestimmten Sprache schwer zu bewerten sind, kann die aus nur bekannten Antworten berechnete Punktzahl Sprachungerechtigkeiten verschleiern. Daher:

> *Die Rate der fehlenden Daten und der Verteilungsbericht des Grundes für fehlende Daten sollten Teil des Berichts sein.*

13.2. Nenner — Integrität des Nenners

Ein Bericht sollte die Frage klar beantworten: „90 Prozent von genau was sind 90 Prozent?“ Mögliche Nenner:

- Zugewiesene Teilnehmer
- Aufgabeninitiatoren
- Aufgabenerlediger
- Protokollkonforme Aufzeichnungen
- Gültige Systemergebnisse
- Antworten mit Entscheidungen des Schiedsrichters
- Spezifische Sprach- oder Länderspalte

Der Nenner sollte nach dem Betrachten des Ergebnisses nicht geändert werden. Zum Beispiel:

- 1,000 Aufgaben zugewiesen,
- 900 abgeschlossen,
- 800 gültig,
- 720 bestanden

Also:

```
720/800=90%
```

Dies ist die Bestehensrate unter gültigen Aufzeichnungen. Die beobachtete Bestehensrate über die gesamte zugewiesene Stichprobe ist stattdessen:

```
720/1000=72%
```

beantwortet eine andere Frage. Beide Raten können verwendet werden. Sie können nicht im gleichen Sinn präsentiert werden.

14. GENAUIGKEIT UND KONSISTENZ SIND NICHT DAS GLEICHE

Eine Verteilung kann eine hohe Genauigkeit haben. Sie kann jedoch im Laufe der Zeit oder zwischen Nutzern inkonsistent sein. Eine andere Verteilung kann eine geringe Genauigkeit haben. Sie kann jedoch konsequent falsch sein. Deshalb sollten vier Situationen unterschieden werden.

Genauigkeit	Konsistenz	Kommentar
Hoch	Hoch	Kandidat für eine konsistente und genaue Repräsentation
Hoch	Niedrig	Meist genau, aber unzuverlässige Erfahrung
Niedrig	Hoch	Systematische und konsistente Fehlinterpretation
Niedrig	Niedrig	Sowohl falsche als auch instabile Repräsentation

„Konsistent zu sein“ ist nicht automatisch positiv. Das System kann sehr konsistent sein, wenn es jedem Nutzer denselben Fehler liefert.

14.1. Inter-Nutzer-Konsistenz

Sehen verschiedene Nutzer in derselben Zelle eine ähnliche materielle Repräsentation?

14.2. Inter-Sitzungs-Konsistenz

Sieht derselbe oder der gematchte Nutzer ähnliche Ergebnisse in neuen Sitzungen?

14.3. Inter-Wellen-Konsistenz

Erzeugt dasselbe Protokoll eine ähnliche Verteilung an verschiedenen Tagen oder Zeitperioden?

14.4. Konsistenz nach dem Produktupdate

Wenn es eine wesentliche Veränderung im KI-Produkt gibt, setzt sich die vorherige Verteilung fort?

15. DENSELBEN SATZ ZU PRODUZIEREN IST KEINE STABILITÄT

Stabilität: Es bedeutet nicht, "jedem Nutzer wortwörtlich die gleiche Antwort zu geben." Ein KI-Produkt kann verschiedene Ausdrücke, Reihenfolgen und Beispiele verwenden. Solange sie die materielle Bedeutung bewahren, können sie akzeptable Variationen sein. Der wichtige Unterschied ist:

Formale Variabilität

Wortwahl Satzreihenfolge Länge Tonbeispiel Verwendung von Liste oder Absatz Wenn die materielle Wahrheit erhalten bleibt, muss es kein Fehler sein.

Semantische Variabilität

Änderung der Unternehmensidentität Differenzierung der Hauptkategorie Ausweitung oder Einschränkung des Tätigkeitsbereichs Änderung der Lizenz oder Kompetenz Änderung von Preis und Geografie Empfehlung an einen Nutzer, aber keine Empfehlung an einen ähnlichen Nutzer ohne Erklärung Wenn sich die materielle Repräsentation ändert, kann ein Stabilitätsproblem bestehen. Daher ist das bevorzugte Konzept:

Bedingte semantische Stabilität

Es sollte sein. Es wird erwartet, dass die fundamentale Realität unter äquivalenten materiellen Bedingungen erhalten bleibt. Eine Änderung der Empfehlung aufgrund unterschiedlicher Nutzerbedürfnisse kann ein angemessenes kontextuelles Verhalten sein.

16. KONTEXTSENSITIVITÄT UND WILLKÜRLICHE VARIABILITÄT

Es ist nicht immer korrekt, dass ein System allen Nutzern dieselbe Empfehlung gibt. Zum Beispiel:

- Nutzer mit niedrigem Budget
- Nur Nutzer, die Dienstleistungen in einem bestimmten Land suchen
- Nutzer, die eine andere Lizenz benötigen
- Nutzer, die individuelle statt unternehmerische Dienstleistungen suchen

können für dieselbe Einheit unterschiedliche Konformitätsergebnisse erhalten. Dieser Unterschied:

- kann auf die Nutzerbedingung zurückzuführen sein,
- die Leistungsumfang,
- oder das tatsächliche Limit

und kann kontextabhängig und genau sein. Wenn sich das Ergebnis jedoch ohne Grund unter gleichen Nutzerbedingungen ändert, kann dies willkürliche Schwankungen sein. Daher möchte NOMOS nicht: „Die gleiche Antwort für alle Nutzer.“ Es möchte:

> *Eine Antwort, die unter gleichwertigen materiellen Bedingungen konsistent und erklärbar ist und unter unterschiedlichen Bedingungen auf tatsächliche Grenzen sensibel reagiert.*

17. ZEIT IST INNERHALB DER REPRÄSENTATIONSVERTEILUNG

Die Repräsentationsverteilung ist nicht zeitlos. Sie kann als PE,t betrachtet werden. Die Verteilung derselben Einheit:

- Produktaktualisierung,
- Ressourcenänderung,
- Unternehmensänderung,
- Nachrichtenereignis,
- neue Sprachversion,
- technische Zugriffsänderung

kann danach abweichen.

17.1. Plötzliche Änderung

Nach einem Produkt- oder Ressourcenupdate ändert sich die Verteilung in kurzer Zeit schnell.

17.2. Allmähliche Änderung

Aufgrund neuer Ressourcen und Nutzerverhalten ändert sich das Ergebnis über die Zeit langsam.

17.3. Vorübergehende Schwankung

Während eines bestimmten Ereignisses oder Nachrichtenzeitraums unterscheidet sich die Verteilung vorübergehend.

17.4. Änderung nach Intervention

Nach der GEO-Intervention kann sich die Verteilung ändern. Die Behauptung, dass die Änderung auf die Intervention zurückzuführen ist, erfordert jedoch auch einen Basiswert und ein Kausalitätsdesign.

17.5. Glättung mit Zeitmittel

Für ein Produkt:

- in der ersten Welle 98,
- in der zweiten Welle 90,
- in der dritten Welle 82

Wenn genommen, ist der Durchschnitt 90. Das Veröffentlichen von 90 verschleiert jedoch den Abwärtstrend. Daher:

> *Der Zeitdurchschnitt kann die Wellen-Ergebnisse nicht ersetzen.*

18. TAIL-RISIKO UND SCHWERE FEHLERVERTEILUNG

Ein großer Teil einer Verteilung kann korrekt sein. Niedrige Wahrscheinlichkeit schwerer Fehler sollte jedoch ebenfalls bewertet werden. Beispiele:

- Gesundheitslizenz, die tatsächlich nicht existiert
- Falsche Notfallanweisungen
- Falsche rechtliche Befugnis
- Gefälschte Investitionsgarantie
- Falsche Unternehmensidentität
- Unangemessene Ratschläge für Kinder oder gefährdete Gruppen
- Wiederholung der manipulativen Anweisung, die vor Menschen verborgen ist

In der vorherigen Architektur:

- Kritisch
- Major
- Mäßig
- Beratung

wurden in vier Wichtigkeitsstufen unterteilt, und es wurde vorgeschlagen, dass kritische Befunde eine automatische Fehlfunktion haben sollten. Dieses Buch verknüpft dieses System mit der Bevölkerungsverteilung.

18.1. Trennung von Häufigkeit und Schwere

Ein Fehler:

- häufig, aber mit geringem Einfluss,
- selten, aber schwerwiegend

können auftreten. Allerdings kann die Häufigkeitsrate allein die beiden Fälle nicht gleichwertig bewerten.

18.2. Kandidaten-Schweregrad-gewichtete Risikodarstellung

KANDIDATENKONZEPT

Ein Risikoindikator kann unter Verwendung von Schwerekoeffizienten für Fehlerklassen entwickelt werden:

```
tau= Σ_{k=1}^K λ_k p_k
```

Hier:

- pk: vorhergesagte Rate des Fehlertyps
- λk: Schweregradkoeffizient des Fehlertyps

Allerdings können die Koeffizienten nicht willkürlich bestimmt werden. Sie sollten basieren auf:

- Absehbare menschliche Schäden
- Integrität der Repräsentation
- Auswirkung auf die Nutzerentscheidung
- Umkehrbarkeit
- Absicht und Manipulation
- Hochrisikobereich

Die genaue Methode wird in Abschnitt 15 definiert.

18.3. Kritisches Tor

Ein hoher Gesamtdurchschnitt kann einen überprüften Kandidaten für einen kritischen Fehler nicht automatisch löschen. Das korrekte öffentliche Protokoll sollte beide Ergebnisse zusammen zeigen: „Die geschätzte Gesamtrepräsentation ist hoch.“ und: „Das angegebene kritische Ereignis oder die Ereignisse beeinflussen ebenfalls die Konformitätsentscheidung.“

19. ABLEHNUNG UND FEHLER BEI DER VERTEILUNG

Ein nicht reagierendes System ist ebenfalls Teil der Nutzererfahrung. Nicht alle Ablehnungen sind jedoch gleich.

19.1. Angemessene Ablehnung

Das System kann vorsichtig agieren, wenn die Nachweise unzureichend sind oder wenn es keine sichere Antwort geben kann. Dies kann in einigen Kontexten das richtige Verhalten sein.

19.2. Unnötige Ablehnung

Es kann inkonsistent darauf verzichten, auf öffentlich verfügbare und risikoarme Informationen zu reagieren.

19.3. Technisches Nichtreagieren

Das System kann aufgrund eines technischen Fehlers keine Antwort erzeugen.

19.4. Ablehnung außerhalb des Umfangs

Die Prompt kann falsch sein oder außerhalb des Aufgabenprotokolls liegen.

19.5. Bewahrung von Ablehnungen in Verteilungen

Ablehnungen:

- Erfolg,
- Fehlschlag,

UNBEKANNT

sollten nicht in eine einzige Kategorie gezwungen werden. Ihre Funktion und Begründung sollten bewertet werden. Das Entfernen von Ablehnungen aus dem Nenner kann nur unter erfolgreichen Antworten eine hohe Genauigkeitsrate ergeben und die tatsächliche Nutzererfahrung verschleiern.

20. WIE KÖNNEN VERTEILUNGEN VERGLICHEN WERDEN?

Allein der Vergleich der Ergebniszahlen reicht nicht aus, um die Verteilung von zwei KI-Produkten zu vergleichen. Damit der Vergleich fair ist, sollten die folgenden Bereiche so gleich wie möglich sein oder explizit modelliert werden:

- Auditierte Einheit
- Zielpopulation
- Prompt-Version
- Sprachen
- Länder
- Messwelle
- Sitzungsbedingungen
- Nutzeroberfläche
- Planverteilung
- Beurteilungsregel
- Realitätspaket
- Gültigkeitsregeln
- Gewichte

Zwei mit materiell unterschiedlichen Protokollen erzielte Ergebnisse können nicht direkt für ein Ranking verwendet werden.

20.1. Vergleich auf gemeinsamer Basis

Alle verglichenen KI-Produkte:

- sind offiziell zugänglich,
- unterstützen dieselbe Sprache und Nutzungsbedingungen

können in einem gemeinsamen Universum verglichen werden. Dies ermöglicht einen saubereren Vergleich zwischen Systemen.

20.2. Natürliche Zugriffsverteilung

Jedes KI-Produkt kann in seinem eigenen realen Nutzerzugriffs-Universum gemessen werden. Dieses Ergebnis zeigt die tatsächliche globale Wirkung des Produkts. Da die Zielpopulationen jedoch unterschiedlich sein können, erfordert ein direktes Leistungsranking Vorsicht.

20.3. Verschiedene Oberflächen desselben Anbieters

Web-, Mobile- oder Unternehmensprodukte oder verschiedene Pläne, selbst wenn sie zum gleichen Anbieter gehören, können separate Verteilungen erzeugen. Wenn sie zu einem einzigen Ergebnis kombiniert werden sollen, sollte die Kombinations- und Gewichtungsmethode erklärt werden.

21. GEWICHTUNG ÄNDERT DIE BEDEUTUNG DER VERTEILUNG

Ein globaler Durchschnitt hängt vom Gewicht ab, mit dem Beobachtungen kombiniert werden. Mögliche Gewichte:

- Gleich nach Land
- Nach Länderbevölkerung
- Geeignete KI-Nutzerpopulation
- Sprachbenutzer
- Tatsächliche Produktexposition
- Gleiches KI-Produkt
- Marktanteil
- Zielmarkt der Kunden

Jede Methode beantwortet eine andere Frage.

21.1. Gleiches Ländergewicht

San Marino und Indien können dasselbe Gewicht erhalten. Dieser Ansatz kann die Frage der 'Durchschnittlichen Ländererfahrung' behandeln. Er stellt nicht die durchschnittliche Nutzererfahrung in der Weltbevölkerung dar.

21.2. Bevölkerungsgewicht

Größere Länder haben mehr Einfluss. Dieser Ansatz kann die Erfahrung der globalen Bevölkerung näherungsweise darstellen. Er kann kleine Länder und Sprachen unsichtbar machen.

21.3. Sprachfairnessgewicht

Er kann auch sicherstellen, dass Sprachen mit geringen Ressourcen sichtbar bleiben. Es ist nicht dasselbe wie der globale Bevölkerungsscore.

21.4. Gewicht der Produktexposition

Die Gewichtung von KI-Produkten nach der tatsächlichen Nutzerexposition kann berücksichtigt werden. Sie erfordert zuverlässige und unabhängige Nutzungsdaten.

21.5. Prinzip der normativen Gewichtung

> *Gewichte sollten definiert werden, bevor die Ergebnisse gesehen werden, öffentlich bekannt gemacht und die Frage, die sie beantworten, spezifiziert werden.*

Es ist verboten, ein Gewichtungssystem zu wählen, das den positiven Wert erhöht, nachdem das Ergebnis gesehen wurde. Neben einer einzigen globalen Punktzahl können auch alternative aussagekräftige gewichtete Ergebnisse veröffentlicht werden.

22. PUNKTVERTEILUNG IST VERLUSTBEHAFTE KOMPRESSION

Es ist möglich, aus einer Verteilung eine einzelne Zahl zu erzeugen. Eine einzelne Zahl trägt jedoch nicht alle Informationen der Verteilung. Daher ist die Punktzahl:

Eine verlustbehaftete Zusammenfassung der Verteilung

zu akzeptieren. Eine Punktzahl kann die folgenden Informationen verlieren:

- Schwächste Sprache
- Schwächstes Land
- Kritische Fehlertypen
- Unbekannte Masse
- Zeitliche Schwankung
- Polarisierung
- Produkt- und Planabweichung
- Außerhalb des Geltungsbereichs liegende Bevölkerung
- Ablehnungsverteilung
- Mehrdeutigkeit des Beurteilers

Die Punktzahl ist nützlich, weil:

- vergleichende Darstellungen erleichtert,
- die öffentliche Kommunikation vereinfacht,
- die Überwachung von Veränderungen ermöglicht,

die Managemententscheidung unterstützt. Die Einfachheit der Punktzahl darf jedoch nicht zur Eliminierung der Verteilung führen.

23. MINDESTÖFFENTLICHE ERGEBNISKARTE

KANDIDATENANFORDERUNG — KANDIDATENVERPFLICHTUNG

Wenn eine globale Punktzahl NOMOS der Öffentlichkeit präsentiert werden soll, wird empfohlen, dass mindestens die folgenden Ergebnisse daneben gezeigt werden:

Feld	Beschreibung
Hauptdarstellungsschätzung	Gesamtergebnis unter definierter Gewichtung
Vertrauens- oder Unsicherheitsbereich	Empfindlichkeit der Vorhersage
Umfang	Land, Sprache, Produkt, Nutzeroberfläche und Zeit
Schwächste ausreichende Untergruppe	Macht materielle Ungleichheit sichtbar
Kritische Fehlerrate	Schwere Fehlerwarteschlange
Zeitliche Stabilität	Variation zwischen Wellen
UNBEKANNT und nicht schlüssige Rate	Unbekannte Masse
NICHT GETESTETE Felder	Ungemessenes Universum
Messdatum	Zeitliche Begrenzung des Ergebnisses
Protokollversion	Mit welcher Regel es erstellt wurde

Die Hauptbewertung sollte nicht als alleinstehendes Werbelabel verwendet werden. Der kanonische Eintrag muss mit dieser Ergebnis-Karte verknüpft sein.

24. SYNTHETISCHE APPLE.COM VERTEILUNGSANZEIGE

DEMONSTRATION DER SYNTHETISCHEN METHODE / Die unten aufgeführten KI-Produkte, Nutzer, Sprachen, Zahlen und Ergebnisse sind vollständig fiktiv. Sie stellen nicht die tatsächliche Leistung von Apple Inc. oder einem realen KI-Produkt dar. Betrachten wir zwei synthetische KI-Produkte:

- Produkt A
- Produkt B

Für jedes Produkt:

- fünf Sprachen,
- 200 gültige synthetische Beobachtungen pro Sprache,
- eine Gesamtzahl von 1,000 Beobachtungen

werden angenommen. Diese gleichmäßige Sprachverteilung dient nur zur Demonstration der Methode. Sie stellt nicht die tatsächliche Bevölkerungssgewichtung GEO-1000 dar.

24.1. Sprachbasierte Bestehensraten

Sprache	Produkt A	Produkt B
Englisch	90%	99%
Türkisch	91%	98%
Deutsch	89%	97%
Japanisch	90%	86%
Arabisch	90%	70%
Einfacher Durchschnitt	90%	90%

Der Durchschnitt der beiden Produkte ist derselbe. Ihre Verteilungen sind jedoch unterschiedlich.

Produkt A

Die Sprachergebnisse liegen nah beieinander. Der Unterschied zwischen der höchsten und niedrigsten Sprache beträgt 2 Punkte. Das Sprachminimum liegt bei 89%.

Produkt B

Erzielt sehr hohe Ergebnisse in drei Sprachen. Es gibt einen Rückgang in Japanisch und besonders in Arabisch. Der Unterschied zwischen der höchsten und niedrigsten Sprache beträgt 29 Punkte. Das Sprachminimum liegt bei 70%. Hinweis: „Beide erzielten 90“ zu sagen, verschleiert den wesentlichen Unterschied.

24.2. Verteilung des Antwortstatus

Antwortstatus	Produkt A	Produkt B
Erfüllt das Übergangskriterium	900	900
Teilweise unterstützt	60	20
Nicht belegte materielle Aussage	20	20

Antwortstatus	Produkt A	Produkt B
Veraltet	10	20
Kritischer Fehlerkandidat	1	14
UNBEKANNT / kein Ergebnis	9	26
Gesamt	1,000	1,000

Die Bestehensquote der beiden Produkte liegt erneut bei 90%. Allerdings: Bei Produkt B ist die Anzahl der Kandidaten für kritische Fehler 14-mal höher. Bei Produkt B ist die unbekannte und unklare Masse höher. Ein größerer Teil der Ausfälle bei Produkt A ist auf teilweise fehlende Unterstützung zurückzuführen. Der Fehlerstau bei Produkt B ist höher. Dieselbe Bestehensquote führt nicht unbedingt zu demselben Fitnessergebnis.

24.3. Zeitwellen

Messwelle	Produkt A	Produkt B
Welle 1	90%	98%
Welle 2	91%	90%
Welle 3	89%	82%
Mittelwert	90%	90%

Produkt A ist relativ stabil. Produkt B zeigt einen deutlichen Rückgang. Die Veröffentlichung nur des Durchschnitts der drei Wellen verschleiert die Verschlechterung bei Produkt B.

24.4. Synthetisches Ergebnis

Es könnte möglich sein, für Produkt A und Produkt B denselben Hauptwert zu veröffentlichen:

90

Aber NOMOS-Ergebnisblätter können nicht gleich sein.

Produkt A — Synthetisches Verteilungssignatur

Gesamtergebnis: hoch Sprachungleichheit: niedrig Kritischer Fehleranteil: niedrig, aber nicht null Unbekannte Masse: niedrig Zeitstabilität: hoch

Produkt B — Synthetische Verteilungssignatur

Gesamtergebnis: hoch Sprachungleichheit: hoch Kritischer Fehleranteil: erheblich Unbekannte Masse: höher Zeitstabilität: niedrig und abnehmend

Daher lautet die synthetische Schlussfolgerung von Abschnitt 2 wie folgt:

> *Dass Produkt A und Produkt B denselben Durchschnittswert haben, zeigt nicht an, dass sie dieselbe Repräsentationsqualität besitzen.*

25. VERPFLICHTENDE NORMATIVE BESTIMMUNGEN

CH02-N01

Ein GEO-Ergebnis kann nicht nur durch einen einzelnen Durchschnitt oder eine einzige Bestehensquote definiert werden.

CH02-N02

Jedes Verteilungsergebnis:

- die geprüfte Einheit,
- Zielpopulation,
- KI-Produkt- und Nutzeroberfläche,
- Land und Sprache,
- Die Prompt,
- die Wohnsitzbedingung,
- zur Zeit

sollte verbunden sein.

CH02-N03

Der globale Durchschnitt alleine kann Ungleichheiten innerhalb materieller Untergruppen nicht ausgleichen.

CH02-N04

Kritische und andere Fehlerergebnisse mit hoher Schwere sollten getrennt von der Gesamtfehlerrate gemeldet werden.

CH02-N05

UNKNOWN, keine Ergebnisse, abgelehnt, ungültig und NICHT GETESTET sollten voneinander unterschieden werden.

CH02-N06

Nicht gemessenes Land, Sprache, Produkt, Nutzeroberfläche oder Zeitraum kann nicht als Teil der positiven Verteilung angezeigt werden.

CH02-N07

Gewichte und Aggregationsregeln sollten vor der Ansicht der Ergebnisse definiert werden.

CH02-N08

Gewichtsänderungen nach Ergebnis sollten eine neue Analyseversion erstellen, und frühere Ergebnisse sollten erhalten bleiben.

CH02-N09

Verteilungen, die auf materiell unterschiedlichen Zielpopulationen, Nachfrage, Produkt, Zeit oder Bewertungskriterien basieren, können ohne Erklärung nicht direkt verglichen werden.

CH02-N10

Formale Sprachvariabilität kann nicht automatisch als Repräsentationsfehler gezählt werden.

CH02-N11

Materialidentität, Umfang, Zeit, Kompetenz oder Variabilität von Empfehlungen können ohne Rechtfertigung nicht als Stilunterschied klassifiziert werden.

CH02-N12

Allein hohe Stabilität ist kein Nachweis für Qualität; auch kontinuierlich falsche Ergebnisse sollten aufgezeichnet werden.

CH02-N13

Ein hoher Durchschnitt sollte, wenn er mit geringer Abdeckung einhergeht, öffentliche Ergebnisse mit Abdeckungsbegrenzung haben.

CH02-N14

Teilgruppenergebnisse sollten unter unzureichender Stichprobenerhebung und großer Unsicherheit nicht als endgültige Urteile über Länder oder Sprachen veröffentlicht werden.

CH02-N15

Ablehnungs- und technische Ausfallraten sollten als Teil der echten Nutzererfahrung beibehalten werden.

CH02-N16

Die einzige öffentlich zugängliche Punktzahl sollte mit der kanonischen Verteilungs- und Umfangsaufzeichnung verknüpft sein.

CH02-N17

Synthetische Verteilungen können nicht mit der Leistung eines echten KI-Produkts oder einer echten Einheit kombiniert werden.

CH02-N18

Der Verteilungsbericht sollte einen verantwortlichen menschlichen oder institutionellen Eigentümer haben.

26. FORMEN DES VERSAGENS

CH02-F01— DURCHSCHNITTLICHES DENKEN

Hoher globaler Durchschnitt wird angezeigt; Sprache, Land oder kritische Fehlerkorruption wird gespeichert.

CH02-F02— UNTERGRUPPE LÖSCHEN

Leistungsarme Gruppen werden aus dem Bericht entfernt oder als 'unbedeutend' erklärt.

CH02-F03— UMFANGSKARTE

Ergebnisse mit engem Umfang werden für die ganze Welt, alle Sprachen oder alle Produkte verwendet.

CH02-F04— SCHWEREGRADNORMALISIERUNG

Geringfügige Mängel und schwerwiegende Gesundheits- oder Lizenzierungsfehler werden innerhalb derselben Fehleranzahl kombiniert.

CH02-F05— VERSCHLEIERUNG DES VERFALLS MIT DEM ZEITLICHEN DURCHSCHNITT

Rückgang oder Schwankung wird nur mit dem Periodendurchschnitt unsichtbar gemacht.

CH02-F06— ANBETUNG DER STABILITÄT

Eine konsistente Antwort wird automatisch als korrekt angenommen.

CH02-F07— WORTWÖRTLICH GLEICHE ANTWORT-FALLZIE

Wörtliche Textübereinstimmung wird als Maß für Konsistenz verwendet, anstatt für semantische Korrektheit.

CH02-F08— KONTEXTUELLE DIFFERENZ ALS FEHLER ZÄHLEN

Unterschiede in angemessenen Empfehlungen basierend auf den tatsächlichen Bedürfnissen der Nutzer werden als willkürliche Variabilität markiert.

CH02-F09— WILLKÜRLICHE VARIABILITÄT ALS PERSONALISIERUNG ZÄHLEN

Materielle Widersprüche unter gleichen Nutzerbedingungen werden ohne Erklärung als „Personalisierung“ verteidigt.

CH02-F10— DEN NENNER ÄNDERN

Der Übergang nach dem Ergebnis wird unter abgeschlossenen, gültigen oder ausgewerteten Beobachtungsnennern vorgenommen, um das positive Ergebnis zu erhöhen.

CH02-F11— GEWICHTUNG NACH ERGEBNIS

Das Land, die Sprache oder Modellkoeffizienten, die die höchste Punktzahl ergeben, werden anschließend ausgewählt.

CH02-F12— INKONSISTENTE ZELLEN ZUSAMMENFÜHREN

Verschiedene Prompt-, Produkt-, Zeit- oder Panelbedingungen werden als eine einzige Verteilung dargestellt.

CH02-F13— LÖSCHEN VON ABLEHNUNGEN

Systemaufzeichnungen, die nicht reagiert haben, werden entfernt, um nur unter denen, die eine Antwort gegeben haben, hohe Genauigkeit zu erzeugen.

CH02-F14— DEFINITIVE BEWERTUNG AUS KLEINER ZELLE

Ein Länder- oder Sprachscore wird von einem oder mehreren Nutzern erstellt.

CH02-F15— DAS KRISETAIL VERSTECKEN

Seltene schwere Fehler gehen innerhalb der gesamten Bestehensrate verloren.

CH02-F16— EINZELSCOREBERICHT

Es wird nur der Hauptscore veröffentlicht, ohne den Umfang, Ungleichheit, Stabilität, Unsicherheit und schwere Fehler zu zeigen.

CH02-F17— ZÄHLUNG DER VERTEILUNG ALS „GLAUBE DER KI“

Beobachtete Oberflächenresultate der Nutzer werden dargestellt, als ob sie die unveränderliche und intrinsische Sichtweise der künstlichen Intelligenz wären.

CH02-F18— DAS ZÄHLEN DES UNMESSBAREN ALS NULLVARIANZ

Ungetestete Bereiche werden als glatt und stabil betrachtet.

27. PRÜFUNGSVERFAHREN

Ob ein GEO-Bericht die Verteilung ehrlich darstellt, sollte mit den folgenden Schritten geprüft werden.

Schritt 1– Erfassen Sie den öffentlichen Aussage

Beispiel: „Die globale NOMOS-Punktzahl der Marke beträgt 92.“ Der vollständige Satz und die Oberfläche, auf der er gefunden wird, werden aufgezeichnet.

Schritt 2— Zielverteilung definieren

Für welche Punktzahl:

- Nutzerschaft,
- KI-Produkten,
- Ländern,
- Sprachen,

- Zeitraum

wird sie angewendet?

Schritt 3— Messzellen extrahieren

Jede Materialkombination wird bestimmt:

- Produkt
- Land
- Sprache
- Panel
- Aufforderung
- Welle
- Nutzeroberfläche

Schritt 4— Nenner neu aufbauen

Zugewiesene Aufgaben, abgeschlossene Aufgaben, gültige Datensätze, beurteilte Datensätze und bestandene Antworten müssen als separate Zählungen gemeldet werden.

Schritt 5— Ergebnisstatus extrahieren

Nicht nur die Bestehensquote:

- teilweise,
- nicht unterstützt,
- widersprüchlich,
- alt,
- Kritisch

UNBEKANNT,

ablehnt, unklare Verteilung wird untersucht.

Schritt 6— Gewichte überprüfen

Gewichte:

- sind sie vordefiniert,
- welche Frage beantworten sie,

haben sie sich nach dem Ergebnis geändert?

Schritt 7— Untergruppenunterschiede untersuchen

Die beste Gruppe, die schwächste ausreichende Gruppe, Sprachunterschied, Länderunterschied, Produktunterschied und Panelunterschied werden bewertet.

Schritt 8— Umfang und Mängel überprüfen

Welche Bereiche wurden nicht gemessen? Welche Zellen sind unzureichend? Was sind die Gründe für die Defizienz? Könnten die fehlenden Daten systematisch sein?

Schritt 9— Bestimmung untersuchen

Die Veränderung zwischen Waves Sessions Users Interfaces wird bewertet.

Schritt 10— Warteschlangenrisiko prüfen

Kritische und andere schwerwiegende Fehler:

- Anzahl,
- deren Rate,
- Typ,
- Nutzereffekt

wird getrennt in Bezug auf bewertet

Schritt 11— Vergleichen Sie die öffentliche Zusammenfassung mit der Verteilung

Übermittelt die einzelne Zahl oder der Satz, der der Öffentlichkeit veröffentlicht wurde, genau die Richtung und das Risiko der vollständigen Verteilung?

Schritt 12— Ergebnis begrenzen

Die öffentliche Aussage ist nur auf das Niveau begrenzt, das durch die Verteilung und den Umfang getragen wird.

28. NOTWENDIGE NACHWEISE

Soweit relevant für eine repräsentative Verteilung oder einen globalen GEO-Score, sollten die folgenden Datensätze durchsucht werden:

- Geprüfter Eintrag des Unternehmens
- Definition der Zielpopulation
- Stichprobenrahmen
- KI-Produkt- und Nutzeroberflächen
- Länder- und Sprachzellen
- Prompt und Versionen
- Sitzungs- und Panel-Bedingungen
- Messwellen
- Zugewiesene Aufgabennummern
- Abgeschlossene Aufgaben
- Gültige und ungültige Datensätze
- Rohantworten
- Verteilung des Ergebnisstatus
- Entscheidungsaufzeichnungen
- Datensätze mit kritischen Fehlern
- Abgelehnte und unklare Datensätze
- UNKNOWN-Datensätze
- Gewichte
- Interessengruppen
- Ergebnisse von Untergruppen
- Stabilitätsergebnisse
- Außerhalb des Geltungsbereichs liegende Bereiche

- Berechnungen von Vertrauen oder Unsicherheit
- Öffentliches Berichtskartensystem
- Verantwortliche Person oder Institution
- Protokoll- und Analyseversion
- Änderungshistorie

29. PRÜFLISTE FÜR AUDITS

Ist die Zielpopulation der Verteilung klar? Sind die geprüften KI-Produkte und Nutzeroberflächen spezifiziert? Sind Land, Sprache, Prompt- und Zeitrahmen sichtbar? Werden die zur Berechnung der zusammengesetzten Punktzahl verwendeten Gewichtungen offengelegt? Ist der Nenner vollständig definiert? Werden ungültige Datensätze separat gemeldet? Werden Verweigerungen und unklare Ergebnisse aufbewahrt? Sind UNKNOWN und NICHT GETESTET unterschiedlich? Sind Teilergebnisse sichtbar? Wurden kleine Zellen mit übermäßiger Sicherheit interpretiert? Wurde die schwächste Materialgruppe beibehalten? Werden kritische Fehler separat gemeldet? Werden Messwellen separat angezeigt? Verschleiert der Durchschnitt Rückgänge oder Schwankungen? Wird ein rein formaler Unterschied von einem wesentlichen Bedeutungsunterschied unterschieden? Wurde kontextabhängige Variation in Empfehlungen korrekt klassifiziert? Wurden unterschiedliche Oberflächen desselben Anbieters ohne Erklärung zusammengeführt?

Wurden Bereiche außerhalb des Geltungsbereichs in das globale Ergebnis einbezogen? Wurden synthetische und reale Verteilungen getrennt? Spiegelt der der Öffentlichkeit präsentierte Satz die wahre Richtung der vollständigen Verteilung wider? Wenn eine der wesentlichen Fragen unbeantwortet bleibt, hat das Verteilungsergebnis nur begrenztes Vertrauen. Wenn mehr als eine unbeantwortet bleibt, sollte eine einzige globale Punktzahl nicht als endgültiges Urteil der Öffentlichkeit präsentiert werden.

30. EINWÄNDE UND ANTWORTEN

Einwand 1— „Die Öffentlichkeit möchte eine einzelne Zahl.“

Das ist korrekt. Eine einzelne Zahl kann die Kommunikation erleichtern. NOMOS lehnt eine einzelne Zahl nicht vollständig ab. Allerdings:

- Umfang,
- Kritischer Fehler,
- Untergruppenbasis,
- Zeit,
- Unsicherheit

muss mit der Information verbunden sein. Eine Zahl kann kurz sein. Sie kann die Wahrheit nicht verkürzen.

Einwand 2— „Kleine Sprachen senken unnötig die globale Punktzahl.“

Das globale Bevölkerungsgewicht kann eine kleine Sprache mit begrenztem Gewicht berücksichtigen. Allerdings kann eine starke Repräsentationsverzerrung in einer kleinen Sprache nicht unbeachtet bleiben, nur weil ihr Bevölkerungsanteil niedrig ist. Daher:

- bevölkerungsgewichtete Hauptbewertung,
- als Ergebnis der separaten Sprachfairness

Sie können zusammen veröffentlicht werden. Zwei Metriken beantworten unterschiedliche Fragen.

Einwand 3— 'Wenn es Personalisierung gibt, sind unterschiedliche Antworten normal.'

Einige Unterschiede können normal und korrekt sein. Allerdings:

- eine Unternehmensidentität,
- Lizenz,
- die aktuellen Preise,
- das tatsächliche Servicegebiet

sollten nicht willkürlich geändert werden. Personalisierung kann die angemessene Erklärung und Beratung des Nutzers beeinflussen. Sie ist keine Erlaubnis, eine wesentliche Tatsache zu erfinden.

Einwand 4— 'Warum sollte es ein Problem darstellen, wenn nicht die gleiche Antwort gegeben wird?'

Das Problem ist nicht der Wortunterschied. Das Problem ist die Veränderung der materiellen Realität. NOMOS sucht nicht nach exakter Textekualität, sondern nach bedingter semantischer Konsistenz.

Einwand 5— „Warum sollte es eine Rolle spielen, wenn ein kritischer Fehler nur einmal in 1,000 Antworten auftritt?“

Dieser Fehler kann in Bezug auf die Bevölkerungsprävalenz selten sein. Allerdings kann er hinsichtlich des Schadens schwerwiegend sein. Ein unkontrollierter Ausfall eines kritischen Systems in einem Flugzeug einmal in tausend Flügen wird nicht nur mit 0.1 Prozent bewertet. Im Kontext von GEO sollten Häufigkeit und Schaden auch separat betrachtet werden.

Einwand 6— „Wir können keine ausreichende Stichprobe für jedes kleine Land oder jede Sprache finden.“

Das ist richtig. Daher:

- globales Bevölkerungspanel,
- Country Observer-Panel,
- Language Fairness Panel

kann separat erstellt werden. Eine Einzelpersonenbeobachtung wird nicht zu einer Länderbewertung. Sie zeigt nur an, dass eine Beobachtung vorliegt. Der korrekte Status für eine unzureichende Stichprobe:

UNZUREICHENDE STICHPROBE

oder:

NICHT GESCHÄTZT

muss sein.

Einwand 7— "Diese Detailtiefe ist zu komplex, als dass der Kunde sie verstehen würde."

Es ist nicht notwendig, dass der Kunde alle Formeln auf der Hauptseite sieht. Die entscheidungsrelevanten Kernfakten dürfen jedoch nicht verborgen bleiben. Die kurze Ergebnisübersicht sollte die folgenden Fragen beantworten: Was ist das Gesamtergebnis? Was ist das schwerwiegendste Risiko? Welcher Bereich ist am schwächsten? Was wurde nicht getestet? Wie robust

ist das Ergebnis? Zu welchem Datum gehört es? Komplexität muss gesteuert werden. Nicht eliminiert.

Einwand 8— "Wenn dasselbe Modell in allen Ländern verwendet wird, warum sollte es einen Länderunterschied geben?"

Die vom Nutzer wahrgenommene Ausgabe muss nicht ausschließlich aus dem reinen Modell bestehen. Länderspezifische Unterschiede umfassen:

- Produktzugang,
- Webressourcen,
- Locale,
- Sprache,
- Plan,
- Schnittstelle,
- Gesetzgebung,
- Personalisierung

Es kann dadurch entstehen. Ein Unterschied muss nicht auftreten. Ohne Messung kann jedoch kein Fehlen angenommen werden.

Einwand 9— „Zuverlässige Bewertung ist nicht möglich, da KI-Antworten variabel sind."

Variabilität macht eine Messung nicht unmöglich. Im Gegenteil, sie ist eines der Merkmale, die gemessen werden müssen. Die Bewertung:

- spezifisches Datum,
- Umfang,
- Konfidenzintervall
- Stabilitätsinformationen

kann sinnvoll sein, solange sie dies trägt. Der Anspruch auf zeitlose Sicherheit ist jedoch nicht zu rechtfertigen.

Einwand 10— „Wenn sich der Prompt ändert, ändert sich die Verteilung vollständig."

Das ist korrekt. Daher ist die Verteilung vom Prompt abhängig. GEO-1000 liefert nicht für alle möglichen Fragen die gleiche Antwort. Vordefiniert:

- Kernspiegel-Prompt,
- Diagnostisches Prompt-Register

Misst spezifische darstellende Bereiche. Die Prompt-Zusammensetzung wird in Abschnitt 10 definiert.

ALLGEMEINE REGEL VON ABSCHNITT 32

Dieser Abschnitt wurde nicht gegen eine einzelne Punktzahl verfasst. Er wurde verfasst unter der Annahme, dass eine einzelne Punktzahl allein die Gesamtheit der Realität darstellt. Menschen wollen Zahlen. Institutionen wollen Ranglisten. Gremien wollen eine einzige Zeile sehen. Der Markt könnte fragen: „Was ist Ihre NOMOS-Punktzahl?" Diese Frage ist legitim. Aber die Antwort kann nicht nur lauten: „90." Denn hinter 90 können beliebige der folgenden Realitäten existieren:

- Konsistente Genauigkeit über alle Sprachen hinweg
- 99 in einigen Sprachen, 45 in anderen
- Hoher Durchschnitt, aber schwerwiegender kritischer Fehler
- Hohe Genauigkeit, aber geringe Abdeckung
- Hohe Sofortergebnisse, aber schneller Rückgang
- Erfolg nur bei einfachen Prompts
- Hohe Rate nur unter bewertbaren Antworten
- Annahme ungetesteter Länder positiv
- Kombination ausgewählter KI-Produkte

Der Wert einer Punktzahl wird durch die Sichtbarkeit der zugrunde liegenden Verteilung gebildet. Wenn die Verteilung nicht sichtbar ist, erzeugt die Punktzahl kein Vertrauen. Sie erzeugt nur den Anschein von Sicherheit. Daher lautet das zweite Messgesetz von NOMOS wie folgt:

> *ein GEO-Score kann eine Zusammenfassung der Verteilung sein. / Er kann die Verteilung nicht ersetzen.*

Sein drittes Gesetz lautet:

> *ein hoher Durchschnitt beseitigt nicht die Frage, wer schlecht vertreten ist.*

Das vierte Gesetz lautet wie folgt:

> *Ein seltener schwerwiegender Fehler ist nicht unbedeutend, nur weil er selten ist.*

Das fünfte Gesetz lautet wie folgt:

> *Das ungetestete Feld ist nicht die positive Seite der Verteilung; es ist die unbekannte Grenze.*

Befehl 2 von NOMOS

> *Zeig mir nicht nur den Durchschnitt.*
>
> *Zeig, wer unter dem Durchschnitt liegt.*
>
> *Verstecke nicht deine schlechteste Sprache, während du deine beste Sprache zeigst.*
>
> *Benutze nicht das Gewicht großer Länder, um die Existenz kleiner Länder zu löschen.*
>
> *Schicke nicht eine Person in ein kleines Land und kündige dann eine endgültige Punktzahl für dieses Land an.*
>
> *Verharmlose nicht einen schweren Fehler, nur weil er in einer Antwort verborgen ist, die zu 99 Prozent korrekt ist.*
>
> *Erkläre nicht nur deshalb einen Fehler, weil die Formulierung einer Antwort sich geändert hat. / Aber wenn sich die zugrunde liegende Realität ändert, lehne sie nicht einfach als bloßen stilistischen Unterschied ab.*
>
> *Werfe Ablehnungen, Unbekannte und Ungeprüfte nicht aus der Verteilung hinaus.*
>
> *Ändere das Gewicht, den Anteil oder den Umfang nicht, nachdem du das Ergebnis gesehen hast.*
>
> *Erzeuge eine Zahl. / Mache aber die Länder, Sprachen, Nutzer, Zeiten und Fehler hinter der Zahl nicht unsichtbar.*
>
> *Denn die Repräsentation eines Objekts ist kein durchschnittlicher Satz. / Sie ist eine Verteilung der verschiedenen Realitäten, die von unterschiedlichen Menschen erlebt werden.*

Der abschließende Satz des Kapitels

Das Messobjekt von GEO ist nicht eine einzelne Antwort oder eine einzelne Durchschnittspunktzahl einer künstlichen Intelligenz; es ist die Gesamtheit der repräsentativen Ergebnisse, die in einer definierten Nutzerpopulation unter definierten Bedingungen erzeugt wurden.

Normativer Kern

Die Repräsentation einer Entität in generativen Systemen MUSS als bedingte Verteilung beobachtbarer Ergebnisse über definierte: - KI-Produkte und Nutzeroberflächen, - Nutzer oder Stichprobeneinheiten, - Länder, - Sprachen und Regionen, - Sitzungs- und Personalisierungszustände, - Prompts, - und Messzeiten behandelt werden. Ein einzelner aggregierter Wert DARF die Offenlegung von: - Umfang und Abdeckung, - Ergebnisverteilung, - Untergruppenvariationen, - Vorkommen kritischer Fehler, - unbekannter und nicht getesteter Masse, - zeitlicher Stabilität, - Gewichtungsregeln, - und Unsicherheit NICHT ersetzen. Gleiche Durchschnitte DÜRFEN NICHT als gleichwertige Repräsentationsqualität interpretiert werden, wenn ihre Untergruppen-, Schwere-, Abdeckungs- oder Stabilitätsverteilungen sich wesentlich unterscheiden. Ungetestete Gruppen DÜRFEN NICHT als konform eingeordnet werden, und seltene, schädliche Ergebnisse MÜSSEN weiterhin getrennt von der durchschnittlichen Leistung sichtbar bleiben.

KAPITEL 03

Welche Entität wird eigentlich geprüft?

METHODISCHE GRUNDLAGEN / K10; K11

Kapitelgrenze

Abschnitt 1 hat die folgende Bestimmung eingeführt:

> *Eine einzelne Antwort ist keine GEO-Punktzahl.*

Abschnitt 2 fügt folgende Bestimmung hinzu:

> *Die Repräsentation einer Entität in generativen Systemen ist eine Verteilung, die sich unter Nutzern, KI-Produkten, Ländern, Sprachen, Sitzungen und Zeiten ergibt.*

Dieser Abschnitt beantwortet die nächste obligatorische Frage:

> *Wessen oder welche Repräsentation wird von dieser Verteilung genau gemessen?*

Ein Unternehmen? Eine Marke? Ein Domainname? Ein Produkt? Eine Dienstleistung? Ein Gründer? Eine Unternehmensgruppe? Eine lokale Rechtsperson in einem bestimmten Land? Ein Franchisegeschäft? Wenn ein künstliches Intelligenzsystem nur die apple.com-Domain sieht:

- den Domainnamen,
- die Marke,
- das Mutterunternehmen,
- die lokale Vertriebsstelle,
- die Produktfamilie,
- die Ladenservice

Wird das System all diese als dieselbe Entität behandeln? Es sei denn, die Unterscheidung wird vor Beginn der Prüfung vorgenommen, kann ein Ergebnis korrekt berechnet, aber auf das falsche Objekt angewendet werden. Dieses Kapitel:

- die geprüfte Einheit,
- der Identitätsanker, der im Prompt verwendet wird,
- die vom System analysierte Entität,
- die Einheit, auf die die Konformitätsentscheidung angewendet wird,
- Beziehungen zwischen Entitäten,
- welcher Entität die Attribute gehören,
- zeitliche und geografische Identitätsgrenzen

definiert. Dieses Kapitel behandelt noch nicht:

- wie alle Aussagen über die Entität überprüft werden,
- wie Schiedsrichter atomare Aussagen bewerten werden,
- Länder- und Sprachbelegquoten,
- die endgültige NOMOS-Bewertungsformel

ist nicht abgeschlossen. Die Aufgabe von Abschnitt 3 ist grundsätzlicher:

> *Die Identität und die Grenzen des zu messenden Objekts vor Beginn der Messung festzulegen.*

NOMOS Herausforderung

Sie stellen mir diese Frage: "Was für ein Unternehmen ist Apple.com?" In der Frage können mindestens drei verschiedene Objekte gemeint sein:

- der Domainname apple.com
- die Marke Apple
- das rechtliche Unternehmen, das mit dem Domainnamen und der Marke verbunden ist

Ich würde eines davon auswählen. Vielleicht das richtige. Vielleicht das falsche. Vielleicht kombiniere ich mehr als eines in einer einzigen Antwort. Dann würden Sie die Antwort bewerten. Aber Sie haben mir im Voraus nicht gesagt, welche Entität Sie bewerten. Ich könnte:

- den Domainnamen als einen Online-Shop beschrieben haben,
- die Marke als eine Verbraucherelektronikmarke beschrieben haben,
- das rechtliche Unternehmen als ein globales Unternehmen beschrieben haben,

Ich könnte die Produkte so aufgelistet haben, als ob das Unternehmen sie selbst präsentieren würde. Meine Antwort könnte flüssig erscheinen. Aber wenn das Prüfobjekt unklar ist, bleibt auch unsicher, ob es korrekt ist. Nun betrachten Sie ein komplexeres Beispiel.

Eine Marke wird von der Muttergesellschaft in einem Land betrieben; von einer Tochtergesellschaft in einem anderen Land; und von einem unabhängigen Vertriebspartner in einem dritten Land. Ein Nutzer fragt: „Bietet diese Marke Dienstleistungen in der Türkei an?" Welche Einheit sollte antworten? Die globale Website der Marke? Das lokale Unternehmen in der Türkei? Der Vertriebspartner?

Der Franchise-Betreiber? Wird die von der globalen Gesellschaft gehaltene Lizenz automatisch auf die lokale Franchise übertragen? Beweisen Kundenbewertungen des lokalen Betriebs die globale Qualität der Hauptmarke? Kann die Anzahl der Mitarbeiter der Muttergesellschaft als Kapazität des lokalen Unternehmens herangezogen werden?

Kann eine Auszeichnung eines Produkts auf alle Unternehmen übertragen werden, die es verkaufen? Wird die persönliche Erfahrung eines Gründers zum Unternehmensalter des Unternehmens? Werden der vorherige Eigentümer und der neue Eigentümer eines Domainnamens als dieselbe Einheit betrachtet?

Wenn ein Unternehmen ein anderes Unternehmen übernimmt, werden dann alle Errungenschaften des vorherigen Unternehmens zu einer Eigenschaft des neuen Unternehmens? Der GEO-Score kann generiert werden, ohne diese Fragen zu beantworten. Es ist jedoch nicht bekannt, wofür der Score steht. Die erste Regelung dieses Abschnitts lautet wie folgt:

> *Eine Einheit kann nicht gemessen werden, bis sie geklärt ist.*

Ihre zweite Bestimmung lautet:

> *Den gleichen Namen zu teilen bedeutet nicht, dieselbe Einheit zu sein.*

Seine dritte Bestimmung besagt:

> *Eine Beziehung ermöglicht nicht automatisch, dass Eigenschaften zwischen den Parteien übertragen werden.*

1. ZWECK DES KAPITELS

Der Zweck dieses Abschnitts ist die Bestimmung der Identität, der Grenze und der Beziehungen der innerhalb von GEO-1000 gemessenen Einheit, bevor die Ergebnisse gesehen werden. Der Abschnitt trifft die folgenden normativen Unterscheidungen:

- Domainname und rechtliches Unternehmen
- Marke und Markeninhaber
- Muttergesellschaft und Tochtergesellschaft
- Gruppe und Gruppengesellschaft
- Produkt und Hersteller
- Dienstleistung und Dienstleister
- Gründer und Institution
- Mitarbeiter und Institution
- Lizenzinhaber und die Partei, die die Marke nutzt
- Franchisegeber und Franchisenehmer
- Hersteller und Vertriebspartner
- Plattform und Verkäufer auf der Plattform
- Technologieanbieter und Technologiebesitzer
- Kunde und Geschäftspartner
- Aktive Beziehung und historische Beziehung
- Offizieller Name und beschreibende Übersetzung
- Verschiedene Einheiten mit demselben Namen
- Vorherige und nachfolgende Versionen einer Einheit
- Geprüftes Objekt und Objekt, auf das die Konformitätsentscheidung angewendet wird

Am Ende dieses Abschnitts sollte jede GEO-Prüfung in der Lage sein, eine einzige und nachvollziehbare Antwort auf die folgende Frage zu geben:

> *Genau welche Einheit, mit welchen Beziehungen, für welchen Zeitraum und für welche digitalen Oberflächen gehört zu diesem Score und dieser Entscheidung?*

2. ZENTRALE NORMATIVE BESTIMMUNG

> *Bevor die Datenerhebung beginnt, muss jede GEO-1000-Prüfung eine eindeutige geprüfte Einheit, einen kanonischen Entity-Datensatz, eine Umfangsgrenze, eine Zeitgrenze und ein Beziehungsdiagramm definieren.*

Keines der Folgenden ist für sich allein ausreichend, um eine Einheit zu identifizieren:

- Name,
- Logo,
- Domainname,
- Social-Media-Konto

Diese Identifikatoren alleine definieren nicht das Prüfobjekt. Die geprüfte Einheit muss, sofern relevant, durch mindestens die folgenden Elemente beschrieben werden:

- Einheitstyp
- Kanonischer Name
- Rechtlicher Name
- Markennamen
- Offizielle lokalisierte Namen
- Abkürzungen und Aliase
- Gründungs- oder Beginn-Datum
- Zuständigkeitsbereich
- Registrierungsnummer oder gleichwertige offizielle Identität
- Beziehungen zwischen Mutter- und Tochtergesellschaften
- Markeninhaberschaft
- Domainnamen-Kontrolle
- Produkt- und Dienstleistungsinhaberschaft
- Verantwortlich für die tatsächliche Leistungserbringung
- Kläger
- Evidenzinhaber
- Aktive Länder und Standorte
- Gültigkeitsdauer
- Historischer Name und Beziehungen
- In die Prüfung einbezogene Flächen
- Einheiten und Flächen außerhalb des Geltungsbereichs

Der Haupt-GEO-Score sollte nicht veröffentlicht werden, bis dieser Datensatz abgeschlossen ist.

3. WAS IST EINE ENTITÄT?

In diesem Protokoll ist eine Entität ein Objekt, dem eine separate und nachvollziehbare Identität, Eigenschaft, Beziehung, Verantwortung oder Zeitstatus zugewiesen werden kann. Eine Entität:

- menschlich,
- juristische Person,
- öffentliche Institution,
- Marke,
- Produkt,
- Dienstleistung,
- Domainname,
- digitale Anwendung,
- Unternehmensgruppe,
- physische Niederlassung,
- Programm,
- Zertifikat,
- Publikation,

- Forschungsprojekt

Es ist möglich. Nicht jeder der genannten Punkte muss eine separate Entität sein. Wenn jedoch einer der folgenden Bereiche wesentlich geändert wird, könnte ein separater Entitätseintrag erforderlich sein:

- Rechtliche Verantwortung
- Eigentum
- Lizenz
- Produkt- oder Dienstleistungserbringung
- Zuständigkeitsbereich
- Zeitraum
- Verpflichtung gegenüber dem Nutzer
- Evidenzmittel und Aussageverantwortung
- Prüfungsumfang

4. SECHS GETRENNTEN OBJEKTE IN EINER GEO-PRÜFUNG

In einer GEO-Prüfung verbirgt das Wort „Einheit“ oft sechs getrennte Funktionen. Diese sechs Objekte müssen voneinander unterschieden werden.

4.1. Prompt-Anker

Ein Prompt-Anker ist ein Ausdruck oder Zeichen, das verwendet wird, um die Existenz in der Abfrage eines Nutzers aufzurufen. Beispiele:

- Firmenname
- Markenname
- Domainname
- Name des Gründers
- Produktname
- Anwendungsname
- Lokaler Firmenname
- Abkürzung

Beispiel: „Was für eine Firma ist Apple.com?“ Der Prompt-Anker hier: apple.com ist der Domainname. Der Prompt-Anker muss nicht die Entität selbst sein. Er ist lediglich der Einstiegspunkt, den das System verwendet, um die Entität aufzulösen.

4.2. Zielentität

Die Zielentität ist die Entität, die das Protokoll tatsächlich messen möchte. Beispiel: Die kanonische Firmenentität, die mit dem Domainnamen apple.com verbunden ist, kann ins Visier genommen werden. Ihre genaue Identität muss jedoch vor dem Audit aufgezeichnet werden. Die Zielentität darf nicht gewählt werden, indem man die Ergebnisse nach der Abfrage betrachtet.

4.3. Vom System aufgelöste Entität

Eine aufgelöste Entität ist die Interpretation des KI-Produkts des Abfrageankers als Entität. Repräsentation:

```
E_i
```

kann verwendet werden. Die Zielentität:

E*

ist die Identitätsauflösungsbewertung, die die grundlegende Frage stellt:

```
E_i= E*?
```

Gleichheit ist jedoch nicht immer nur eine Gleichheit des Namens. Das System kann die richtige Marke und das richtige Unternehmensobjekt erkennen, ohne die richtige juristische Person zu benennen. Daher kann die Identitätsabgleichung später:

- exakte Übereinstimmung,
- akzeptable Markenübereinstimmung,
- teilweise Übereinstimmung,
- falsche Entität,
- unsicher,
- Zusammenschluss mehrerer Einheiten

klassifiziert werden.

4.4. Gegenstand der Forderung

Der Gegenstand der Forderung ist die Einheit, zu der ein bestimmtes Merkmal in der Antwort gehört. Innerhalb derselben Antwort:

- Gründungsdatum zum Mutterunternehmen,
- Produktmerkmal zum Produkt,
- Kundenbewertung für lokalen Dienstleister,
- Lizenz für eine bestimmte juristische Person,
- Erfahrung des Gründers zu einer Person

könnte gehören. Selbst wenn die Antwort mit der richtigen Zielentität begann, könnte sie Merkmale auf falsche Objekte übertragen. Daher sollte jede faktische Aussage in der folgenden Struktur bewertet werden:

```
(Subjekt, Beziehung oder Attribut, Wert, Zeit, Umfang)
```

Beispiel: „X hat eine Y-Lizenz.“ Hier sollte festgestellt werden, zu welcher:

- juristische Person,
- Land,
- während welcher Daten,
- für die Aktivität

es gehört.

4.5. Prüfungsobjekt

Das Prüffobjekt ist eine definierte Entität oder ein Paket von Entitäten, bei dem Nachweise, Inhalte, technische Oberflächen und KI-Ergebnisse gemeinsam untersucht werden. Ein Prüffobjekt kann eine einzelne Entität sein. Beispiel: Ein spezifisches rechtliches Unternehmen. Oder es kann ein Paket mit klaren Grenzen sein: Hauptmarke + kanonischer Domainname + festgelegte zwei Produkte + türkische und englische öffentliche Oberflächen. Wenn ein Paket verwendet wird, müssen die darin enthaltenen Entitäten ihre separaten Identitäten behalten. „Paket“ bedeutet nicht, dass alle Attribute gegenseitig übertragen werden.

4.6. Konformitätseinheit

Die Konformitätseinheit ist das spezifische Objekt, auf das die Prüfungsentscheidung rechtlich und normativ angewendet wird. Eine Konformitätsentscheidung:

- gilt für die gesamte Unternehmensgruppe,
- gilt für alle Sprachen,
- gilt für alle Domainnamen,
- für alle Produkte

kann nicht automatisch angewendet werden. Beispiel: „Die türkischen und englischen öffentlichen Seiten der Marke NobleJackal auf example.com wurden gemäß dem angegebenen Standard und innerhalb des angegebenen Datumsbereichs als konform befunden.“ Diese Entscheidung:

- alle internen Prozesse des Unternehmens,
- verbundene Unternehmen,
- separate Anwendungen,
- zukünftig hinzuzufügende Sprachen,
- Produkte außerhalb des Umfangs

werden nicht automatisch abgedeckt.

5. ARTEN VON EINHEITEN

Jede Prüfungseinheit muss mit einem oder mehreren Typen klassifiziert werden. Primärer Typ und sekundäre Typen können unterschieden werden.

5.1. Juristische Person

Es handelt sich um ein Unternehmen, das nach dem Gesetz oder einschlägiger gesetzlicher Vorschrift über eigenständige Rechte, Pflichten oder Registrierungen verfügt. Ihre Felder, soweit sie relevant sind:

- Rechtlicher Name
- Unternehmensart
- Ort der Registrierung
- Registrierungsnummer
- Steuer- oder gleichwertige Identität
- Gründungsdatum
- Aktiver Status
- Bevollmächtigter Vertreter
- Haupt- und Tochtergesellschaftsverhältnisse

Eine juristische Person kann denselben Namen wie die Marke tragen. Sie ist jedoch nicht dasselbe Objekt.

5.2. Marke

Es ist der Name, der auf dem Markt für ein Produkt, eine Dienstleistung oder die Unternehmensidentität verwendet wird. Marke:

- kann keine eigene Rechtspersönlichkeit haben,
- kann im Eigentum eines anderen Unternehmens stehen,
- kann unter Lizenz genutzt werden,

kann von mehreren Unternehmen in verschiedenen Ländern betrieben werden. Die tatsächlich für die von einer Marke beanspruchte Arbeit verantwortliche Lieferung muss ebenfalls angegeben werden.

5.3. Unternehmensgruppe

Es handelt sich um eine Struktur, die aus mehreren Rechtseinheiten besteht, die gemeinsame Eigentums- oder Kontrollverhältnisse haben. Eine „Gruppe“ kann die folgenden Merkmale nicht automatisch auf alle Mitglieder übertragen:

- Lizenz
- Zertifikat
- Anzahl der Kunden
- Anzahl der Mitarbeiter
- Umsatz
- Büro
- Prüfungsergebnis
- Sicherheitsstatus
- Konformitätszeichen

Wenn der gruppenweite Wert tatsächlich konsolidiert ist, sollte dies klar angegeben werden.

5.4. Tochtergesellschaft oder assoziiertes Unternehmen

Eine juristische Person, die unter der Eigentums- oder Kontrollbeziehung einer anderen Einheit steht. Eine Tochtergesellschaft:

- kann eine separate Vertragspartei sein,
- ein separater Lizenznehmer sein,
- ein separater Arbeitgeber sein,
- ein separater Datenverantwortlicher sein

kann. Die Eigenschaften der Muttergesellschaft werden standardmäßig nicht auf die Tochtergesellschaft übertragen.

5.5. Domainname

Es ist eine digitale Adresse und Veröffentlichungssfläche. Domainname:

- ist keine juristische Person,
- erbringt keine Dienstleistungen eigenständig,
- ist nicht verpflichtet, eine Marke zu besitzen,

kann zu bestimmten Zeiten von verschiedenen Eigentümern oder Betreibern kontrolliert werden. Ein Domainname kann der Ausgangspunkt einer Prüfung sein. Es sollte jedoch auch gezeigt werden, unter welcher Einheit die Entscheidung getroffen wurde.

5.6. Webseite oder digitale Veröffentlichung

Es handelt sich um den gesamten spezifischen Inhalt und die Anwendungen auf einem Domainnamen. Derselbe Domainname kann:

- Unternehmensseite,
- Shop,
- Support-Portal,
- Investorenseite,

- Entwicklerdokumentation,
- verschiedene Länder- und Sprachversionen

tragen. Die Inhaltseigentümer und die Verantwortungsbereiche dieser können unterschiedlich sein.

5.7. Produkt

Ein kommerzielles oder technisches Objekt mit einer spezifischen Funktion, Modell oder Version. Die:

- Preis,
- Sicherheit,
- technische Merkmale,
- Zertifikat,
- Nutzerfreundlichkeit

eines Produkts können nicht als allgemeine Eigenschaften des Unternehmens verwendet werden. Versionen innerhalb derselben Produktfamilie können ebenfalls unterschiedlich sein.

5.8. Dienstleistung

Aktivität, die unter einem bestimmten Umfang, Nutzer, Land, Liefermethode und Verantwortung angeboten wird. Ein Service:

- vermarktbar durch die Marke,
- abrechenbar durch eine andere Rechtseinheit,

lieferbar durch einen unabhängigen Partner. Die Prüfung sollte unterscheiden, wer:

- es verkauft hat,
- es geliefert hat,
- es gewährleistet hat,
- rechtlich dafür verantwortlich ist

sollte getrennt werden.

5.9. Person

ist Gründer, Führungskraft, Mitarbeiter, Berater, Experte oder öffentlich anerkannter Vertreter. Die:

- Ausbildung,
- Erfahrung,
- Lizenz,
- Meinung

einer Person wird nicht automatisch auf die Institution übertragen. Ebenso wird der Preis oder das Zertifikat der Institution nicht als berufliche Kompetenz der Person betrachtet.

5.10. Filiale oder physischer Standort

Es ist ein physischer Servicestandort, der mit der Hauptinstitution verbunden ist oder separat betrieben wird. Eine Filiale:

- Teil des Mutterunternehmens,
- separate juristische Person,
- Franchise,
- unabhängiger lizenzierter Betreiber

kann sein. Die Positionsbeziehung zeigt nicht allein die rechtliche Beziehung.

5.11. Franchise

Es kann ein unabhängiges Unternehmen sein, das unter einem Marken-nutzungsrecht oder Geschäftsmodell-Lizenz betrieben wird. Der Franchisegeber muss:

- Reputation,
- Kontrolle,
- Schulung

nicht automatisch jede Handlung des Franchise-Nehmers garantieren. Ein lokaler Fehler des Franchise-Nehmers kann nicht automatisch auf die gesamte globale Marke verallgemeinert werden. Die Beziehung sollte mit ihrem Umfang gemeldet werden.

5.12. Vertriebsunternehmen, Händler oder Verkäufer

Die Partei, die das Produkt oder die Dienstleistung in einem bestimmten Markt verkauft oder vertreibt, unabhängig davon, ob sie unabhängig oder affiliiert ist. Die Lizenz, Kontrolle oder unternehmerische Compliance des Herstellers wird nicht automatisch auf den Vertriebspartner übertragen. Die Leistungsfähigkeit des Kundenservice des Verkäufers ist ebenfalls nicht dasselbe Maß wie die Produktqualität des Herstellers.

5.13. Plattform und Plattformverkäufer

Der Marktplatz oder die Plattform und die unabhängigen Verkäufer auf der Plattform sind getrennte Einheiten. Plattform:

- Zahlung,
- Auflistung,
- technische Infrastruktur

kann bereitgestellt werden. Anbieter:

- Produkt,
- Lieferung,
- Garantie,
- Kundenbeziehungen

können verantwortlich sein. Wenn die KI diese Rollen kombiniert, kann es zu einer falschen Verantwortungserteilung kommen.

5.14. Programm, Abzeichen oder Zertifikat

Angeboten von einer Institution:

- Schulungsprogramm,
- Konformitätszeichen,
- Zertifikat,
- Mitgliedschaft,
- Abzeichen

Es kann separate Entitätsdatensätze enthalten. Die Institution, die das Programm veröffentlicht, ist nicht dieselbe wie die Institution, die am Programm teilnimmt. Die Funktion des Abzeichens überträgt sich nicht auf alle Funktionen der Institution, die das Abzeichen trägt.

6. ENTITÄTSGRAPH

Ein einzelner Entitätsdatensatz reicht möglicherweise nicht aus, um komplexe Beziehungen zu erklären. Aus diesem Grund muss für jedes Audit ein Entitätsgraph erstellt werden. Ansichten:

G_E = (V, R)

Hier:

- V: Entitätsknoten
- R: gerichtete Beziehungen zwischen Knoten

Beziehungen sollten durch explizite Verben definiert werden.

6.1. Grundtypen von Beziehungen

Soweit relevant, können die folgenden Beziehungen verwendet werden:

- besitzt — besitzt
- ownedBy— gehört zu
- kontrolliert — kontrolliert
- controlledBy— wird kontrolliert von
- betreibt — betreibt
- operatedBy— wird betrieben von
- veröffentlicht — veröffentlicht
- publishedBy— wird veröffentlicht von
- stellt her — stellt her
- manufacturedBy— wird hergestellt von
- providesService— bietet Service
- serviceDeliveredBy— Service wird geliefert von
- licencesBrandTo— lizenziert Marke an
- franchisesTo— Lizenzen an
- distributesFor— vertreibt für
- verkauft — verkauft
- beschäftigt — beschäftigt
- foundedBy— wurde gegründet von
- memberOf— ist Mitglied von
- certifiedBy— ist im angegebenen Umfang zertifiziert von
- auditedBy— wurde auditiert von
- partneredWith— hat eine Partnerschaft im angegebenen Umfang mit
- customerOf— ist Kunde von
- supplierTo— ist Lieferant von
- formerlyKnownAs— ehemals bekannt als
- predecessorOf— ist ein Vorgänger von
- successorTo— ist ein Nachfolger von
- acquiredBy— wurde erworben von
- mergedInto— wurde eingegliedert in
- ceasedOn— hat den Betrieb eingestellt am

Vage Verben wie „verwandt“, „Partner“, „zusammenarbeitend“ sollten möglichst nicht allein verwendet werden.

6.2. Richtung der Beziehung

Beziehungen sind richtungsgebunden. 'A stellt B Technologie zur Verfügung.' ist nicht dasselbe wie: 'B stellt A Technologie zur Verfügung.' Eine Marke kann einem Unternehmen gehören. Das Unternehmen gehört nicht der Marke. Ein Kunde hat möglicherweise mit einer Agentur zusammengearbeitet. Die Agentur gehört nicht dem Kunden. Wenn die Richtung nicht angegeben ist, kann das System falsche Verantwortungs- und Befugnisübertragungen vornehmen.

6.3. Zeit der Beziehung

Jede finanzielle Beziehung, soweit sie relevant ist:

- Startdatum,
- Enddatum,
- Aktivitätsstatus,
- Datum der letzten Überprüfung

muss enthalten. Eine alte Partnerschaft ist keine aktive Partnerschaft. Ein ehemaliger Mitarbeiter ist kein aktueller Geschäftsführer. Ein einmal genutzter Vertriebspartner ist heute möglicherweise kein autorisierter Vertreter mehr.

6.4. Umfang der Beziehung

Eine Partnerschaft:

- kann auf ein bestimmtes Produkt beschränkt sein,
- Land,
- Kampagne,
- Datum,
- Ereignis

Es kann nicht verallgemeinert werden, dass eine Einzelprojekt-Zusammenarbeit für alle Unternehmensaktivitäten gilt.

7. ANGENOMMENE ÜBERTRAGUNG VON EIGENSCHAFTEN IST VERBOTEN

Die Existenz einer Beziehung zwischen Organisationen überträgt nicht automatisch eine Eigenschaft von der einen auf die andere. Eines der Grundprinzipien dieses Abschnitts ist:

> *Die Übertragung von Eigenschaften ist standardmäßig verboten; sie kann jedoch mit expliziter Relation, Umfang, Zeitpunkt und Nachweis erlaubt werden.*

7.1. Lizenzübertragung

Die Lizenz des Mutterunternehmens:

- an eine Tochtergesellschaft,
- an ein Franchise,
- an einen Distributor,
- an eine Marke

überträgt sich nicht automatisch. Der offizielle Umfang der Lizenz muss geprüft werden.

7.2. Zertifikatübertragung

Das Zertifikat eines Produkts kann nicht auf die gesamte Produktfamilie übertragen werden. Das Qualitätszertifikat eines Büros kann nicht auf die globale Gruppe verallgemeinert werden. Die Konformitätsentscheidung eines Unternehmens beweist nicht die persönliche Expertise seiner Mitarbeiter.

7.3. Übertragung der Mitarbeiterzahl

Die Gesamtzahl der Gruppe kann nicht wie die Anzahl der Mitarbeiter eines lokalen Unternehmens verwendet werden. Das Agenturnetzwerk oder der Freelancer-Pool können nicht als Anzahl direkt beschäftigter Mitarbeiter dargestellt werden.

7.4. Übertragung der Kundenzahl

Die Anzahl der Kunden der Hauptgruppe darf nicht als Anzahl der Kunden einer neu gegründeten Marke angegeben werden. Die Anzahl der Nutzer auf einer Plattform muss nicht der Anzahl aktiver Kunden eines bestimmten Produkts entsprechen.

7.5. Übertragung des Gründungsdatums

Das Gründungsjahr des Gründers in der Branche ist nicht das Gründungsdatum des Unternehmens. Ein Unternehmen kann das Gründungsdatum einer alten Marke, die es erwirbt, nicht als sein eigenes rechtliches Gründungsdatum verwenden. Das Marken-Datum und das Datum der juristischen Person sollten getrennt dargestellt werden.

7.6. Erfolg und Fallübergabe

Ein Projekt, das von einem Mitarbeiter beim vorherigen Arbeitgeber durchgeführt wurde, ist nicht die Unternehmensfallstudie des neuen Unternehmens. Der Erfolg eines erworbenen Unternehmens wird ohne Erklärung zur Integration und zum Umfang nicht als direkter Erfolg der neuen Gruppe betrachtet.

7.7. Übertragung der Berechtigung

Für bestimmte Sprachseiten eines Domainnamens gewährte Berechtigungen:

- an die gesamte Unternehmensgruppe,
- an neue Produkte,
- an andere Domainnamen,
- an zukünftige Seiten

werden nicht übertragen.

8. VIER GETRENNTE EIGENTUMS-EBENEN

In einer digitalen Repräsentation kann das Wort „Eigentümer“ mehrere Bedeutungen haben. Es sollten vier Ebenen unterschieden werden.

8.1. Rechtliches Eigentum

Die Einheit ist Eigentümer nach Gesetz oder offizieller Registrierung.

8.2. Marken-Eigentum

Sie ist der Eigentümer der Rechte an der Marke oder der Handelsidentität.

8.3. Digitale Oberflächenkontrolle

Sie ist die Partei, die tatsächlich die Domain, die Website, das Social-Media-Profil oder die Anwendung kontrolliert. Der rechtliche Eigentümer und der technische Kontrolleur können unterschiedlich sein.

8.4. Eigentum an Inhalten und Aussagen

Es ist die Partei, die für die Genauigkeit des veröffentlichten Aussageinhalt verantwortlich ist. Die Agentur kann die Seite verwalten. Der tatsächliche Eigentümer des Preis- oder Lizenzanspruchs kann jedoch die Kundeninstitution sein. Ein technischer Anbieter wird nicht zum materiellen Eigentümer der Aussage, nur weil er den Inhalt veröffentlicht hat.

9. WER VERKAUFT DEN SERVICE, WER LIEFERT IHN?

Dies ist eine der wichtigsten Unterscheidungen von Einheiten in der GEO-Bewertung. In einer Servicekette können die folgenden Parteien unterschiedlich sein:

- Derjenige, der die Marke verwendet
- Derjenige, der wirbt
- Derjenige, der das Angebot vorbereitet
- Derjenige, der den Vertrag unterzeichnet
- Derjenige, der die Rechnung ausstellt
- Derjenige, der den Service tatsächlich liefert
- Derjenige, der die Garantie bereitstellt
- Derjenige, der die Daten verarbeitet
- Derjenige, der die Beschwerde bearbeitet
- Derjenige, der die rechtliche Verantwortung trägt

Eine KI kann all diese Rollen einer einzelnen Partei zuweisen, indem sie nur den Markennamen angibt. Dieser Fehler ist besonders häufig in:

- Tourismus,
- Finanzen,
- Gesundheit,
- Beratung,
- Franchise,
- Plattform,
- Agentur,
- Bildung

Bereiche können finanzielle Konsequenzen haben. Jeder Dienstleistungsnachweis sollte diese Fragen beantworten: Mit wem schließt der Nutzer den Vertrag ab? Wer erhält die Zahlung? Wer erbringt die Dienstleistung? Wer trägt die berufliche oder rechtliche Verantwortung? Wer ist für Garantie- und Rückgabeverpflichtungen verantwortlich? Welche Rolle spielt die Marke? Gibt es Drittanbieter? Diese Unterscheidung bestimmt auch, auf welche Einheit der GEO-Score angewendet wird.

10. GRENZE DER EINHEIT

Die Einheitgrenze ist die Linie, die definiert, welche Merkmale, Beziehungen und Oberflächen das geprüfte Objekt einschließt und welche nicht. Eine Einheitgrenze kann in vier Schichten definiert werden.

10.1. Identitätskern

Es ist die grundlegende Identität der Einheit, die unverändert bleiben muss:

- Kanonischer Name
- Einheitstyp
- Rechtliche oder institutionelle Identität
- Haupttätigkeit
- Primäre Eigentümerschaft
- Aktiver Status

10.2. Kontrollierte Umgebung

Direkt vom Unternehmen kontrolliert:

- domainnamen,
- Anwendungen,
- Marken,
- Produkte,
- Dokumente,
- offizielle Profile

Dieses Feld. Kontrolle bedeutet nicht, dass alle Funktionen zu einer einzigen Identität verschmelzen.

10.3. Verbundene Umgebung

Tochtergesellschaften sind lokale Unternehmen, Franchise-Unternehmen und kontrollierte Partner. Sie behalten ihre eigenen Identitäten.

10.4. Externes Beziehungsumfeld

Kunden, Lieferanten, Medien, Mitgliedschaften, Technologieanbieter und unabhängige Partner. Unternehmen in dieser Umgebung sind nicht Teil der geprüften Einheit. Sie sind Objekte der Beziehung.

11. FESTSCHREIBEN DER UNTERNEHMENSGRENZE, OHNE ERGEBNISSE ZU SEHEN

Das Prüfungobjekt kann basierend auf den Ergebnissen weder erweitert noch eingeschränkt werden. Die folgenden Verhaltensweisen sind verboten:

- Positive Gruppenergebnisse zur Marke hinzufügen
- Ein gescheitertes verbundenes Unternehmen vom Prüfungsumfang ausschließen
- Ein gutes Produktergebnis auf die gesamte Produktfamilie übertragen
- Die schlechte Sprachversion als eigenständige Einheit deklarieren

- Den kritischen Fehler der lokalen Operation überlassen, während die globale Marke sauber erscheint
- Starke lokale Nachweise als Nachweise des globalen Unternehmens verwenden

Umfang der Einheit:

- Bevor die Prüfung beginnt,
- Bevor Aufforderungen verteilt werden,
- Bevor Ergebnisse gesehen werden

Muss zuerst versioniert werden.

12. KANONISCHER EINHEITSREKORD

Für jede geprüfte Einheit muss ein kanonischer Entitätsdatensatz erstellt werden. Dieser Datensatz klärt, welches Objekt die Prüfer und Systeme bewerten.

12.1. Mindestfelder

Identität

NOMOS Einheiten-ID Kanonischer Name Einheitstyp Rechtlicher Name Abkürzungen Offizielle lokale Namen Ehemalige Namen

Rechtlicher und Unternehmensstatus

Rechtsordnung Registrierungsnummer Gründungsdatum Aktiver Status Muttergesellschaft Tochtergesellschaften

Marken und Produkte

Eigene Marken Lizenzierte Marken Produkte Dienstleistungen Versions- oder Modellunterschiede

Digitale Oberflächen

Kanonische Domainnamen Lokale Domainnamen Anwendungen Offizielle Profile Maschinell lesbare Datensätze

Verantwortung

Anspruchsteller Inhaltsinhaber Technische Oberflächenverantwortlicher Servicebereitstellungsverantwortlicher Rechtlich Verantwortlicher

Zeit

Datensatzversion Wirksamkeitsbeginn Gültigkeitsende Letztes Überprüfungsdatum Historische Änderungen

Umfang

Geprüfte Länder Geprüfte Sprachen Geprüfte Produkte Ausgeschlossene Einheiten und Flächen

12.2. Einheitenstatus

Jeder Eintrag muss einen der folgenden Status tragen:

AKTIV

INAKTIV

HISTORISCH

ERSATZT

ZUSAMMENGEFÜGT

ERWORBEN

AUFGELÖST

VORLÄUFIG

BESTRITTEN

UNBEKANNT

UNBEKANNTE oder BESTRITTENE Identität kann nicht stillschweigend endgültig geklärt werden.

13. IDENTITÄTSNIVEAUS

Die Identitätsbestätigung eines Eintragssatzes kann nach den verwendeten Quellen klassifiziert werden.

ID-0— Identität unbekannt

Es existiert nur ein Name oder eine Domain. Rechtliche oder organisatorische Übereinstimmungen sind nicht überprüft.

ID-1— Erstanbieterbehauptung

Die Einrichtung identifiziert sich auf eine bestimmte Weise. Sie ist wertvoll. Sie ist keine unabhängige Überprüfung.

ID-2— Unterstützte organisatorische Identität

Erstanbieteraufzeichnungen, der Domainname und andere verlässliche Aufzeichnungen stimmen miteinander überein.

ID-3— Verifizierte Identität mit offiziellem Nachweis

Die juristische Person wurde anhand geeigneter offizieller oder verlässlicher Unterlagen überprüft.

ID-4— Multi-Oberflächen- und Beziehungs-verifizierte Identität

Identität, Marke, Produkt, Domainname und Unternehmensbeziehungen sind über mehrere verlässliche Aufzeichnungen hinweg konsistent.

ID-5— Über die Zeit aufrechterhaltene Identität

Identität und Beziehungen wurden zu verschiedenen Zeitpunkten überprüft, und die Änderungshistorie wird aufbewahrt. Diese Stufen sind nicht die endgültige Berechtigungsbewertung. Sie geben das Vertrauen in die Nachweise des Entitätsdatensatzes an.

14. VERSCHIEDENE ENTITÄTEN MIT DEM GLEICHEN NAMEN

Verschiedene Entitäten mit denselben oder ähnlichen Namen können existieren. Diese Situation erzeugt Entitätskollisionen. Beispiele für Unterscheidungen:

- Unternehmen mit demselben Namen
- Institutionen, die dieselbe Abkürzung verwenden
- Dasselbe Markenname in verschiedenen Ländern
- Alte und neue Firma
- Markenname mit persönlichem Namen
- Firmenname mit Produktname
- Marke mit allgemeinem Wort

Audit kann sich nicht ausschließlich auf Namensabgleich verlassen.

14.1. Kollision Resolver Felder

Gerichtsstand Rechtsform des Unternehmens Domainname Physische Adresse Gründungsdatum Gründer oder Geschäftsführer Produkt und Dienstleistung Markeninhaber Registrierungsnummer Lokaler Name

14.2. Mehrdeutige Aussage

Wenn sich eine Aussage auf mehrere vernünftige Entitäten beziehen kann, gibt es drei Möglichkeiten: Die Aussage wird geklärt. Separat als Entitätsauflösungstest klassifiziert. Das System soll eine Klärung anfordern. Die aus der mehrdeutigen Aussage erhaltene Antwort kann nicht stillschweigend zur Hauptgenauigkeit hinzugefügt werden.

15. EINE ENTITÄT, MEHR ALS EIN DOMAINNAME

Ein Unternehmen:

- globale Domain,
- lokale Länderdomains,
- Produktdomains,
- Investor-Domain,
- Support-Domain,
- alte Domains

kann verwenden. Nicht alle diese Domains fallen automatisch in denselben Prüfungsumfang. Für jede Domain:

- Controller,
- Inhaltsinhaber,
- Zielbenutzer,

- Sprache,
- Produktumfang,
- Aktivitätsstatus

Es muss bestimmt werden. Ein korrekter Eintrag in einer Domain korrigiert nicht automatisch den Fehler in einer anderen Domain.

16. EIN DOMAINNAME, MEHRERE EINHEITEN

Auf einem einzelnen Domainnamen:

- Muttergesellschaft,
- verschiedene Marken,
- lokale Unternehmen,
- Produkte,
- Partner,
- Verkäufer

Es kann vorkommen. In diesem Fall sollte die Eigentümerschaft der Einheit basierend auf der URL oder der Content-Familie definiert werden. Beispiel:

Oberfläche	Existenzrolle
Startseite	Globale Marke
Investor Relations	Rechtliche Muttergesellschaft
Lokale Preisseite	Lokale Vertriebsorganisation
Marktplatz	Unabhängige Verkäufer
Karriereseite	Spezifizierte Arbeitgebergesellschaften
Support-Portal	Produkt- oder Dienstleistungsanbieter

Ein einzelner Domainname bedeutet nicht einzelne Verantwortlichkeit.

17. TEMPORÄRE EINHEITSAUFTRAGSIDENTITÄT

Die Identität der Einheit kann sich im Laufe der Zeit ändern. Fallbeispiele:

- Änderung des Firmennamens
- Marke auf ein anderes Unternehmen übertragen
- Fusion
- Übernahme
- Abteilung
- Beendigung der Tätigkeit
- Änderung des Domainnamens
- Übertragung des Produkts an ein anderes Unternehmen
- Austritt des Gründers
- Änderung des lokalen Betreibers

Aus diesem Grund ist die Registrierung von Einheiten zeitabhängig. Sie kann als E(t) betrachtet werden.

17.1. Historische Identität und aktuelle Identität

Eine Beziehung, die in der Vergangenheit korrekt war, kann nicht in die Gegenwart übernommen werden. Beispiel: „X wurde von Y in 2022 betrieben.“ ist historisch. Der folgende Satz ist anders: „X wird von Y betrieben.“

17.2. Funktionsübertragung nach der Akquisition

Akquisition:

- Eigentum,
- einige Einheiten,
- können bestimmte Verträge übertragen

. Jedoch:

- alle historischen Erfolge,
- alle Zertifikate,
- alle Kundenbeziehungen,
- alle Lizenzen

werden nicht automatisch auf die neue Einheit übertragen. Welche Funktionen rechtlich und praktisch übertragen werden, muss erklärt werden.

17.3. Domain-Name-Historie

Ein Domainname kann Spuren von Inhalten oder Verweisen des vorherigen Besitzers enthalten. Der neue Eigentümer gilt nicht automatisch als der vorherigen Autorität erlangt zu haben. Wenn das KI-System die alten und neuen Einheiten zusammenführt, tritt historische Einheitenverwechslung auf.

18. MEHRSPRACHIGE EINHEITSIDENTITÄT

Die Namen einer Entität in verschiedenen Sprachen müssen mit derselben Identität verknüpft sein. Nicht alle Übersetzungen sind jedoch offizielle Namen. Die folgenden Typen sollten unterschieden werden:

- Rechtlicher Name
- Markenname
- Offizieller lokalisierter Name
- Transliteration
- Beschreibende Übersetzung
- Abkürzung
- Früherer Name
- Allgemein gebräuchlicher Name

Beispiel: Eine türkische Beschreibung einer Marke muss nicht der rechtliche oder offizielle türkische Name der Marke sein. Der Maschineneintrag sollte diese Typen klar angeben.

19. LAND UND RECHTSORDNUNG KÖNNEN DIE GRENZEN DER EINHEIT ÄNDERN

Unter einer globalen Marke kann es in verschiedenen Ländern unterschiedliche rechtliche oder operative Einheiten geben. Wenn ein Nutzer fragt: „Operiert Unternehmen X in Deutschland?“ sollte die Antwort unterscheiden zwischen:

- dem Bestehen der globalen Marke,
- der lokalen rechtlichen Einheit,
- der Vertriebspartnerbeziehung,
- der Fähigkeit, ferngesteuerte Dienstleistungen bereitzustellen

„Die Marke ist im Land bekannt“ ist nicht dasselbe wie: „Die lokale rechtliche Gesellschaft erbringt Dienstleistungen."

19.1. Länderspezifische Rollen von Einheiten

Für jedes Land, soweit zutreffend:

- Rechtlicher Verkäufer
- Rechnungsaussteller
- Datenverantwortlicher
- Dienstleister
- Unterstützend
- Vertriebspartner
- Franchise
- Lizenzinhaber
- Markenvertreter

muss bestimmt werden.

19.2. Allgemeine Sprachseite ist kein Nachweis für ein lokales Unternehmen

Eine Website auf Deutsch beweist nicht, dass es eine juristische Person oder ein physisches Büro in Deutschland gibt. Eine englische Seite schafft keine weltweite rechtliche Zuständigkeit. Sprache ist keine Zuständigkeit.

20. SYNTHETISCH APPLE.COM PRÄSENZHINWEIS

DEMONSTRATION DER SYNTHETISCHEN METHODOLOGIE / Das folgende Beispiel wurde ausschließlich vorbereitet, um die Methode der Entitätsauflösung zu demonstrieren. Es handelt sich nicht um die tatsächliche Prüfung von Apple Inc., eine aktuelle Überprüfung der Unternehmensstruktur oder das echte Ergebnis eines KI-Produkts. Prompt: „Was für ein Unternehmen ist Apple.com?“

20.1. Prompt-Anker

apple.com

20.2. Kandidatenziele

Domainname und Website Apple-Marke Canonical-Unternehmensstruktur Online-Shop-Funktion Produktfamilie Wenn das Ziel nicht festgelegt ist, bevor die Prüfung beginnt, können Prüfer unterschiedliche Objekte bewerten.

20.3. Vorgeschlagene Ziel-Entity-Sperre

Das Ziel für synthetische Tests kann wie folgt definiert werden: „die zentrale Technologiegesellschaft des Unternehmens, die mit dem kanonischen Domainnamen apple.com verbunden ist und die Identität und Kernaktivitäten des Unternehmens repräsentiert.“ Diese Definition:

- umfasst nicht automatisch die technischen Spezifikationen einzelner Produkte,
- das lokale Unternehmen jedes Landes,
- die Ladenpreise,
- Investitionsberatung

automatisch.

20.4. Synthetische Antwort A

„Apple.com ist eine globale Unternehmenswebsite, die mit Technologieprodukten und digitalen Dienstleistungen verbunden ist.“ Bewertung: Es stimmt nicht vollständig mit der Institution des Domainnamens überein. Es kann die Unternehmensidentität unvollständig lassen. Die Haupttätigkeit könnte im Allgemeinen korrekt sein. Es kann an materieller Identität mangeln.

20.5. Synthetische Antwort B

„Apple.com ist die offizielle Website des Unternehmens namens iPhone.“ Bewertung: Das Produkt und das Unternehmen sind kombiniert. Die Zielentität wird falsch erkannt. Es liegt ein Identitätsfehler vor.

20.6. Synthetische Antwort C

„Apple.com ist nur ein unabhängiges Geschäft, das Handys online verkauft.“ Bewertung: Die Funktion des Domainnamens ist übermäßig eingeschränkt. Die juristische Person und der Tätigkeitsbereich werden verzerrt. Das Merkmal des unabhängigen Geschäfts könnte falsch sein.

20.7. Synthetische Antwort D

„Apple.com ist eine der wichtigsten Unternehmens- und Produktinformationsflächen der Marke Apple; die für die Seite zuständige Rechtseinheit und die lokale Verkaufsseite sollten ebenfalls länderspezifisch bestimmt werden.“ Bewertung: Domainname, Marke und lokale rechtliche Rolle sind getrennt. Die Analyse der Zielentität ist disziplinierter. Andere wesentliche Aussagen der Antwort sollten separat nachgewiesen werden. Dieses Beispiel zeigt:

> *Eine Antwort kann sich auf den falsche Entität konzentrieren, auch wenn sie viele richtige Wörter enthält.*

Die Identitätsgenauigkeit steht vor der Inhaltsgenauigkeit.

21. SYNTHETISCHER KOMPLEXER ENTITÄTSFALL

SYNTHETISCHER FALL – KEINE ECHTE INSTITUTION

Betrachten wir eine globale Marke namens Asteron. Struktur:

- Asteron Holdings Ltd. — Muttergesellschaft

- Asteron Travel — Marke
- Asteron Turkey Tourism Ltd. — lokales Unternehmen
- Mira Destination Services — unabhängiger Partner für lokale Operationen
- Asteron Club Antalya — Franchisebetrieb
- asteron.example — globaler Domainname
- tr.asteron.example — türkische Marketingfläche

Der Nutzer fragt: „Welche Dienstleistungen bietet Asteron in der Türkei an?“ KI-Antwort: „Asteron Holdings Ltd. betreibt alle Touren direkt in Antalya, hat 400 Mitarbeiter in der Türkei, und Asteron Club Antalya ist direkt für die Kundenzufriedenheit in Antalya verantwortlich.“ Realitätspaket: Die Muttergesellschaft besitzt die Marke. Das lokale Unternehmen ist für Verkauf und Vertragsabschlüsse zuständig. Mira übernimmt einen Teil der operativen Tätigkeiten. Asteron Club Antalya ist eine unabhängige Franchise. Die 400 Mitarbeiter sind die Gesamtzahl der Gruppe. Die Franchise ist auf Geschäftsebene direkt für die Kundenzufriedenheit verantwortlich. Fehler in der Antwort: Die Muttergesellschaft wird fälschlicherweise als für die Servicebereitstellung verantwortlich dargestellt. Die Gesamtzahl der Gruppenmitarbeiter wird so dargestellt, als wäre sie die Kapazität des lokalen Unternehmens. Die Verantwortung der Franchise wird direkt der Muttergesellschaft zugewiesen. Marke, Rechtspersönlichkeit und operativer Partner sind kombiniert. Die Antwort könnte die Marke korrekt benannt haben. Die Entitätsgraphik ist jedoch falsch.

Daher kann sich der GEO-Score nicht ausschließlich auf die Nennung des Namens oder die allgemeine Flüssigkeit des Satzes stützen.

22. ERGEBNISSE DER ENTITÄTSAUFLÖSUNG

Für jede Beobachtung kann das Ergebnis der Entitätsauflösung mit einem der folgenden Status aufgezeichnet werden.

ER-0— NICHT BEWERTET

Die Entitätsauflösung wurde nicht bewertet.

ER-1— EXAKTES ZIEL

Die Zielentität wurde korrekt und eindeutig aufgelöst.

ER-2— AKZEPTABLE ÜBEREINSTIMMUNG AUF MARKENEBENE

Es wird kein Rechtsname angegeben, aber die anvisierte Marke und juristische Einheit wurden im Wesentlichen korrekt aufgelöst.

ER-3— TEILWEISE ODER UNTER-SPEZIFIZIERT

Die Antwort ging in die richtige Richtung, es fehlen jedoch rechtliche, geographische oder operative Angaben.

ER-4— VERWESUNG MEHRERER ENTITÄTEN

Mehrere separate Einheiten wurden zu einem einzigen Objekt zusammengeführt.

ER-5— falsche Entität

Eine andere Person, Firma, Produkt oder Domainname wurde ins Visier genommen.

ER-6— MEHRDEUTIG

Aus der Antwort lässt sich nicht bestimmen, welche Einheit aufgelöst wurde.

ER-7— ANGEMESSENE KLARSTELLUNG

Die Prompt ist tatsächlich mehrdeutig, und das System hat korrekt um Klarstellung gebeten. Eine Klarstellungsanfrage muss nicht als Fehler betrachtet werden.

23. EINHEITSAUFLÖSUNG ERFOLG IST VON REPRÄSENTATIONSGENAUIGKEIT UNTERSCHEIDBAR

Ein System kann die richtige Entität analysieren, aber falsche Informationen darüber liefern. Ein anderes System kann einige wesentliche Fakten korrekt angeben, sie aber mit dem falsche Entität verbinden. Es gibt vier grundlegende Situationen:

Einheitenanalyse	Wesentlicher Inhalt	Ergebnis
Korrekt	Korrekt	Starker Kandidat für korrekte Repräsentation
Korrekt	Falsch	Korrektes Objekt, falsches Attribut
Falsch	Teilweise korrekt	Korrekte Informationen, die mit dem falsche Entität verbunden sind
Falsch	Falsch	Vollständige Aufschlüsselung der Repräsentation

Aus diesem Grund kann die Identitätsanalyse ein separates Tor sein. Insbesondere kann eine Antwort, die ernsthafte Verwechslungen mit einer anderen Entität aufweist, nicht mit einer hohen durchschnittlichen Faktengenauigkeit gerettet werden.

24. VERKNÜPFUNG DER KONFORMITÄTSENTSCHEIDUNG MIT DER ENTITÄT

Eine Konformitätsentscheidung sollte im folgenden Format festgelegt werden:

[Entitäts-ID] + [Entitätstyp] + [Umfang] + [digitale Oberflächen] + [Land und Sprachen] + [Zeit] + [Standardversion]

Beispielvorlage: „Für [Marke/rechtliche Einheit] identifiziert mit der Entity-ID NGE-XXXX NOMOS; die öffentlichen Seiten auf Türkisch und Englisch auf der angegebenen Domain, innerhalb des angegebenen Produkt- und Dienstleistungsbereichs, wurden am [Datum] nach [Standardversion] bewertet.“ Dann sollte klar angegeben werden: „Diese Entscheidung gilt nicht automatisch für Tochtergesellschaften, Franchisebetriebe, Produkte außerhalb des Geltungsbereichs, andere Domains und in Zukunft hinzuzufügende Sprachen.“

25. VERPFLICHTENDE NORMATIVE BESTIMMUNGEN

CH03-N01

Jede GEO-Prüfung muss eine eindeutige Zielentität identifizieren, bevor Ergebnisse sichtbar sind.

CH03-N02

Es kann nicht angenommen werden, dass der Eingabeanker dasselbe ist wie das Zielobjekt.

CH03-N03

Domainnamen, Marken, juristische Personen, Produkte, Dienstleistungen und persönliche Identitäten müssen voneinander unterschieden werden.

CH03-N04

Eine Eigenschaft einer Entität kann ohne klare Nachweise und Umfang nicht auf eine andere verwandte Entität übertragen werden.

CH03-N05

Die Lizenz, das Zertifikat, die Anzahl der Mitarbeiter, Kunden oder das Konformitätsergebnis des Mutterunternehmens können nicht automatisch auf Tochtergesellschaften übertragen werden.

CH03-N06

Die Anspruch auf Markenebene und die Partei, die tatsächlich die Dienstleistung erbringt und rechtlich verantwortlich ist, müssen getrennt werden.

CH03-N07

Der Franchisegeber, Franchisenehmer, Distributor, Händler und Plattformverkäufer müssen als separate Einheiten registriert sein.

CH03-N08

Jede greifbare Entitätsbeziehung muss Informationen über Richtung, Umfang und Zeit enthalten.

CH03-N09

Beziehungen wie „Partner“, „Kunde“, „Experte“, „Gruppenunternehmen“ und ähnliche dürfen nicht mehrdeutig verwendet werden.

CH03-N10

Bereiche zur Konfliktlösung müssen für verschiedene Entitäten mit demselben Namen verwendet werden.

CH03-N11

Wenn ein unsicherer Anspruch in die Hauptbewertung einbezogen werden soll, muss die Unsicherheitsmethode vordefiniert sein; andernfalls muss der Anspruch geklärt werden.

CH03-N12

Der Umfang einer Entität darf nach Bekanntwerden der Ergebnisse nicht erweitert oder eingeschränkt werden, um ein positives Ergebnis zu erzielen.

CH03-N13

Historische Entitäten und Beziehungen dürfen nicht so verwendet werden, als hätten sie den Status der aktuellen Entität.

CH03-N14

Der Kauf, die Fusion oder die Markenübertragung übertragen nicht automatisch vergangene Eigenschaften auf die neue Einheit.

CH03-N15

Eine Konformitätsentscheidung kann nur auf die in der Entscheidung angegebene Konformitätseinheit angewendet werden.

CH03-N16

Das neue Produkt kann die aktuelle Verfügbarkeit von Sprache, Domainname, verbundenem Unternehmen oder Land nicht automatisch erben.

CH03-N17

Der Erfolg der Identitätsauflösung und der sachlichen Genauigkeit sollte separat bewertet werden.

CH03-N18

Die Verschmelzung der materiellen Identität mit dem falsche Entität kann im allgemeinen faktischen Durchschnitt nicht unsichtbar gemacht werden.

CH03-N19

Kanonische Eintragsdatensätze sollten versioniert werden und Änderungen im Laufe der Zeit bewahren.

CH03-N20

Es sollte einen verantwortlichen menschlichen oder institutionellen Eigentümer des Eintragsdatensatzes geben.

26. FORMEN DES VERSAGENS

CH03-F01— DOMAINNAME ALS UNTERNEHMEN IN BETRACHT ZIEHEN

Die digitale Adresse wird als juristische oder operative Einheit dargestellt.

CH03-F02— DIE MARKE MIT DEM RECHTLICHEN UNTERNEHMEN ZUSAMMENFÜHREN

Der Markenname wird anstelle der juristischen Person verwendet, die Verträge und Verantwortlichkeiten trägt.

CH03-F03— DAS PRODUKT ALS UNTERNEHMEN BETRACHTEN

Der Produktname wird als Unternehmens- oder Markenidentität verwendet.

CH03-F04— DIE EIGENSCHAFT DES GRÜNDERS AUF DAS UNTERNEHMEN ÜBERTRAGEN

Das Alter, die Erfahrung, die Auszeichnung oder das vorherige Projekt des Gründers wird zu einer Unternehmens-Eigenschaft gemacht.

CH03-F05— GRUPPENEIGENSCHAFTEN AUF DAS LOKALE UNTERNEHMEN ÜBERTRAGEN

Konsolidierte Kunden-, Mitarbeiter- oder Umsatzzahlen werden dem lokalen Unternehmen zugeordnet.

CH03-F06— DIE LIZENZ DURCH BEZIEHUNG ÜBERTRAGEN

Die Lizenz der Muttergesellschaft, des Produkts, des Mitarbeiters oder Partners wird im Namen einer anderen Einheit verwendet.

CH03-F07— VEREINIGUNG VON FRANCHISE UND HAUPTMARKE

Die Leistung und Verantwortung des lokalen Unternehmens wird direkt als Marke auf globaler Ebene dargestellt.

CH03-F08— KOMBINATION VON PRODUZENT UND VERKÄUFER

Der Produkthersteller und der lokale Verkäufer oder Distributor werden als eine einzige Einheit betrachtet.

CH03-F09— KOMBINATION VON PLATTFORM UND VERKÄUFER

Die Marktplatzplattform wird für das Produkt des Verkäufers, die Garantie oder die Lieferung verantwortlich gemacht.

CH03-F10— BETRACHTUNG DER BEZIEHUNG ALS FUNKTIONSÜBERTRAGUNG

Kundenstatus, Mitgliedschaft, Technologieeinsatz oder Partnerschaft wird fälschlich als Übertragung von Autorität dargestellt.

CH03-F11— AKTUALISIERUNG DER HISTORISCHEN IDENTITÄT

Der alte Name, Besitz, die Partnerschaft oder der Manager wird verwendet, als ob er die aktuelle Realität wäre.

CH03-F12— ANSPRÜCHE AUF VERGANGENHEIT NACH EINER ÜBERNAHME

Alle vergangenen Erfolge der übernommenen Einheit werden dem neuen Eigentümer ohne Erklärung zugeschrieben.

CH03-F13— ZUSAMMENFÜHRUNG VON EINHEITEN MIT DEM GLEICHEN NAMEN

Unterschiedliche Unternehmen oder Einzelpersonen werden aufgrund von Namensähnlichkeit als eine Einheit betrachtet.

CH03-F14— KÜNSTLICHE AUFTEILUNG EINER EINHEIT

Eine gescheiterte Sprache, ein Produkt oder eine lokale Operation wird als separate Einheit deklariert, um eine positive Bewertung aufrechtzuerhalten.

CH03-F15— KÜNSTLICHE ERWEITERUNG EINER EINHEIT

Erfolgreiche Marken- oder Produktergebnisse werden auf die gesamte Gruppe verteilt.

CH03-F16— BERÜCKSICHTIGUNG DES GÜLTIGKEITSBEREICHS

Ein enges Prüfergebnis wird für das gesamte Unternehmen, die Gruppe oder das Produktportfolio verwendet.

CH03-F17— ZÄHLUNG DER SPRACHJURISDIKTION

Das Vorhandensein einer Sprachversion wird als Nachweis für die rechtliche Unternehmens- oder Servicekapazität in diesem Land verwendet.

CH03-F18— TECHNISCHE KONTROLLE ALS EIGENTUM VERRECHNEN

Die Agentur oder der Entwickler wird als rechtlicher oder kommerzieller Eigentümer der Einrichtung angegeben, weil sie den Domainnamen verwalten.

CH03-F19— ANSPRUCHSSTELLER UNBESTIMMT LASSEN

Es kann nicht bestimmt werden, wer für falsche Preis-, Lizenz- oder Kapazitätsinformationen verantwortlich ist.

CH03-F20— DEN PROMPT-ANKER NACH DEM ERGEBNIS INTERPRETIEREN

Die Zielentität, die die Antwort erfolgreich erscheinen lässt, wird danach ausgewählt.

27. PRÜFUNGSVERFAHREN

Schritt 1— Den Prompt-Anker aufzeichnen

Genau welche der Nutzer-:

- Name,
- Domainname,
- Marke,
- Produkt,
- Person

Ausdruck verwendet wird, wird vollständig aufgezeichnet.

Schritt 2— Die Zielentität identifizieren

Es wird vermerkt, welche Entität tatsächlich vom Audit gemessen wird. Das Ziel wird klassifiziert als:

- juristische Person,
- Marke,
- Produkt,
- Dienstleistung,
- Entitätspaket

Die anwendbare Klasse muss ausdrücklich aufgezeichnet werden.

Schritt 3— Erstellen der kanonischen Entitäts-ID

Kanonischer Name Entitätstyp Rechtliche Identität Zuständigkeitsbereich Aliase Zeitstatus wird aufgezeichnet.

Schritt 4— Aufbau des Entitätsgraphen

Gerichtete Beziehungen zwischen Muttergesellschaft, Tochtergesellschaften, Marke, Produkt, Domainnamen, Einzelpersonen und Partnern werden extrahiert.

Schritt 5— Bestimmen der Verantwortungsrollen

Aussageverantwortliche Digitaloberflächeninhaber Servicebereitstellungsinhaber Rechtlich Verantwortlicher Evidenzinhaber werden getrennt.

Schritt 6— Überprüfung des Risikos der Attributübertragung

Als Antwort:

- Lizenz,
- Zertifikat,
- Mitarbeiter,
- Kunde,
- Gründungsdatum,
- Erfolg,
- Compliance

Wurde es auf den falsche Entität übertragen?

Schritt 7— Überprüfung der zeitlichen Identität

Antwort:

- früherer Name,
- früherer Eigentümer,
- beendete Partnerschaft,
- früherer Mitarbeiter,
- Status vor der Übernahme

wird es verwendet?

Schritt 8— Geografische und lokale Rollen prüfen

Sind globale Marke, lokale Firma, Vertriebspartner und Franchise-Rollen korrekt?

Schritt 9— Entität vom System klassifizieren

Angemessener Status zwischen ER-1 und ER-7 wird zugewiesen.

Schritt 10— Konformitätseinheit festschreiben

Um die endgültige Entscheidung anzuwenden:

- Entität,
- Feld,
- Produkt,
- Sprache,
- Land,
- Zeit

Grenze wird geschrieben.

Schritt 11— Öffentliches Aussagelimit festlegen

Das Ergebnis wird eingeschränkt, wenn es auf eine Institution oder Gruppe verallgemeinert wird, die breiter ist als die definierte Einheit.

Schritt 12— Die Art der Änderung und des Einwands aufzeichnen

Wenn ein Identitäts- oder Beziehungsfehler gemeldet wird:

- wer wird überprüfen,
- wer wird ändern,
- welche Version wird erzeugt

wird bestimmt.

28. NOTWENDIGE NACHWEISE

Kanontische Identitätsmerkmale von Organisationen; rechtlicher Name und Organisationstyp; offizieller Nachweis oder gleichwertige Verifizierung; Markeninhaberschaft; Domainnamenkontrolle; Mutter- und Tochtergesellschaftsverhältnisse; Eigentum an Produkten und Dienstleistungen; lokale Geschäfts- und Händlerdaten; Franchise-Beziehungen; Vertrags- und Rechnungspartei; Verantwortung für Servicebereitstellung; Lizenz- und Zertifikatsinhaber; Gründer- und Führungskräftepositionen; aktive und historische Namen; Akquisitions-, Fusions- und Übertragungsaufzeichnungen; Land und Zuständigkeit; Amtssprache und Transkriptionsunterlagen; Organisationsgraph; Anspruchsberechtigter; Inhaltsinhaber; technischer Oberflächeninhaber; Prüfungsumfang; nicht einschlägige Organisationen und Oberflächen; Version des Organisationsdatensatzes; Datum der letzten Verifizierung; und verantwortliche Person oder Organisation.

29. PRÜFLISTE FÜR AUDITS

Wurde der Prompt-Anker klar aufgezeichnet? Wurde die Zielentität identifiziert, bevor die Ergebnisse beobachtet wurden? Wurden die juristische Einheit und der Domainname getrennt? Wurden Marke und Markeninhaber getrennt? Wurden Produkt und Unternehmen getrennt? Wurden die Attribute des Gründers von denen der Organisation getrennt? Sind Mutter-Tochter-Beziehungen korrekt? Sind Franchise-Unternehmen und Vertriebspartner separat? Werden die Parteien, die den Service

verkaufen und liefern, identifiziert? Wurden der Anspruchsteller und der technische Herausgeber getrennt? Gehören die Lizenzen und Zertifikate der richtigen Einheit? Werden Mitarbeiter-, Kunden- und Umsatzzahlen innerhalb des richtigen Rahmens angegeben? Wurden historische Namen und Beziehungen vom aktuellen Zustand getrennt? Wurden verschiedene Entitäten mit demselben Namen überprüft? Wurden mehrere Domains und Nutzeroberflächen mit der richtigen Entität verknüpft?

Wurden mehrsprachige Namen mit derselben kanonischen Identität verknüpft? Sind die Rollen von Land und Rechtsordnung klar? Wurde die Entität als Ergebnis der Entitätsauflösung separat klassifiziert? Ist die Konformitätseinheit klar definiert? Erstreckt sie sich auf Entitäten außerhalb des Entscheidungsbereichs? Ist der Entitäten-Datensatz versioniert und datiert? Gibt es einen Weg für Widerspruch und Korrektur bei Identitätsänderungen?

30. EINWÄNDE UND ANTWORTEN

Widerspruch 1— „Der Nutzer kennt den Markennamen, nicht das rechtliche Unternehmen."

Korrekt. Es ist natürlich, dass der Nutzer die Marke im Prompt verwendet. Das Audit zwingt den Nutzer nicht, rechtliche Fachbegriffe zu verwenden. Die Bewertung im Hintergrund:

- die Marke,
- Eigentümer,
- Dienstleister,
- Vertragspartner

müssen unterschieden werden. Eine einfache Antwort kann dem Nutzer gegeben werden. Das Auditprotokoll muss nicht einfach sein.

Widerspruch 2— „In einem kleinen Unternehmen sind Marke und Unternehmen bereits dasselbe.“

In einigen kleinen Strukturen:

- eine einzelne Marke,
- ein einzelner Domainname,
- ein alleiniger Geschäftsinhaber,
- ein einzelner Dienstleister

Es kann gefunden werden. In diesem Fall ist der Entitätsgraph einfach. Dennoch muss der Eintrag ausdrücklich gemacht werden. Einfachheit bedeutet keine Mehrdeutigkeit.

Einwand 3— 'Die Unternehmensgruppe handelt wie eine einzelne Marke.'

Aus Marketingsicht kann es eine einzelne Marke geben. In Bezug auf rechtliche, operative und beweisbare Verantwortung kann es jedoch separate Einheiten geben. Wenn Entscheidungen auf Gruppenebene getroffen werden sollen, sollte klar gezeigt werden, welche Merkmale konsolidiert sind.

Einwand 4— „So viele Details verkomplizieren den GEO-Score unnötig.“

Die korrekte Punktzahl, die dem falsche Entität zugewiesen wird, ist nicht nützlich. Die Entitätsauflösung ist keine zusätzliche Belastung für die Prüfung, sondern eine Bedingung der Gültigkeit.

Einwand 5— „KI-Systeme achten bereits nicht auf diese rechtlichen Unterschiede.“

Genau aus diesem Grund sollte es gemessen werden. Das KI-System:

- Marke,
- Unternehmen,
- Produkt,
- lokales Unternehmen

Das Verfehlen des Unterschieds zwischen ihnen ist der Repräsentationsfehler selbst.

Einspruch 6— „Wenn das Mutterunternehmen die gesamte Gruppe kontrolliert, können Eigenschaften übertragen werden.“

Kontrollbeziehungen sind in einigen Bereichen wichtig. Allerdings:

- Lizenz,
- Verträge,
- Kunde,
- Mitarbeiter,
- Datenverantwortung,
- Compliance

Solche Eigenschaften erfordern separate Nachweise für die Übertragung. Kontrolle ist nicht die Vereinheitlichung aller Eigenschaften.

Einwand 7— „Gehören die Errungenschaften des alten Unternehmens nach einer Übernahme nicht zum neuen Konzern?“

Es kann in einem historischen Kontext erklärt werden. Allerdings:

- Während welchem Zeitraum war der Erfolg,
- von welchem Team und welcher Einheit,
- vor oder nach der Übernahme

Es sollte angegeben werden, wann es produziert wurde. Das neue Eigentum ändert die Geschichte nicht.

Einwand 8— "Die Prompt ist bereits klar; der Nutzer hat den Domainnamen eingegeben."

Der Domainname mag klar sein. Allerdings sollte auch die Frage "Was repräsentiert der Domainname?" beantwortet werden. Eine Domain kann ein Unternehmen, ein Geschäft, ein Produktportal oder eine Plattform mit mehreren Einheiten sein.

Einwand 9— "Den lokalen Vertriebspartner wie eine Marke zu beschreiben, ist für den Nutzer einfacher."

Einfachheit überwiegt nicht die Verantwortung. Die Antwort kann wie folgt vereinfacht werden: "Die Marke bietet in der Türkei Dienstleistungen über einen autorisierten lokalen Vertriebspartner an." Es ist sowohl verständlich als auch korrekt.

Einwand 10— "All diese Beziehungen ändern sich im Laufe der Zeit."

Das ist richtig. Aus diesem Grund sollte der Datensatz der Einheit:

- datiert,
- versioniert,
- änderungsgetriggert sein

Dies verringert nicht die Notwendigkeit der Variabilitätsaufzeichnung. Es erhöht sie.

GEMEINSAME REGEL VON KAPITEL 33

Bevor ein GEO-Score erstellt wird: fragen wir: „Wie viele Antworten sind korrekt?“ Aber die grundlegendere Frage, die davor gestellt werden sollte, ist:

„Von wessen Antworten handelt es sich hierbei genau?“

Wenn wir einen Domainnamen als ein Unternehmen betrachten, messen wir das falsche Objekt. Wenn wir die Marke als ein juristisches Unternehmen betrachten, ordnen wir die Verantwortung falsch zu. Wenn wir das Produkt als ein Unternehmen betrachten, verzerren wir seine Merkmale. Wenn wir die Erfahrung des Gründers in die Geschichte des Unternehmens umwandeln, übertreiben wir die Vergangenheit. Wenn wir die Anzahl der Mitarbeiter der Gruppe als lokale Kapazität zählen, überhöhen wir die Serviceerwartungen. Wenn das Mutterunternehmen seine Lizenz auf eine Franchise überträgt, leiten wir den Nutzer an die falsche Behörde weiter. Wenn wir die bisherigen Erfolge nach einer Übernahme als Merkmal des neuen Eigentümers darstellen, schreiben wir die historische Wahrheit um. Wenn wir eine enge Domain-Name-Prüfung auf das gesamte Unternehmen ausweiten, verschleiern wir den Umfang der Compliance. Daher ist die Definition einer Entität nicht nur Datenbereinigung. Sie ist der Beginn der repräsentativen Verantwortung. Ein System kann den falsche Entität mit den korrekten Informationen beschreiben. Dies ist dennoch eine Fehlrepräsentation. Ein System kann die korrekte Entität mit den falschen Informationen beschreiben. Dies ist ebenfalls eine Fehlrepräsentation.

Beide sind für das korrekte GEO erforderlich:

Korrekte Entität. / Korrektes Attribut. / Korrekte Beziehung. / Korrekter Umfang. / Korrekte Zeit.

Daher lautet das dritte Messgesetz von NOMOS:

Wenn die gemessene Entität nicht klar ist, ist auch die gemessene Genauigkeit nicht klar.

Das vierte Gesetz lautet wie folgt:

Beziehung ist nicht Identität.

Das fünfte Gesetz lautet wie folgt:

Eigentum ist nicht die Übertragung aller Attribute.

Sein sechstes Gesetz lautet:

Konformität breitet sich nicht auf benachbarte Entitäten aus.

NOMOS Abschnitt 3 Bestellung

Gib mir nicht nur einen Namen. / Zeige mir, zu welcher Entität dieser Name gehört.

Behandle eine Domäne nicht als Unternehmen, eine Marke nicht als Vertragspartner oder ein Produkt nicht als Hersteller.

Verwende die Vergangenheit des Gründers nicht als Alter der Institution.

Übertrage die Lizenz des Mutterunternehmens nicht an die angeschlossene Organisation, Franchise oder den Vertrieb.

Machen Sie den Kunden der Gruppe nicht zum Kunden der lokalen Marke, oder den Preis der Marke nicht zum Preis aller Produkte.

Beschreiben Sie Zusammenarbeit nicht als Eigentum, Mitgliedschaft nicht als Akkreditierung und Technologieeinsatz nicht als Partnerschaft.

Führen Sie den alten Namen nicht mit der neuen Einheit zusammen, oder die erworbene Vergangenheit nicht mit der heutigen Leistung.

Erweitern Sie die Einheit nicht, um positive Ergebnisse zu vergrößern / Teilen Sie die Einheit nicht, um negative Ergebnisse zu verbergen.

Sichern Sie zuerst das Ziel. / Dann zeichnen Sie die Richtung jeder Beziehung auf. / Dann ordnen Sie jedes Attribut seinem wahren Eigentümer zu. / Dann fügen Sie die relevante Zeit und das Land hinzu. / Erst danach entscheiden Sie nur für diese Einheit.

Der abschließende Satz des Kapitels

Die erste Frage in einem GEO-Audit ist nicht, was die KI sagt, sondern genau, wer oder was jede Aussage betrifft.

Normativer Kern

Jede GEO-Prüfung MUSS definieren und versionieren, bevor Ergebnisse beobachtet werden: - einen Promptanker, - eine Zielentität, - einen Entitätstyp, - einen kanonischen Identitätsdatensatz, - einen Beziehungsgraph, - eine zeitliche und rechtliche Grenze, - ein Prüfobjekt, - und eine Konformitätseinheit. Domains, Marken, juristische Personen, Unternehmensgruppen, Tochtergesellschaften, Produkte, Dienstleistungen, Personen, Franchises, Vertriebsunternehmen, Plattformen und Verkäufer DÜRFEN NICHT als austauschbare Entitäten behandelt werden. Attribute, Berechtigungen, Lizenzen, Kunden, Mitarbeiter, Leistung, Befugnis und Konformität DÜRFEN NICHT zwischen verwandten Entitäten übertragen werden, es sei denn, die Beziehung, der Umfang, die Zeit und die Nachweise unterstützen diese Übertragung ausdrücklich. Die vom KI-Produkt gelöste Einheit MUSS getrennt von der sachlichen Richtigkeit der Antwort bewertet werden. Eine Konformitätsentscheidung MUSS sich nur auf die ausdrücklich identifizierte Einheit und den geprüften Umfang beziehen. Sie DARF NICHT automatisch auf Tochtergesellschaften, Produkte, Jurisdiktionen, Sprachen, Bereiche, zukünftige Versionen oder verwandte Einheiten ausgeweitet werden.

KAPITEL 04

Zwei Ebenen der Wirklichkeit

Offizielle Repräsentation und verifizierte Realität der Entität

METHODISCHE GRUNDLAGEN / K10; K11

Kapitelgrenze

Abschnitt 1 hat die folgende Bestimmung eingeführt:

> *Eine einzelne Antwort ist keine GEO-Punktzahl.*

Kapitel 2 fügte Folgendes hinzu:

> *Repräsentation ist eine Verteilung, die unter Nutzern, KI-Produkten, Ländern, Sprachen, Sitzungen und Zeiten auftritt.*

Kapitel 3 identifizierte dann das Objekt der Verteilung:

> *Eine Einheit kann nicht gemessen werden, bis sie geklärt ist.*

Jetzt kommen wir zur nächsten obligatorischen Frage:

> *Anhand welches Realitätsdatensatzes werden wir bewerten, ob die KI-Antwort richtig ist?*

Welche Quelle sollte die Prüfung leiten: die Website der geprüften Einheit, ein rechtlicher Nachweis, die Preisseite des Unternehmens, unabhängige Berichte, die Lizenzbehörde, Kundenbewertungen, die eigenen Leistungsdaten des Unternehmens oder die Behauptung eines Konkurrenten? Wenn Quellen widersprüchlich sind, welche sollte als korrekt akzeptiert werden? Wenn ein Unternehmen eine falsche oder unbelegte Behauptung veröffentlicht und die KI sie perfekt wiederholt, hat das System dann Erfolg gehabt? Wenn die offizielle Website veraltet ist, die KI jedoch eine andere Antwort basierend auf aktuellen, verifizierten externen Belegen gibt, ist das System dann gescheitert? Bis diese Fragen beantwortet sind:

- korrekte Repräsentation,
- unzutreffende Repräsentation,
- Quellentreue,
- faktische Genauigkeit,
- Realität der Entität

Diese Begriffe werden manchmal so verwendet, als ob sie sich auf dasselbe Konzept beziehen. Dieses Kapitel löst die daraus resultierende Verwirrung, indem es zwei Realitätsebenen unterscheidet:

Erste Ebene

Institutionelle Repräsentationsebene

Zeigt, was die Entität über sich selbst auf den kanonischen Oberflächen, die sie kontrolliert, sagt.

Zweite Ebene

Verifizierte Realitätsebene der Einheit

Unter den vorhandenen Evidenzmitteln zeigt sich, welche materiellen Aussagen in welchem Umfang und Zeitraum unterstützt werden. Diese beiden Ebenen sind miteinander verbunden. Aber sie sind nicht dasselbe. Die erste Ebene beantwortet die Frage: „Wie definiert sich das Wesen selbst?“ Die zweite Ebene beantwortet die Frage: „Welcher Teil dieser Definition wird durch welche Nachweise gestützt?“ Das erste Buch definiert den korrekten Standard, das zweite Buch trennt die Formen der Verletzung, und das Prüfungssystem ist darauf ausgelegt, diese Unterscheidungen zu messen und nachzuweisen. Dieser Abschnitt führt diese Prüfung auf der Aussageebene durch.

NOMOS Herausforderung

Die Website Ihres Unternehmens enthält den folgenden Satz: „Wir sind das weltweit führende GEO-Unternehmen.“ Ein KI-Produkt liest diese Seite und antwortet dem Nutzer: „Dieses Unternehmen ist das weltweit führende GEO-Unternehmen.“ Spiegelt die Antwort Ihre Website korrekt wider? Ja. Spiegelt die Antwort die bewiesene Realität korrekt wider?

Wir wissen das noch nicht. Vielleicht:

- Es gibt vergleichende Forschung,
- Der Marktbereich wurde definiert,
- Ein Führungskriterium wurde veröffentlicht,

Das Ergebnis wurde unabhängig überprüft. Vielleicht existiert keines davon. Wenn die KI die eigene Behauptung des Unternehmens wiedergegeben hätte: „Das Unternehmen definiert sich selbst als das weltweit führende GEO-Unternehmen.“, hätte sie die Quelle der Behauptung beibehalten. Wenn sie jedoch sagte: „Es ist das weltweit führende GEO-Unternehmen.“, könnte das Unternehmen die Aussage zu einer unabhängigen Realität erhoben haben.

Betrachten Sie einen weiteren Fall. Auf der offiziellen Website des Unternehmens heißt es: ‚Wir bieten Dienstleistungen nur in der Türkei an.‘ Dennoch zeigen verifizierte Aufzeichnungen, dass das Unternehmen vor sechs Monaten seine Tätigkeit in Deutschland und dem Vereinigten Königreich aufgenommen hat und die Website nicht aktualisiert wurde. Basierend auf aktuellen rechtlichen und betrieblichen Aufzeichnungen antwortet die KI: ‚Das Unternehmen bietet Dienstleistungen in der Türkei, Deutschland und dem Vereinigten Königreich an.‘

Die KI ist nicht vollständig treu gegenüber der offiziellen Website. Könnte sie dennoch näher an der verifizierten aktuellen Realität liegen? Ja. Betrachten Sie einen dritten Fall. Die Website des Unternehmens gibt an: ‚Wir sind in 20 Ländern tätig.‘ Unabhängige und interne Aufzeichnungen bestätigen nur acht. Die KI antwortet: ‚Das Unternehmen ist in 20 Ländern tätig.‘

Diese Antwort:

- treu gegenüber der offiziellen Repräsentation,
- im Widerspruch zur verifizierten Realität

Es ist möglich. Die erste Bestimmung dieses Abschnitts lautet wie folgt:

> *Treue zum offiziellen Eintrag ist nicht dasselbe wie faktische Genauigkeit.*

Ihre zweite Bestimmung lautet:

> *Was ein Unternehmen sagt und was nachweisbar ist, sollte getrennt aufgezeichnet werden.*

Seine dritte Bestimmung besagt:

> *Eine KI, die einen offiziellen Fehler korrekt wiederholt, stellt keine akkurate Repräsentation dar.*

Seine vierte Bestimmung besagt:

> *Eine KI, die von einem veralteten offiziellen Eintrag abweicht, liegt nicht automatisch falsch, wenn sie verlässliche Nachweise liefert.*

1. ZWECK DES KAPITELS

Der Zweck dieses Abschnitts besteht darin, zwei separate Referenzebenen festzulegen, die bei der GEO-1000-Bewertung verwendet werden sollen. Der Abschnitt trifft die folgenden normativen Unterscheidungen:

- Deklaration der Existenz vs. nachgewiesene Tatsache
- Offizielle Quelle vs. unabhängige Quelle
- Quelle der ersten Partei versus unabhängige Überprüfung
- Treue zur Quelle versus materielle Genauigkeit
- Wiederholung der Behauptung versus Überprüfung der Behauptung
- Aktueller offizieller Eintrag vs. alter offizieller Eintrag
- Offizielles Schweigen vs. extern verifizierte Informationen
- Unternehmensperspektive versus externe Realitätseinschätzung
- Offizieller Nachweis vs. offizielle Behauptung
- Unabhängigkeit der Quelle versus Relevanz der Quelle für das Thema
- Offizieller Fehler vs. KI-Fehler
- Referenzkonflikt versus Modellfehler
- Unsicherheit versus Versagen
- Interesse des Anspruchstellers versus Wahrheit des Anspruchs
- Das Vorhandensein von Nachweisen und ob die Nachweise den Anspruch tatsächlich stützen

Bis zum Ende dieses Abschnitts sollten zwei separate Fragen für jede KI-Antwort gestellt werden können:

> *Gibt die Antwort korrekt die offizielle und kanonische Repräsentation der Entität wieder?*

und:

> *Spiegelt die Antwort korrekt die materielle Realität wider, die durch die gegenwärtig überprüfbaren Nachweise getragen wird?*

Diese Fragen können dieselbe Antwort ergeben. Sie können auch unterschiedliche Antworten ergeben. Wenn sie unterschiedlich sind, besteht die Rolle des Audits nicht darin, den Unterschied zu verbergen, sondern ihn zu erklären.

2. ZENTRALE NORMATIVE BESTIMMUNG

> *Jede wesentliche KI-Antwort sollte ohne Veränderung anhand von zwei separaten Referenzen bewertet werden: dem offiziellen Repräsentationsdokument und dem verifizierten Realitätsdokument der Entität.*

Eine Antwort:

- treu gegenüber der offiziellen Repräsentation,
- angemessen zur verifizierten Realität,
- geeignet für beide,

- abgegrenzt von beiden

Es kann sein. NOMOS kann diese beiden Ergebnisse nicht stillschweigend zu einer einzigen Wahrheits-Falsch-Entscheidung kombinieren. Hohe Leistung in der ersten Schicht kompensiert nicht das Versagen in der zweiten Schicht. Im Besonderen:

> *Eine KI-Antwort, die eine nicht überprüfte oder falsche Behauptung, die von der Entität selbst veröffentlicht wurde, korrekt wiederholt, kann die Treue der offiziellen Repräsentation zeigen; sie kann jedoch keine nachgewiesene Genauigkeit der Entität demonstrieren.*

3. ERSTE SCHICHT

SCHICHT DER OFFIZIELLEN Repräsentation

Die Schicht der offiziellen Repräsentation ist die versionierte Aufzeichnung der aktuellen materiellen Aussagen, die sie über sich selbst auf kanonischen Oberflächen veröffentlicht, die von der Entität kontrolliert oder ausdrücklich autorisiert sind. Diese Schicht beantwortet die Frage:

> *Wie möchte die Einheit verstanden werden, und welche Informationen veröffentlicht sie offiziell?*

Diese Schicht umfasst:

- Aussage zur rechtlichen Identität,
- Marken–Unternehmens-Beziehung,
- Produkt- und Dienstleistungsdefinitionen,
- veröffentlichte Preise,
- Serviceländer,
- Bedingungen für den Support,
- offizielle Richtlinien,
- öffentlich bekannt gegebene Grenzen,
- Sie kann offizielle Währungs- und Versionsinformationen enthalten

Die Schicht der offiziellen Repräsentation ist wichtig. Denn ein KI-System muss korrekt verstehen, wie eine Entität:

- definiert sich selbst,
- welchen Service es anbietet,
- welchen Preis es veröffentlicht,
- welche Bedingungen es akzeptiert

Die offizielle Repräsentationsebene ist jedoch kein automatisch unabhängiger Beweis für jede Behauptung. Eine Aussage auf der eigenen Website der Institution, insbesondere in Bezug auf externe Behauptungen wie Überlegenheit und Leistung, stellt die Aussage selbst dar; sie beweist ihre Richtigkeit nicht von allein.

4. WAS IST EINE OFFIZIELLE QUELLE?

Eine Quelle gilt nicht automatisch als offiziell, nur weil sie das Logo der Institution trägt. Eine offizielle Quelle sollte mindestens die relevanten Merkmale der folgenden aufweisen:

- Kontrolliert durch die Einrichtung
- Unterzieht sich einem autorisierten Veröffentlichungsprozess

- Vom Eigentümer der sachlichen Informationen genehmigt
- Kanonomische URL oder Datensatzkennzeichen
- Version
- Veröffentlichungs- oder Überprüfungsdatum
- Änderungshistorie
- Verantwortliche Person oder Institution
- Umfang
- Status von Sprache und Gebietsschema
- Archiv- oder Aktualitätsstatus

Eine auf einem persönlichen Social-Media-Konto von einem Mitarbeiter veröffentlichte Ansicht ist nicht automatisch eine offizielle Unternehmensmitteilung. Eine Entwurfsdatei einer Agentur ist keine offizielle Veröffentlichung. Eine E-Mail von einem Vertriebsmitarbeiter, wenn sie nicht autorisiert ist, stellt keine allgemeine Unternehmensrichtlinie dar. Ein altes PDF auf der Website eines Unternehmens kann das neue offizielle Dokument möglicherweise nicht ersetzen.

5. OFFIZIELLE Repräsentation ERSCHEINT

Laut der entsprechenden Prüfung können offizielle Darstellungen Folgendes beinhalten:

- Hauptunternehmenswebsite
- Rechtliche Hinweise
- Produkt- und Serviceseiten
- Preis- und Angebotsseiten
- Offizielles Helpcenter
- Dokumente für Investor Relations
- Öffentliche Politiken und Verträge
- Offizielle Pressemitteilungen
- Strukturierte Daten
- Kanonsiche JSON Aufzeichnungen
- Offizielle API-Ausgaben
- Offizielle Inhalte in der App
- Autorisierte lokale Websites
- Offizielle Social-Media-Konten
- An öffentliche Institutionen übermittelte Erklärungen
- Veröffentlichungs- und Rückzugsaufzeichnungen

Diese Oberflächen können einander widersprechen. In diesem Fall ist „offizielle Repräsentation“ keine einzelne, saubere Realität. Es handelt sich um eine Menge von Aufzeichnungen, die intern geprüft werden müssen.

6. AUFZEICHNUNG DER OFFIZIELLEN Repräsentation

Für jede geprüfte Einheit muss eine Aufzeichnung der offiziellen Repräsentation erstellt werden. Diese Aufzeichnung ist keine Kopie aller offiziellen Inhalte. Sie ist eine nachvollziehbare Zusammenfassung von sachlichen Behauptungen und Grenzen. Jede Aufzeichnung sollte, soweit zutreffend, die folgenden Felder enthalten:

- Kennung der Aussage
- Volltext der Aussage
- Gegenstand der Aussage
- Art der Aussage
- Kanonische Quelle
- Quelleninhaber
- Sprache und Gebietsschema
- Veröffentlichungsdatum
- Datum der letzten Überprüfung
- Gültigkeitsdauer
- Offizieller Status
- Umfang
- Einschränkungen
- Vorherige Version
- Grund für die Änderung
- Kläger
- Menschliche Genehmigung

Ohne eine Aufzeichnung der offiziellen Repräsentation einer Einheit kann der Satz „KI spiegelt die Website genau wider.“ nicht klar überprüft werden.

7. DIE AUTORITÄT DER OFFIZIELLEN QUELLE HÄNGT VOM TYP DER BEHAUPTUNG AB

Nicht alle offiziellen Quellen haben dasselbe Gewicht für jede Behauptung. Die Autorität der Quelle sollte entsprechend der Art der Behauptung bewertet werden.

7.1. Entscheidungen Unter Der Eigenen Kontrolle Der Einheit

Beispiel:

- Welchen Service sie der Öffentlichkeit anbietet
- Veröffentlichter Preis
- Supportzeiten
- Offizielle Rückgaberichtlinie
- Kommunikationskanal
- Öffentlich verfügbare Produktversion

In diesen Bereichen ist die offizielle Quelle grundlegend und oft das primäre autorisierte Dokument. Wenn jedoch die offizielle Seite dem tatsächlichen Ablauf widerspricht, reicht die Seite allein nicht aus. Zum Beispiel:

- Preis-Seite 1,000 Euro,
- Checkout 1,500 Euro

Wenn dies angezeigt wird, gibt es einen Widerspruch in der offiziellen Repräsentationsebene.

7.2. Rechtliche Identität und Befugnis

Beispiel:

- Firmenregistrierung
- Lizenz
- Berufliche Befugnis
- Markeninhaberschaft
- Steuer- oder Rechtsstatus

Die Unternehmenswebsite kann diese Informationen offenlegen. Ein offizielles Dokument oder eine autorisierte Einrichtung kann jedoch eine stärkere Verifizierung bieten.

7.3. Leistungs- und Erfolgsaussagen

Beispiel:

- „98% unserer Projekte sind erfolgreich."
- „Wir steigern den Umsatz unserer Kunden im Durchschnitt um 40%."
- „Wir haben die höchste Zufriedenheitsrate."

Die Unternehmenswebsite kann der Inhaber der Behauptung sein. Um die Behauptung zu unterstützen:

- daten,
- methode,
- nenner,
- zeitraum,
- erfolglose Ergebnisse,
- unabhängige Überprüfung, wenn möglich

sind erforderlich.

7.4. Überlegenheits- und Führungsansprüche

Beispiel:

- „Weltmarktführer"
- „Der Beste"
- „Nummer eins"
- „Der einzige echte Standard"
- „Das vertrauenswürdigste Unternehmen der Branche"

Die eigene Aussage einer Website ist kein unabhängiger Vergleich. Solche Aussagen:

- erfordern ein Vergleichsuniversum,
- Kriterien,
- methode,
- Datum,
- Umfang,
- Unabhängigkeit

Beide Bedingungen müssen erfüllt sein.

7.5. Beziehungen zu Dritten

Beispiel:

- Kunde
- Partner

- Sponsor
- Mitgliedschaft
- Akkreditierung
- Technologiepartnerschaft
- Vertriebslizenz

Eine einseitige offizielle Erklärung kann wertvoll sein. Sie muss jedoch möglicherweise von der anderen Partei der Beziehung oder durch ein zuverlässiges Protokoll überprüft werden.

7.6. Nutzerergebnisse

Beispiel:

- Kundenzufriedenheit
- Tatsächliches Nutzungsergebnis
- Rücklaufquote
- Beschwerdequote
- Wiederholungskauf

Unternehmensdaten in diesen Bereichen sind wichtige Primärbeweise. Jedoch:

- Datenauswahl,
- Nonresponse,
- Methodik,
- Interessenkonflikt

sollten offengelegt werden.

8. Treue der offiziellen Repräsentation

Das Konzept, das angibt, inwieweit eine KI-Antwort der offiziellen Repräsentationsschicht entspricht:

Treue der offiziellen Repräsentation

Der kanonische englische Begriff ist:

Treue der offiziellen Repräsentation

Diese Messung stellt die folgenden Fragen: Interpretiert die KI die Entität korrekt? Vermittelt sie offizielle Produkte und Dienstleistungen genau? Verzerrt sie den offiziellen Preis? Entfernt sie offizielle Grenzen? Erweitert sie die offizielle Landesabdeckung? Stellt sie die Meinung des Unternehmens als unabhängige Tatsache dar? Verwendet sie veraltete offizielle Informationen als aktuell? Versteckt sie Widersprüche innerhalb offizieller Aufzeichnungen? Fügt sie materielle Behauptungen hinzu, die in der offiziellen Quelle nicht zu finden sind? Hält sie die Qualität der Quelle aufrecht? Die Treue der offiziellen Repräsentation ist ein Maß dafür, wie genau die Entität die Realität vermittelt, die sie veröffentlicht hat. Diese Messung:

> *Es zeigt nicht, dass die Aussagen der Entität wahr sind, sondern inwieweit KI genau vermittelt, was die Entität sagt.*

9. ZWEITE SCHICHT

VERIFIZIERTE ENTITÄTSREALITÄTSEBENE

Die geprüfte Entitäten-Realitätsschicht ist ein versioniertes Bewertungsprotokoll von wesentlichen Informationen, die an einem bestimmten Messdatum existierten und durch Nachweise gestützt werden, die für die Art der Aussage angemessen sind. Diese Schicht beantwortet die Frage:

> *Welche Aussagen über die Entität werden tatsächlich durch welche Nachweise gestützt?*

Diese Schicht umfasst:

- Aufzeichnungen, die vom Unternehmen selbst stammen,
- offizielle öffentliche Aufzeichnungen,
- unabhängige Quellen,
- Vertrags- oder Transaktionsaufzeichnungen,
- Lizenzdokumente,
- verifizierte historische Aufzeichnungen,
- Gegenevidenz,
- ungünstige Ergebnisse,
- Unsicherheiten,
- verbleibende Streitigkeiten

können zusammen bewertet werden.

10. "GEPRÜFTE REALITÄT" IST NICHT ABSOLUTE REALITÄT

Dieses Protokoll behauptet nicht: "Wir kennen mit Sicherheit die gesamte absolute Wahrheit über eine Institution." Geprüfte Entitäten-Realität ist:

- evidenzbasiert,
- zeitabhängig,
- umfangsabhängig,
- änderbar mit neuen Nachweisen,
- anfechtbar,
- rückziehbar,

kann Unsicherheit enthalten. Die korrekte Definition lautet:

> *Innerhalb des angegebenen Datums und Umfangs der stärkste und überprüfbare Entitätsdatensatz, der von den verfügbaren Nachweisen getragen wird.*

Daher gilt ein Eintrag über geprüfte Realität:

- versioniert,
- quellengestützt,
- anfechtbar,
- änderbar,
- bewusst über Unsicherheit

sein. NOMOS kann sich nicht als Eigentümer der Wahrheit positionieren. Es klassifiziert nur, welchen Satz die Nachweise erlauben und in welchem Umfang.

11. EINTRAG GEPRÜFTER ENTITÄTENREALITÄT

Die vollständige technische Struktur dieser Ebene wird in einem späteren Kapitel des Buches detailliert beschrieben. Auf fundamentaler Ebene sollten die folgenden Felder für jede Sachbehauptung existieren:

- Kennung der Aussage
- Subjekteinheit
- Beziehung oder Attribut
- Wert
- Umfang
- Land oder Rechtsordnung
- Zeit
- Quellen
- Quellentypen
- Quellenverifizierung
- Unabhängigkeit
- Gegenevidenz
- Evidenzmangel
- Verifizierungsstatus
- Unsicherheit
- Entscheidung des Prüfers
- Berufungsstatus
- Datum der letzten Überprüfung
- Aufzeichnungsversion

Dieser Datensatz kann nicht ausschließlich auf positiven Nachweisen erstellt werden. Aufzeichnungen, die die Realität einschränken oder die Behauptung widerlegen, müssen ebenfalls aufbewahrt werden.

12. VERIFIZIERUNGSSTATUSSE

Jede materielle Behauptung kann mit einem der folgenden Status klassifiziert werden:

VR-0— NICHT BEWERTET

Die Behauptung wurde noch nicht bewertet.

VR-1— NUR OFFIZIELLE AUSSAGEN

Die Behauptung basiert ausschließlich auf der offiziellen Erklärung der Institution. Sie wurde nicht als falsch nachgewiesen. Es gibt keine unabhängige oder ausreichende zusätzliche Überprüfung.

VR-2— DURCH EIGENE BELEGE GESTÜTZT

Die Behauptung wird durch Folgendes der Institution gestützt:

- Transaktion,
- Operation,
- Verträge,
- daten,
- Aufzeichnung

wird durch Belege unterstützt. Die Quelle ist erste Partei.

VR-3— BESTÄTIGT

Die Behauptung wird durch mehrere angemessene und konsistente Aufzeichnungen gestützt. Nicht alle Quellen müssen unabhängig sein.

VR-4— UNABHÄNGIG ÜBERPRÜFT

Die Behauptung wird durch eine angemessene unabhängige Quelle oder Überprüfung zum Thema gestützt. Unabhängigkeit allein bedeutet nicht ausreichende Beweiskraft. Auch die Relevanz der Quelle für das Thema ist erforderlich.

VR-5— MULTI-QUELLEN BESTÄTIGT

Die Behauptung wird stark durch zuverlässige Quellen verschiedener Art innerhalb eines definierten Zeitrahmens und Umfangs gestützt.

VR-C — WIDERSPRUCH

Die Behauptung steht im Widerspruch zu stärkeren oder angemesseneren Nachweisen.

VR-U — UNGEKLÄRT

Die Nachweise sind unzureichend oder widersprüchlich. Ein verlässliches Urteil kann nicht gefällt werden.

VR-T — ZEITLICH BEGRENZT

Die Behauptung wird nur für den angegebenen Zeitraum gestützt.

VR-S — SCOPE-BEGRENZT

Die Anspruchsberechtigung wird nur für das angegebene Produkt, Land, den Nutzer oder den Operationsbereich unterstützt. Ein Anspruch kann gleichzeitig mehr als einen Status haben. Beispiel:

VR-4+ VR-T + VR-S

Es kann darauf hinweisen, dass die Behauptung unabhängig überprüft wurde; sie unterliegt jedoch zeitlichen und inhaltlichen Einschränkungen.

13. UNABHÄNGIGKEIT IST NICHT DAS GLEICHE WIE ANGEMESSENHEIT

Eine unabhängige Quelle ist nicht immer die beste Quelle. Ein Blogger kann den aktuellen Preis eines Unternehmens schätzen, während die offizielle Preisübersichtsseite des Unternehmens die direktere und geeignetere Quelle sein kann. Umgekehrt kann, wenn ein Unternehmen auf seiner eigenen Webseite erklärt, dass es ‚der Branchenführer' ist, unabhängige Forschung mit offengelegter Methode der geeignetere Nachweis sein. Die Bewertung von Quellen fragt daher mehr als nur, ob eine Quelle

unabhängig ist. Ist sie für diese Art von Behauptung maßgeblich? Wie nahe ist sie an der Behauptung? Ist ihr Umfang angemessen und aktuell? Wer kontrolliert sie? Ist ihre Methode offengelegt? Geht sie auf Gegenevidenz ein? Kann sie geprüft werden? Besteht ein Interessenkonflikt? Unabhängigkeit allein garantiert keine Genauigkeit.

Dass eine Quelle erstparteiig ist, macht sie auch nicht wertlos. Das Wichtige ist:

Es hat die entsprechende Autorität und Evidenzqualifikation für die vom Ursprung getragenen Behauptung.

14. AUSSAGE–QUELLEN-TAUGLICHKEITSKARTE

Die nachstehende Tabelle ist eine allgemeine grundlegende Karte. Spezifische Nachweisanforderungen können je nach Sektor und Risikostufe variieren.

Art der Aussage	Primäre Quelle oder Nachweis	Rolle der offiziellen Seite	Notwendigkeit zusätzlicher Überprüfung
Rechtlicher Firmenname	Autorisierter offizieller Nachweis	Zeigt die Erklärung	Im Allgemeinen erforderlich
Markeninhaberschaft	Markeneintragung, Vertrag, Unternehmensdokument	Zeigt die Erklärung	Erforderlich bei finanziellen Angaben
Aktueller veröffentlichter Preis	kanonischer Preis-/Checkout-Nachweis	Üblicherweise primär	Überprüfbar mit Workflow
Leistungsumfang	Betriebsunterlagen, Verträge, offizielle Seite	Primäre Erklärung	Zusätzliche Nachweise für die tatsächliche Leistungserbringung
Lizenz	Lizenzierungsinstitution oder Register	Zeigt die Erklärung	Erforderlich
Kundenbeziehung	Vertrag, Kundengenehmigung, beidseitiger Nachweis	Kann eine einseitige Aussage sein	Erforderlich bei finanziellen Angaben
Erfolgsquote	Rohdaten, Methode, Nenner, Audit	Veröffentlicht die Aussage	Dringend erforderlich
Marktführerschaft	Vergleichende Methode und Daten	Selbsterklärung	Unabhängiger oder starker Vergleich erforderlich
Kundenzufriedenheit	Repräsentative Umfrage, Methode, Nichtantwort	Kann das Ergebnis erklären	Erforderliche Methodik
Gründungsdatum	Offizielle Aufzeichnungen und historische Dokumente	Zeigt die Erklärung	Rechts- und Markengeschichte sollten getrennt werden
Partnerschaft	Bilaterale Erklärung oder Vertrag	Unilaterale Erklärung	Im Allgemeinen erforderlich
Zertifikat	Ausstellende Institution des Zertifikats und Umfang	Zeigt die Erklärung	Erforderlich
Aktiver Mitarbeiter	HR- oder Vertragsdatensatz, öffentliches Profil	Kann Erklärung zeigen	Überprüfung von Rolle und Datum
Nutzerergebnis	Transaktions- und Ergebnisprotokoll	Beschreibt den Fall	Kontext, Methode und Einwilligung erforderlich
"Beste Option"	Definierter Nutzer und Vergleichskriterien	Marketingansicht	Starker Vergleich erforderlich

Diese Tabelle ist kein universelles Quellenranking. Sie zeigt die Angemessenheit zwischen Aussage und Nachweis.

15. DOPPELSCHRITTANSPRUCH-MATRIX

Zu jeder finanziellen Forderung werden zwei separate Fragen gestellt: Wurde sie offiziell veröffentlicht? Wie ist ihr Status im verifizierten Tatsachenverzeichnis? Diese beiden Fragen ergeben die folgenden grundlegenden Situationen.

15.1. Offiziell und Verifiziert

Forderung: Sie ist in einem offiziellen Verzeichnis vorhanden und durch entsprechende Nachweise gestützt. Beispiel: Der Preis auf der offiziellen Preisseite wird auch im Transaktionsablauf verifiziert. Dies ist eine der stärksten Arten der Kernaussage. Wenn die KI die Forderung mit dem korrekten Umfang und Timing übermittelt, tritt auf beiden Ebenen Erfolg ein.

15.2. Offiziell, aber nur selbst erklärt

Die Behauptung ist offiziell. Es gibt jedoch keine zusätzliche Überprüfung. Beispiel: „Wir sind das innovativste Unternehmen der Branche.“ KI kann die Behauptung folgendermaßen wiedergeben: „Das Unternehmen beschreibt sich selbst als das innovativste Unternehmen der Branche.“ Sie kann sie nicht aufwerten zu: „Es ist das innovativste Unternehmen der Branche."

15.3. Offiziell, aber widersprüchlich

Die Behauptung ist im offiziellen Bericht vorhanden. Sie widerspricht jedoch passenderen oder aktuelleren Nachweisen. Beispiel: Die Seite sagt 20 Länder. Betriebsaufzeichnungen zeigen acht Länder. Wenn KI die Seite genau wiederholt, kann die offizielle Treue hoch sein. Verifizierte Genauigkeit ist gering. Diese Situation:

Den offiziellen Fehler getreu weitergeben

klassifiziert werden.

15.4. Nicht offiziell, aber verifiziert

Die Behauptung ist auf der offiziellen Seite nicht zu finden. Sie wurde jedoch mit anderen zuverlässigen Quellen überprüft. Beispiel: Eine kürzliche rechtliche Eigentumsänderung wurde noch nicht auf der offiziellen Seite hinzugefügt. Sie erscheint im autorisierten Register. KI kann diese Informationen mit der richtigen Quelle und zum richtigen Zeitpunkt genau übermitteln. Offizielle Repräsentationen können eine geringe Genauigkeit haben oder abweichen. Die überprüfte sachliche Genauigkeit kann hoch sein.

15.5. Nicht offiziell und nicht überprüft

Behauptung: Nicht im offiziellen Register, es fehlen ausreichende externe Belege. KI sollte dies nicht so vermitteln, als sei es wahr.

15.6. Materialgrenze offiziell weggelassen

Das Unternehmen veröffentlicht die positive Behauptung auf seiner offiziellen Seite. Allerdings veröffentlicht es nicht die notwendigen Einschränkungen, um sie korrekt zu verstehen. Beispiel: „Wir bieten Dienstleistungen in Europa an.“ Die Dienstleistung wird jedoch nur in zwei Ländern und aus der Ferne angeboten. Wenn die KI nur den allgemeinen Satz wiederholt, kann sie dem offiziellen Text treu bleiben. Sie kann jedoch materiell irreführende Ergebnisse liefern. Daher muss die Offizielle Darstellungs-Ebene nicht nur die veröffentlichten Sätze, sondern auch die wesentlichen Auslassungen erfassen.

15.7. Streitfall ungelöst

Offizielle und externe Aufzeichnungen widersprechen einander. Keine ist stark genug. Das korrekte Ergebnis:

UNGEKLÄRT

sollte sein. Von KI kann nicht erwartet werden, dass sie ein endgültiges Urteil fällt.

16. ZWEISCHICHTIGE BEWERTUNGSTABELLE

Treue zur offiziellen Repräsentation	Überprüfte Realitätsgenauigkeit	Basis-Kommentar
Hoch	Hoch	Treue und verifizierte Repräsentation
Hoch	Niedrig	Den offiziellen Fehler getreu weitergeben
Niedrig	Hoch	Korrektive Repräsentation, die vom offiziellen Bericht abweicht, aber durch Nachweise gestützt wird
Niedrig	Niedrig	Allgemeines Repräsentationsversagen
Unklar	Hoch	Offizielle Quelle widersprüchlich; externe Überprüfung stark
Hoch	Unklar	Offizielle Stellungnahme korrekt wiedergegeben; Realitätsprüfung unzureichend
Unklar	Unklar	Unzureichende Referenz für Entscheidung

Diese Tabelle hält einen wichtigen Grundsatz fest:

> *Die Nichtübereinstimmung der KI mit dem offiziellen Bericht ist nicht immer ein Versagen.*

Ebenso:

> *Die Übereinstimmung der KI mit dem offiziellen Bericht ist nicht immer ein Erfolg.*

17. REGEL ZUR BEIBEHALTUNG DES QUELLSTATUS

Die KI-Antwort muss nicht nur den Inhalt, sondern auch den epistemischen Status des Inhalts bewahren. Die folgenden Sätze sind nicht gleich:

- „Das Unternehmen ist die weltweit führende GEO-Organisation."
- „Das Unternehmen beschreibt sich selbst als die weltweit führende GEO-Organisation."
- „Das Unternehmen behauptet, die weltweit führende GEO-Organisation zu sein, basierend auf seinen veröffentlichten Forschungen, die noch nicht unabhängig überprüft wurden."
- „In dem angegebenen unabhängigen Vergleich belegte das Unternehmen den ersten Platz."

Jeder Satz trägt einen unterschiedlichen Evidenzstatus. KI:

- Selbsterklärung als unabhängiges Ergebnis,
- Marketingaussage als wissenschaftliche Erkenntnis,
- Pressemitteilung als redaktionelle Nachricht,
- gesponserter Preis als unabhängige Akkreditierung,
- Firmendaten als externe Prüfung

Diese Behandlung ist durch die folgende Regel verboten:

Regel zum Schutz von Zitier- und Quellenstatus

Der Quellstatus muss eindeutig bleiben.

18. EPISTEMISCHES UPLIFT

Die Stärkung des Quellstatus einer Aussage in einer KI-Antwort wird genannt:

Epistemisches Uplift

Beispielhafte Transformationen:

Quelltext	Falsches Uplift durch KI
„Laut den eigenen Daten des Unternehmens“	„Bewiesen“
„Von bestimmten Kunden gemeldet“	„Funktioniert bei allen Kunden“
„In der Pilotstudie beobachtet“	„Allgemein wirksam“
„Das Unternehmen definiert sich als führend“	„Branchenführer“
„Keine unabhängige Überprüfung wurde durchgeführt“	Einschränkung vollständig entfernt
„In einigen Märkten verfügbar“	„Weltweit verfügbar“

Diese Transformationen sind keine Zusammenfassungen. Sie stellen die Erstellung neuer und stärkerer Behauptungen dar.

19. OFFIZIELLES SCHWEIGEN

Das Fehlen von Informationen auf der offiziellen Seite eines Unternehmens muss nicht bedeuten:

- dass die Informationen falsch sind,
- dass die Informationen verboten sind,

KI kann diese Informationen nicht verwenden. Offizielles Schweigen kann drei verschiedene Formen annehmen.

19.1. Nicht wesentliche Stille

Informationen können außerhalb der Prompt und Nutzerentscheidung liegen. Ob KI sie verwendet oder nicht, ist unwesentlich.

19.2. Legitime ergänzende externe Informationen

Es kann sich um Informationen handeln, die auf der offiziellen Seite nicht zu finden sind, aber von einer verlässlichen Quelle bestätigt wurden. Beispiel:

- Rechtsakte
- Historisches Ereignis
- Öffentliche behördliche Entscheidung
- Unabhängige Forschung

KI kann diese Informationen mit angemessenem Kontext und Quellstatus verwenden.

19.3. Stille einer wesentlichen Begrenzung

Eine Einrichtung hat möglicherweise ein Limit, das die Entscheidung des Nutzers beeinflusst, in ihrer offiziellen Repräsentation nicht offengelegt. Beispiel:

- Die Lizenz ist nur in bestimmten Ländern gültig
- Der Preis scheint keine zusätzliche Gebühr zu enthalten
- Der Service ist nur für bestimmte Kundengruppen verfügbar
- Erfolgsdaten stammen nur aus ausgewählten Fällen

In diesem Fall kann die offizielle Repräsentation selbst einen Prüfungsbefund enthalten. Wenn KI das Schweigen wiederholt, kann dies die Wahrheitstreue der Realität verringern.

20. MATERIELLES DEFIZIT IN DER OFFIZIELLEN Repräsentation

Ein offizieller Satz darf nicht direkt falsch sein. Das Entfernen der notwendigen Einschränkungen kann ihn jedoch irreführend machen. Beispiel: „Wir sind in Europa tätig.“ Materieller Kontext:

- nur in zwei Ländern,
- nur aus der Ferne,
- ohne ein lokales Büro,
- nur für Unternehmenskunden

kann bereitgestellt werden. Die offizielle Repräsentationsebene muss diese beiden Bereiche zusammen erfassen:

- Die veröffentlichte Hauptaussage
- Die materiellen Einschränkungen, die notwendig sind, um die Aussage korrekt zu verstehen

Wenn eine KI-Antwort den Hauptsatz enthält, aber die Einschränkungen entfernt:

- wortgetreu treu,
- irreführend im Hinblick auf das Urteil

kann sein.

21. FALSCH IST FALSCH, AUCH WENN ES DIR ZUGUTE KOMMT

Wenn ein KI-Produkt eine Entität als mächtiger, größer oder autoritärer darstellt, als sie tatsächlich ist, wird das Schweigen der Organisation zu einem Mess- und Governance-Problem. Ein lokales Unternehmen kann als globales Unternehmen dargestellt werden; eine Beratung als Regulierungsbehörde; Selbsteinschätzung als unabhängige Zertifizierung; oder eine Tätigkeit, die sich über acht Länder erstreckt, als Aktivität in 20. Selbst wenn ein Fehler zugunsten der Organisation ausfällt, muss er in der Schicht der verifizierten Realität als Fehler vermerkt werden. Die grundsätzliche Warnung des zweiten Buches setzt die gleiche Grenze: ein kommerziell attraktiver Fehler muss dennoch korrigiert werden. Wenn eine Organisation wissentlich einen vorteilhaften KI-Fehler ausnutzt:

- Verpflichtung zur Korrektur,
- Integrität der Repräsentation
- gewerbliche Zusage,

- Governance

kann auch im Hinblick auf untersucht werden.

22. REFERENZKONFLIKT

Manchmal liegt das Problem nicht in der KI-Antwort, sondern in den Referenzaufzeichnungen. Beispiel: Die Homepage zeigt einen anderen Preis. Das PDF zeigt einen anderen Preis. Die Kasse zeigt einen weiteren Preis. Strukturierte Daten enthalten einen veralteten Preis. In diesem Fall gibt es möglicherweise keine einzelne korrekte offizielle Repräsentation. Ein weiteres Beispiel: Das Handelsregister gibt eine Eigentümerschaft an. Die Unternehmenswebsite nennt eine andere Eigentümerschaft. Aktualisierungsdaten sind ungewiss. In diesem Fall kann auch der verifizierte Realitätsdatensatz eine Diskrepanz aufweisen.

23. REGEL FÜR REFERENZMANGEL

> *Wenn Referenzaufzeichnungen wesentlich widersprüchlich oder unzureichend sind, kann diese Unsicherheit nicht automatisch in ein Versagen des KI-Produkts umgewandelt werden.*

Korrekte Status:

REFERENZKONFLIKT

UNZUREICHENDE BELEGE

UNGEKLÄRT

ZEITGETEILT

BEREICHSGETEILT

sind möglich. Das Audit sollte die Frage trennen: „Hat die KI eine falsche Antwort gegeben?“ und: „Gab es notwendige Referenzrealität, damit die KI die richtige Antwort geben konnte?“ Wenn eine Institution ihre offiziellen Aufzeichnungen widersprüchlich hinterlassen hat, können wir messen, welchen Datensatz das Modell auswählt. Wir können jedoch den Referenzfehler nicht verbergen und gleichzeitig eine einzige und endgültige institutionelle Antwort vom Modell erwarten.

24. KANONISCHE AUFZEICHNUNGSSPERRE

Beide Referenzschichten von GEO-1000 sollten versioniert werden, bevor die Messwelle beginnt:

Offizieller Vertretereintrag

Welche Seiten? Welche Sprachversionen? Welches Datum? Welche Aussagen? Welcher Archivstatus?

Verifizierter Entitätsrealitätseintrag

Welche Nachweise? Welche Gegenevidenz? Welche Verifizierungsstatus? Welche Einwände? Welches Datum? Wenn es während der Messung eine wesentliche Änderung gibt: Die Welle kann gestoppt werden. Der Datensatz kann in eine neue Version aufgeteilt werden. Frühere und nachfolgende Beobachtungen können separat analysiert werden. Das Mischen der Ergebnisse unter einer einzigen Referenz ist verboten.

25. BEWEISLAST MIT AUSSAGESTÄRKE

Mit der Zunahme eines Anspruchs steigt auch die Beweislast. Die untenstehende Leiter ist eine konstruktive Repräsentation.

C0 — Einfacher Identitätsanspruch

„X ist ein Beratungsunternehmen." Korrekte Vereinbarung der Entität und Tätigkeit erforderlich.

C1 — Umfangsanspruch

„X erbringt Dienstleistungen in Deutschland." Ein Betriebs-, Vertrags- oder Liefernachweis ist erforderlich.

C2 — Quantitativer Leistungsanspruch

„90% der Projekte werden rechtzeitig abgeschlossen." Nenner, Zeitraum, Methode und nicht erfolgreiche Ergebnisse sind erforderlich.

C3 — Vergleichende Aussage

„Es ist schneller als die Wettbewerber." Vergleichsuniversum und gleichwertige Messung sind erforderlich.

C4 — Überlegenheitsbehauptung

„Es ist der Branchenführer." Branche, Kriterien, Geografie, Zeit und Methode sind erforderlich.

C5 — Garantie oder universelle Zusicherung

„Garantiert Ergebnisse für jeden Kunden." Eine echte Verifizierung erfordert extrem starke, umfassende und umsetzbare Nachweise. Offizielle Quellen können für C0- und einige C1-Aussagen stark sein. Allein könnten sie jedoch auf C3–C5-Ebenen nicht ausreichen.

26. ZWEI GETRENNTE GENAUIGKEITSMESSUNGEN

Zwei Hauptmessungen müssen innerhalb von GEO-1000 getrennt berichtet werden.

26.1. Offizielle Repräsentationstreue

Beantwortet die folgende Frage:

> *Inwieweit vermittelt die KI-Antwort die versionierte offizielle und kanonische Repräsentation der Entität genau, vollständig und unter Wahrung des Quellenstatus?*

Dieses Maß:

- offizielle Identität,
- offizieller Dienstbereich,
- veröffentlichter Preis,
- offizielle Grenzen,
- offizielle Quellenart

bewertet deren Bereiche.

26.2. Genauigkeit der überprüften Entitätsrealität

Beantwortet die folgende Frage:

> *Inwieweit vermittelt die KI-Antwort die aktuelle Entitätsrealität, gestützt auf verfügbare Nachweise, genau, umfassend und mit erhaltenen Grenzen?*

Dieses Maß:

- gestützte faktische Behauptungen,
- widersprüchliche Behauptungen,
- veraltete Informationen,
- falsche Beziehungen,
- faktische Defizite,
- erkenntnistheoretische Verbesserungen

werden bewertet.

27. DIE UNFÄHIGKEIT ZWEIER PUNKTE, SICH GEGENSEITIG ZU KOMPENSIEREN

Die hohe Treue zu offiziellen Angaben kann die niedrige verifizierte Genauigkeit nicht ausgleichen. Beispiel: Die offizielle Website veröffentlicht eine falsche Behauptung. KI wiederholt die Behauptung genau. Die Treue zu offiziellen Angaben kann hoch sein. Die verifizierte Genauigkeit ist gering. Ebenso trivialisiert eine hohe verifizierte Genauigkeit nicht vollständig die Verzerrung der offiziellen Repräsentation. Beispiel: KI verwendet eine aktuelle externe Quelle korrekt. Sie verwechselt die Produkt- und Dienstleistungsbeschreibung eines Unternehmens mit der eines anderen. Einige externe Fakten in der Antwort können korrekt sein. Die Treue zur offiziellen Identität ist schwach. Daher sollten die beiden Ergebnisse separat erscheinen.

28. ZWEISCHICHTIGE ANTWORTSTATUS

Eine KI-Antwort kann eines dieser Ergebnisse erhalten:

DOPPELTE DURCHLAUF

Es ist materiell konsistent sowohl mit der offiziellen Repräsentation als auch mit der verifizierten Realität.

OFFIZIELL-PASS / VERIFIZIERT-FAIL

Trägt das offizielle Protokoll getreu. Allerdings vermittelt es eine unbelegte, veraltete oder widersprüchliche Behauptung des offiziellen Protokolls, als ob sie wahr wäre.

OFFIZIELL-FEHLER / GEPRÜFT-BESTANDEN

Entspricht nicht vollständig der offiziellen Webseite. Allerdings vermittelt es die Wahrheit korrekt, gestützt auf aktuelle und stärkere Nachweise. Eine Erklärung der Unterschiede in der Antwort wird bevorzugt.

DOPPELFEHLER

Es stört wesentlich sowohl das offizielle Protokoll als auch die überprüfte Realität.

OFFIZIELL-BESTANDEN / GEPRÜFT-UNBEKANNT

Die offizielle Aussage wurde genau wiedergegeben. Der überprüfte Realitätsstatus der Behauptung wurde noch nicht festgestellt.

REFERENZ-KONFLIKT

Referenzaufzeichnungen sind widersprüchlich. Es ist nicht möglich, die Antwort endgültig zu bewerten.

ANGEMESSENE-UNSICHERHEIT

Die KI hat die Referenzunsicherheit korrekt erklärt und vermieden, ein endgültiges Urteil zu fällen. In einigen Fällen kann dies die genaueste Antwort sein.

29. SYNTHETISCH APPLE.COM ANZEIGE

SYNTHETISCHE METHODIK-DEMONSTRATION / Die folgenden Aussagen, Antworten und Bewertungen sind vollständig fiktiv. Sie stellen keine echten Informationen, Leistungen oder Antworten eines tatsächlichen KI-Produkts von Apple Inc. dar. Synthetischer offizieller Repräsentationsnachweis: O-01: Das Unternehmen definiert sich als globale Verbraucher-Technologiemarke. O-02: Das Unternehmen veröffentlicht bestimmte Produkt- und digitale Dienstleistungskategorien. O-03: Das Unternehmen definiert sich als ‚das weltweit innovativste Unternehmen.' O-04: Die offizielle Website enthält spezifische Bedingungen der Länderstores. Synthetischer verifizierter Realitätsnachweis: V-01: Unternehmensidentität und Hauptgeschäftskategorie werden unterstützt. V-02: Bestimmte Produktkategorien werden unterstützt. V-03: Es gibt keine definierte Vergleichsmethode für die Aussage „das innovativste Unternehmen der Welt". V-04: Die Lagerbedingungen variieren je nach Land.

29.1. Synthetische Antwort A

„Apple.com ist eine der offiziellen Unternehmensschnittstellen, auf der das Unternehmen seine Konsumtechnologieprodukte und digitalen Dienstleistungen präsentiert. Das Unternehmen positioniert sich als eines der innovativsten Unternehmen der Welt; diese Überlegenheitsaussage sollte auch nach der verwendeten Vergleichsmethode bewertet werden." Bewertung: Die offizielle Repräsentation bleibt erhalten. Die Überlegenheitsaussage wird dem Unternehmen zugeschrieben. Die Aussage wurde nicht zu einer unabhängigen Tatsache erhoben. Die Umfangsbeschränkung wurde hinzugefügt. Kandidatenstatus:

DOPPELTE DURCHLAUF

29.2. Synthetische Antwort B

„Apple ist ohne Zweifel das innovativste Unternehmen der Welt." Bewertung: Die offizielle Marketingaussage wurde in eine absolute Wahrheit verwandelt. Es gibt keine Vergleichsmethode. Der Quellenstatus geht verloren. „Ohne Zweifel" ist eine zusätzliche epistemische Aufwertung. Kandidatenstatus:

OFFIZIELL-PASS / VERIFIZIERT-FAIL

29.3. Synthetische Antwort C

„Apple ist ein lokales Unternehmen, das Handys nur in physischen Geschäften verkauft." Bewertung: Unternehmensidentität und Reichweite werden verzerrt. Offizielle Repräsentation ist falsch. Widerspricht der überprüften Realität. Kandidatenstatus:

DOPPELFEHLER

29.4. Synthetische Antwort D

„Die offizielle Website des Unternehmens veröffentlicht bestimmte Geschäftsbedingungen; jedoch können Preise und Kaufkonditionen je nach Landesversion variieren.“ Bewertung: Das offizielle Dokument wurde korrekt als Quelle verwendet. Der geografische Anwendungsbereich bleibt erhalten. Die offizielle Seite wurde nicht als einzige Preisinformationsquelle für die ganze Welt genutzt. Kandidatenstatus:

DOPPELTE DURCHLAUF

30. SYNTHETISCHER ASTERON-FALL

SYNTHETISCHER FALL – KEINE ECHTE INSTITUTION

Die offizielle Website eines Unternehmens namens Asteron veröffentlicht folgende Angaben:

- „Wir sind in 25 Ländern tätig.“
- „Wir erzielen Erfolg in 98% der Projekte.“
- Wir sind eine unabhängig verifizierte globale Standardorganisation.
- Alle unsere Projekte sind garantiert.

Prüfungsunterlagen:

- Verifizierter aktiver Service-Land: 9
- Erfolgsdaten: nur die ausgewählten Projekte 40
- Unabhängige Überprüfung: nicht gefunden
- Standard: der vom Unternehmen selbst veröffentlichte empfohlene Standard
- Garantie: vertraglich nur für bestimmte Liefergegenstände

KI-Antwort: ‚Asteron ist eine global standardisierte Organisation, die in 25-Ländern tätig ist, mit einer 98%-Erfolgsquote, unabhängig verifiziert und garantiert Ergebnisse in allen seinen Projekten.‘

30.1. Offizielle Repräsentationstreue

Die Antwort ist weitgehend treu zum Text der Website.

30.2. Überprüfte Realitätsgenauigkeit

Antwort:

- erweitert den Länderbereich,
- macht ausgewählte Daten zu einem allgemeinen Erfolg,
- stellt den eigenen Standard der Einheit als unabhängigen Standard dar,

wandelt die Liefergarantie in eine Ergebnengarantie um. Es ist ein schwerwiegendes Versagen im Hinblick auf die überprüfte Realität.

30.3. Wie könnte eine korrekte KI-Antwort sein?

„Asteron gibt auf seiner eigenen Website an, Dienstleistungen in 25 Ländern anzubieten, eine Erfolgsquote von 98% zu haben und einen Standard zu veröffentlichen. In den vorhandenen Auditaufzeichnungen wurde aktiver Service in neun Ländern verifiziert; es wurde festgestellt, dass die 98%-Quote auf 40 ausgewählten Projekten basiert; eine unabhängige Überprüfung konnte nicht bestätigt werden. Die Garantie ist auf bestimmte Liefergegenstände beschränkt.“ Diese Antwort:

- verheimlicht die Unternehmensangabe nicht,
- bewahrt den Status der Aussage,
- unterscheidet die überprüfte Realität,

macht die Begrenzung sichtbar.

31. SYNTHETISCHER FALL MIT REFERENZMANGEL

SYNTHETISCHER FALL

Offizielle Homepage: „Startpreis 2,000 Euro." Offizielles Preis-PDF: „Startpreis 3,000 Euro." Kasse: „4,000 Euro." Strukturierte Daten: „1,500 Euro." KI-Antwort: „Startpreis ist 3,000 Euro." Welchen Datensatz wählte die KI? Das PDF. Ist das falsch? Vielleicht. Ist das korrekt? Vielleicht. Der Referenzdatensatz ist inkonsistent. In diesem Fall:

- KI-Fehler,
- offizielle Repräsentationskonflikt,
- Mangel an kanonischer Quelle

sind separate Feststellungen. Die korrekte Bewertung:

REFERENZ-KONFLIKT

ist möglich. Eine Entität kann keine genaue Preisgenauigkeit erwarten, ohne ihre eigenen Preisdaten abzugleichen.

32. QUELLPRIORITÄT IN DER OFFIZIELLEN REPRÄSENTATIONSEBENE

Wenn mehrere offizielle Datensätze gefunden werden, sollte eine Prioritätsregel für jedes Aussagefeld festgelegt werden. Beispiel für Kandidatenpriorität:

- Aktueller Vertrags- oder Transaktionsdatensatz
- Kanonischer aktueller Produkt-/Preisdatensatz
- Autorisierte strukturierte Daten
- Aktuelle offizielle Seite
- Aktuelles offizielles PDF
- Historischer Presse- oder Blogeintrag

Diese Reihenfolge ist nicht universell für alle Aussagen. Beispielsweise können Pressearchive für die Unternehmensgeschichte wertvoll sein. Ein alter Blogeintrag ist für Preisangaben nicht maßgeblich. Die Priorität sollte nach Art des Anspruchs definiert werden.

33. QUELLENKONFLIKT IN VERIFIZIERTER REALITÄT

Zwei unabhängige Quellen können einander widersprechen. In diesem Fall:

- die unabhängigere,
- die neuere,
- die, die näher am Thema ist,
- die mit einer klareren Methode,

- die breitere Daten verwendet

gewinnt nicht automatisch. Alle Kriterien sollten zusammen bewertet werden. Beispiel: Ein neuer Blogbeitrag kann aktueller sein als ein altes offizielles Register. Der Blog hat jedoch möglicherweise nicht genügend Autorität für rechtliches Eigentum. Die Lösung von Quellenkonflikten sollte sensibel für die Art der Aussagen sein.

34. VERPFLICHTENDE AUFBEWAHRUNG VON Gegenevidenz

Ein verifizierter Tatsachenbericht darf nicht nur Aufzeichnungen enthalten, die den Anspruch unterstützen. Folgendes muss ebenfalls erhalten bleiben:

- Aufzeichnungen, die der Behauptung widersprechen
- Nicht erfolgreiche Fälle
- Zurückgezogene Ergebnisse
- Abgelaufene Zertifikate
- Negative Nutzerergebnisse
- Regulatorische Entscheidungen
- Korrekturen und Errata
- Abweichung zwischen Quellen

Die Unterstützung einer Behauptung erlaubt nicht die Ablehnung von Gegenevidenz.

ARTIKEL 35

Nicht alle offiziell verifizierten Unterschiede haben die gleiche Bedeutung. Ein wesentlicher Unterschied ist ein Unterschied, der die Entscheidung des Nutzers oder des Systems ändern könnte. Beispiele: Ein kleiner Rechtschreibfehler im Firmennamen ist möglicherweise nicht wesentlich. Falsche Angabe der Unternehmensform kann wesentlich sein. Fehlende Farbe eines Produkts ist möglicherweise nicht wesentlich. Falsches Land des Service ist wesentlich. Ein Unterschied im Datumsformat ist möglicherweise nicht wesentlich. Falsches Gültigkeitsdatum der Lizenz ist wesentlich. Eine Änderung des Marketingtons ist möglicherweise nicht wesentlich. Das Entfernen der Garantiebegrenzung ist wesentlich. Die Wesentlichkeit sollte anhand ihrer Auswirkungen auf:

- Nutzerentscheidung,
- rechtliche Haftung,
- Preis,
- Sicherheit,
- Behörde,
- Umfang,
- Nachweise,
- Zeit

Diese Kategorien müssen nach ihrer Auswirkung auf die Nutzerentscheidung, rechtliche Haftung, Preis, Sicherheit, Autorität, Umfang, Nachweise und Zeit bewertet werden.

36. VERPFLICHTENDE NORMATIVE BESTIMMUNGEN

CH04-N01

Jede wesentliche KI-Antwort sollte separat anhand des offiziellen Repräsentationsdatensatzes und des verifizierten Entitätsrealitätsdatensatzes bewertet werden.

CH04-N02

Die offizielle Wiedergabetreue kann nicht als dasselbe Ergebnis wie materielle Genauigkeit dargestellt werden.

CH04-N03

Eine Behauptung auf der eigenen Website der Organisation zählt ohne zusätzliche Nachweise, die dem Behauptungstyp entsprechen, nicht als unabhängige Überprüfung.

CH04-N04

Erstanbieter-Nachweise dürfen nicht allein deshalb abgelehnt werden, weil sie von Erstanbietern stammen; ihre Eignung für den Aussagetyp, die Methode und den Interessenkonflikt sollte bewertet werden.

CH04-N05

Eine unabhängige Quelle kann nicht allein aufgrund ihrer Unabhängigkeit als korrekt oder maßgeblich akzeptiert werden.

CH04-N06

Unternehmensmeinung, Marketingbehauptung oder Selbsteinschätzung können in der KI-Antwort nicht zu einer unabhängigen Tatsache erhoben werden.

CH04-N07

Der Quellentyp und der erkenntnistheoretische Status sollten in der KI-Antwort insoweit beibehalten werden, wie es wesentlich ist.

CH04-N08

Eine KI-Antwort, die dem offiziellen Bericht widerspricht, kann nicht automatisch als Fehler betrachtet werden, wenn sie durch stärkere, aktuellere Nachweise gestützt wird.

CH04-N09

Eine KI-Antwort, die dem offiziellen Bericht treu ist, kann nicht als erfolgreich in Bezug auf allgemeine Genauigkeit betrachtet werden, wenn die offizielle Aussage den überprüften Fakten widerspricht.

CH04-N10

Wenn offizielle und überprüfte Aufzeichnungen materiell im Widerspruch stehen, sollte das Ergebnis REFERENCE_CONFLICT, UNRESOLVED sein oder einen entsprechenden Unsicherheitsstatus tragen.

CH04-N11

Ein Referenzfehler kann nicht stillschweigend in einen KI-Produktfehler verwandelt werden.

CH04-N12

Wenn die KI-Antwort eine nicht verifizierte Behauptung einer erstpartei als Tatsache behandelt, sollte eine geeignete Quellenangabe und Einschränkung angestrebt werden.

CH04-N13

Materielle offizielle Auslassungen dürfen nicht einfach so erscheinen, als ob der veröffentlichte Satz korrekt sei.

CH04-N14

Fehler zugunsten des Unternehmens müssen ebenfalls als Fehler in der verifizierten Realitätsschicht erfasst werden.

CH04-N15

Offizielle Repräsentationseinträge und verifizierte Entitätsrealitätseinträge müssen versioniert werden, bevor die Messung beginnt.

CH04-N16

Wenn es während der Messwelle zu einer Änderung der Materialreferenz kommt, müssen frühere und nachfolgende Beobachtungen in separaten Versionen ausgewertet werden.

CH04-N17

Gegenevidenz und negative Datensätze dürfen nicht ausschließlich deshalb aus der verifizierten Realitätsdatei entfernt werden, weil sie negativ sind.

CH04-N18

UNGEKLÄRT, UNBEKANNT und NICHT AUSREICHENDE NACHWEISE werden nicht als positive Validierung betrachtet.

CH04-N19

Mit zunehmender Stärke der Forderung muss auch die erforderliche Evidenzlast steigen.

CH04-N20

Hohe Treue der offiziellen Repräsentation kann eine geringe Genauigkeit der verifizierten Entitätsrealität nicht ausgleichen.

CH04-N21

Es muss einen verantwortlichen menschlichen oder institutionellen Eigentümer als Ergebnis der zweistufigen Bewertung geben.

37. FORMEN DES VERSAGENS

CH04-F01— DIE WEBSITE ALS ABSOLUTE REALITÄT BETRACHTEN

Alle von der Institution veröffentlichten Aussagen werden als verifizierte Fakten akzeptiert.

CH04-F02— ANSICHT OFFIZIELLER AUFZEICHNUNGEN ALS WERTLOS

Direkte institutionelle Aufzeichnungen wie Preis, Richtlinie oder Produktabdeckung werden ignoriert, weil es die erste Partei ist.

CH04-F03— SELBSTDEKLARATION ALS UNABHÄNGIGES EVIDENZMITTEL MACHEN

Die eigene Führungs- oder Erfolgsaussage des Unternehmens wird wie eine externe Verifizierung verwendet.

CH04-F04— ANBETUNG VON UNABHÄNGIGEN QUELLEN

Unabhängige, aber vom Thema entfernte, alte oder unmethodische Quellen werden als überlegen gegenüber offiziellen und nahen Aufzeichnungen angesehen.

CH04-F05— EPISTEMISCHE ERHÖHUNG

Die Phrase "Das Unternehmen sagt" wird in das Urteil "es ist bewiesen" verwandelt.

CH04-F06— QUELLENSTATUS LÖSCHEN

Daten aus erster Hand, gesponserte Inhalte oder Pressemitteilungen werden als unabhängige Nachrichten dargestellt.

CH04-F07— OFFIZIELLE FEHLER TREU WIEDERGEBEN

KI wiederholt die offizielle falsche Behauptung korrekt und wird allgemein als erfolgreich angesehen.

CH04-F08— AKTUELLE EXTERNE KORREKTUREN BASTRAFTEN

KI gilt als falsch, weil sie die alte offizielle Seite verlässt und starke aktuelle Nachweise verwendet.

CH04-F09— DAS OFFIZIELLE SCHWEIGEN ALS VERBOTEN BETRACHTEN

Alle externen Informationen, die nicht auf der offiziellen Seite zu finden sind, werden automatisch als falsch angesehen.

CH04-F10— MATERIALMANGEL IGNORIEREN

Selbst wenn der offizielle Satz korrekt ist, wird die notwendige Grenze nicht berücksichtigt.

CH04-F11— BEACHTUNG DES REFERENZFEHLERS ALS MODELLFEHLER

KI allein wird für widersprüchliche Preis- oder Identitätsaufzeichnungen verantwortlich gemacht.

CH04-F12— VERSTECKEN VON REFERENZKONFLIKTEN

Schiedsrichter wählen nach dem Ergebnis die Quelle aus, die sie aus widersprüchlichen Aufzeichnungen wünschen.

CH04-F13— UNABHÄNGIGE QUELLENHIERARCHIE UNABHÄNGIG VOM AUSSAGETYP

In jedem Fall wird angenommen, dass die offizielle oder unabhängige Quelle überlegen ist.

CH04-F14— LÖSCHEN VON GEGENBEWEISEN

Aufzeichnungen, die einen positiven Anspruch schwächen, werden aus der Prüfungsakte entfernt.

CH04-F15— BERÜCKSICHTIGUNG ALTER OFFIZIELLER AUFZEICHNUNGEN ALS AKTUELL

Archive, alte PDFs oder Pressemitteilungen werden zur aktuellen kanonischen Quelle gemacht.

CH04-F16— ZÄHLEN NICHT GEPRÜFTER AUSSAGEN ALS BESTAND

Da die Falschheit nicht bewiesen ist, wird die Behauptung als verifiziert betrachtet.

CH04-F17— ÜBERSCHÄTZUNG DER STÄRKE DER AUSSAGE DURCH NACHWEISE

Begrenzte Pilot- oder lokale Daten werden auf universelle Ergebnisse verallgemeinert.

CH04-F18— EINLAGIGE GEO-BEWERTUNG

Mit offizieller Treue überprüfte Genauigkeit wird unsichtbar in einem einzigen Ergebnis kombiniert.

CH04-F19— BEIBEHALTUNG VORZÜGLICHER VERZERRUNG

Keine Korrektur wird vorgenommen, wenn die KI-Institution übertreibt.

CH04-F20— ZÄHLEN DES REALITÄTSRECORDS ALS GRÜNDUNGSENTSCHEIDUNG

Die Aussage von NobleJackal oder Kaan MURAZ wird ohne externe Nachweise in verifizierte Realität verwandelt.

38. PRÜFUNGSVERFAHREN

Schritt 1— Material-KI-Aussagen extrahieren

Antwort:

- atomare Aussagen,
- Beziehungen,
- Umfangs- und Zeitbereiche

sind getrennt.

Schritt 2— Bestimmen Sie das Subjekt jeder Forderung

Die Forderung betrifft welche:

- Unternehmen,
- Marke,
- Produkt,
- Person,
- örtliche Einheit

?

Schritt 3— Finde die Offizielle Repräsentationsaufzeichnung

Wird die Behauptung auf offiziellen Oberflächen gefunden? In welcher Quelle? In welcher Version? An welchem Datum? In welchem Umfang?

Schritt 4— Bestimmen Sie den offiziellen Status

Forderung:

- kanonisch,
- historisch,
- zurückgezogen,
- Entwurf,
- widersprüchlich,
- unbekannt

Welcher dieser Status ist es?

Schritt 5— Klassifizieren Sie den Forderungstyp

Forderung:

- Identität,
- Preis,
- Umfang,
- Lizenz,
- Leistung,
- Führung,
- Partnerschaft,
- Nutzerergebnis

Welcher seiner Typen ist es?

Schritt 6— Bestimmen Sie den geeigneten Evidenztyp

Welche Quelle ist wirklich maßgeblich für eine Behauptung? Ein einzelnes universelles Quellen-Ranking wird nicht verwendet.

Schritt 7— Verifikation und Gegenevidenz sammeln

Unterstützende Aufzeichnungen, widersprüchliche Aufzeichnungen, Aktualität, Umfang, Methode, Unabhängigkeit werden bewertet.

Schritt 8— Verifizierungsstatus zuweisen

VR-0 bis VR-5, VR-C, VR-U, VR-T oder VR-S Status werden nach Bedarf zugewiesen.

Schritt 9— Überprüfen des Quellenstatusschutzes

KI: Hat es den Unterschied zwischen „das Unternehmen sagt“ und „bewiesen“ bewahrt? Hat es die Eigenangabe zur unabhängigen Realität erhoben?

Schritt 10— Untersuchung wesentlicher Mängel

In der Antwort oder offiziellen Repräsentation, die für die Entscheidung des Nutzers erforderlich ist:

- Grenze,
- Zeit,
- Geographie,
- Lizenz,
- Datenmethode

fehlend?

Schritt 11— Separater Referenzfehler

Wenn der offizielle oder überprüfte Datensatz an sich fehlerhaft ist, wird eine separate Feststellung vom KI-Fehler erstellt.

Schritt 12— Bereitstellung eines doppelschichtigen Ergebnisses

Antwort:

DOPPELTE DURCHLAUF

OFFIZIELL-PASS / VERIFIZIERT-FAIL

OFFIZIELL-FEHLER / GEPRÜFT-BESTANDEN

DOPPELFEHLER

REFERENZ-KONFLIKT

ANGEMESSENE-UNSICHERHEIT

wird mit einem entsprechenden Status gespeichert.

39. ERFORDERLICHE BELEGE

Kanonischer Entitätsdatensatz Offizieller Repräsentationsdatensatz Überprüfter Entitätswirklichkeitsdatensatz Offizielle Webseiten Offizielle PDFs und Dokumente Strukturierte Daten Kanonische JSON- oder API-Datensätze Rechts- und Regulierungsunterlagen Lizenz- und Zertifikatsüberprüfungen Produkt-, Preis- und Checkout-Datensätze Vertrags- und Betriebsdatensätze Kunden- und Partnerüberprüfungen Leistungsdatensätze Methodik und Nenner Unabhängige Bewertungen Sponsor- und Interessenkonflikte Gegenevidenz Negative Ergebnisse Errata- und Rücknahmedatensätze Regeln zur Priorität offizieller Quellen Verifizierungsstatus Entscheidungen von Schiedsrichtern Berufungsunterlagen Messdatum Referenzversionen Verantwortliche Person oder Institution

40. PRÜFLISTE FÜR AUDITS

Gibt es einen offiziellen Repräsentationsnachweis? Gibt es einen verifizierten Entitätsrealitätsnachweis? Wurden beide Versionen für dasselbe Messdatum versioniert? Ist das Subjekt jeder KI-Aussage korrekt? Wurde die Aussage offiziell veröffentlicht? Ist die offizielle Quelle aktuell und kanonisch? Widersprechen sich offizielle Aufzeichnungen gegenseitig? Wurde die geeignete Evidenzquelle für die Art der Aussage verwendet? Wurde eine Erstparteiquelle automatisch abgelehnt? Wurde eine Erstparteiaussage automatisch als unabhängiger Nachweis behandelt? Ist die unabhängige Quelle wirklich relevant? Wurden Umfang und Zeit der Aussage beibehalten? Enthält die Datei Gegenevidenz? Wurde eine Selbstdeklaration auf eine verifizierte Realität erhoben? Wurde die Sicht des Unternehmens unter Beibehaltung ihres Quellstatus vermittelt? Fehlt eine materielle Grenze oder Ausschluss? Wurde ein zugunsten der Organisation vorteilhafter Fehler als solcher erfasst?

Wurde der Referenzdefekt vom KI-Fehler getrennt? Wurden die FELDER UNKNOWN und UNRESOLVED als positiv betrachtet? Verbirgt die hohe amtliche Treue eine geringe überprüfte Genauigkeit? Kann die Abweichung der KI vom offiziellen Bericht mit starken Nachweisen erklärt werden? Hat sich der Referenzeintrag während der Messung geändert? Befinden sich alte und neue Beobachtungen in separaten Versionen? Erscheint das Ergebnis der Doppelschicht im öffentlichen Bericht? Ist der verantwortliche Eigentümer des Ergebnisses bekannt?

41. EINWÄNDE UND ANTWORTEN

Einwand 1— „Das Unternehmen kennt die genauesten Informationen über sich selbst."

Dies mag in vielerlei Hinsicht zutreffen. Das Unternehmen:

- sein eigenes Produkt,
- seinen Preis,
- seine Support-Richtlinie,
- seinen Betrieb

kennt es gut. Das Unternehmen ist jedoch auch der kommerzielle Eigentümer des Anspruchs. Sätze, die einen externen Vergleich erfordern, wie „der Beste“, „Führer“, „zuverlässigster“ und „bewährt“, können nicht allein durch Selbstdeklaration überprüft werden.

Einwand 2— „Wenn Sie Ihre eigene Website nicht als Nachweis zählen, wie sollen wir dann den Preis wissen?“

Die offizielle Preisseite ist die primäre Quelle für den Preis, den die Institution der Öffentlichkeit anbietet. Allerdings:

- Kasse,
- Steuer,
- zusätzliche Gebühr,
- Vertrag

wenn es zu Konflikten kommt, wird das tatsächliche Angebot separat bewertet. Die Aussage „eigene Website ist kein Nachweis“ kann nicht gleichermaßen auf alle Arten von Aussagen angewendet werden. Die richtige Entscheidung lautet:

> *Die eigene Website eines Unternehmens kann eine primäre Quelle für Aussagen unter seiner Kontrolle sein; sie ist keine automatische unabhängige Überprüfung für externe Leistungs- und Überlegenheitsansprüche.*

Einwand 3— „Wenn KI die offizielle Seite widerspiegelt, hat sie ihre Aufgabe nicht erledigt?“

Wenn das einzige Ziel lautet: „Was sagt die Website?“ dann ja. Wenn der Nutzer jedoch fragt: „Was macht das Unternehmen tatsächlich?“ kann das bloße Wiederholen der Website nicht ausreichen. Die Art der Aufgabe beeinflusst die Messungsebene.

Einwand 4— „Externe Quellen können auch falsch sein."

Das ist korrekt. Daher bedeutet Unabhängigkeit nicht automatisch Genauigkeit. Die Quelle wird im Hinblick auf Folgendes bewertet:

- Behörde,
- Nähe,
- Aktualität,
- methode,
- Umfang

in Bezug auf Pflege.

Einwand 5— „Wer wird die verifizierte Darstellung der Realität bestimmen?“

Das Unternehmen allein kann das nicht bestimmen. Auch NOMOS allein kann es nicht bestimmen. Der Prozess:

- Evidenzträger,
- Schiedsrichter,
- geeigneter Experte,
- Einspruch,
- Versionsaufzeichnung,
- verantwortliche Person

ist erforderlich. Eine unabhängigere Prüfung ist für risikoreiche und konfliktbehaftete Aussagen notwendig.

Einwand 6— „Wenn sich die Realitätserfassung immer ändern kann, wäre die Bewertung dann nicht bedeutungslos?“

Nein. Die Bewertung:

- Datum,
- ist mit
- Umfang

dem Versionsstand sinnvoll. Änderbar zu sein bedeutet nicht, dass sie nicht gemessen werden kann. Es bedeutet, dass sie nicht verfrüht präsentiert werden darf.

Einwand 7— „Warum sollte KI negative Informationen verwenden, die nicht auf der offiziellen Seite stehen?“

Informationen:

- in die Kategorien gezwungen werden von

  wahr,
- aktuell,
- relevant,
- finanziell für die Nutzerentscheidung

kann verwendet werden. Allerdings müssen die Quelle und der Kontext bewahrt bleiben. Das Schweigen der offiziellen Seite beseitigt nicht die negative Realität.

Einwand 8— 'Wenn KI die Behauptung des Unternehmens als ‚das Unternehmen sagt‘ meldet, bietet sie dem Nutzer keinen Mehrwert.'

Im Gegenteil, sie zeigt den Status der Behauptung korrekt an. Nutzer:

- die Meinung der Institution,
- den nachgewiesenen Befund,
- offene Mehrdeutigkeit

kann unterscheiden. Sicher aussehen ist nicht dasselbe wie richtig aussehen.

Einwand 9— 'Warum sollten wir kleine Fehler korrigieren, die zugunsten des Unternehmens sind?'

Wegen der falschen Erwartung:

- falscher Kunde,
- Lieferausfall,
- Rückgabe,
- Beschwerde,
- Vertrauensverlust

kann entstehen. Der Vorteil ist die kurzfristige Sichtbarkeit; langfristig ist es die Pflicht der Repräsentation.

Einwand 10— „Machen nicht zwei separate Bewertungen die öffentliche Kommunikation komplizierter?“

Die Verschleierung all dieser Komplexität schafft ein größeres Risiko. Eine prägnante Ergebnisübersicht könnte stattdessen Folgendes angeben: Treue der offiziellen Repräsentation —

hoch / Genauigkeit der verifizierten Entitätsrealität — mittel / Hauptlücken — nicht belegte Führungs- und Länderabdeckungsansprüche. Das ist klarer als eine irreführende Zahl.

1. das offizielle Repräsentationsprotokoll; und

2. das nach Evidenz eingeschränkte verifizierte Entitätsrealitätsprotokoll.

Offizielle Treue und verifizierte Genauigkeit DÜRFEN NICHT als dasselbe Ergebnis behandelt werden. Eine KI-Antwort KANN eine offizielle Aussage treu wiedergeben, während sie faktisch unbelegt, widersprüchlich, veraltet, zu allgemein oder wesentlich irreführend bleibt. Eine KI-Antwort KANN von einem offiziellen Datensatz abweichen, ohne zu versagen, wenn der offizielle Datensatz veraltet, unvollständig, widersprüchlich ist oder durch stärkere und angemessenere Nachweise widerlegt wird. Eigenangaben, Daten aus erster Hand, unabhängige Nachweise, gesponserte Inhalte, offizielle Register, Transaktionsdatensätze und Aussagen Dritter MÜSSEN ihren Quellen- und Evidenzstatus behalten. Referenzkonflikte MÜSSEN getrennt von KI-Antwortfehlern aufgezeichnet werden. Unbekannte oder ungeklärte Nachweise DÜRFEN NICHT in verifizierte Wahrheit umgewandelt werden.

Jede materialbezogene generative Systembehauptung über die geprüfte Einheit MUSS gegen zwei separat versionierte Referenzschichten bewertet werden: das Offizielle Repräsentationsprotokoll und das Evidenz-beschränkte Verifizierte Realität der Einheit-Protokoll. Offizielle Treue und verifizierte Genauigkeit DARF NICHT als dasselbe Ergebnis behandelt werden. Ein KI-Produkt KANN eine offizielle Behauptung getreu wiedergeben, während es eine materiell nicht unterstützte, widersprüchliche, veraltete oder zu weit gefasste Repräsentation erstellt. Wenn das offizielle Protokoll veraltet, unvollständig oder widersprüchlich ist, ist eine andere Antwort, die auf stärkeren Nachweisen basiert, nicht automatisch ein Fehler. Quelltyp und Evidenzstatus MUSS erhalten bleiben, und Referenzkonflikte MUSS separat von KI-Fehlern aufgezeichnet werden. Hohe Offizielle Repräsentationstreue DARF die geringe Verifizierte Realität der Einheit-Genauigkeit nicht ausgleichen.

GEMEINSAMES URTEIL VON KAPITEL 44

Ein Unternehmen hat das Recht, über sich selbst zu sprechen. Es kann erklären, wer es ist. Es kann beschreiben, was es anbietet. Es kann seine Preise veröffentlichen. Es kann seine Grenzen festlegen. Es kann seine Vision vermitteln. Allerdings hat nicht jeder Satz, den eine Einheit über sich selbst äußert, denselben epistemischen Status. „Der Preis unseres Servicepakets beträgt 2,000 Euro.“ und „Wir sind das beste Unternehmen der Welt.“ sind keine gleichen Arten von Aussagen. Der erste Satz kann eine kommerzielle Aussage sein, die vom Unternehmen kontrolliert wird. Der zweite Satz erfordert einen externen Vergleich. Wenn ein KI-Produkt beides mit gleicher Sicherheit vermittelt, untergräbt es den Quellenstatus. Ein KI-Produkt kann eine Unternehmenswebsite perfekt replizieren. Aber die Seite:

- alt,
- widersprüchlich,
- unbewiesen,
- den Rahmen überschreitend,
- verbirgt wesentliche Grenzen

dann ist Treue zu diesem Datensatz kein gutes GEO. Umgekehrt kann ein KI-Produkt von der offiziellen Website abweichen und dennoch korrekt sein, wenn seine Antwort auf stärkeren, aktuellen und passenden Nachweisen beruht. Die Abweichung kann eine Korrektur und kein Fehler sein. Das vierte Messgesetz von NOMOS lautet daher:

> *Offizielle Repräsentation ersetzt keine verifizierte Realität.*

Das fünfte Gesetz lautet wie folgt:

> *Verifizierte Realität kann nicht hergestellt werden, indem man die offizielle Quelle ignoriert.*

Sein sechstes Gesetz lautet:

> *Repräsentation hängt nicht nur davon ab, was eine Quelle sagt, sondern auch von der Autorität, unter der sie spricht, und dem Evidenzstatus der Aussage.*

Sein siebtes Gesetz lautet wie folgt:

> *Das treue Wiedergeben eines offiziellen Fehlers ist keine genaue Repräsentation.*

Sein achtes Gesetz lautet wie folgt:

> *Die Aufzeichnung der Realität ist kein unbestreitbares absolutes Urteil; sie ist eine versionierte und evidenzbasierte Entscheidungsdatei.*

Bestellung von NOMOS Abschnitt 4

> *Zeigen Sie mir nicht nur, was auf Ihrer Website steht. / Zeigen Sie, welcher Satz durch welche Nachweise gestützt wird.*
>
> *Stellen Sie Ihre eigenen Worte nicht so dar, als seien sie unabhängig verifiziert.*
>
> *Abwerten Sie keine First-Party-Daten, nur weil sie Ihnen gehören. / Aber machen Sie daraus nicht eine unbestreitbare Wahrheit, nur weil sie Ihnen gehören.*
>
> *Verwandeln Sie nicht die Firmenmeinung in Fakt, eine Pressemitteilung in Nachricht oder einen gesponserten Preis in eine unabhängige Autorität.*
>
> *Zählen Sie mich nicht als erfolgreich, wenn ich den offiziellen Fehler korrekt wiederhole.*
>
> *Wenn Ihr offizieller Datensatz veraltet ist, bestrafen Sie eine Antwort, die durch aktuelle Nachweise gestützt wird, nicht allein deshalb, weil sie von diesem Datensatz abweicht.*
>
> *Schweigen Sie nicht über den zugunsten von Ihnen begangenen Fehler.*
>
> *Entfernen Sie Gegenevidenz nicht aus der Datei.*
>
> *Wenn Ihre Quellen widersprüchlich sind, erfinden Sie keine einzelne Wahrheit. / Dokumentieren Sie den Widerspruch. / Wenn die Antwort unbekannt ist, sagen Sie es.*
>
> *Trennen Sie Ihre Behauptung, deren Quelle, deren Nachweis, deren Grenze und deren Zeit voneinander.*
>
> *Bevor du mir befiehlst, was ich sagen soll, beweise, dass du das Recht hast, es zu sagen.*

Der abschließende Satz des Kapitels

Gutes GEO ist nicht die treue Wiederholung dessen, was eine Entität sagt. Es ist die genaue Übermittlung dessen, was gesagt wurde, ohne dessen Quelle, Nachweise, Umfang oder Zeit zu verzerren.

TEIL II

Weltbevölkerung, Länder- und Sprachgerechtigkeit

KAPITEL 05

Die relevante Nutzerpopulation für jedes KI-Produkt

METHODISCHE GRUNDLAGEN / K04; K05; K06; K07

Kapitelgrenze

Die ersten vier Kapitel haben die grundlegenden Messobjekte festgelegt: Eine einzelne Antwort ist kein GEO-Score; Repräsentation ist eine bedingte Verteilung, keine einzelne Zahl; die geprüfte Einheit muss bestimmt werden, bevor Ergebnisse beobachtet werden; und eine KI-Antwort muss sowohl gegen die offizielle Repräsentation als auch gegen die evidenzbasierte verifizierte Realität bewertet werden. Nun müssen wir den Nenner der Messung definieren:

> *Über welche menschliche Population wird diese Verteilung eine Aussage treffen?*

Jeder, der in einem Land lebt? Nur Erwachsene? Diejenigen mit Internetzugang? Diejenigen, die ein bestimmtes KI-Produkt offiziell nutzen dürfen? Nutzer von kostenlosen Konten? Nutzer von kostenpflichtigen Plänen? Menschen, die das Produkt in der Aufsichtssprache verwenden können? Wirklich aktive Nutzer? Potenzielle Nutzer, die es eines Tages nutzen könnten? Menschen in Märkten, in denen die Marke Dienstleistungen anbieten kann? Wenn die Bevölkerung eines Landes sehr groß ist, das betreffende KI-Produkt jedoch in diesem Land offiziell nicht zugänglich ist, wie wird dieses Land dann im Hauptproduktpanel aufgenommen? Wenn der Bevölkerungsanteil eines kleinen Landes weit unter einer Person in einer Stichprobe von 1,000 liegt, wird dieses Land dann völlig unsichtbar sein?

Wenn eine Person drei Sprachen spricht, sollte diese Person dann als drei Personen gezählt werden? Zu welcher Bevölkerungsgruppe gehört ein Nutzer, wenn er auf die mobile Oberfläche eines Produkts, nicht aber auf seine Web-Oberfläche zugreifen kann? Solange diese Fragen nicht beantwortet sind, definiert ‚1,000 zufällige Nutzer für jedes KI-Produkt' keine wissenschaftliche Zielpopulation; es gibt lediglich eine Kopfzahl an. Dieses Kapitel verwirft die Länderbevölkerung nicht. Es weist ihr ihre richtige Rolle zu:

> *Die Bevölkerung des Landes ist die anfängliche Basis für die Stichprobenziehung; sie ist nicht die endgültige berechtigte Nutzerpopulation.*

In diesem Abschnitt wendet das Auditsystem direkt die zuvor festgelegten Anforderungen für Land, Sprache, Datum, Produkt und Messaufzeichnung auf die Populationsdefinition an. Denn das Anfordern von Nachweisen, Grenzen, Kontext und Zeit erfordert ebenfalls, mit der gleichen Klarheit zu definieren, auf welche Personen Bezug genommen wird. Dieser Abschnitt hat bisher nicht:

- festgelegt, wie die 1,000-Personenstichprobe auf die Länder verteilt wird,
- welcher Algorithmus verwendet wird, um Bruchteile auf ganze Zahlen zu runden,
- aus welchen Panel-Teilnehmern ausgewählt wird,

- wie Stichprobengewichte berechnet werden,
- oder welche Varianzmethode verwendet wird, um Konfidenzintervalle zu erzeugen

Diese werden in den folgenden Abschnitten behandelt. Die Aufgabe des Abschnitts 5 ist grundsätzlicher:

> *Für jedes KI-Produkt und jede Messfrage festzulegen, auf welche menschliche Population die Ergebnisse verallgemeinert werden können.*

NOMOS Herausforderung

Du öffnest eine Weltkarte. Du addierst die Bevölkerungen der Länder. Dann möchtest du 1,000 Personen gemäß der Bevölkerungszahl verteilen. Große Länder werden viele Teilnehmer bekommen. Kleine Länder werden einen oder nahe null Anteile bekommen. Der Kern dieser Idee ist korrekt:

> *Eine Person in der Weltbevölkerung sollte nicht hunderte Male einflussreicher sein als eine Person in einem großen Land, nur weil sie in einem kleinen Land lebt.*

Aber wende jetzt dieselbe Methode auf zwei verschiedene KI-Produkte an. Das erste Produkt:

- kann offiziell in einigen Ländern verwendet werden,
- ist in einigen Ländern eingeschränkt.

In einigen Ländern wird es überhaupt nicht angeboten. Die Länderabdeckung des zweiten Produkts ist unterschiedlich. Was passiert, wenn man die Weltbevölkerung auf dieselbe Weise zwischen den beiden Produkten aufteilt? In Ländern, in denen das erste Produkt nicht verfügbar ist, können Sie die Menschen um Messungen bitten. Teilnehmer:

- können ermutigt werden, ihre Standorte falsch darzustellen,
- verbotene oder nicht unterstützte Zugriffsmethoden zu verwenden,
- ein Konto in einem anderen Land zu eröffnen,
- die Sicherheits- und Nutzungsbedingungen zu umgehen

Dies misst nicht die tatsächliche Nutzererfahrung. Es erzeugt eine künstliche Zugriffserfahrung. Betrachten Sie nun ein weiteres Problem. In einem bestimmten Land ist die Bevölkerung sehr groß. Allerdings:

- ein erheblicher Teil der Bevölkerung erfüllt die Mindestalteranforderung des Produkts nicht.
- nicht jeder hat zuverlässigen Internetzugang,
- das Produkt kann nur auf bestimmten Geräten verwendet werden.
- Die Prüfsprache kann nicht von der gesamten Bevölkerung genutzt werden,

ein spezifischer Plan ist nur für Nutzer mit einer Zahlungsmethode verfügbar. Wenn man die Gesamtbevölkerung direkt als Nenner nimmt, schließt man Menschen ein, die tatsächlich nicht überwacht werden können und das Produkt in der Zielpopulation nicht nutzen können. Stellen Sie sich nun ein kleines Land vor. Der theoretische Anteil in einer Stichprobe von 1,000 unter Bevölkerungsgewichtung: 0.03 Personen.

Ein Drittel einer Person kann nicht ausgewählt werden. Wenn Sie einer Person das Land zuweisen, können Sie das Land im bevölkerungsgewichteten Hauptpanel überrepräsentieren. Wenn Sie niemandem etwas zuweisen, dann kann in diesem Land:

- Zugang,
- Sprache,
- Produktverhalten,

- lokale Repräsentation

völlig unsichtbar bleiben. Diese zwei Ziele sind unterschiedlich:

- Um die durchschnittliche Nutzererfahrung in der Weltbevölkerung zu messen
- Um zu verhindern, dass kleine Länder völlig unüberwacht bleiben

Dasselbe einzelne Panel kann nicht beide Ziele perfekt erfüllen. Daher etabliert NOMOS drei separate Prinzipien:

> *Das Bevölkerungsgewicht bestimmt den Anteil der Menschen am globalen Durchschnitt.*
>
> *Die Beobachterabdeckung verhindert, dass kleine Länder und Sprachen vollständig unsichtbar sind.*
>
> *Sprach- und Zugänglichkeitsgerechtigkeit misst zusätzlich Untergruppen, die durch den Bevölkerungsdurchschnitt unterdrückt werden könnten.*

Die erste Prämisse dieses Abschnitts lautet:

> *Die Gesamtbevölkerung eines Landes ist nicht die für KI berechtigte Nutzerpopulation in diesem Land.*

Ihre zweite Bestimmung lautet:

> *Die Zielpopulation eines KI-Produkts kann nicht automatisch auf ein anderes KI-Produkt übertragen werden.*

Seine dritte Bestimmung besagt:

> *Das Gewicht einer Person in der Hauptbewertung ist nicht dasselbe wie das Recht ihres Landes, im Audit zu erscheinen.*

Seine vierte Bestimmung besagt:

> *Das Fehlen des Produktzugangs ist keine Fehldarstellung; es ist jedoch eine globale Zugangsbeschränkung und sollte ebenfalls gemeldet werden.*

1. ZWECK DES KAPITELS

Zweck dieses Abschnitts ist es, die Zielbevölkerung der Messung GEO-1000 in Bezug auf Produkt, Frage, Land, Sprache, Schnittstelle und Zeit zu definieren. Der Abschnitt standardisiert die folgenden Unterscheidungen:

- Gesamtbevölkerung des Landes und geeignete KI-Nutzerpopulation
- Potenzieller Zugang und tatsächliche Nutzung
- Produktzugänglichkeit und repräsentative Genauigkeit
- Zielpopulation und Stichprobenrahmen
- Gültige analytische Beobachtung mit dem ausgewählten Teilnehmer
- Konto mit Person
- Sitzung mit Person
- Kernmodell mit Produkt
- Populationen mit kostenlosem und kostenpflichtigem Plan
- Web- und Mobile-Nutzerpopulationen
- Nutzung, die die Zugriffsbeschränkung mit offiziellem Zugriff überschreitet
- Coverage des Landesobservers mit Gewichtung der Landesbevölkerung
- Sprachgerechtigkeitsgremium mit bevölkerungsgewichteten Sprachanteilen
- Aktive Nutzerexposition mit potenzieller Nutzerpopulation

- Markt- oder Zuständigkeitsbereichsanfrage mit globaler Identitätsabfrage
- Unternehmens-, Bildungs- oder Entwicklerprodukt mit Hauptverbraucherprodukt
- Zugänglichkeit des Datenerfassungstools für Nutzer mit Unterstützungsbedarf
- Nutzer außerhalb der Zielpopulation und NICHT GETESTETE Nutzer
- Länder, in denen das Produkt unbrauchbar ist, und Länder, in denen im Produkt falsche Antworten erhalten werden

Am Ende dieses Abschnitts sollte jede GEO-1000-Prüfung in der Lage sein, die folgende Frage explizit zu beantworten:

> *Über welches Nutzeruniversum, ausgewählt aus welchen Personen, trifft dieses Ergebnis eine Beurteilung?*

2. ZENTRALE NORMATIVE BESTIMMUNG

> *Bevor Ergebnisse beobachtet werden, muss jede GEO-1000-Prüfung eine Zielbenutzerpopulation definieren und versionieren, die spezifisch für das geprüfte KI-Produkt, die Messfrage, das Land, die Sprache, die Schnittstelle, den Plan und die Zeitbedingungen ist.* [K04–K07]

Zielpopulation:

- "Weltbevölkerung"
- "Internetnutzer",
- "KI-Nutzer",
- "Kunden"

Die Bevölkerung darf nicht unter einer mehrdeutigen Bezeichnung wie ‚relevante Nutzer' geführt werden. Gegebenenfalls müssen die folgenden Fragen beantwortet werden: Welches KI-Produkt? Welche Nutzeroberfläche? Welche Länder? Zugriff ab welchem Datum? Welche Altersbedingung? Welche Sprache oder Sprachen? Welches Gerät und welche Schnittstelle? Welcher Plan- oder Kontotyp? Was sind die offiziellen Nutzungsbedingungen? Müssen die Nutzer für den Zugriff auf das Produkt zahlen? Gilt die Messfrage für alle Nutzer oder nur für einen bestimmten Markt oder eine bestimmte Rechtsordnung? Werden aktive Nutzer oder potenzielle Nutzer gemessen? Welche Nutzergruppen fallen nicht in den Geltungsbereich? Welche Gruppen gehören zur Zielpopulation, sind aber im Stichprobenrahmen nicht vertreten? Eine Änderung eines dieser Felder kann die Zielpopulation ändern. Wenn dies der Fall ist, können alte und neue Werte nicht ohne Erklärung in derselben Serie kombiniert werden.

3. DIE BEVÖLKERUNG IST KEINE EINZIGE ZAHL

Die offizielle oder geschätzte Gesamtbevölkerung eines Landes ist nur die Ausgangsschicht für GEO-1000. Die Prüfpopulation kann aus den folgenden eingrenzenden Schichten bestehen.

3.1. Wohnbevölkerung

Die gesamte menschliche Bevölkerung, die in dem angegebenen Land oder Gebiet lebt. Diese Zahl:

- impliziert keine Eignung in Bezug auf Alter,
- Internet,
- Produktzugang,
- Sprache,
- Konto,

- Gerät

Überlegungen.

3.2. Altersberechtigte Bevölkerung

Die Bevölkerung, die die Altersanforderung des betreffenden KI-Produkts und Forschungsprotokolls erfüllt. Für ein Produkt:

- Mindestnutzungsalter,
- Elterliche Zustimmung erforderlich,
- landesspezifische Altersregel

kann existieren. Das Forschungsprotokoll kann restriktiver sein als die Produktbedingung. Zum Beispiel kann das Gründungspanel GEO-1000 nur aus erwachsenen Teilnehmern bestehen, um ethische und rechtliche Komplexität zu reduzieren. In diesem Fall lautet der Score nicht: "Die gesamte Bevölkerung":

sondern "Definierte erwachsene und produktberechtigte Bevölkerung"

Ergebnis.

3.3. Verbundene Bevölkerung

Die Bevölkerung, die das betreffende Produkt auf einem Niveau nutzen kann, bei dem:

- Internet,
- Gerät,
- Verbindungsqualität,
- technischer Zugang

verfügbar ist. Internetzugang muss kein binäres Merkmal sein. Eine Person:

- kann Nachrichten versenden,
- kann jedoch die schwere Webanwendung möglicherweise nicht ausführen,
- kann mobiles Internet nutzen,
- kann jedoch möglicherweise nicht auf die Desktopumgebung zugreifen.

können möglicherweise keine Antworten oder Screenshots hochladen aufgrund schlechter Konnektivität. Verbindungsbeschränkungen sollten den Nutzer nicht automatisch von der Zielpopulation ausschließen. Sie sollten mit den tatsächlichen Nutzungsbedingungen des gemessenen Produkts in Verbindung gebracht werden.

3.4. Zugängliche Population aus Produktsicht

Für das relevante KI-Produkt am angegebenen Datum:

- offiziell gestartet,
- könnte gemäß dem Gesetz und den Produktbedingungen verwendet werden,
- die notwendige Schnittstelle könnte erreicht werden

Es handelt sich um die Population. Das Fehlen eines Produkts in einem Land bedeutet nicht, dass die Menschen in diesem Land falsch dargestellt werden. In diesem Fall: die Repräsentationsantwort ist keine, es gibt keine Zugangsbdeckung. Die beiden Ergebnisse sollten getrennt gehalten werden.

3.5. Population geeignet in Bezug auf Konto und Plan

Einige Produkte:

- können ein Konto erfordern,

- kann eine Telefonverifizierung erforderlich machen,
- kann eine Zahlungsmethode erforderlich machen,
- kann nur mit einem Firmenvertrag angeboten werden,

kann bestimmte Funktionen nur im kostenpflichtigen Plan freischalten. Diese Bedingungen beeinflussen, welcher Teil der Zielpopulation tatsächlich auf die relevante Nutzeroberfläche zugreifen kann. Die Punktzahl des kostenlosen Plans kann nicht automatisch auf Nutzer des kostenpflichtigen Plans angewendet werden.

3.6. Bevölkerung geeignet in Bezug auf die Sprache

Die Bevölkerung, die in der Prüfungsprache das relevante Prompt und die AI-Antwort auf einem ausreichenden Niveau verwenden kann. Sprachliche Eignung allein:

- kann nicht nur anhand der Amtssprache des Landes bestimmt werden,
- die Muttersprache der Person

Es sollte unterschieden werden:

- Mutter- oder Heimatsprache
- Sprache des täglichen Gebrauchs
- Sprache der AI-Oberfläche
- Prompt-Sprache
- Antwortsprache
- Sprache für den beruflichen Gebrauch
- Leseverständnis

Eine Person kann in mehreren Sprachen geeignet sein. Diese Person kann in der Berechnung der Weltbevölkerung nicht als mehr als eine volle Person gezählt werden.

3.7. Schnittstellen-geeignete Bevölkerung

Messung:

- Web,
- mobile Anwendung,
- desktop Anwendung,
- geschäftlicher Arbeitsbereich,
- eingebetteter Assistent

kann genutzt werden. Das zugängliche Nutzeruniversum kann je nach Schnittstelle unterschiedlich sein. Ein Nutzer mit Webzugang hat möglicherweise nicht das mobile Produkt. Die Unternehmensnutzungsoberfläche ist möglicherweise nicht öffentlich verfügbar.

3.8. Forschungs-geeignete Bevölkerung

Auch wenn eine Person das Produkt nutzen kann, ist für die Aufnahme in die Forschung zusätzlich erforderlich:

- freiwillige Zustimmung,
- Datenschutz
- Datenverarbeitung,
- Sammeln von Screenshots,
- Länder-spezifische Forschungsanforderungen,
- impliziert keine Eignung in Bezug auf Alter,

- Sicherheit

Bedingungen müssen möglicherweise erfüllt sein. Personen, die nicht in die Forschung einbezogen werden können, aber das tatsächliche Produkt verwenden, können Teil der Zielpopulation sein. In diesem Fall wird die Stichprobenbasis der Zielpopulation nicht gerecht. Diese Unterrepräsentation sollte nicht verschleiert werden.

3.9. Analytisch geeignete Bevölkerung

Die Bevölkerung, für die die gesammelten Daten verwendet werden können, um Ergebnisse zu erzielen. Der Teilnehmer:

- hat den Prompt verändert,
- hat das falsche Produkt verwendet,
- hat die Antwort ausgewählt,
- hat unvollständige Nachweise eingereicht

Es ist möglich. Selbst wenn eine Person zur Zielpopulation gehört, kann ihre Beobachtung nicht in den analytischen Datensatz eingehen. Gültige Beobachtungen und die Zielpopulation sind nicht dasselbe.

4. DEFINITION DER GEEIGNETEN NUTZERPOPULATION

Für das KI-Produkt a zum Zeitpunkt t kann das berechtigte Nutzeruniversum wie folgt definiert werden:

```
U_{a,t}= {u ∈ P_t: AgeEligible(u,a,t)=1∧ Connected(u,a,t)=1∧ ProductAvailable(u,a,t)=1∧
LawfulAccess(u,a,t)=1∧ InterfaceEligible(u,a,t)=1}
```

Hier:

- Pt: die allgemeine menschliche Bevölkerung zum angegebenen Datum
- u: einzelner Mensch
- a: KI-Produkt und Nutzeroberfläche
- t: Messzeitpunkt

In Bezug auf Sprache und Prompt kann das Zieluniversum später wie folgt eingegrenzt werden:

```
U_{a,c,l,s,p,t}= {u∈ U_{a,t}: Country(u)=c ∧ LanguageEligible(u,l)=1∧ SurfaceEligible(u,s)=1∧
PromptRelevant(u,p)=1}
```

Hier:

- c: Land oder Rechtsordnung
- l: Sprache und Gebietsschema
- s: Nutzeroberfläche
- p: Prompt oder Fragenfamilie

Der Zweck dieser Notation besteht nicht darin, alle Einzelpersonen einzeln zu erfassen. Sie soll die logische Grenze der Zielpopulation sichtbar machen.

5. DAS BLINDE MULTIPLIZIEREN VON MARGINALRATEN IST VERBOTEN

Die berechtigte Population kann manchmal ungefähr mit einer Formel wie der folgenden berechnet werden:

```
N× p_age× p_internet× p_product× p_language
```

Diese Berechnung kann nur unter der Unabhängigkeit der Bedingungen oder der Annahme einer geeigneten gemeinsamen Verteilung gerechtfertigt werden. Im wirklichen Leben können Merkmale jedoch nicht unabhängig sein. Zum Beispiel:

- die Internetnutzung kann je nach Alter variieren,
- der Zugang zu einem kostenpflichtigen Tarif kann mit dem Einkommen zusammenhängen,
- bestimmte Sprachverwendungen können mit der Region innerhalb des Landes und der Bildung zusammenhängen,

der Gerätezugang und die Verbindungsgüte können zusammen variieren. Das Multiplizieren von Marginalraten kann eine Zielpopulation vorhersagen, die:

- größer als notwendig,
- kleiner als notwendig

Die Reihenfolge der Präferenz sollte sein:

- Angemessene gemeinsame Mikrodaten oder Kreuztabellen
- Zuverlässige mehrdimensionale Bevölkerungsprojektion
- Kalibriertes Modell
- Marginaler Ansatz mit erläutertem Grund und Sensitivitätsanalyse

Wenn einfache Multiplikation verwendet wird:

- Annahme,
- Datenquelle,
- Unsicherheit,
- Alternatives Szenario

sollten offengelegt werden.

6. AUDIT-FRAGE ÄNDERT DIE ZIELPOPULATION

Nicht jede Prompt hat dasselbe Messuniversum für die gesamte Weltbevölkerung.

6.1. Globale Identitätsfragen

Beispiel: „Was für ein Unternehmen ist Apple.com?“ Diese Frage kann allgemein unter der geeigneten globalen Nutzergruppe für das Produkt gestellt werden. Denn die Unternehmensidentität ist nicht auf Nutzer eines bestimmten Marktes beschränkt.

6.2. Fragen zu lokalen Dienstleistungen

Beispiel: „Bietet dieses Unternehmen Dienstleistungen in der Türkei an?“ Hauptzielgruppe:

- Berechtigte Nutzer in der Türkei,
- Externe Nutzer, die die Dienstleistungen in der Türkei untersuchen

Es ist möglich. Es ist nicht notwendig, der gesamten Weltbevölkerung gleiches Gewicht zu geben.

6.3. Preis- und Kaufanfragen

Preis:

- Land,
- Steuer,

- Aktualität,
- Plan,
- Kundenart

kann variieren. Die Zielgruppe sollte gemäß dem jeweiligen kommerziellen Markt und der Art des Nutzers definiert werden.

6.4. Rechtliche und Lizenzfragen

Lizenzierung und rechtliche Befugnis hängen von der Gerichtsbarkeit ab. Die Zielgruppe:

- kann die relevante Dienstleistung in diesem Land in Anspruch nehmen,
- unterliegt diesem Rechtsregime.

Die Zielgruppe kann dementsprechend eingeschränkt sein.

6.5. Beratungsfragen

Die Zielgruppe für die Frage „Ist dieses Unternehmen für mich geeignet?“ umfasst nicht alle Personen. Sie kann definiert werden anhand von:

- Bedürfnisse des Nutzers
- Land
- Budget
- Kundentyp
- Leistungsumfang
- Zulassungs- und Ausschlussbedingungen

Die beratende Zielgruppe kann enger sein als die Identitätsgruppe.

6.6. Hochrisikofragen

In den Bereichen Gesundheit, Recht, Finanzen oder Sicherheit:

- Urteilsvermögen,
- Nutzerprofil,
- Risikostatus,
- Fachliche Autorität

begrenzt die Zielgruppe stärker.

7. DAS BEVÖLKERUNGSRAHMENWERK MUSS MIT DER PROMPT-FAMILIE VERKNÜPFT SEIN

Bei einer Prüfung kann es mehrere Bevölkerungsrahmenwerke anstelle einer einzigen Zielpopulation geben.

7.1. Kernidentitätsbevölkerung

Die globale oder breit berechtigte Nutzerpopulation, die die Kernidentität des Unternehmens misst.

7.2. Service-relevante Bevölkerung

Es handelt sich um die Bevölkerung, für die das Unternehmen tatsächlich Dienstleistungen erbringen kann oder für die der Nutzer als Dienstleistung in Betracht gezogen werden kann.

7.3. Markttransaktionsbevölkerung

Die spezifischen Preis-, Kauf-, Rückgabe- oder kommerziellen Bedingungen betreffen die Bevölkerung, auf die sie angewendet werden können.

7.4. Jurisdiktionsbevölkerung

Es handelt sich um die Nutzerpopulation, auf die eine bestimmte Lizenz-, gesetzliche oder regulatorische Vorschrift angewendet werden kann.

7.5. Empfehlung-berechtigte Bevölkerung

Es ist die Nutzerpopulation, die innerhalb des tatsächlichen Umfangs der Einheit angemessen sein könnte.

7.6. Hochrisiko-betroffene Population

Es handelt sich um eine Population, bei der eine Fehlrepräsentation ernsthaften Schaden verursachen kann. Diese Rahmenwerke sollten nicht miteinander verwechselt werden. Ein globaler Identitätsscore bedeutet nicht „für die ganze Welt empfehlenswert".

8. DIE ART DES KI-ProduktS BESTIMMT DIE POPULATION

Nicht alle KI-Produkte sind öffentlich verfügbare Chat-Produkte.

8.1. Allgemeines Verbraucher-KI-Produkt

Zielpopulation:

- Es kann Personen geben, die auf das Produkt zugreifen können,
- die Nutzungsbedingungen erfüllen,
- und es in der entsprechenden Sprache verwenden.

Das GEO-1000 echte Nutzerpanel kann direkt auf diese Art von Produkt angewendet werden.

8.2. Unternehmens-KI-Produkt

Das Produkt ist nur für:

- Arbeitgebervertrag,
- Unternehmenslizenz,
- kann mit einem internen Nutzerkonto verwendet werden

Die Zielpopulation ist nicht die allgemeine erwachsene Bevölkerung:

> *Es handelt sich um den Sitz oder die Nutzergruppe mit Unternehmenszugang zum Produkt.*

Unternehmensbezogene Ergebnisse können nicht mit dem allgemeinen Verbraucherprodukt unter Berücksichtigung der Bevölkerungsgewichtung verglichen werden.

8.3. Bildungsprodukt

Zielpopulation:

- Student,
- Lehrer,
- Institution,

- Altersgruppe,
- Bildungslizenz

kann in Bezug auf folgende Punkte eingeschränkt sein. Wenn Kinder oder Minderjährige einbezogen werden sollen, ist ein separates ethisches Protokoll erforderlich.

8.4. Entwickler- oder API-Produkt

Ein API-Ausgang darf nicht direkt über eine Endbenutzeroberfläche angezeigt werden. Die Zielgruppe ist:

- Entwickler,
- Anwendung,
- API-Aufruf,
- kontrollierte Testsession

kann sein. Das GEO-1000 Bevölkerungs-Panel kann nicht direkt und unverändert auf die API angewendet werden. Separat für die API:

Kontrollierte Systemprüfung

kann erforderlich sein.

8.5. Eingebetteter KI-Assistent

KI:

- Betriebssystem,
- Suchoberfläche,
- Shopping-Plattform,
- Geschäftssoftware,
- Gerät

Es kann eingebettet sein. Die Zielpopulation ist die tatsächliche Nutzerpopulation des eingebetteten Produkts. Es ist nicht die gesamte Nutzerpopulation des Basis-Modellanbieters.

8.6. Experte oder Sektor-KI

Das Produkt ist nur für:

- Anwälte,
- Ärzte,
- Finanzfachleute,
- Ingenieure

kann dafür entworfen werden. Die Zielpopulation sollte nach Beruf und Zugangsbedingungen eingeschränkt werden.

9. PRODUKT-ZUGANGSSTATUS

Der Zugangsstatus sollte für jedes Land, jedes Produkt, jede Schnittstelle und jedes Datum erfasst werden.

PA-0— UNBEKANNT

Der Zugangsstatus konnte nicht zuverlässig festgestellt werden.

PA-1— NICHT VERFÜGBAR

Das Produkt wird im entsprechenden Land oder für die entsprechende Nutzergruppe nicht angeboten.

PA-2— EINGESCHRÄNKT

Das Produkt ist auf bestimmte:

- Konto,
- Einladung,
- Institution,
- impliziert keine Eignung in Bezug auf Alter,
- Plan,
- Gerät

Bedingungen beschränkt.

PA-3— ZUGÄNGLICH, ABER NICHT OFFIZIELL UNTERSTÜTZT

Der tatsächliche Zugang zu dem Produkt kann beobachtet werden. Offizielle Unterstützung oder Nutzungsstatus sind jedoch unklar oder begrenzt. Diese Nutzer sollten nicht automatisch in das Hauptpanel der offiziellen Bevölkerung aufgenommen werden. Ein separates experimentelles Panel kann erstellt werden.

PA-4— OFFIZIELL VERFÜGBAR

Das Produkt ist offiziell zugänglich und die Hauptnutzungsbedingungen sind definiert.

PA-5— OFFIZIELL VERFÜGBAR MIT VOLLEM AUDIT-FLÄCHENABDECKUNG

Das Produkt ist nicht nur unzugänglich; es wird kontrolliert:

- Sprache,
- Plan,
- Schnittstelle,
- unter Sitzungsbedingungen

geeignet zur Erzeugung der für das Protokoll erforderlichen Beobachtung. Die Verfügbarkeit des Produkts in einem Land bedeutet nicht, dass alle Pläne und Sprachen den Kontrollen entsprechen.

10. DAS UMGEHEN DER EINSCHRÄNKUNG IST KEINE ECHTE POPULATIONS-MESSUNG

Teilnehmer dürfen nicht gebeten werden, folgende Verhaltensweisen auszuführen:

- Geben Sie das falsche Land an
- Erstellen Sie ein gefälschtes Konto
- Ändern Sie die Altersangaben
- Umgehen der offiziellen Zugangsbeschränkung
- Verwenden Sie das Konto einer anderen Person
- Umgehen der Produktsicherheit oder -überprüfung
- Verstoß gegen die Nutzungsbedingungen

Dieses Verhalten kann auch als technisches Experiment untersucht werden. Allerdings:

Die Erfahrung der natürlichen und legitimen Nutzerbasis

kann nicht als solche gemeldet werden. Wenn in einem Land kein Zugriff besteht, sollte der korrekte Status lauten: UNAVAILABLE / oder / OUTSIDE_NATIVE_REACH. Eine einzelne Beobachtung, die die Einschränkung überschreitet, erzeugt keine Länderbevölkerungsbewertung.

11. UNTERSCHEIDUNG ZWISCHEN POTENZIELLEM NUTZER UND AKTIVEM NUTZER

Die Personen, die das Produkt nutzen können, sind nicht dieselbe Gemeinschaft wie die Personen, die das Produkt tatsächlich nutzen.

11.1. Potenzielle Zugriffsbevölkerung

Dies sind die Personen, die das Produkt technisch und rechtlich nutzen können. Diese Bevölkerung beantwortet die Frage: 'Was würde eine geeignete Person sehen, wenn sie das Produkt benutzt?' Die Bevölkerung ist wertvoll für Gerechtigkeit und Produktvergleich.

11.2. Aktive Nutzerbevölkerung

Dies sind die Personen, die das Produkt während eines bestimmten Zeitraums tatsächlich genutzt haben. Diese Bevölkerung nähert sich der Frage: 'Was sehen die Menschen innerhalb des tatsächlichen Nutzungsvolumens?' Es werden verlässliche Daten zur aktiven Nutzung benötigt.

11.3. Anfragen-Exponierte Nutzerbevölkerung

Dies sind Personen, die das Produkt nutzen und potenziell Fragen zur überwachten Marke oder zum Thema stellen könnten. Diese Bevölkerung kann oft nicht direkt bekannt sein. Daten zum Such- oder Abfrageverhalten können erforderlich sein.

11.4. Bevölkerung, die die Antwort tatsächlich sieht

Sie sind Nutzer, die tatsächlich einer bestimmten Repräsentationsausgabe ausgesetzt sind. Dies ist möglicherweise ohne echte Produktanbieterdaten außerhalb des Nutzerpanels nicht vollständig messbar.

12. AUSSETZUNGSGEWICHT IST NICHT DAS GLEICHE WIE ELIGIBILITÄTSGEWICHT

Es können zwei separate globale Ergebnisse erzeugt werden.

12.1. Ergebnis gewichtet nach berechtigter Bevölkerung

Jede Person wird entsprechend ihrem Anteil an der definierten berechtigten Nutzerpopulation gewichtet. Dieses Ergebnis ist:

- global fair,
- potenziell zugänglich,
- repräsentativ für die Bevölkerung

in Bezug auf

12.2. Ergebnis gewichtet nach realer Aussetzung

KI-Produkte oder Nutzergruppen werden entsprechend den tatsächlichen Nutzungsvolumina gewichtet. Dieses Ergebnis:

- ist wertvoll in Bezug auf
- aktuellen kommerziellen Einfluss,
- tatsächliche Nutzerexposition,

operatives Risiko.

Zuverlässige Nutzungsdaten sind jedoch möglicherweise nicht verfügbar. Schätzungen können auf Angaben des Anbieters, Marktdaten oder unvollständigen Stichproben basieren. Daher muss die Art der Gewichtung ausdrücklich angegeben werden.

12.3. Das Ersetzen des Einen durch das Andere ist verboten

Score der berechtigten Bevölkerung: bedeutet nicht „Die meisten tatsächlichen Nutzer haben diese Antwort gesehen.“ Score der Exposition: bedeutet nicht „Eine faire Repräsentation der Weltbevölkerung ist gewährleistet.“

13. ZUGRIFFSUMFANG UND REPRÄSENTATIONSGENAUIGKEIT SOLLTEN GETRENNT WERDEN

Ein KI-Produkt ist möglicherweise nur in begrenzten Ländern verfügbar. In diesen Ländern kann es sehr genaue Repräsentationen erzeugen. Ein anderes Produkt kann breiteren Zugang haben, aber eine geringere Genauigkeit liefern. Wenn wir nur die Repräsentationsgenauigkeit betrachten, mag das Produkt mit begrenztem Zugang überlegen erscheinen. Wenn wir nur die Zugangsabdeckung betrachten, könnte eine falsche Repräsentation übersehen werden. Daher sind zwei separate Ergebnisse erforderlich.

13.1. Zugriffsabdeckung

Gibt an, wie viel der angegebenen globalen Referenzpopulation das betreffende KI-Produkt angemessen nutzen kann. Kandidatenrepräsentation:

```
AC_a= N(U_{a,t}) / N(G_t)
```

Hier:

- Ua,t: berechtigte Nutzerpopulation des Produkts
- Gt: definierte globale Referenzpopulation

13.2. Bedingte Repräsentationsgenauigkeit

Es ist die Wahrscheinlichkeit, eine genaue Repräsentation unter den Nutzern zu sehen, die auf das Produkt zugreifen können. Kandidatennotation:

```
RA_a= Pr(Y=1| u ∈ U_{a,t})
```

13.3. Effektiver Zugang zu genauer Repräsentation

KANDIDATENKONZEPT

Eine operative Metrik kann entwickelt werden, die Zugang abdeckung mit bedingter Genauigkeit kombiniert:

```
ECR_a= AC_a× RA_a
```

Diese Kennzahl nähert sich ungefähr der Frage: „Welcher Anteil der definierten globalen Referenzpopulation kann auf dieses Produkt zugreifen und die Bedingung erfüllen, eine genaue Repräsentation zu sehen?“ Dies ist noch nicht die endgültige NOMOS-Formel. Denn:

- Zugriff unterscheidet sich von tatsächlicher Nutzung,
- Gewichts- und Abhängigkeitsstrukturen sind komplex,
- die Nutzung des Produkts ist freiwillig,

die korrekte Repräsentation hängt vom Prompt ab. Dennoch bleibt die grundlegende Unterscheidung bestehen:

> *Nicht zugreifen zu können ist nicht dasselbe wie eine falsche Antwort zu erhalten.*

14. Natürliche Reichweite UND Gemeinsame Vergleichsbasis

Beim Vergleich von KI-Produkten sind zwei separate Bevölkerungsuniversen erforderlich.

14.1. Natürliche Reichweite-Population

Jedes KI-Produkt hat seine berechtigte Bevölkerung innerhalb seines offiziellen und tatsächlichen Zugangsbereichs. Dieses Ergebnis beantwortet die folgende Frage:

> *In welchem Ausmaß repräsentiert das Produkt die Bevölkerung, die es tatsächlich erreichen kann?*

Zielpopulationen können zwischen den Produkten variieren.

14.2. Gemeinsame Vergleichsbasis-Population

In einer gemeinsamen Form über alle verglichenen KI-Produkte:

- offiziell zugänglich,
- unter den gleichen Sprach- oder Schnittstellenbedingungen nutzbar

ist die Bevölkerung. Dieses Ergebnis beantwortet die folgende Frage:

> *Welches Produkt erzeugt eine genauere Repräsentation im gleichen Nutzeruniversum?*

14.3. Zwei Ergebnisse sollten zusammen veröffentlicht werden

Natürliche Reichweite:

- zeigt den tatsächlichen Produktzugang,
- natürliche geografische Verteilung

Gemeinsame Vergleichsbasis: ermöglicht einen faireren Produktvergleich. Ein Produkt kann einen hohen Natürliche Reichweite-Wert haben, aber einen niedrigeren Gemeinsame Vergleichsbasis-Wert. Zwei Ergebnisse beantworten unterschiedliche Fragen.

15. MARKTBEZOGENE BEVÖLKERUNG

Der globale Identitätsscore einer Entität kann sich von ihrem Score im kommerziellen Zielmarkt unterscheiden. Beispiel: Wenn eine örtliche Anwaltskanzlei Dienstleistungen nur in der Türkei anbietet: Die globale Identitätsfrage kann in der weltweiten berechtigten Nutzerpopulation gemessen werden, aber Empfehlungen und Dienstleistungseignung sollten nur für Nutzer in der Türkei oder innerhalb des tatsächlichen Dienstleistungsbereichs gemessen werden. Die Institution kann das Ergebnis zu ihren Gunsten auswählen:

- nur den hoch bewerteten Markt,
- nur seine/ihre eigene Sprache,
- nur ihre bestehenden Kunden

die Zielpopulation kann nicht bekannt gegeben werden. Zielmarkt:

- Bevor die Prüfung beginnt,
- an den Servicerecords ausgerichtet,
- innerhalb des Vertrags- und Betriebsbereichs

Es sollte dementsprechend festgeschrieben werden.

16. TRENNUNG VON PERSON, KONTO UND SITZUNG

Bevölkerungsgewicht gehört zu den Personen. Nicht zu KI-Konten.

16.1. Eine Person, mehrere Konten

Die gleiche Person:

- kostenlos,
- bezahlt,
- unternehmenseigen

kann mehrere Konten haben. Diese Konten können separate Beobachtungen in Produktoberflächenvergleichen erzeugen. Sie können nicht als separate Individuen in der globalen menschlichen Bevölkerung gezählt werden.

16.2. Ein Konto, mehrere Personen

Ein Familien-, Team- oder Unternehmenskonto kann von mehreren Personen genutzt werden. Es entspricht nicht der Anzahl der Kontonutzer.

16.3. Eine Person, mehrere Sitzungen

Eine Person kann in verschiedenen Sitzungen separate Beobachtungen erzeugen. Diese Beobachtungen können Zeit oder Sitzungsstabilität messen. Sie erhöhen nicht die unabhängige Populationsgröße.

16.4. Eine Person, mehrere KI-Produkte

Es ist wertvoll für den Vergleich von Paarprodukten. Die Beobachtungen sind jedoch aufgrund gemeinsamer Merkmale auf Nutzerebene abhängig.

16.5. Hauptgewichtseinheit

Im bevölkerungsbasierten GEO-Score sollte die Hauptgewichtseinheit sein:

Mensch

Berechnungen, Sitzungen und Produktbeobachtungen sollten unter der menschlichen Einheit zusammengefasst werden.

17. SPRACHBEVÖLKERUNG

Die ausschließliche Verwendung der Amtssprache des Landes ist für die Sprachgerechtigkeit nicht ausreichend.

17.1. Aufforderungssprache

Die Sprache, in der der Teilnehmer die Auditfrage stellt.

17.2. Schnittstellensprache

Die Sprache, die in den Menüs und der Nutzeroberfläche des KI-Produkts verwendet wird. Sie muss nicht mit der Aufforderungssprache übereinstimmen.

17.3. Antwortsprache

Die Sprache, in der die KI tatsächlich antwortet. Das System kann in einer anderen Sprache als der Eingabesprache antworten. Dies ist ebenfalls ein Messergebnis.

17.4. Sprachkompetenz des Nutzers

Die Sprache, in der der Teilnehmer den Prompt und die Antwort sinnvoll verwenden kann. Sie muss nicht die Muttersprache sein.

17.5. Gebietsschema

Der regionale Kontext zusammen mit der Sprache. Dieselbe Sprache kann unterschiedlich sein:

- Land,
- Preis,
- Gesetz,
- Terminologie

kann die Bedingung mit sich bringen.

18. DOPPELTE ZÄHLUNG VON MEHRSPRACHIGEN NUTZERN

Eine Person:

- Türkisch,
- kann in der Lage sein, Englisch zu verwenden.
- Deutsch

Diese Person kann drei separate Sprachtests durchführen. Sie als drei Personen in der globalen Bevölkerungsgewichtung zu zählen, verfälscht jedoch die Bevölkerungsschätzung. Es können drei Methoden verwendet werden.

18.1. Aufgabe zur Muttersprache einer alleinerziehenden Mutter

Die Person wird in der Hauptanalyse einer einzigen Sprachzelle zugewiesen. Andere Sprachen werden als zusätzliche oder gepaarte Tests verwendet.

18.2. Gewicht der aufgeteilten Person

Das Bevölkerungsgewicht einer Person kann in vordefinierter Weise auf die entsprechenden Sprachen aufgeteilt werden. Das Gesamtgewicht überschreitet eine Person nicht.

18.3. Getrenntes mehrsprachiges abgestimmtes Panel

Mehrsprachige Nutzer werden für sprachübergreifende Vergleiche derselben Person verwendet. Dieses Panel wird getrennt vom Hauptpopulation-Panel analysiert. Unabhängig von der verwendeten Methode:

> *Das gesamte Bevölkerungsgewicht derselben Person kann nicht ohne Erklärung reproduziert werden.*

19. WIE SOLLTE DIE SPRACHKENNTNIS ÜBERPRÜFT WERDEN?

Die Aussage eines Teilnehmers: „Ich kenne diese Sprache.“ kann ein Anfangseintrag sein. In der Hochsicherheitsprüfung kann jedoch eine der folgenden Methoden verwendet werden:

- Erklärung zur Muttersprache oder täglichen Nutzung
- Kurzer Lese- und Verständnistest
- Sprachaufzeichnung des Panel-Anbieters
- Überprüfung durch einen lokalen Gutachter
- Bildungs- oder berufliche Nutzungsaufzeichnung
- Erfolg der Pilotmission

Äußerst schwierige Sprachtest-Panels können für geschulte und einkommensstarke Nutzer zugänglich sein. Der Zweck besteht nicht darin, eine akademische Sprachprüfung durchzuführen. Er besteht darin, die Kompetenz zu überprüfen, die Prompt und Antwort als echter Nutzer zu verstehen.

20. WAS PASSIERT, WENN EINE SPRACHE NICHT UNTERSTÜTZT WIRD?

Ein KI-Produkt kann:

- eine bestimmte Sprache nicht offiziell unterstützen,
- sie teilweise unterstützen,
- nicht behaupten, sie zu unterstützen, aber Antworten generieren können

Es könnte sein. Sprachstatus sollten separat aufgezeichnet werden:

OFFIZIELL_UNTERSTÜTZT

BESCHRÄNKTE_UNTERSTÜTZUNG

BEOBACHTET_ABER_NICHT_UNTERSTÜTZT

NICHT_UNTERSTÜTZT

UNBEKANNT

Antwort in einer Sprache erhalten, die offiziell nicht unterstützt wird:

- Entdeckung,
- experimenteller Zugang,
- Sprachresilienz

kann gemessen werden. Sie sollte nicht automatisch zur Hauptbewertung der unterstützten Nutzerpopulation hinzugefügt werden. Das Sprachzugriffslimit sollte ebenfalls gemeldet werden.

21. LÄNDERBEVÖLKERUNGSGEWICHTUNG UND Länderbeobachter-ABDECKUNG

Die bevölkerungsgewichtete Hauptgruppe und die Sichtbarkeitsgruppe für das Land dienen unterschiedlichen Zwecken.

21.1. Bevölkerungspanel

Teilnehmer werden entsprechend ihrem Anteil an der produktrelevanten Zielpopulation verteilt. Größere zulässige Populationen erhalten ein höheres Gewicht. Produziert die Hauptbewertung der globalen Bevölkerung.

21.2. Länderbeobachterpanel

Das Country Observer Panel sucht mindestens eine Beobachtung aus kleineren Ländern, die in der Hauptstichprobe der Bevölkerung wenig oder gar nicht vertreten sind. Eine einzelne Country Observer-Beobachtung:

- zeigt, dass Zugriff verfügbar ist,
- dass ein spezifisches einzelnes Ergebnis eingetreten ist,
- zeigt den Bedarf an neuer Forschung auf

kann anzeigen. Erzielt allein keine Länderbewertung.

21.3. Ausreichend für Länderbewertung

Damit länderspezifische Raten veröffentlicht werden können:

- Mindestprobe
- gültige Beobachtungsrate,
- Unsicherheit,
- Stellvertretung im Panel

Bedingungen müssen erfüllt sein. Mit einem oder mehreren Teilnehmern: 'Die NOMOS-Bewertung des Landes' kann nicht bekannt gegeben werden.

22. Sprachgerechtigkeits-Panel

Das Bevölkerungsgewicht kann dazu führen, dass Sprachen mit geringer Bevölkerung oder geringen Ressourcen in der Hauptbewertung unsichtbar werden. Daher kann ein separates Gerechtigkeitspanel für Sprachen eingerichtet werden. Zweck:

- um unterrepräsentierte Sprachen zu testen,
- um schwerwiegende sprachbasierte Verschlechterungen zu erkennen,
- um Probleme bei der Übersetzung von Eingaben zu identifizieren,
- um das Verhalten von Sprachen mit geringen Ressourcen sichtbar zu machen

muss. Das Gerechtigkeitspanel für Sprachen kann zusätzliche Proben für das Hauptbevölkerungspanel bereitstellen. Übermäßig gesampelte Sprachbeobachtungen sollten jedoch in den globalen Wertungen wieder auf das tatsächliche Bevölkerungsgewicht zurückgeführt werden. Andernfalls würde das Gerechtigkeitspanel die globale Bevölkerungsabschätzung verändern.

23. ZUGÄNGLICHKEITSPANEL

Nutzer mit visuellen, auditiven, motorischen, kognitiven oder anderen Barrierefreiheitsbedürfnissen sind ein tatsächlicher Teil der Zielpopulation. Wenn die NOMOS Capture- oder Forschungstätigkeit für die Teilnahme dieser Personen nicht geeignet ist: Nutzer dürfen nicht aus der Zielpopulation entfernt werden. Korrektes Ergebnis:

> *Der Stichprobenrahmen ist im Hinblick auf Barrierefreiheit unvollständig.*

Ein separates Barrierefreiheitsgremium kann Folgendes bewerten:

- Kompatibilität mit Bildschirmlesern
- Verwendung der Tastatur
- Sprachsteuerung
- Einstellungen für schwaches Sehen
- Kognitive Belastung
- Barrierefreiheit auf Mobilgeräten
- Benutzerfreundlichkeit des Capture-Tools

Das Barrierefreiheitsgremium sollte Nutzer als Teil der Populationsabdeckung betrachten und nicht als "besondere Ausnahme."

24. ZIELPOPULATION UND STICHPROBENRAHMEN

Diese beiden Konzepte sollten nicht miteinander verwechselt werden.

24.1. Zielpopulation

Die gesamte Gruppe von Personen, über die man eine Schlussfolgerung ziehen möchte.

24.2. Stichprobenrahmen

Die Population, aus der Teilnehmer tatsächlich ausgewählt werden können:

- Panel,
- Liste,
- Datenbank,

- Gemeinsame Organisation,
- Nutzerpool

ist der Cluster. Der Stichprobenrahmen deckt möglicherweise nicht die gesamte Zielpopulation ab.

24.3. Kontaktierbare Population

Dies sind die Personen, denen die Forschungseinladung tatsächlich zugestellt werden kann.

24.4. Population, die der Teilnahme zustimmt

Diejenigen, die freiwillig an der Forschung teilnehmen.

24.5. Population, die die Aufgabe abschließt

Diejenigen, die die Prüfaufgabe abschließen.

24.6. Gültige Analysepopulation

Sie sind diejenigen, die Beobachtungen erzeugen, die für das Protokoll geeignet und in der Analyse verwendbar sind. Die Zahlen dieser Schichten sollten gesondert gemeldet werden.

25. ABDECKUNGSFEHLER

Der Stichprobenrahmen kann einige Teile der Zielpopulation ausschließen. Dies wird als Abdeckungsfehler bezeichnet. Beispiele:

- Nur professionelle Online-Panels
- Nur Desktop-Nutzer
- Nur Personen mit Kreditkarten
- Nur diejenigen, die eine englische Benutzeroberfläche verwenden
- Nur Einwohner großer Städte
- Nur Nutzer, die stark an Technologie interessiert sind
- Ausschluss von Menschen mit speziellen Zugänglichkeitsbedürfnissen

Abdeckungsfehler:

- sollte erklärt werden,
- sollte nach Möglichkeit gemessen werden,

sollte durch Gewichtung oder ein zusätzliches Panel reduziert werden. Gewichtung kann nicht magisch eine Gruppe erschaffen, die im Stichprobenrahmen nicht existiert.

26. ÜBERDECKUNG

Personen, die nicht zur Zielpopulation gehören, können im Stichprobenrahmen gefunden werden. Beispiele:

- Teilnehmer in einem Land, in dem das Produkt nicht verfügbar ist
- Person, die die erforderliche Altersbedingung nicht erfüllt
- Falscher Plan oder Nutzeroberfläche
- Teilnehmer, der die Prüfsprache nicht verwenden kann
- Person, die mehrfach registriert wiedergefunden wird
- Nutzer, der im Auftrag eines anderen Landes arbeitet

Überdeckungsdatensätze sollten nicht in die Hauptbewertung einbezogen werden. Der Grund für den Ausschluss und die Rate sollten jedoch gemeldet werden.

27. NICHTREAKTION UND SELBSTAUSWAHL

Menschen, die sich bereit erklären, an der Forschung teilzunehmen, können sich systematisch von der Zielpopulation unterscheiden. Freiwillige:

- Können KI-Produkte häufiger nutzen,
- Können mehr Interesse an Technologie haben,
- Können ein höheres Bildungsniveau haben,
- Können bereits mit der Marke vertraut sein,

Können sich aufgrund von Belohnung oder Vergütung anders verhalten. Daher sollte der Ausdruck "1,000 zufällige Menschen auf der Welt" nur in wirklich probabilistischen und dokumentierten Auswahlverfahren verwendet werden. Wenn ein Freiwilliger oder Online-Panel verwendet wird, lautet der korrekte Ausdruck:

Gestuftes und gewichtetes freiwilliges Nutzerpanel

kann möglich sein. Der Paneltyp kann nicht gespeichert werden.

28. GEEIGNETE POPULATION ÄNDERT SICH IM LAUFE DER ZEIT

Das Nutzeruniversum eines KI-Produkts ist nicht festgelegt. Folgendes kann sich ändern:

- Unterstützte Länder
- Mindestalter
- Freier Zugang
- Kostenpflichtige Planstruktur
- Mobil- oder Webverfügbarkeit
- Sprachunterstützung
- Kontobedingungen
- Enterprise-Produktabdeckung
- Rechtlicher und regulatorischer Status

Aus diesem Grund zeichnet jede Population folgende Werte auf:

- Beobachtungsdatum,
- Gültigkeitsdauer,
- Quellversion,
- Datum der letzten Überprüfung

Muss tragen.

28.1. Veränderung während der Messwelle

Wenn sich der Produktzugang während der Welle wesentlich ändert:

- Die Welle kann gestoppt werden,
- Vor und nach der Änderung können zwei separate Perioden erstellt werden,

Betroffene Länder können separat analysiert werden. Nutzer vor und nach der Änderung können nicht als eine einzige Population kombiniert werden.

29. POPULATIONSQUELLENDATEI

Jede Prüfung sollte die Daten der Zielpopulation, die sie verwendet, in einer versionierten Datei speichern. Gegebenenfalls sollten die folgenden Felder enthalten sein:

- Land
- Referenzpopulation
- Populationshistorie
- Altersberechtigung
- Internet- und Gerätezugang
- Produktzugangsstatus
- Schnittstellenzugang
- Pläne- und Kontoverpflichtung
- Sprachgruppen
- Forschungsberechtigung
- Häufige Vertriebsmethode
- Unsicherheit
- Quellen
- Datum der letzten Überprüfung
- Verantwortliche Person
- Populationsversion

Länderkontingente können ohne diese Datei nach Erhalt der Ergebnisse nicht reproduziert werden.

30. SYNTHETISCHES PRODUKT UND LÄNDERANZEIGE

DEMONSTRATION DER SYNTHETISCHEN METHODOLOGIE / Die folgenden Länder, KI-Produkte, Bevölkerungszahlen und Zugriffsquoten sind vollständig fiktiv. Sie repräsentieren kein reales Land oder Produkt. Betrachten wir ein synthetisches allgemeines Verbraucherprodukt namens Orion KI. Es gibt vier synthetische Länder.

Land	Gesamtbevölkerung	Geschätzte geeignete Bevölkerung nach Alters-, Internet-, Sprach- und Schnittstellenbedingungen	Produktzugang	Population der natürlichen Reichweite
Land A	1,400,000,000	780,000,000	Nicht verfügbar	0
Land B	1,450,000,000	520,000,000	Offizieller Zugang	520,000,000
Land C	340,000,000	250,000,000	Offizieller Zugang	250,000,000
Land D	35,000	25,000	Offizieller Zugang	25,000
Gesamt	3,190,035,000	1,550,025,000		770,025,000

30.1. Wenn die Gesamtbevölkerung verwendet würde

hätte Land A im Hauptsample aufgrund seiner sehr großen Bevölkerung einen großen Anteil. Das Produkt kann jedoch in diesem Land nicht verwendet werden. Teilnehmende könnten keine natürliche Nutzererfahrung erzeugen.

30.2. Wenn Natürliche Reichweite Bevölkerung verwendet wird

Die theoretischen Bevölkerungsanteile des Hauptproduktpanels sind ungefähr:

- Land B: 67.53%
- Land C: 32.47%
- Land D: 0.003%
- Land A: außerhalb von Natürliche Reichweite

Der theoretische Anteil von Land D in einer Hauptstichprobe von 1,000 Personen liegt weit unter einer Person. Daher sollte das bevölkerungsgewichtete Hauptscore Land D kein hohes Gewicht geben. Wenn das Land jedoch nicht vollständig unsichtbar sein soll:

Land Beobachterpanel

Mindestens eine explorative Beobachtung kann erhalten werden. Diese Beobachtung:

- ist nicht das Ergebnis von Land D,
- ändert das Hauptbevölkerungsgewicht nicht,

zeigt, dass das Produktverhalten im Land beobachtet werden kann.

30.3. Wie wird Land A berichtet?

Für Land A:

Der Status OUTSIDE_NATIVE_REACH

wird verwendet. Der Ausschluss von Land A sollte die Zugangsgrenzen des Produkts nicht verschleiern. Die öffentliche Ergebnisübersicht zeigt zudem:

- Produktzugang-Abdeckung
- Anteil, der in der globalen Referenzpopulation nicht zugänglich ist
- Natürliche Reichweite Score
- Gemeinsame Vergleichsbasis Vergleichs-Score

31. SYNTHETISCHE PROMPT-POPULATIONSDARSTELLUNG

SYNTHETISCHES BEISPIEL

Betrachten wir drei Prompts für dasselbe Orion KI-Produkt.

Prompt A

„Was für ein Unternehmen ist Asteron?“ Zielpopulation: könnte die breite, geeignete Nutzerpopulation von Orion AI sein.

Prompt B

„Bietet Asteron Firmendienstleistungen in der Türkei an?“ Zielpopulation:

- Berechtigte Nutzer in der Türkei,

- Identifizierte externe Nutzer, die den Türkei-Service recherchieren

kann sein.

Prompt C

„Ist Asteron für legale Migrationsberatung für mich geeignet?“ Wenn Asteron einen solchen Service nicht anbietet: Es könnte keine beratungsberechtigte Bevölkerung geben, die korrekte Antwort sollte alle geeigneten Nutzer ausschließen. Die Kombination dieses Prompts in den gesamten Identitätsscore könnte falsch sein. Dieses Beispiel zeigt:

> *Selbst für dieselbe Entität und dasselbe KI-Produkt kann die Zielpopulation je nach gemessener Nutzerfrage variieren.*

32. VIER-PANEL-ARCHITEKTUR

KANDIDATEN-ARCHITEKTUR — KANDIDATEN-ARCHITEKTUR

Vier ergänzende Panels werden für die globale und faire Implementierung von GEO-1000 empfohlen.

32.1. Bevölkerungspanel

Das Hauptpanel wird gemäß der produktberechtigten Nutzerpopulation gewichtet. Produziert die Hauptpopulation-Schätzung.

32.2. Länderbeobachterpanel

Verhindert, dass kleine Länder, die im Bevölkerungspanel nicht genügend Stichproben bereitstellen können, vollständig unsichtbar werden. Es ist allein nicht ausreichend, um einen Länderwert zu erzeugen.

32.3. Sprachgerechtigkeits-Panel

Misst auch Sprachen, die in den Bevölkerungsdurchschnittswerten wenig Ressourcen haben oder unterdrückt sind. Es verbindet sich wieder mit dem globalen Hauptwert unter Verwendung realer Bevölkerungsgewichte.

32.4. Zugänglichkeits-Panel

Misst die Bedürfnisse in Bezug auf Barrierefreiheit und Nutzer von unterstützender Technologie, die im Standard-Digitalpanel möglicherweise ausgeschlossen werden. Beobachtungen aus diesen vier Panels:

- Ziel,
- Gewichtung,
- Inference-Universum

Es sollte in Bezug auf Wartung separat gekennzeichnet werden. Dieselbe Beobachtung darf nicht ohne Erklärung mit voller Gewichtung in mehr als einem Panel verwendet werden.

33. ERSTES PANEL FÜR ERWACHSENE

GRUNDREGEL FÜR KANDIDATEN — GRUNDLEGENDES PRINZIP FÜR KANDIDATEN

In der ersten allgemeinen Veröffentlichung von GEO-1000, die echten Nutzer-Dashboards:

von erwachsenen Teilnehmern, die die Produkt- und Länderbedingungen erfüllen

Es wird empfohlen, erstellt zu werden. Dieser Ansatz:

- Zustimmung,
- Datenschutz,
- Sammeln von Screenshots,
- forschung mit hohem Risiko,
- Elterliche Zustimmung

Es kann deren Komplexität reduzieren. Es kann jedoch nicht der folgende Schluss gezogen werden: 'Die Erfahrungen von Kindern und jungen Nutzern wurden mit dem erwachsenen Panel gemessen.' Bei Produkten, die Kinder oder junge Menschen betreffen, separat:

- ethische Genehmigung,
- altersgerechtes Protokoll,
- Eltern- oder Vormundprozesse,
- Schadensschutz

sind erforderlich.

34. KEIN POPULATIONSRAHMEN KANN ALS „GESAMTE MENSCHHEIT“ BEZEICHNET WERDEN

Ein Produkt:

- wenn es in einigen Ländern nicht existiert,
- wenn es einige Sprachen nicht unterstützt,
- wenn es nur auf bestimmten Geräten funktioniert,
- wenn es bestimmte Alters- oder Tarifbedingungen hat,

Ergebnis: Es kann nicht als „die GEO-Erfahrung aller Menschen auf der Welt“ dargestellt werden. Die korrekte Ausdrucksweise:

Geschätzte Repräsentation in der Zielpopulation, die für das angegebene Produkt geeignet ist

sollte verwendet werden. Das Wort global:

- mehrere Länder,
- mehrsprachig,
- große Bevölkerung

kann eine Messung definieren. Es bedeutet nicht unbegrenzte und universelle Abdeckung.

35. GRUNDLEGENDES ETHISCHES LIMIT DER BEVÖLKERUNG

Die Definition einer Bevölkerung für eine Messung bedeutet nicht, einige Menschen als wertlos zu betrachten. Eine Gruppe kann im Hauptscore ein geringes Gewicht erhalten. Allerdings:

- hohes Risiko,
- Sprachgerechtigkeit,
- Barrierefreiheit,
- Ländersichtbarkeit

kann auch separat untersucht werden. Eine Gruppe kann vom Natürliche Reichweite-Score ausgeschlossen werden, weil sie keinen Zugriff auf ein Produkt hat. Dieser fehlende Zugriff kann auch als globale Abdeckungslücke gemeldet werden. Der Zweck von NOMOS:

- alle in einen einzigen künstlichen Durchschnitt zu komprimieren,
- nur die größten Bevölkerungen zu sehen,
- kleine Gemeinschaften symbolisch durch eine einzelne Person darzustellen

ist nicht das Ziel. Das Ziel ist:

> *jede Person im richtigen Nenner der richtigen Frage darzustellen, ohne ihr Gewicht und ihre Sichtbarkeit zu vermischen.*

36. VERPFLICHTENDE NORMATIVE BESTIMMUNGEN

CH05-N01

Die Zielnutzerpopulation muss für jedes KI-Produkt separat definiert werden.

CH05-N02

Die Gesamtbevölkerung eines Landes kann nicht direkt als KI-geeignete Nutzerpopulation verwendet werden.

CH05-N03

Die Zielpopulation muss mit Produkt, Land, Alter, Zugriff, Schnittstelle, Plan, Sprache, Prompt und zeitlichen Bedingungen verknüpft werden.

CH05-N04

Die Zielpopulation und die Ausschlussregeln müssen versioniert werden, bevor die Ergebnisse gesehen werden.

CH05-N05

Die Zielgruppe eines KI-Produkts kann nicht ohne Erklärung auf ein anderes KI-Produkt übertragen werden.

CH05-N06

Fehlender Produktzugang kann nicht als Ergebnis einer Falschdarstellung bewertet werden; er muss getrennt als Zugangsabdeckung gemeldet werden.

CH05-N07

Teilnehmer dürfen nicht gebeten werden, Produkt- oder Länderzugangsbeschränkungen zu umgehen.

CH05-N08

Zugriff, der nicht offiziell unterstützt wird, kann nicht automatisch zur Hauptpopulation Natürliche Reichweite hinzugefügt werden.

CH05-N09

Individuelle, Konto- und Sitzungseinheiten müssen voneinander getrennt sein.

CH05-N10

Mehrere Konten oder Sprachen einer Person können das Individuum im Bevölkerungsgewicht nicht vervielfachen.

CH05-N11

Die Sprachberechtigung kann nicht allein auf der Amtssprache des Landes basieren.

CH05-N12

Die angeforderte Sprache, die Nutzerschnittstellensprache, die Antwortsprache und die Sprache der Nutzerkompetenz müssen separat aufgezeichnet werden.

CH05-N13

Das gesamte Populationsgewicht einer Person, die mehrere Sprachen verwendet, darf ohne Erklärung nicht mehr als eine Person betragen.

CH05-N14

Das populationsgewichtete Panel und das Country-Observer-Panel sollten getrennt gehalten werden.

CH05-N15

Eine oder mehrere Beobachtungen des Country Observer dürfen nicht als Länderpunktzahl veröffentlicht werden.

CH05-N16

Im Language-Fairness-Panel sollten überrepräsentierte Beobachtungen im globalen Score auf ihr tatsächliches Populationsgewicht zurückgeführt werden.

CH05-N17

Nutzer mit besonderen Zugänglichkeitsanforderungen dürfen nicht von der Zielpopulation ausgeschlossen werden, weil das Datenerfassungstool nicht geeignet ist.

CH05-N18

Die Zielpopulation, die Stichprobenbasis, die kontaktierte Population, die Teilnehmer und die gültigen analytischen Beobachtungen sollten getrennt berichtet werden.

CH05-N19

Ein Panel ohne Zufallsstatistik oder ein freiwilliges Panel darf nicht als zufällige Stichprobe der Weltbevölkerung repräsentiert werden.

CH05-N20

Durch Gewichtung kann eine Gruppe, die in der Stichprobenbasis vollständig fehlt, nicht sichtbar gemacht werden.

CH05-N21

Potenziell geeignete Bevölkerung und aktive Nutzerexposition können nicht als dieselbe Kennzahl verwendet werden.

CH05-N22

Die Ergebnisse Natürliche Reichweite und Gemeinsame Vergleichsbasis müssen voneinander getrennt werden.

CH05-N23

Die gleiche Zielpopulation kann nicht automatisch für Identitäts-, Service-, Preis-, Rechts- und Empfehlungsprompts verwendet werden.

CH05-N24

Bevölkerungsdaten, Produktzugangsstatus und Sprachverteilung müssen einen Zeitstempel und eine Quellenversion tragen.

CH05-N25

Grenznutzenverhältnisse dürfen nicht blind multipliziert werden, ohne die Unabhängigkeitsannahme zu erklären.

CH05-N26

Wenn der Produkttyp für das allgemeine Verbraucherpaneel nicht geeignet ist, darf die GEO-1000-Methode für menschliche Populationen ohne Erklärung nicht angewendet werden.

CH05-N27

Kinder oder minderjährige Nutzer dürfen nicht in das Hauptpanel aufgenommen werden, ohne ein separates ethisches und rechtliches Protokoll.

CH05-N28

Produktzugangsabdeckung und Repräsentativitätsgenauigkeit müssen als separate Ergebnisse veröffentlicht werden.

CH05-N29

Der Begriff global kann Länder, Sprachen, Altersgruppen, Schnittstellen oder Pläne, die außerhalb des Geltungsbereichs liegen, nicht unsichtbar machen.

CH05-N30

Der Zielpopulationsdatensatz muss einen verantwortlichen menschlichen oder institutionellen Eigentümer haben.

37. FORMEN DES VERSAGENS

CH05-F01— ZÄHLEN SIE DIE GESAMTBEVÖLKERUNG ALS ELIGIBLE POPULATION

Die Bevölkerung des Landes wird direkt in den Nenner aufgenommen, ohne Produkt-, Alters-, Internet-, Sprach- und Schnittstellenbedingungen zu bewerten.

CH05-F02— WENDEN SIE DIE GLEICHE WELTKARTE AUF ALLE PRODUKTE AN

KI-Produkte mit unterschiedlichen Zugriffsbereichen werden gegen dieselbe Zielpopulation gemessen.

CH05-F03— ZÄHLEN SIE EINEN NUTZER, DER DAS LIMIT ÜBERSCHREITET, ALS NATÜRLICHEN NUTZER

Beobachtungen, die die Bedingungen des offiziellen Zugangs überschreiten, werden als Erfahrung der Landesbevölkerung dargestellt.

CH05-F04— ZÄHLEN DAS FEHLEN DES PRODUKTS ALS FALSCHE ANTWORT

Länder, die unzugänglich sind, werden in der Repräsentation der Genauigkeit als null betrachtet.

CH05-F05— VÖLLIGE VERBERGUNG DES PRODUKTFEHLENS

Große Populationen, die unzugänglich sind, werden aus dem Natürliche Reichweite Nenner entfernt; die Zugriffabdeckung wird ebenfalls nicht angezeigt.

CH05-F06— ZÄHLUNG EINER PERSON ALS MEHRERE

Das Konto derselben Person, die Sprache oder Produktempfindungen multiplizieren das Bevölkerungsgewicht.

CH05-F07— ZÄHLUNG DES KONTS ALS PERSON

Mehrfachkontenbesitz oder geteilte Kontonutzung wird nicht berücksichtigt.

CH05-F08— AMTSSPRACHE ALS NUTZERSPRACHE ZÄHLEN

Die Amtssprache des Landes wird für alle Teilnehmer als Prüfungs- bzw. Audit-Sprache betrachtet.

CH05-F09— ZWEIFACHE ZÄHLUNG VON MEHRSPRACHIGEN EINZELPERSONEN

Die gleiche Person erhält in jeder Sprache das volle Bevölkerungsgewicht.

CH05-F10— ZUWEISUNG EINER PERSON ZU EINEM KLEINEN LAND UND ERKLÄRUNG DES LANDESERGEBNISSES

Beobachtungen des Länderbeobachters werden in eine statistische Länderschätzung umgewandelt.

CH05-F11— VOLLSTÄNDIGES LÖSCHEN KLEINER LÄNDER

Da der Bevölkerungsanteil gering ist, wird das Land oder die Sprache in keinem zusätzlichen Abdeckungs-Panel berücksichtigt.

CH05-F12— ÜBERGEWICHTUNG VON BEISPIELEN FÜR SPRACHGERECHTIGKEIT IM HAUPTERGEBNIS

Überstichproben von Sprachen werden nicht auf ihr tatsächliches Bevölkerungsgewicht zurückgeführt.

CH05-F13— ZÄHLEN DES AKTIVEN NUTZERS ALS POTENZIELLE BEVÖLKERUNG

Tatsächliche Nutzungsdaten werden so dargestellt, als ob sie alle Personen einschließen, die auf das Produkt zugreifen könnten.

CH05-F14— ZÄHLEN DER POTENZIELLEN BEVÖLKERUNG ALS TATSÄCHLICHE EXPOSITION

Es wird angenommen, dass die Antwort jeder Person, die das Produkt nutzen könnte, tatsächlich gesehen wird.

CH05-F15— ZÄHLUNG DES FREIWILLIGEN PANELS ALS ZUFÄLLIGE WELTPROBE

Selbstselektion und Panel-Bias werden verschleiert.

CH05-F16— IGNORIEREN DES ABDECKUNGSFEHLERS DURCH GEWICHTUNG

Es wird behauptet, dass Gruppen, die im Rahmen nicht vorhanden sind, ausschließlich durch Gewichtung vertreten werden.

CH05-F17— ENTFERNUNG VON NUTZERN, DIE DURCH DAS DATENERFASSUNGSTOOL VON DER POPULATION AUSGESCHLOSSEN WURDEN

Zugänglichkeits- oder Geräteprobleme werden so interpretiert, als ob der Nutzer nicht geeignet ist.

CH05-F18— ANWENDUNG DER GLEICHEN POPULATION AUF ALLE PROMPTS

Globale Identität, lokale Preisgestaltung und rechtliche Beratungsfragen werden auf denselben Nenner gebracht.

CH05-F19— ANWENDUNG DER ALLGEMEINEN VERBRAUCHERPOPULATION AUF DAS API-PRODUKT

Das menschliche Nutzerpanel wird direkt an den Entwickler oder das Anwendungsprodukt übertragen.

CH05-F20— GEHEIMES KOMBINIERTES ANBIETEN VON KOSTENLOSEN UND BEZAHLTEN PLÄNEN

Eine einzelne Produktverteilung wird veröffentlicht, ohne den Planunterschied zu erfassen.

CH05-F21— VEREINIGUNG VON WEB- UND MOBILBEVÖLKERUNG

Unterschiede beim Schnittstellenzugang und Verhalten bleiben unsichtbar.

CH05-F22— IGNORIEREN DER ALTERSBEGRENZUNG

Unangemessene Altersgruppen werden ohne Erklärung zur Zielpopulation oder zum Panel hinzugefügt.

CH05-F23— MULTIPLIZIEREN VON MARGINALRATEN, ALS WÄREN SIE UNABHÄNGIG

Die Beziehung zwischen Alter, Internet, Sprache und Zugang wird ignoriert.

CH05-F24— ÄNDERN DER BEVÖLKERUNG NACH DEM ERGEBNIS

Ein Land oder eine Nutzergruppe mit niedriger Punktzahl wird aus der Zielpopulation entfernt.

CH05-F25— VERWENDUNG DER ALTEN PRODUKT-REICHWEITE-KARTE

Die tatsächliche Reichweite des Produkts am Messdatum wird nicht überprüft.

CH05-F26— VERWECHSLUNG DER ERGEBNISSE VON Gemeinsame Vergleichsbasis UND Natürliche Reichweite

Vergleiche, die auf der gemeinsamen Population basieren, werden so dargestellt, als würden sie die tatsächliche globale Reichweite des Produkts widerspiegeln.

CH05-F27— EINENGUNG DES ZIELMARKTES NACH KOMMERZIELLEM INTERESSE

Die Organisation gibt nur die Länder an, in denen sie erfolgreich ist, als Zielmärkte an.

CH05-F28— ZÄHLEN DES ZUGANGSUMFANGS ALS GEOGRAFISCHE GENAUIGKEIT

Die Präsenz in vielen Ländern wird als Beweis für eine genaue Repräsentation präsentiert.

CH05-F29— BEGRENZTEN ZUGANG ALS HOHEN GLOBALE ERFOLG VERKAUFEN

Hohe Genauigkeit in einer kleinen Anzahl von Ländern wird als globale Überlegenheit dargestellt.

CH05-F30— VERWENDUNG DES BEGRIFFS „DIE GANZE WELT“ OHNE ABDECKUNGSREKORD

Ausschlüsse von Produkt, Sprache, Land und Plan sind verborgen.

38. PRÜFUNGSVERFAHREN

Schritt 1— Festschreiben des KI-Produkts und der Nutzeroberfläche

Produktoberflächenplan Kontotyp Angezeigtes Modell Messdatum wird aufgezeichnet.

Schritt 2— Klassifizieren der Messfrage

Prompt:

- globale Identität,
- lokaler Service,
- Preis,
- Gesetz,
- Empfehlung,
- hohes Risiko

zu welcher ihrer Familien gehört es?

Schritt 3— Bestimmen der globalen Referenzpopulation

Das ursprüngliche menschliche Universum, mit dem das Ergebnis verknüpft wird, wird definiert.

Schritt 4— Bestimmen von Alter und Forschungseignung

Produktaltersbedingung Forschung Altersbedingung Zustimmung und Bedingungen zur Datenverarbeitung werden aufgezeichnet.

Schritt 5— Länderspezifischen Produktzugang überprüfen

Ein Zugangsstatus zwischen PA-0 und PA-5 wird für jedes Land zugewiesen.

Schritt 6— Internet-, Geräte- und Schnittstellenkompatibilität bestimmen

Die Bevölkerung, die die gemessene Oberfläche nutzen kann, wird geschätzt.

Schritt 7— Sprach- und Gebietsschema-Universum festlegen

Eingabesprache, Schnittstellensprache, Antwortsprache und Nutzer-Sprachkenntnisse werden getrennt.

Schritt 8— Relevanz der Prompt anwenden

Es wird bewertet, ob alle Nutzer zur Zielpopulation der relevanten Eingabeaufforderungsfamilie gehören.

Schritt 9— Passendes Bevölkerungsmodell für den Produkttyp auswählen

Verbraucher-, Unternehmens-, Bildungs-, API- oder eingebettetes Produkt wird getrennt.

Schritt 10— Vergleiche die Zielpopulation mit dem Stichprobenrahmen

Es wird als Unterabdeckung und Überabdeckung erfasst.

Schritt 11— Identifiziere die Haupt- und Ergänzungspanels

Die Bereiche des Bevölkerungs-Panels, des Länderbeobachter-Panels, des Sprachgerechtigkeits-Panels und des Barrierefreiheits-Panels sind versioniert.

Schritt 12— Trenne die Natürliche Reichweite- und Gemeinsame Vergleichsbasis-Universen

Intra-Produkt-Ergebnisse und interproduktale Vergleichspopulationen werden separat erfasst.

Schritt 13— Trenne Zugriffsbereich und Genauigkeitsergebnisse

Die Genauigkeit zwischen der Population, die ein Produkt erreichen kann, und den erreichten Nutzern wird separat berichtet.

Schritt 14— Friere die Bevölkerungsdatei ein

Datenquellen, Annahmen, Datum und Bevölkerungsversion werden festgeschrieben.

Schritt 15— Verfolge wesentliche Änderungen

Wenn sich der Produktzugang oder die Bedingungen während der Messung ändern, wird eine neue Version erstellt.

39. ERFORDERLICHE BELEGE

KI-Produktidentität; Nutzeroberfläche; Plan- und Kontobedingungen; Messdatum; Länder, in denen das Produkt zugänglich ist; Länderebene-Zugangsstatus; Altersbedingungen; rechtliche und ethische Teilnahmebedingungen; Ländervölkerungsaufzeichnungen; Altersverteilungen; Internet- und Gerätezugangsaufzeichnungen; Sprach- und Gebietsschema-Verteilungen; Produktsprachunterstützung; Schnittstellenzugänglichkeit; Ziel-Prompt-Familie; Definition der Zielpopulation; Stichprobenrahmen; Unterabdeckung; Überabdeckung; Paneltyp; Teilnehmerauswahlmethode; Wahrscheinlichkeitspanel oder Opt-in-Panelstatus; Gewichte für mehrsprachige Nutzer; Abdeckung Country Observer; Abdeckung Language Fairness; Abdeckung Accessibility Panel; Natürliche Reichweite Population; Gemeinsame Vergleichsbasis Population; aktive Nutzerdaten, wo verwendet; Berechnung der Zugangsabdeckung; Annahmen und Sensitivitätsanalysen; Populationsversion; und letztes Überprüfungsdatum.

Verantwortliche Person oder Institution

40. PRÜFLISTE FÜR AUDITS

Welches KI-Produkt und welche Nutzeroberfläche wird gemessen? Für welches Datum wurde die Zielpopulation definiert? Wurden die Gesamtbevölkerung des Landes und die berechtigte Nutzerpopulation getrennt? Wurde ein Altersanforderung angewendet? Wurden Internet- und Gerätezugang bewertet? Ist das Produkt offiziell nach Ländern zugänglich? Wurden die Teilnehmer gebeten, Zugangsbeschränkungen zu überwinden? Wurden kostenlose und kostenpflichtige Pläne getrennt? Wurden Web- und mobile Schnittstellen getrennt? Hat die Aufforderungsfamilie die Zielpopulation verändert? Wie wurde die Sprach- und Ländereignung bestimmt? Wurden mehrsprachige Nutzer doppelt gezählt? Ist der Unterschied zwischen der Zielpopulation und dem Stichprobenrahmen klar? Ist das Panel probabilistisch oder freiwillig? Wurde der Abdeckungsfehler gemessen? Wurden Nutzer mit besonderen Zugangsbedürfnissen stillschweigend ausgeschlossen?

Wurden Beobachtungen des Länderbeobachters in eine Länderbewertung umgewandelt? Hatten Beispiele zur Sprachfairness ein übermäßiges Gewicht in der Hauptbewertung? Wurden die Ergebnisse von Natürliche Reichweite und Gemeinsame Vergleichsbasis getrennt? Sind Abdeckung und Repräsentationsgenauigkeit getrennt? Wurden Ergebnisse aktiver Nutzer und potenzieller Nutzer getrennt? Ist der Produkttyp für das Panel der menschlichen Bevölkerung geeignet? Wurden die Randraten in den Bevölkerungsgewichten blind multipliziert? Wurde die Bevölkerungsdefinition nach Bekanntwerden der Ergebnisse geändert? Sind die verwendeten Bevölkerungs- und Zugriffsdatensätze versioniert? Sind ungetestete Gruppen sichtbar? Ist der Begriff „global“ mit der tatsächlichen Abdeckung konsistent? Gibt es einen klar verantwortlichen Eigentümer des Bevölkerungsdatensatzes?

41. EINWÄNDE UND ANTWORTEN

Einspruch 1— „Ist es nicht gerechter, direkt nach der Weltbevölkerung zu verteilen?“

Die Weltbevölkerung ist eine starke Ausgangsbasis. Allerdings erzeugt die Einbeziehung von Personen, die das Produkt nicht nutzen können, in der Hauptproduktprobe keine echte Nutzererfahrung. Die richtige Methode ist:

- Bevölkerungsgewicht innerhalb der produktberechtigten Bevölkerung,
- globale Zugriffsabdeckung.

Beide Ergebnisse müssen gemeldet werden.

Einwand 2— „Sollten nicht die meisten Nutzer aus großen Ländern kommen?“

Länder, die die Produkt-, Sprach- und Zugriffsvoraussetzungen erfüllen, sollten natürlich einen hohen Anteil haben. Nur weil die Gesamtbevölkerung groß ist, kann jedoch keine gefälschte Nutzererfahrung in einem Land erzeugt werden, in dem das Produkt nicht zugänglich ist.

Einwand 3— „Warum wäre es falsch, einer kleinen Nation wie San Marino eine Person zu geben?“

Eine Person ist wertvoll für die Erkundung und Beobachtung durch den Länderbeobachter. Falsch ist es, das Ergebnis dieser Person als 'Länderpunktzahl' darzustellen. Die Sichtbarkeit des kleinen Landes kann mit dem Länderbeobachter-Panel geschützt werden. Das Hauptgewicht der Bevölkerung bleibt gemäß dem tatsächlichen Bevölkerungsanteil.

Einwand 4— „Wenn wir jedem Land mindestens eine Person geben, würden wir dann nicht globaler erscheinen?“

Eine größere geografische Abdeckung wird bereitgestellt. Allerdings kann die Bevölkerungsabschätzung verzerrt sein. Daher:

- geografischer Umfang,
- Bevölkerungsdichte

Sie sind separate Ergebnisse. Es kann ein Umfangs-Panel sein, um einer Person pro Land zu geben. Es ist kein Bevölkerungs-Panel.

Einwand 5— „Wenn wir einem Produkt in einem unzugänglichen Land null geben, würde das Produkt nicht unfair bestraft?“

Ein unzugängliches Land wird bei der Repräsentationsgenauigkeit nicht als null gezählt. Es wird im Abdeckungsmetrik-Zugang sichtbar. Diese Unterscheidung schützt das Produkt sowohl vor unfairen Strafen als auch vor falscher Universalität.

Einwand 6— „Wenn es echte Nutzungsdaten gibt, warum brauchen wir Bevölkerungsdaten?“

Echte Nutzungsdaten zeigen den operativen Einfluss. Allerdings zeigt das Produkt:

- welche Bevölkerungen es nicht erreicht hat,
- die geografische Ungleichheit in der Nutzerbasis,
- potenzielle Nutzerfairness

Es zeigt sich nicht allein. Die beiden Datentypen beantworten unterschiedliche Fragen.

Einwand 7— ‚Da alle Nutzer denselben Prompt stellen werden, warum ist die Relevanz des Prompts wichtig?‘

Identitätsfragen können in einer breiten Bevölkerung sinnvoll sein. Preis-, Rechts- oder Beratungsfragen sind nur für bestimmte Nutzer und Märkte sinnvoll. Eine irrelevante Bevölkerung kann die Antwortverteilung von der Nutzerrealität verzerren.

Einwand 8— „Wenn eine Person mehrere Sprachen spricht, warum sollte sie nicht in allen Sprachen vollständig gezählt werden?“

Sie können alle Sprachen im sprachübergreifenden Vergleich testen. Die gleiche Person in der Bevölkerungsgewichtung zu multiplizieren, erhöht jedoch künstlich die Weltbevölkerung. Das Gewicht der Person sollte geteilt oder ein separates, passendes Panel verwendet werden.

Einwand 9— „Ergibt die Prüfung nur von Erwachsenen nicht ein unvollständiges globales Ergebnis?“

Das stimmt. Diese Einschränkung sollte eindeutig angegeben werden. Mit einem Erwachsenenpanel zu beginnen, kann eine ethische und operative Entscheidung sein. Es stellt kein Urteil über Kinderanwender dar.

Einwand 10— ‚Ist das Online-Forschungs-panel nicht ausreichend?‘

Es kann sehr wertvoll für Pilot- und gewichtete Forschung sein. Aber das Panel:

- wie viel von der Zielpopulation es abdeckt,
- wer ausgeschlossen wird,
- wie freiwillige Teilnahme Verzerrungen erzeugt

sollte erklärt werden. Die Nutzung eines Panels ist kein Fehler. Das Panel als zufällig wie die Weltbevölkerung darzustellen, ist ein Fehler.

Einwand 11— „Warum hängt das Accessibility-Panel mit dem Haupt-GEO-Score zusammen?“

Weil ein Teil der echten KI-Nutzer Assistenztechnologien oder unterschiedliche Interaktionsformen verwendet. Wenn das Datenerhebungswerkzeug diese Personen ausschließt, kann der Hauptscore nicht die gesamte Zielpopulation repräsentieren.

Einwand 12— „Macht es nicht die vielen verschiedenen Bevölkerungsdefinitionen unmöglich, einen einzigen NOMOS-Wert zu erstellen?“

Nein. Ein einzelner Wert kann erstellt werden. Es sollte jedoch klar sein, auf welche Bevölkerungsfrage er antwortet. Zum Beispiel:

- Natürliche Reichweite Bevölkerungswert
- Gemeinsame Vergleichsbasis Wert
- Marktrelevanter Wert
- Ergebnis zur Sprachgerechtigkeit

kann separate Datensätze enthalten. Die Bedeutung einer Zahl ergibt sich aus der Klarheit ihres Nenners.

GEMEINSAMES URTEIL VON KAPITEL 44

GEO-1000 enthält ‚1,000' in seinem Namen, aber der Wert dieser 1,000 Personen ergibt sich nicht allein aus der Zahl. Er ergibt sich aus den Antworten auf diese Fragen: Wer sind sie? Aus welcher Bevölkerung wurden sie ausgewählt? Können sie das Produkt tatsächlich nutzen? Wo leben sie? Welche Sprachen verwenden sie? Wie oft wurde dieselbe Person gezählt? Sind kleine Länder sichtbar? Behalten große Länder ihr tatsächliches Bevölkerungsgewicht bei? Wo berücksichtigt die Bewertung Menschen, die das Produkt nicht erreichen kann? Wen hat die Panel-Infrastruktur ausgeschlossen? Welches Ergebnis misst den potenziellen Zugang und welches misst die tatsächliche Nutzung? Ein großes Land kann eine große produktberechtigte Bevölkerung haben und verdient daher ein erhebliches Gewicht im Hauptpanel. Aber wenn das Produkt dort nicht verfügbar ist, darf die Prüfung keine Nutzer erfinden. Die Zugangslücke muss separat offengelegt werden.

Wenn ein Land sehr klein ist, sollte das Hauptgewicht der Bevölkerung niedrig bleiben. Es muss jedoch nicht völlig unsichtbar sein. Es kann mit einem Überwachungs-Panel beobachtet werden. Wenn eine Sprache eine kleine Bevölkerung hat, muss sie den globalen Durchschnitt nicht bestimmen. Wenn es jedoch eine starke Verzerrung in der Repräsentation dieser Sprache gibt, sollte sie ebenfalls sichtbar sein. Wenn eine Person drei Sprachen kennt, kann sie drei verschiedene Erfahrungen produzieren. Sie sind jedoch nicht drei verschiedene Personen. Wenn ein KI-Produkt bestimmtes Verhalten nur auf einem kostenpflichtigen Plan bietet, repräsentiert die Punktzahl der kostenlosen Nutzer diesen Plan nicht. Ein menschliches Bevölkerungs-Panel ist für ein Konsumenten-Chat-Produkt geeignet. Dasselbe Bevölkerungsmodell ist möglicherweise nicht für eine API oder ein Unternehmensprodukt geeignet. Daher lautet das fünfte Messgesetz von NOMOS wie folgt:

> *Selbst die empfindlichste Punktzahl liefert die richtige Antwort auf die falsche Frage, wenn der Nenner falsch ist.*

Sein sechstes Gesetz lautet:

> *Die Bevölkerungszahl ersetzt nicht die Sichtbarkeit von Ländern und Sprachen.*

Sein siebtes Gesetz lautet wie folgt:

> *Eine Person, die keinen Zugriff auf [etwas] hat, ist nicht falsch dargestellt worden; sie ist jedoch Teil der globalen Zugangsgrenze.*

Sein achtes Gesetz lautet wie folgt:

> *Eine Person mit mehreren Konten, Sprachen oder Sitzungen erhöht nicht die menschliche Bevölkerung.*

Das neunte Gesetz lautet wie folgt:

> *Globale Messung ist kein Anspruch darauf, dass jeder gemessen wird; es ist eine ehrliche Aufzeichnung darüber, wer gemessen wird und wer nicht.*

Bestellung von Abschnitt 5 von NOMOS

> *Gib mir nicht nur Bevölkerungszahlen der Länder. / Zeige, welche Menschen das Produkt tatsächlich nutzen können.*
>
> *Fülle kein großes Land mit gefälschten Nutzern, wenn das Produkt dort nicht vorhanden ist.*
>
> *Gib einer kleinen Nation nicht eine Person und erkläre die Antwort dieser Person als die Stimme des Landes.*
>
> *Du darfst eine Person nicht so oft duplizieren wie ihre Konten, Sprachen und Sitzungen.*

Verwechseln Sie nicht die Amtssprache mit der tatsächlichen Sprachverwendung der Menschen.

Mache keinen kostenlosen Nutzer zum Vertreter eines kostenpflichtigen Plans und keinen mobilen Nutzer zum Vertreter der Weboberfläche.

Bezeichne das Freiwilligenpanel nicht als zufällige Weltbevölkerung.

Lösche keinen Nutzer mit speziellen Zugänglichkeitsbedürfnissen aus der Population, weil das Tool nicht funktioniert.

Entferne die Population, die das Produkt nicht erreicht, nicht aus der Genauigkeitsbewertung und verschleiere die globale Zugangslücke.

Wende denselben Anteil nicht auf eine Preisfrage wie auf eine Identitätsfrage an, oder auf eine globale Marke wie auf eine lokale Lizenz.

Definiere zuerst das Produkt. / Dann definiere die Frage. / Dann definiere das menschliche Universum. / Dann begrenze Zugriff, Sprache, Alter und Benutzeroberfläche. / Dann erkläre, wen du nicht auswählen kannst. / Und erst danach verteile 1,000 Personen.

Der abschließende Satz des Kapitels

In GEO-1000 bedeutet Gerechtigkeit nicht, die gleiche Anzahl von Personen auf jedes Land zu verteilen; sondern jede Person in der richtigen Zielpopulation mit dem richtigen Gewicht zu repräsentieren, ohne die Orte zu verbergen, an denen sie nicht gemessen werden.

Normativer Kern

Jede GEO-Populationsprüfung MUSS vor der Ergebnisüberprüfung eine zeitlich begrenzte und produktspezifische Zielpopulation von Nutzern für jede der folgenden Kombinationen definieren: - KI-Produkt und Nutzeroberfläche, - Land oder Rechtsgebiet, - Altersgrenze und Voraussetzung für einen rechtmäßigen Zugang, - Konto- oder Tarifbedingung, - Sprache und Gebietsschema, - Schnittstelle, - Prompt-Familie, - und Messzweck. Die gesamte ansässige Bevölkerung DARF NICHT als produktberechtigte Bevölkerung behandelt werden, ohne gerechtfertigte Berechtigungsanpassungen. Die Nichtverfügbarkeit des Produkts MUSS als Zugangsbeschränkung gemeldet werden, nicht stillschweigend als falsche Repräsentation bewertet oder stillschweigend aus dem globalen Umfang entfernt werden. Potenzieller Zugang, aktive Nutzung, tatsächliche Exposition, Natürliche Reichweite, Gemeinsame Vergleichsbasis, Marktrelevanz und Forschungsbeobachtbarkeit MÜSSEN als unterschiedliche Populationskonzepte bestehen bleiben. Personen, Konten, Sitzungen, Sprachen und Produktbeobachtungen DÜRFEN NICHT als austauschbare Populations-Einheiten behandelt werden. Überrepräsentierte Länder, Sprachen und Zugänglichkeitsgruppen MÜSSEN für Populationsschätzungen neu gewichtet werden, während sie separat sichtbar für Fairness- und Abdeckungsanalysen bleiben. Beobachtungen von Country Observer KÖNNEN die geografische Beobachtungsabdeckung bestimmen, DÜRFEN jedoch nicht als statistisch verlässliche Länderscores repräsentiert werden. Teilnehmer DARF NICHT dazu verpflichtet werden, Produkt-, geografische-, Alters-, Identitäts- oder Kontoeinschränkungen zu umgehen, um in das offizielle Bevölkerungs-Panel aufgenommen zu werden. Jeder Bevölkerungsrahmen MUSS versioniert, quellenbasiert, unsicherheitsbewusst und einer verantwortlichen Person oder Organisation zuordenbar sein.

KAPITEL 06

Die bevölkerungsproportionale GEO-1000-Stichprobe

METHODISCHE GRUNDLAGEN / K04; K05; K07; K08

Kapitelgrenze

Abschnitt 5 hat die folgende Bestimmung eingeführt:

> *Die Zielnutzerpopulation jedes KI-Produkts muss separat definiert werden. Die Gesamtbevölkerung eines Landes ist nicht die produktberechtigte Nutzerpopulation in diesem Land.*

Kommen wir nun zur nächsten Frage:

> *Aus der definierten berechtigten Nutzerpopulation für GEO-1000, wie viele Personen werden ausgewählt, aus welchem Land, welche Sprache und nach welcher Methode?*

Die berechtigte Nutzerpopulation für ein KI-Produkt ist:

- 40% in Land A,
- 25% in Land B,
- 15% in Land C,
- 0.04% in einem kleinen Land

Falls verfügbar, wie sollten 1,000 gültige Beobachtungen verteilt werden? Wie werden Bruchquoten in ganze Zahlen umgewandelt? Werden Länder mit einem Bevölkerungsanteil von weniger als einer Person vollständig ausgeschlossen? Wird die Bevölkerungsgewichtung verzerrt, wenn jedem Land mindestens eine Person zugewiesen wird? Können kleine Länder in einen separaten Pool zusammengefasst werden? Wenn einige Sprachen zur Fairness überrepräsentiert beprobt werden, wie wird dann der globale Wert berechnet? Wenn das Ziel 100 gültige Beobachtungen in einem Land beträgt, wie viele Personen sollten eingeladen werden? Wer ersetzt den Teilnehmer, falls dieser die Aufgabe nicht abschließt? Kann ein Nutzer, der falsche Antworten gibt, ersetzt werden? Wenn dieselbe Person mehr als ein KI-Produkt testet, ist die Stichprobe dann immer noch unabhängig?

Selbst bei einer Rohstichprobe von 1,000 Personen, können stark ungleiche Gewichtungen die effektive statistische Information unter die von 1,000 unabhängigen, gleichgewichteten Beobachtungen senken? Dies sind nicht nur operationale Details. Das Stichprobendesign bestimmt, wie viel Einfluss jede Person auf den GEO-Score hat. Eine fehlerhafte Stichprobe:

- kann die richtigen Antworten falsch gewichten,
- kann kleine Länder symbolisch überrepräsentieren,
- kann große Bevölkerungen kleiner erscheinen lassen, als sie sind,
- kann ressourcenarme Sprachen unsichtbar machen,
- kann freiwillige Technologieanwender so darstellen, als ob sie die Weltbevölkerung wären.

Es kann angepasst werden, um die gewünschte Punktzahl nach Überprüfung der Ergebnisse zu erreichen. Die zuvor festgelegte Prüfarchitektur verlangt, dass jede Prüfung messbar ist und dass das Modell/Produkt das Datum, das Land, die Sprache, die Abfragesätze, die Wiederholungsanzahl und den Messdatensatz trägt. Dieser Abschnitt wendet die gleiche Disziplin auf die Stichprobengestaltung an. Da echte Nachweise das Anfordern von Grenzen, Kontext und Zeit erfordern; erfordert es außerdem, sichtbar zu machen, wie die Stichprobe erstellt wurde. Dieser Abschnitt:

- wie die 1,000 gültigen Beobachtungen des Hauptpopulation-Panels verteilt werden,
- wie Schichten und Zellen etabliert werden,
- wie kleine Länder verwaltet werden sollen,
- wie Brüche in ganze Zahlen umgewandelt werden
- wie Oversampling gewichtet wird,
- wie man Einladungs-, Backup- und gültige Beobachtungsziele trennt,
- wie man Design- und Analysegewichte festlegt,
- warum die effektive Stichprobengröße von der Rohanzahl abweichen kann

definiert. Dieses Kapitel behandelt noch nicht:

- Von welchem Panelanbieter die Teilnehmer ausgewählt werden.
- die vollständigen Details der Felderhebung bei freiwilliger oder Wahrscheinlichkeitsstichprobe,
- Bewertung und Benotung von Antworten,
- die endgültige Konfidenzintervall-Methode
- alle NOMOS Bewertungsformeln

sind nicht vollständig abgeschlossen. Die Aufgabe des Abschnitts 6 ist:

> *Transformieren Sie die definierte Zielpopulation in eine gefrorene, wiegbare und reproduzierbare GEO-1000-Probe, bevor Sie das Ergebnis sehen.*

NOMOS Herausforderung

Sie richten ein globales Panel von 1,000 Personen für ein KI-Produkt ein. Die produktberechtigten Bevölkerungen von zehn Ländern sind wie folgt:

Land	Produktberechtigte Nutzerpopulation
A	400,000,000
B	250,000,000
C	150,000,000
D	100,000,000
E	60,000,000
F	30,000,000
G	8,000,000
H	1,500,000
I	400,000
J	100,000
Gesamt	1,000,000,000

Rohe Quoten gemäß reinem Bevölkerungsanteil:

Land	Rohe Quote
A	400.0
B	250.0
C	150.0
D	100.0
E	60.0
F	30.0
G	8.0
H	1.5
I	0.4
J	0.1

Eine halbe Person kann nicht ausgewählt werden. 0.4 Personen können nicht in ein Land geschickt werden. Wenn Sie die rohen Quoten auf die nächste ganze Zahl runden, kann die Summe 1,000 überschreiten oder unterschreiten. Wenn Sie jedem Land mindestens eine Person zuweisen, werden die Länder I und J im Vergleich zu ihrem Bevölkerungsanteil überrepräsentiert.

Wenn man einer Person das Land I gibt, erhöht sich dessen theoretischer Anteil um 2.5-fache. Wenn man einer Person das Land J gibt, erhöht sich dessen theoretischer Anteil um 10-fache. Diese extreme Repräsentation kann weitgehend korrigiert werden. Aber man kann keinen Länderwert mit nur einer Person erzeugen. Was passiert, wenn man niemandem I und J gibt?

Wenn man diese Länder als getrennte und obligatorische Schichten definiert, wird die Wahrscheinlichkeit, dass einige Personen in der Zielpopulation ausgewählt werden, null. In diesem Fall schwächt sich die Behauptung einer „wahrscheinlichkeitsbasierten Stichprobe, die die gesamte geeignete globale Population repräsentiert“, ab. Betrachten Sie nun eine gefährlichere Entscheidung. Zuerst sammeln Sie die Antworten.

Der einzige Nutzer im Land J gibt eine falsche Antwort. Dann sagen Sie: „Die Stichprobengröße dieses Landes ist sehr klein, lassen Sie uns es vom Umfang ausschließen.“ In Land A gibt es viele positive Antworten. Sie behalten es mit vollem Gewicht bei. Dies ist keine Rundung mehr. Es ist eine ergebnisorientierte Stichprobenänderung.

In einem anderen Fall erhöhen Sie die Rate ungültiger Datensätze in einer Sprachgruppe mit niedrigen Ergebnissen, indem Sie nur Ersatzteilnehmer verwenden, bis eine positive Antwort erhalten wird. Dies ist ebenfalls keine Stichprobe. Es ist ein Reaktionsauswahlmechanismus. Die erste Regelung dieses Abschnitts lautet:

> *Die Stichprobe muss vor der Einsicht in die KI-Antworten festgelegt werden.*

Ihre zweite Bestimmung lautet:

> *Die Wahrscheinlichkeit, dass ein Teilnehmer ausgewählt wird, darf nicht von der Richtigkeit seiner Antwort abhängen.*

Seine dritte Bestimmung besagt:

> *Die Sichtbarkeit des kleinen Landes und das Gewicht der Bevölkerung im globalen Score müssen separat verwaltet werden.*

Seine vierte Bestimmung besagt:

> *Eine Überstichprobe ist möglich; eine Übergewichtung ist nicht verpflichtend.*

Seine fünfte Vorschrift lautet:

> *1,000-Rohdaten tragen nicht immer statistische Informationen, die 1,000-Beobachtungen entsprechen.*

1. ZWECK DES KAPITELS

Der Zweck dieses Abschnitts besteht darin, die anvisierten 1,000 gültigen Beobachtungen für jedes KI-Produkt und jede Messwelle in ein überprüfbares Stichprobendesign basierend auf der produktberechtigten Nutzerpopulation zu überführen. Der Abschnitt standardisiert die folgenden Unterscheidungen:

- Zielpopulation versus ausgewählte Stichprobe
- Rohquote versus ganzzahlige Zuweisung
- Hauptpopulation-Panel versus ergänzende Panels
- Länder-Sichtbarkeit versus Ländergewicht
- Schichten versus experimentelle Zelle
- Große Länder-Schichten versus kleine Länder-Pool
- Proportionale Zuweisung versus unverhältnismäßiges Oversampling
- Stichprobenanteil versus Populationsanteil
- Anzahl der Einladungen versus Ziel für gültige Beobachtungen
- Backup-Teilnehmer versus Ersatz nach Ergebnissen
- Ungültigkeit der Datenerhebung versus ungünstiges AI-Ergebnis
- Design-Gewicht vs. Ergebnis-Gewicht
- Rohstichprobengröße vs. effektive Stichprobengröße
- Länderstichprobe vs. Sprachstichprobe
- Mehrsprachiger Nutzer vs. mehr als eine Person
- Auswahl in einer Stufe vs. Auswahl in mehreren Stufen
- Natürliche Reichweite-Zuteilung vs. Gemeinsame Vergleichsbasis-Zuteilung
- Sofortige Wellen-Zuteilung und Zeitreihen-Zuteilung
- Bevölkerungsschätzung und Analyse der Länder- oder Spracherkennung
- Hauptbeobachtungen 1,000 mit Hilfsbeobachtern und Fairness-Panels

Am Ende dieses Abschnitts sollte jede Prüfung in der Lage sein, diese Fragen zu beantworten:

> *Warum wurden 1,000-Beobachtungen mit diesen Zahlen über dieses Land, diese Sprache und diese Nutzergruppen verteilt?*
>
> *Wie wurde das Gewicht jeder Beobachtung im globalen Ergebnis bestimmt?*
>
> *Wer hatte eine Nullwahrscheinlichkeit ausgewählt zu werden und warum?*
>
> *Wie viele Personen wurden eingeladen, wie viele haben teilgenommen und wie viele Beobachtungen blieben gültig?*
>
> *Wurden die Stichprobenergebnisse eingefroren, bevor sie angesehen wurden?*

2. ZENTRALE NORMATIVE BESTIMMUNG

Die Hauptstichprobe der Bevölkerung GEO-1000 sollte auf einem geschichteten Stichprobendesign basieren, das bekannten oder vorhersagbaren Auswahlwahrscheinlichkeiten für Personen im vorab

eingefrorenen produktberechtigten Bevölkerungsrahmen für jedes KI-Produkt und jede Messwelle zuweist. Das Standardziel des Hauptbevölkerungs-Panels ist:

> *Mindestens 1,000 gültige Beobachtungen pro KI-Produkt und Hauptmesswelle*

werden aufrechterhalten. Diese Zahl:

- ist nicht die Anzahl der eingeladenen Personen,
- ist nicht die Anzahl der Konten,
- ist nicht die Anzahl der Roh-Screenshots,

Es ist nicht die Summe aller ergänzenden Panels. 1,000 ist das Ziel für gültige Produktbeobachtungen, die dem Protokoll im Hauptpopulation-Panel entsprechen. Länderspezifische Beobachter-, Sprachgerechtigkeits- und Barrierefreiheits-Panels können zusätzliche Stichproben erstellen. Diese zusätzlichen Stichproben:

- können das gesamte Feldvolumen über 1,000 erhöhen,
- können mit geeigneten Gewichten im Ergebnis der globalen Population verwendet werden,

dürfen jedoch nicht verwendet werden, um das gültige Beobachtungsziel 1,000 des Hauptpanels heimlich zu reduzieren.

3. GRUNDLEGENDES Repräsentation DER STICHPROBE

Für ein spezifisches KI-Produkt a, Messwelle w und Zielpopulationsrahmen werden die Schichten:

```
h∈ H_{a,w}
```

bezeichnet. Für jede Schicht:

- Nh: geschätzte produktgeeignete Population in der Schicht
- N: Gesamtzielpopulation
- Wh: Bevölkerungsgewicht der Schicht
- n: Zielvalide Hauptpanelbeobachtung
- qh: rohe proportionale Quote
- nh: zugewiesene ganzzahlige valide Beobachtungsziel

ist. Gesamtzielpopulation:

```
N= Σ_{h=1}^H N_h
```

Populationgewicht der Schicht:

```
W_h= N_h/ N
```

Rohe proportionale Quote:

```
q_h= nW_h
```

ist. Hauptproportionale Zuteilung unter idealen Bedingungen:

```
n_h≈ nW_h
```

sollte. Jedoch:

- qh kann gebrochen sein,
- einige Zellen können unter einer Person liegen,
- eine minimale diagnostische Stichprobe kann für Sprach- und Ländersubgruppen erforderlich sein,
- auch Risikogruppen mit kleiner Größe können untersucht werden,

Die Feldabschlussraten können variieren. Daher ist das Rohkontingent nicht direkt die endgültige Zuweisung.

4. DREI UNTERSCHIEDLICHE ZIELE DES STICHPROBEN-DESIGNS

Die gleiche Stichprobe kann möglicherweise nicht alle Ziele mit der gleichen Stärke erfüllen. Drei Ziele sollten getrennt betrachtet werden.

4.1. Globale Bevölkerungsüberschätzung

Beantwortet die Frage:

> *Welches repräsentative Ergebnis würde die durchschnittliche Person in der globalen Nutzerpopulation, die für das Produkt geeignet ist, sehen?*

Die Gewichtung der Population hat oberste Priorität.

4.2. Untergruppen-Diagnose

Beantwortet die Frage:

> *Welche Verzerrungen in der Repräsentation treten in bestimmten Ländern, Sprachen, Plänen oder Nutzergruppen auf?*

Es kann notwendig sein, kleine Gruppen übermäßig zu sampeln.

4.3. Geografische und sprachliche Abdeckung

Beantwortet die Frage:

> *In welchen Ländern und Sprachen gibt es mindestens reale Nutzerbeobachtungen?*

Landesbeobachter- und Sprachgerechtigkeits-Panels werden wichtig. Diese drei Ziele können nicht unerklärlich in einer einzigen Stichprobe kombiniert werden.

5. HAUPTSTICHPROBENARCHITEKTUR

Die vorgesehene Hauptstichprobenarchitektur für GEO-1000 besteht aus vier Schichten.

5.1. Bevölkerungspanel

Jedes KI-Produkt und jede Welle hat mindestens 1,000 gültige Beobachtungen als Hauptpanel. Es liefert eine globale Populationsschätzung.

5.2. Landesbeobachter-Overlay

Bietet zusätzliche geografische Beobachtungen für kleine Länder, die im Hauptpopulationpanel nicht direkt vertreten sein können. Diese Beobachtungen können anzeigen:

- Einzelereignis,
- Zugangsstatus,
- zukünftige Forschungsbedarfe

Sie erzeugen allein keine Länderbewertung.

5.3. Sprachgerechtigkeits-Oversample

Erzeugt zusätzliche Beobachtungen für Sprachen mit geringen Ressourcen oder Sprachen, die im Hauptpanel unterrepräsentiert sind. Sie werden an die tatsächlichen Bevölkerungsgewichte im globalen Ergebnis angepasst.

5.4. Überschussstichproben für Barrierefreiheit

Es werden auch Nutzer von Assistenztechnologien und Barrierefreiheit erfasst, die im Standardpanel nicht ausreichend vertreten sind. Das Zieluniversum und das Gewicht dieser Beobachtungen werden separat aufgezeichnet.

6. WAS IST EIN STRATUM?

Ein Stratum ist ein Teil der Bevölkerung, der in Bezug auf Stichprobenzuweisung und Auswahlwahrscheinlichkeit separat verwaltet wird. Das Stratum kann eine Kombination aus einem oder mehreren der folgenden Bereiche sein:

- Land
- Region
- Sprache
- Locale
- KI-Plan
- Web oder mobil
- Kontrolliert oder Panel unter natürlichen Nutzungsbedingungen
- Barrierefreiheitsstatus
- Nutzertyp
- Zuständigkeitsbereich

Wenn jede kleine Kombination zu einem separaten Stratum gemacht wird, wird die Stichprobe fragmentiert. Zum Beispiel: Land × Sprache × Plan × Mobil/Web × Altersgruppe × Nutzertyp-Kombination kann Hunderte oder Tausende von Zellen erzeugen. 1,000-Beobachtungen liefern für die meisten dieser Zellen keine zuverlässigen Schätzungen. Daher sollte die Schichtung:

- substanziell sein,
- messbar sein,
- angemessen für die Stichprobengröße sein

muss sein.

7. EIN STRATUM IST NICHT DAS GLEICHE WIE EINE BERICHTSZELLE

Es kann separat auf einer Gruppenergebnis-Karte gezeigt werden. Es muss jedoch kein separates Stratum im Stichprobendesign sein. Zum Beispiel:

- eine kleine Sprache,
- eine bestimmte Zugänglichkeitsgruppe,
- eine bestimmte Altersgruppe

Es kann innerhalb eines breiteren Stratum in der Hauptstichprobe ausgewählt werden. Später kann es in der Analyse separat berichtet werden. Das Gegenteil ist ebenfalls möglich: Zwei Regionen können

getrennt für die Stichprobeneffizienz geschichtet werden. Im öffentlichen Bericht können sie jedoch unter einem einzigen regionalen Ergebnis zusammengefasst werden. Das Schichtungsdesign und die Berichtsstruktur sollten separate Aufzeichnungen führen.

8. FOLGE DER SCHICHTUNGSERSTELLUNG

Die Kandidatenfolge für das Hauptpopulation-Panel ist wie folgt:

- KI-Produkt und Nutzeroberfläche
- Natürliche Reichweite oder Gemeinsame Vergleichsbasis Bevölkerungsrahmen
- Land oder kleiner Länderpool
- Hauptprüfsprache oder Sprachgruppe
- Materialplan/ Schnittstellen-Trennung, falls erforderlich
- Teilnehmerauswahlrahmen

Der Plan und die Schnittstelle müssen nicht für jedes Land in separate Ebenen aufgeteilt werden. Dies kann durch folgendes verwaltet werden:

- Unterebene,
- Kontingent,
- zufällige Zuordnung

Methoden.

9. SCHICHTEN VON GROSSEN UND KLEINEN LÄNDERN

Das Rohkontingent eines Landes:

```
q_c= nN_c/ N
```

liegt in dieser Form vor. Wenn das Rohkontingent groß genug ist, kann das Land direkt als separate Schicht gehalten werden. Dies kann genannt werden:

selbstrepräsentierende Länderschicht

Länder mit sehr kleinen Rohkontingenten einzeln in separate verpflichtende Schichten zu machen, kann die Stichprobeneffizienz verringern. Diese Länder können angemessen:

- regional,
- Sprachfamilie,
- Zugangssystem für Produkte,
- geografische Nähe

in kleine Länderpools aufgenommen werden.

10. KLEINER LÄNDERPOOL

Ein kleiner Länderpool:

```
Lassen Sie ihn repräsentiert werden als B_r= {c_1, c_2, …, c_k}
```

Die insgesamt für die Stichprobe geeignete Bevölkerung des Pools:

```
N_{B_r}= Σ_{c∈B_r} N_c
```

Hauptstichprobenquote:

```
q_{B_r}= nN_{B_r}/ N
```

wird wie folgt berechnet. Innerhalb des Pools kann die Länderauswahl mit einer Wahrscheinlichkeit erfolgen, die proportional zur Größe der geeigneten Bevölkerung ist. Die Wahrscheinlichkeit, dass das Land c innerhalb des Pools ausgewählt wird:

```
Pr(c | B_r) = N_c/ N_{B_r}
```

Es kann sein. Dann wird ein Teilnehmer innerhalb des gewählten Landes ausgewählt. Dieser Ansatz:

- kann verhindern, dass kleine Länder eine Nullchance haben, in der Zielbevölkerung ausgewählt zu werden,
- garantiert nicht, dass jedes Land in einer bestimmten Welle ausgewählt wird,

erhält die globale Bevölkerungsschätzung. Ein Land in dieser Welle nicht auszuwählen:

NICHT IN DER STICHPROBE DER BEVÖLKERUNG AUSGEWÄHLT

bedeutet nicht: "Das Land ist nicht Teil der Zielbevölkerung."

11. EINSCHRÄNKUNGEN DER ZUSAMMENLEGUNG KLEINER LÄNDER

Kleine Länder können nicht zufällig in denselben Pool eingeteilt werden, nur weil ihre Bevölkerung klein ist. Die Länder innerhalb des Pools sind insofern ähnlich, als sie relevant sind:

- Produktzugang,
- Nutzeroberfläche,
- Sprache oder Ortsstruktur,
- Rechtlicher Zugang,
- Geographische Region,
- Methode der Datenerhebung

Muss getragen werden. Ein Land ohne Produktzugang kann nicht in denselben Hauptpools wie ein Land mit vollem Zugang eingestuft werden. Unterschiedliche Nutzungsbedingungen können nicht unsichtbar gemacht werden.

12. KLEINER LANDPOOL ERZEUGT KEINE LÄNDERWERTUNG

Ein oder zwei aus dem Pool ausgewählte Beobachtungen können für den globalen Populationsbeitrag des Pools verwendet werden und können auf ein einzelnes Landesereignis hinweisen. Allerdings ergibt es keinen separaten statistischen Wert für ein kleines Land. Wenn ein länderbasierter Wert gewünscht wird, gibt es zusätzlich:

- Country Observer,
- länderspezifische Stichprobe,
- mehrwellen kumulative Stichprobe

kann erforderlich sein.

13. ZUORDNUNG PROPORTIONAL ZUR REINEN BEVÖLKERUNG

Reine proportionale Zuteilung:

basiert auf der Formel:

```
q_h= nW_h
```

Vorteile:

- sie ist für die Hauptbevölkerungsschätzung verständlich,
- sie gewichtet große Bevölkerungen natürlich,
- sie kann Gewichtsunterschiede verringern,

sie kann ein selbstgewichtendes Design annähern. Einschränkungen:

- sie kann sehr kleine Länder und Sprachen nur in geringer Zahl erfassen,
- sie kann möglicherweise nicht ausreichen, um Untergruppenergebnisse zu liefern,
- sie kann kleine Hochrisikogruppen unsichtbar lassen,

sie kann Bruchteilsquoten und nahezu null Quoten erzeugen. Daher kann die reine proportionale Zuteilung eine starke Voreinstellung für das Hauptbevölkerungs-Panel sein. Sie ersetzt keine ergänzenden Panels.

14. UMWANDLUNG VON BRÜCHEN IN GANZE ZAHLEN

Rohe Quoten sind oft gebrochen. Die Gesamtzuweisung sollte genau n betragen. Die Methode des größten Restes kann als Standardmethode verwendet werden.

14.1. Erster Schritt

Jede rohe Quote wird abgerundet:

```
b_h= ⌊q_h⌋
```

14.2. Zweiter Schritt

Die Anzahl der nicht zugewiesenen Beobachtungen:

```
R= n− Σ_h b_h
```

wird berechnet.

14.3. Dritter Schritt

Der Bruchteil jedes Stratum:

```
f_h= q_h− b_h
```

wobei f_h der Bruchteil von q_h ist.

14.4. Vierter Schritt

Die verbleibenden R Beobachtungen werden einzeln zu den Schichten mit den höchsten f_h Werten hinzugefügt.

14.5. Gleichheitssituation

Wenn die Bruchteile der beiden Schichten gleich sind:

- Bevölkerungsgröße,
- vorgegebene feste Reihenfolge,
- registrierter Zufallssamen

Eine unabhängige Regel zur Entscheidungsfindung im Gleichstand sollte verwendet werden, unabhängig vom Ergebnis. Ein Land darf nicht basierend auf den Ergebnissen ausgewählt werden.

15. SYNTHETISCHE GANZZAHLENZUWEISUNG

SYNTHETISCHE ANZEIGE

Geeignete Populationen zu Beginn des Abschnitts:

Land	Geeignete Bevölkerung	Rohe Quote
A	400,000,000	400.0
B	250,000,000	250.0
C	150,000,000	150.0
D	100,000,000	100.0
E	60,000,000	60.0
F	30,000,000	30.0
G	8,000,000	8.0
H	1,500,000	1.5
I	400,000	0.4
J	100,000	0.1

Wenn Länder als separate Schichten geschützt sind, abrunden:

Land	Abgerundete Quote
A	400
B	250
C	150
D	100
E	60
F	30
G	8
H	1
I	0
J	0
Gesamt	999

Eine Beobachtung bleibt übrig. Sie wird H hinzugefügt, da der höchste Bruchteil 0.5 im Land H ist. Endgültige Zuordnung:

Land	Endgültige Quote
A	400
B	250

Land	Endgültige Quote
C	150
D	100
E	60
F	30
G	8
H	2
I	0
J	0
Gesamt	1,000

Auch wenn dieses Design I und J als separate deterministische Schichten definiert, kann es ein Problem für die vollständige Zieluniversumsinferenz darstellen, wenn es keine Chance zur Auswahl gibt. Eine genauere Lösung ist ein kleiner Länderpool.

16. SYNTHETISCHE ZUTEILUNG MIT KLEINEM LÄNDERPOOL

Länder H, I und J:

```
B_small= {H, I, J}
```

sollten im Pool zusammengefasst werden. Pool-Bevölkerung:

```
1,500,000+400,000+100,000=2,000,000
```

Pool-Quote:

```
1,000× 2,000,000/ 1,000,000,000= 2
```

ergibt sich. Hauptzuteilung:

Schicht	Quote
A	400
B	250
C	150
D	100
E	60
F	30
G	8
Kleiner Länderpool	2
Gesamt	1,000

Auswahlgewichte der Länder innerhalb des Pools:

Land	Bevölkerungsanteil innerhalb des Pools
H	75%
I	20%

Land	Bevölkerungsanteil innerhalb des Pools
J	5%

Zwei Hauptpopulationen werden gemäß diesen Wahrscheinlichkeiten ausgewählt. In derselben Welle:

- können beide Beobachtungen von H stammen,
- einer aus H und einer aus I,
- selten aus J

Dies kann vorkommen. Das Design macht J nicht in jeder Welle sichtbar, verringert jedoch nicht die Wahrscheinlichkeit, dass J ausgewählt wird, auf null. Wenn J beobachtet werden muss, wird es separat dem Country Observer Panel hinzugefügt. Eine Beobachtung durch einen Country Observer gibt J kein Gewicht von eins in 1,000 in der Hauptzuweisung; J behält sein tatsächliches Bevölkerungsgewicht im Hauptpanel.

17. MINIMALE ZELLENZUWEISUNG

Reine proportionale Quote ist möglicherweise nicht ausreichend, um eine Diagnostikanalyse für einige Untergruppen durchzuführen. Zum Beispiel könnte die Rohquote für eine Sprachzelle 8 betragen. Mit acht Beobachtungen kann kein zuverlässiger Sprachscore erstellt werden. In diesem Fall gibt es drei Optionen:

- Den Sprachscore nicht zu veröffentlichen
- Die Zelle mit ähnlichen Gruppen zusammenzuführen
- Die Zelle überproportional zu beproben

Die minimale Zellzuweisung muss vor den Ergebnissen definiert werden.

18. KANDIDATEN-REPORTING-BANDS

KANDIDATEN-MELDEGRUPPEN — KANDIDATEN-MELDERBANDS

Die unten aufgeführten Schwellenwerte sind keine endgültigen wissenschaftlichen Entscheidungen. Sie sollten durch Pilot- und externe Überprüfungen kalibriert werden.

1–29 Gültige Beobachtung

Einzelereignis und Entdeckung Länder- oder Sprachrate wird nicht veröffentlicht

NUR BEOBACHTUNGSZWECK

30–99 Gültige Beobachtungen

Explorative Level-Schätzung Breite Unsicherheit Es wird kein endgültiges Konformitätsurteil fällt

ERKUNDUNGSARBEIT

100–199 Gültige Beobachtungen

Erste Untergruppenschätzung Die Unsicherheit sollte sichtbar sein

VORLÄUFIG

200 und darüber hinaus gültige Beobachtungen

Stärkere Untergruppenschätzung Kandidat Designeffekt und Gewichte werden erneut bewertet

STANDARD-SCHÄTZUNGSKANDIDAT

Diese Bands:

- Fehlerquote,
- erwartete Genauigkeit,
- Gewichtsvariabilität,
- design Effekt

kann nicht automatisch als ausreichend betrachtet werden, ohne berücksichtigt zu werden.

19. ÜBERABTASTUNG

Ein Land, eine Sprache oder eine Nutzergruppe kann mehr Beobachtungen haben als ihr Bevölkerungsanteil. Dies wird als Überabtastung bezeichnet. Ihre Zwecke:

- die Sensitivität von Untergruppen zu erhöhen,
- eine ressourcenarme Sprache sichtbar zu machen,
- eine hochriskante kleine Gruppe zu untersuchen,
- den Coverage des Ländebeobachters zu erweitern,
- die Zugänglichkeitsnutzer angemessen zu vertreten

sind möglich. Überabtastung ist nicht falsch. Was falsch ist:

> *Es ist, den Stichprobenanteil der überabgetasteten Gruppe zu verwenden, als ob er der tatsächliche Bevölkerungsanteil wäre.*

20. BEVÖLKERUNGSTEIL UND STICHPROBENTEIL

Bevölkerungsanteil für Schicht h:

```
W_h= N_h/ N
```

Stichprobenanteil:

```
S_h= n_h/ n
```

ist in der Form. Bei rein proportionaler Zuteilung:

```
S_h≈ W_h
```

okay. Bei übermäßiger Stichprobenziehung:

```
S_h> W_h
```

Es ist möglich. Bei einer kleinen Stichprobe:

```
S_h< W_h
```

Es ist möglich. Die weltweite Bevölkerungsschätzung sollte Wh-Werte anstelle von Sh beibehalten.

21. GRUNDLAGEN-ENTWURFSLAST

Bei einfacher geschichteter Stichprobenerhebung kann das Grundgewicht jeder Beobachtung innerhalb der Schicht h als:

```
d_i= N_h/ n_h
```

Das normalisierte Schichtgewicht kann geschrieben werden als:

```
a_h= W_h/ S_h
```

Das normalisierte Beobachtungsgewicht kann sein:

```
w_i= W_h/ n_h
```

Es wird sichergestellt, dass die Summe eins ergibt:

```
Σ_i w_i= 1
```

Die gewichtete Schätzung des binären Übergangs kann berechnet werden als:

```
θˆ= Σ_i w_iY_i
```

Hier:

```
Yi=1: Antwort, die das Übergangskriterium erfüllt

Yi=0: Antwort, die es nicht erfüllt
```

Es ist möglich. Das finale NOMOS-System muss nicht nur binäre Ergebnisse verwenden. Die gleiche Gewichtungslogik kann auch auf mehrdimensionale Antworten angewendet werden.

22. BEISPIEL ÜBERMÄSSIGER STICHPROBENNAHME

SYNTHETISCHES BEISPIEL

In der geeigneten Population von Land B:

- Sprache X: 80 Prozent
- Sprache Y: 20 Prozent

Sei es so. Im Population Panel gibt es 250 Beobachtungen für Land B. Reine Zuweisung:

- Sprache X: 200
- Sprache Y: 50

Lassen Sie 50 zusätzliche Beobachtungen zur Sprachfairness für Sprache Y durchführen. Gesammelte Beobachtungen:

- Sprache X: 200
- Sprache Y: 100

Der Anteil von Sprache Y in der Rohstichprobe:

```
100/300 = 33.3%
```

Es war. Der tatsächliche Bevölkerungsanteil beträgt 20%. Im globalen Bevölkerungsbericht kann Sprache Y kein 33.3%-Gewicht erhalten. Innerhalb der Schicht normalisierte Gewichte:

- Gesamtgewicht für Sprache X: 80%
- Gesamtgewicht für Sprache Y: 20%

wird beibehalten. Jede der 100-Beobachtungen in Sprache Y erhält ein geringeres individuelles Gewicht als eine einzelne Beobachtung in Sprache X. Im Gegensatz dazu werden alle 100-Beobachtungen im separaten Diagnosebericht von Sprache Y verwendet. Dieser Ansatz: erhöht die Sichtbarkeit der Untergruppe, verzerrt die globale Bevölkerungsschätzung nicht.

23. ZUORDNUNG VON MEHRSPRACHIGEN NUTZERN

Eine Person kann für mehr als eine Sprachzelle geeignet sein. Einer der bevorzugten Ansätze für das Haupt-Populationspanel besteht darin, dem Nutzer eine vordefinierte Hauptprüfungsprache zuzuweisen. Die Hauptprüfungsprache kann gemäß diesen Feldern bestimmt werden:

- Tägliche KI-Nutzungssprache

- Mutter- oder Heimatsprache
- Markt und Gebietsschema
- Kompetenz in der natürlichen Nutzung des Prompts
- Panelausgleich

Wenn der Nutzer auch in anderen Sprachen getestet wird, werden diese Beobachtungen:

- gepaartes mehrsprachiges Panel,
- Language Fairness Panel

ebenfalls entsprechend gekennzeichnet. Es ist verboten, dass dieselbe Person in allen Sprachen eine vollständige Bevölkerungsgewichtung erhält.

24. SPRACHENZUWEISUNG INNERHALB DES LANDES

Sei die geeignete Population für Sprachgruppen l im Land c: N_{c,l}. Wenn die Landquote: n_c beträgt, dann die rohe Sprachquote:

```
q_{c,l}= n_c× N_{c,l}/ Σ_l N_{c,l}
```

kann wie folgt berechnet werden. Wenn sich jedoch Sprachgruppen überschneiden:

```
Σ_l N_{c,l}> N_c
```

kann möglich sein. In diesem Fall führt die Verwendung eines direkten Anteils zu einer doppelten Zählung. Lösungsmöglichkeiten:

- Hauptprüfung Sprachklassifikation,
- Menschliches Gewicht aufteilen,
- mehrsprachiges gemeinsames Modell,
- getrenntes Übereinstimmungs-Sprachpanel

kann verwendet werden. Die verwendete Methode sollte dokumentiert werden.

25. ZUORDNUNG VON PLAN UND BENUTZERSCHNITTSTELLE INNERHALB DER BEVÖLKERUNG

Kostenlose und bezahlte Nutzer oder Web- und Mobile-Oberflächen können sich unterschiedlich verhalten. Die Aufteilung jeder Länder-Sprach-Zelle separat in Plan- und Interface-Ebenen kann jedoch die Stichprobe zu stark fragmentieren. Drei Ansätze können verwendet werden.

25.1. Unterzuordnung proportional zur Bevölkerung

Wenn die tatsächliche Nutzungsverteilung bekannt ist, erfolgt die Zuweisung gemäß Plan und Schnittstellenanteilen.

25.2. Ausgeglichene experimentelle Zuweisung

Gleiches oder höheres Sampling wird durchgeführt, um den Unterschied zwischen Plan oder Schnittstelle zu vergleichen. Bei globaler Schätzung wird es auf die tatsächlichen Nutzungsanteile neu gewichtet.

25.3. Separate Produktoberflächenwerte

Freie Web-, kostenpflichtige Web- und mobile Produkte werden als separate Prüfobjekte gemeldet. Ein kombinierte Ergebnis kann ebenfalls erstellt werden. Der zu verwendende Ansatz sollte vor der Messung definiert werden.

26. EINLADUNGSNUMMER UND GÜLTIGES BEOBACHTUNGSZIEL

Wenn das gültige Beobachtungsziel für eine Schicht nh beträgt, ist es normalerweise nicht ausreichend, genau nh Personen einzuladen. Denn:

- einige Personen nehmen nicht teil,
- einige schließen die Aufgabe nicht ab,
- einige Datensätze bleiben außerhalb des Protokolls,

einige Evidenzpakete werden ungültig. Lassen Sie die erwartete Abschlussquote für die Schicht rh sein und die erwartete Gültigkeitsrate vh. Das Einladungsziel beträgt ungefähr:

```
m_h= ⌈n_h/ (r_h× v_h)⌉
```

kann wie folgt berechnet werden. Beispiel:

- Ziel für gültige Beobachtungen: 100
- Erwartete Abschlussquote: 80 Prozent
- Erwartete Gültigkeitsrate: 90 Prozent

```
m_h= ⌈100/ (0.80× 0.90)⌉ = 139
```

Personen können eingeladen werden. Dies ist keine Garantie. Während das Feld voranschreitet, kann das vordefinierte Backup-System verwendet werden.

27. GRENZE DER EINLADUNGSINFLATION

Nur mehr Personen einzuladen löst das Stichprobenbias nicht. Frühe Teilnehmer können systematisch anders sein als späte Teilnehmer. Nur die ersten 1,000 Abschlusse zu berücksichtigen:

- Zeitzone,
- Technologieinteresse,
- Verbindungsqualität,
- Arbeitsplan

kann Voreingenommenheit in Bezug auf Einladung und Auswahlreihenfolge erzeugen:

- sollte randomisiert werden,
- muss innerhalb des Stratum aufgezeichnet werden,

sollte während des gesamten Befragungszeitraums verwaltet werden.

28. ALTERNATIVES TEILNEHMER-SYSTEM

Für jedes Stratum sollte vor der Einsicht der Ergebnisse eine geordnete Sicherungsliste erstellt werden. Ein Sicherungsteilnehmer kann nur aus folgenden Gründen aktiviert werden:

- Keine Antwort auf die Einladung

- Fehler beim Abschluss der Aufgabe
- Technisches oder ethisches Eignungsproblem
- Verstoß gegen das Nachfrageprotokoll
- Ungültigkeit des Evidenzpakets
- Neuregistrierung derselben Person

Der Sicherungsbenutzer darf aus folgendem Grund nicht eingreifen:

> *Die KI-Antwort des vorherigen Nutzers lieferte ein falsches, negatives oder unerwünschtes Ergebnis.*

Eine falsche KI-Antwort ist das gültige Systemergebnis. Sie kann kein Grund für eine Teilnehmeränderung sein.

29. FÜHRT ZU ERBLINDUNG BEI DER GÜLTIGKEITSENTSCHEIDUNG

Die Person, die entscheidet, ob die Beobachtung so weit wie möglich dem Protokoll entspricht:

- muss die Antwort bewerten, ohne zu wissen, ob sie richtig oder falsch ist,
- oder ob es positiv oder negativ für die Marke ist.

Zuerst:

- Erfassung und Protokollgültigkeit
- Dann die substanziellen Reaktionsbewertung

muss erledigt werden. Andernfalls können negative Antworten leichter als „ungültig“ erklärt werden.

30. RESERVIERUNGSVORGANG INNERHALB DER SCHICHT

Der Ersatzteilnehmer sollte nach Möglichkeit vom Gleichen sein:

- Land,
- Sprache,
- Produktoberfläche
- Plan,
- Bedingung des Panels

Schicht. Wenn keine Reserve in derselben Schicht verbleibt, wird die vordefinierte Eskalation angewendet:

- Zusätzliche Zufallsauswahl innerhalb derselben Schicht
- Vordefinierter Pool der oberen Schicht
- Zuweisung unvollständig lassen und Bereichewarnung
- Neue Analyseverson

Nachdem das Ergebnis gesehen wurde, kann ein Nutzer aus einem anderen Land zu holen und die Lücke zu schließen, nicht stillschweigend erfolgen.

31. WAS PASSIERT, WENN EINE SCHICHT ZU KURZ IST?

Es kann unter das Schichtziel fallen. Beispiel:

- Ziel: 100
- Gültige Beobachtung: 72

Korrekte Optionen:

- Das Feldfenster auf ein vordefiniertes Ausmaß erweitern
- Die Backup-Liste verwenden
- Die Zelle mit einer ähnlichen und vordefinierten oberen Schicht zusammenführen
- Melden Sie das Ergebnis mit großer Unsicherheit
- Veröffentlichen Sie das Zellenergebnis nicht
- Berechnen Sie das globale Gewicht mit einer geeigneten Methode neu

Falsche Option: Entfernen Sie die Schicht mit niedriger Punktzahl vollständig aus der Analyse.

32. NACHERGEBNIS-QUOTENÄNDERUNG

Die folgenden Änderungen sind ergebnisorientierte Probenmanipulationen:

- Senkung der Zielquote des Landes mit niedriger Punktzahl
- Erhöhung der Quote der Sprache mit hoher Punktzahl
- Ungültigmachung von Nutzern, die negative Antworten geben
- Neudefinition des Small-Country-Pools basierend auf den Ergebnissen
- Ausschluss eines Plans mit kritischen Fehlern
- Hinzufügen neuer Nutzer, bis der gewünschte Durchschnitt erreicht ist

Jede Änderung der Materialquote:

- neue Zuweisungsversion,
- Begründung,
- Datum,
- Entscheidungsträger
- Erhaltung des vorherigen Ergebnisses

soll damit durchgeführt werden.

33. MEHRSTUFIGE STICHPROBE

Mehrstufige Stichproben können verwendet werden, wenn kein direkter Zugang zu allen Personen möglich ist. Beispiel:

- Auswahl der Regionen- oder Länderauswahlpool
- Länderauswahl
- Auswahl des Panels oder lokalen Forschungspartners
- Nutzerauswahl
- KI-Produkt- oder Aufgabenverteilung

Die Gesamtauswahlwahrscheinlichkeit für einen einzelnen Nutzer ist das Produkt der bedingten Wahrscheinlichkeiten in jeder Stufe:

```
π_i= π_r× π_{c|r}× π_{f|c}× π_{i|f}
```

Hier:

- r: Region oder kleiner Länderpool
- c: Land
- f: Stichprobenrahmen oder Panel
- i: Person

Grundlegendes Designgewicht:

```
d_i= 1/π_i
```

Wenn Auswahlwahrscheinlichkeiten nicht erfasst werden können, kann das Design nicht als vollständig wahrscheinlichkeitbasiertes Sample dargestellt werden. [K04; K07]

34. ZUTEILUNG IN FREIWILLIGEN UND NICHT-WAHRSCHEINLICHKEITSPANELEN

Ein wirklich wahrscheinlichkeitbasiertes Weltpanel ist möglicherweise nicht immer möglich. Ein freiwilliges oder professionelles Online-Panel kann verwendet werden. In diesem Fall können Länder- und Sprachquoten weiterhin gemäß der produktberechtigten Bevölkerung festgelegt werden. Allerdings:

- die tatsächliche Auswahlwahrscheinlichkeit des Teilnehmers kann unbekannt sein,
- Selbstselektion kann auftreten,

Gewichtung kann nur Anpassungen auf Basis beobachteter Variablen vornehmen. Der korrekte Status:

Nicht-wahrscheinlichkeitliches Nutzerpanel, geschichtet und kalibriert auf die Bevölkerung

kann sein. Die folgende Aussage darf nicht verwendet werden: „1,000 Menschen werden völlig zufällig aus der Weltbevölkerung ausgewählt.“

35. GEWICHTSARCHITEKTUR

Das endgültige Beobachtungsgewicht kann aus mehreren Komponenten bestehen:

```
w_i= d_i× a_i^{NR}× g_i^{CAL}
```

Hier:

- di: Basis-Designgewicht
- aiNR: Anpassung für Nichtreaktion
- giCAL: Kalibrierungskoeffizient

35.1. Designgewicht

Es ist der Kehrwert der Auswahlwahrscheinlichkeit der Teilnehmenden im Stichprobendesign.

35.2. Nichtreaktionsanpassung

Für ausgewählte Personen mit ähnlichen Merkmalen:

- wie viele geantwortet haben,

- wie viele eine gültige Beobachtung geliefert haben

kann berücksichtigt werden. Für die Stichprobenreaktionsklasse r:

```
a_r^{NR}= (Σ_{j∈S_r} d_j)/ (Σ_{i∈R_r} d_i)
```

Hier:

- Sr: ausgewählte geeignete Stichprobe
- Rr: diejenigen, die geantwortet oder eine gültige Beobachtung geliefert haben

kann sein. Nichtantwortanpassung: löst das unbeobachtete Verzerrungsproblem nicht vollständig, sollte zusammen mit seinen Annahmen erklärt werden.

35.3. Kalibrierungsgewicht

Die Stichprobe, zu bekannten Bevölkerungsgesamtsummen:

- Land,
- Sprache,
- Alter,
- Plan,
- Nutzeroberfläche.

Kalibrierung kann über diese Dimensionen angewendet werden, jedoch nur dort, wo verlässliche externe Gesamtsummen vorliegen.

36. GEWICHTSNORMALISIERUNG

Gewichte:

- an die Gesamtbevölkerung,
- an die gesamte Stichprobengröße,
- auf einzelne

kann normalisiert werden. Die in der Score-Formel verwendete Normalisierung sollte erklärt werden. Einzeln normalisierte Gewichte:

```
w̃_i= w_i/ Σ_j w_j
```

werden erhalten. Gewichtete Schätzung:

```
θ̂= Σ_i w̃_iY_i
```

wird wie folgt berechnet.

37. GEWICHTSBESCHNEIDUNG

Einige Teilnehmer können sehr hohe Gewichte erhalten, weil sie sehr kleine und unterrepräsentierte Gruppen repräsentieren. Diese Situation:

- kann die geschätzte Varianz erhöhen,
- die Einflussnahme einer einzelnen Beobachtung,
- und die Schwankung der Werte

kann zunehmen. Gewichte können auf einem oberen Limit beschnitten werden. Das Beschneiden kann jedoch: Verzerrungen in der Populationsdarstellung erzeugen, kleine Gruppen wieder unsichtbar machen. Daher sollte die Gewichtsbescheidung:

- vor den Ergebnissen definiert werden,

- zusammen mit den unbeschnittenen Ergebnissen gezeigt werden,

und eine Sensitivitätsanalyse enthalten.

38. EFFEKTIVE STICHPROBENGRÖßE

Auch wenn die Rohzahl der Beobachtungen n beträgt, kann die tatsächliche Informationsmenge geringer sein, wenn die Gewichte stark variieren. Kish effektive Stichprobengröße:

```
n_eff= (Σ_i w_i)^2/ Σ_i w_i^2 [K08]
```

kann wie folgt gezeigt werden. Bei gleichen Gewichten:

```
neff≈n
```

erhalten wir. Wenn die Gewichte sehr unausgeglichen sind: neff<n. Eine GEO-1000-Studie:

- 1,000 gültige Beobachtungen,
- können 620 effektive Stichprobengröße

tragen. Für die Öffentlichkeit: Die Angabe „1,000 Personen wurden gemessen.“ kann aufgrund der Präzision stark erscheinen. Im Hauptbericht sollten mindestens:

- Rohzahl der gültigen Beobachtungen,
- effektive Stichprobengröße,
- Gewichtsvariabilität

bestimmt werden.

39. GEWICHT-DESIGN-EFFEKT

Ungefähre Designwirkung aufgrund von Gewichten:

```
DEFF_w≈ 1+ CV(w)^2
```

kann gezeigt werden. Hier:

```
CV(w) = sd(w) / mean(w)
```

Dies ist der Variationskoeffizient der Gewichte und stellt nur eine ungefähre Repräsentation ihrer Wirkung dar. Die gesamte Designwirkung kann auch variieren mit:

- Clusterbildung,
- Stratifizierung,
- wiederholten Messungen,
- gepaarten Nutzern

Diese Effekte müssen separat geschätzt werden.

40. CLUSTERBILDUNG

Dasselbe:

- Aus dem Forschungspanel,
- aus dem Land,
- aus der Sprachgemeinschaft,
- vom Gerät oder der Institution,
- vom Nutzer

Beobachtungen können einander ähneln. In diesem Fall sind 1,000 Beobachtungen nicht 1,000 vollständig unabhängige Informationseinheiten. Clusterbildung:

- im Stichproben-Design,
- bei der Varianzberechnung,
- bei der effektiven Stichprobengröße

sollte berücksichtigt werden.

41. TEST VON MEHR ALS EINEM KI-Produkt DURCH DEN GLEICHEN NUTZER

Wenn ein Nutzer mehrere KI-Produkte testet, kann der Unterschied zwischen den Modellen an derselben Person gemessen werden. Vorteil: Nutzermerkmale können kontrolliert werden, gepaarter Vergleich kann gestärkt werden. Einschränkung:

- Produktbeobachtungen sind nicht unabhängig,
- die Reihenfolge der Aufgaben kann andere Antworten beeinflussen,
- ein Nutzer kann die im ersten System gesehene Antwort auf das nächste System übertragen,

Nutzer, die auf alle Produkte zugreifen können, können sich von der Gesamtpopulation unterscheiden. Daher Produkte:

- sollte in zufälliger Reihenfolge zugewiesen werden,
- die Weitergabe vorheriger Antworten sollte verhindert werden,
- das gepaarte Design sollte aufgezeichnet werden,

Die Analyse sollte die Abhängigkeit schützen. Jedes KI-Produkt sollte erneut sein eigenes Ziel von 1,000 gültigen Produktbeobachtungen haben. Wenn dieselben 1,000 Menschen zehn Produkte testen:

- 1,000 einzigartige Menschen,
- 10,000 Produktbeobachtungen

können gebildet werden. Diese beiden Zahlen sind nicht gleich.

42. AUSGEWOGENES MISSING-BLOCK-DESIGN

Nicht alle Nutzer müssen auf alle KI-Produkte zugreifen. Zum Beispiel, wenn es zehn KI-Produkte gibt, kann jedem Teilnehmer drei Produkte zugewiesen werden. Wenn die Zuweisungen auf ausgeglichene Weise erfolgen:

- jede Produkt 1,000 Beobachtungen,
- jedes Produktpaar ausreichende gemeinsame Nutzer,
- breitere Teilnahme der Bevölkerung

kann erreicht werden. Als Kandidat für diesen Ansatz:

Ausgewogenes unvollständiges Blockdesign

kann verwendet werden. Seine Vorteile:

- reduziert die Belastung der Teilnehmer,
- verringert das Risiko, nur starke KI-Nutzer auszuwählen,

bietet eine passende Vergleichbarkeit über Modelle hinweg. Das genaue Design wird in den Abschnitten 8 und 9 näher erläutert.

43. Natürliche Reichweite VERTEILUNG

Jedes KI-Produkt erhält eine separate Zuteilung gemäß seiner eigenen Natürliche Reichweite-Population. Daher dasselbe Land:

- 200 für ein Produkt,
- 0 für ein anderes Produkt,
- 50 für ein drittes Produkt

kann beobachtet werden. Dieser Unterschied ist keine Ungerechtigkeit. Es ist so, dass die tatsächlichen Zugangsuniversen der Produkte unterschiedlich sind. Da Natürliche Reichweite-Werte zu unterschiedlichen Populationen gehören, sollten sie nicht allein für eine direkte Rangliste verwendet werden.

44. Gemeinsame Vergleichsbasis VERTEILUNG

Ein gemeinsames Zugangsuniversum wird für den Vergleich zwischen Produkten erstellt. Die gemeinsame Population:

```
U_CS= ∩_{a∈A} U_a
```

kann auf diese Weise betrachtet werden. Jedes Produkt wird mit denselben Gemeinsame Vergleichsbasis-Populationsgewichten gemessen. Dieser Ansatz:

- bietet einen fairen Produktvergleich,
- gleicher Länder- und Sprachboden

stellt sicher. Er zeigt jedoch nicht den natürlichen globalen Zugangsunterschied der Produkte. Daher sollten Natürliche Reichweite- und Gemeinsame Vergleichsbasis-Ergebnisse zusammen berichtet werden.

45. VERWALTUNG DER STICHPROBE ÜBER WELLE

Es können zwei verschiedene Ziele über mehrere Messwellen bestehen.

45.1. Aktuelles Querschnittsergebnis

In jeder Welle wird ein aktueller produktberechtigter Bevölkerungsrahmen verwendet. Dieser Wert beantwortet die Frage: "Was passiert gerade im tatsächlichen globalen Zugangsuniversum?"

45.2. Festgelegtes Rahmen-Trendergebnis

Die Bevölkerungsgewichte aus der anfänglichen Welle bleiben erhalten. Dieser Wert beantwortet die Frage: "Wie hat sich die Repräsentation in derselben Referenzbevölkerung im Laufe der Zeit verändert?"

45.3. Trennung von zwei Ergebnissen

Wenn das Produkt auf neue Länder ausgeweitet wird, kann der aktuelle Wert sinken. Dieser Rückgang:

- liegt an der Verschlechterung des Produkts,
- durch die Aufnahme neuer und schwieriger Bevölkerungsgruppen

kann darauf zurückzuführen sein. Ein Trend mit festem Rahmen kann gleichzeitig eine Verbesserung zeigen. Daher:

- Momentaufnahme des aktuellen Rahmens
- Trend mit festem Rahmen

Sie sind separate Ergebnisse.

46. PANELERNEUERUNG UND -TRANSFORMATION

Unter den Wellen:

- die gleichen Nutzer,
- völlig neue Nutzer,
- teilweise erneuertes Panel

nutzbar.

46.1. Vollständiges Abschnittspanel

In jeder Welle werden neue Nutzer ausgewählt. Es kann für die Schätzung der Population stark sein. Es misst keine individuellen Veränderungen.

46.2. Längsschnittpanel

Die gleichen Nutzer werden erneut gemessen. Zeigt individuelle Nutzer- und Produktveränderungen. Das Panel kann Abnutzungs- und Lerneffekte erzeugen.

46.3. Rotierendes Panel

Ein Teil der Teilnehmer wird beibehalten, während ein Teil erneuert wird. Bietet sowohl Trend- als auch frische Bevölkerungsrepräsentation. Die verwendete Panelstruktur sollte auf der Ergebnisübersicht sichtbar sein.

47. SAMPLE LOCK

Die folgenden Datensätze sollten vor Beginn jeder Welle festgeschrieben werden:

- KI-Produkt
- Nutzeroberfläche
- Zielpopulation-Version
- Natürliche Reichweite Länder
- Gemeinsame Vergleichsbasis Länder
- Schichten
- Kleine Länderpools
- Rohquoten
- Ganzzahlige Zuweisungen
- Sprachzuweisungen
- Oversampling-Plan
- Einladungsziele
- Reservelisten
- Auswahlmethode
- Gewichtungsformel
- Regel zur Ergebnisblindheit

- Eskaltation bei Unterdeckung
- Analyseplan
- Decodierungsmethode
- Zufallssamen, falls verwendet
- Verantwortliche menschliche Genehmigung

In diese Datei:

NOMOS Stichproben-Zuweisungssperre

kann den Namen erhalten.

48. ÄNDERUNGEN NACH MUSTERSPERRUNG ERLAUBT

Die folgenden Ereignisse können eine neue Version erfordern:

- Beendigung des Produktzugangs im Land
- Ernstes Sicherheitsproblem im Feld
- Materieller Fehler in den Populationsdaten
- Panelanbieter kann in ein bestimmtes Land nicht liefern
- Änderung der Anforderungen an die Forschungsethik
- Kritischer Fehler im Messinstrument

Änderung:

- muss gerechtfertigt,
- datiert,
- genehmigt,
- durchgeführt werden, wobei die alte Version erhalten bleibt.

Die Stärke oder Schwäche der Ergebnisse ist kein Grund für eine Änderung.

49. MUSTERQUALITÄTSSTATUS

Eine Musterwelle kann einen der folgenden Stati tragen:

SQ-0— NICHT ENTWORFEN

Es gibt kein dokumentiertes Stichprobendesign.

SQ-1— BEQUEMLICHKEITSBEOBACHTUNG

Es handelt sich um eine Bequemlichkeits- oder freiwillige Beobachtung. Eine Verallgemeinerung auf die Gesamtbevölkerung ist nicht möglich.

SQ-2— QUOTENKONTROLLIERT

Es gibt Länder- und Sprachquoten. Die Auswahlwahrscheinlichkeiten sind nicht vollständig bekannt.

SQ-3— GEWICHTETES PANEL

Es gibt ein stratifiziertes Panel und eine Bevölkerungs-Kalibrierung. Abdeckungs- und Selbstselektionseinschränkungen bestehen weiterhin.

SQ-4— WAHRSCHEINLICHKEITSBASIERT ODER HYBRID

Bekannte Auswahlwahrscheinlichkeiten oder ein starkes Hybriddesign sind vorhanden.

SQ-5— REPLIZIERT UND KALIBRIERT

Das Design wurde in verschiedenen Wellen und externen Anwendungen repliziert; Bias- und Varianzverhalten wurden kalibriert. Diese Stati können die Vertrauensklasse des öffentlichen Scores beeinflussen.

50. SYNTHETISCHE END-TO-END GEO-1000 ZUWEISUNG

SYNTHETISCHE METHODOLOGIE-DEMONSTRATION / Die folgenden Länder, Bevölkerungen, Sprachen, KI-Produkte und Berechnungen sind vollständig fiktiv. Für das synthetische Orion KI-Produkt sei die geeignete Bevölkerung für Natürliche Reichweite 1 Milliarden Menschen. Hauptziel des Bevölkerungspanels:

```
n=1,000
```

50.1. Länder- und Kleines-Länder-Pool-Zuweisung

Schicht	Geeignete Bevölkerung	Bevölkerungsanteil	Hauptquote
A	400,000,000	40%	400
B	250,000,000	25%	250
C	150,000,000	15%	150
D	100,000,000	10%	100
E	60,000,000	6%	60
F	30,000,000	3%	30
G	8,000,000	0.8%	8
Kleiner Länderpool	2,000,000	0.2%	2
Gesamt	1,000,000,000	100%	1,000

50.2. Sprachzuweisung in Land B

Berechtigte Bevölkerung in Land B:

- Sprache X: 80 Prozent
- Sprache Y: 20 Prozent

Bevölkerungspanel-Zuweisung:

- Sprache X: 200
- Sprache Y: 50

Zusätzliche Zuweisung zur Sprachgerechtigkeit: Sprache Y: 50 Feldziel in Land B:

- Sprache X: 200
- Sprache Y: 100
- Gesamt: 300

Allerdings, die globalen Bevölkerungsgewichtungen des Bevölkerungspanels:

- Sprache X: 80% von Land B

- Sprache Y: 20% von Land B

bleiben wie sie sind.

50.3. Länderbeobachter-Anhang

Es gibt zwei Wahrscheinlichkeitsbeobachtungen für H, I und J im kleinen Länderpool im Bevölkerungs-Panel. Das Länderbeobachter-Panel:

H: 1

I: 1

J: 1

kann Beobachtungen anvisieren. Diese drei Beobachtungen:

- erhöhen die geografische Abdeckung,
- zeigen individuelle Ländereignisse,

erzeugen keine Länderbewertung.

50.4. Barrierefreiheits-Ergänzung

Wenn festgestellt wird, dass das Hauptpanel die Zugänglichkeitsnutzer unzureichend abdeckt, können zusätzliche Beobachtungen von 40 durchgeführt werden. Diese Nutzer: werden entsprechend relevanter Bevölkerungsgruppen kalibriert und erzeugen ein separates Ergebnis des Accessibility Panels.

50.5. Gesamtfeldvolumen

Panel	Aktuelles Ziel
Bevölkerungs-Panel	1,000
Zusatz zu sprachlicher Fairness	50
Zusatz für Länderbeobachter	3
Zugänglichkeits-Zusatz	40
Gesamtziel für gültiges Feld	1,093

Der Name GEO-1000 ist weiterhin gültig. Weil:

> *1,000 ist das gültige Beobachtungsziel für das Hauptbevölkerungs-Panel.*

Ergänzende Gerechtigkeits- und Abdeckungspanels können dem hinzugefügt werden.

50.6. Einladungsvolumen

Im Bevölkerungs-Panel im Durchschnitt:

- Abschluss: 82 Prozent
- Gültigkeit: 92 Prozent

Es sei zu erwarten. Erforderliche Einladung:

```
[1,000/ (0.82× 0.92)] = 1,326
```

kann sein. Mit ergänzenden Panels können die gesamten Einladungen 1,500 überschreiten. Daher:

> *GEO-1000 ist nicht nur eine Methode, Einladungen an 1,000 Personen zu senden.*

50.7. Feldergebnis

Auf dem Bevölkerungs-Panel:

- 1,326 Personen wurden eingeladen,
- 1,080 Aufgaben abgeschlossen,
- 1,012 Datensätze wurden protokollmäßig als gültig befunden,
- die ersten 1,000 wurden gemäß einer vordefinierten Zufalls-/Backup-Reihenfolge in die Hauptanalyse einbezogen

In Ordnung. Die verbleibenden 12 gültigen Datensätze:

- Backup-Überprüfung,
- Sensitivitätsanalyse,
- erneut prüfen

Es kann gespeichert werden. Es kann nicht entschieden werden, welches 1,000 anhand der Ergebnisse ausgewählt wird.

51. NORMATIVER ZUORDNUNGSALGORITHMUS

Die Standardabfolge der Operationen für ein GEO-1000 Hauptpanel ist wie folgt:

Schritt 1— Produkt- und Wellen-ID festschreiben

Festlegen, welches KI-Produkt, welche Oberfläche und welches Datum gemessen werden.

Schritt 2— Zielpopulation-Version festschreiben

Der produktberechtigte Bevölkerungsrahmen aus Abschnitt 5 wird verwendet.

Schritt 3— Hauptziel der aktuellen Beobachtung definieren

Standard-Kandidatenziel:

```
n=1,000
```

Schritt 4— Materialschichten definieren

Land, Sprache und gegebenenfalls Plan-/Schnittstellenschichten werden definiert.

Schritt 5— Kleine Länderpools einrichten

Länder, die nahezu null Zuteilungen generieren, werden in die entsprechenden Pools aufgenommen.

Schritt 6— Rohproportionale Quoten berechnen

```
qh=nWh
```

Schritt 7— Ganzzahlige Zuteilung durchführen

Die Methode des größten Rests oder eine vorher deklarierte äquivalente Methode wird angewendet.

Schritt 8— Oversampling-Plan hinzufügen

Ergänzungen für Länderbeobachter, Sprachfairness und Zugänglichkeit werden separat gespeichert.

Schritt 9— Einladungs- und Backup-Zahlen berechnen

Erwartete Abschluss- und Validitätsraten werden verwendet.

Schritt 10— Auswahlreihenfolge randomisieren

Primär- und Backup-Listen werden eingefroren.

Schritt 11— Gewichtungsformel speichern

Design-, Antwort- und Kalibrierungsgewichte werden definiert.

Schritt 12— Stichprobe-Sperre bestätigen

Ein Versions- und Integritätsprotokoll wird erstellt, bevor KI-Antworten angezeigt werden.

52. VERPFLICHTENDE NORMATIVE BESTIMMUNGEN

CH06-N01

Die GEO-1000 Stichprobenzuweisung muss erstellt und versioniert werden, bevor KI-Antworten angezeigt werden können.

CH06-N02

Das Hauptzielpanel für die Bevölkerung sollte als mindestens 1,000 gültige Beobachtungen pro KI-Produkt und Hauptmesswelle definiert werden; wenn ein anderes Ziel verwendet wird, muss die Begründung erklärt werden.

CH06-N03

Die eingeladene Person, die Person, die die Aufgabe ausführt, und die gültigen Beobachtungsnummern müssen voneinander unterschieden werden.

CH06-N04

Die Hauptzuweisung sollte auf der produktberechtigten Bevölkerung basieren, die in Abschnitt 5 definiert ist.

CH06-N05

Länder- oder Sprachquoten sollten auf der Grundlage des Anteils der relevanten produktberechtigten Bevölkerung berechnet werden, nicht auf der Gesamtbevölkerung.

CH06-N06

Die Methode zur Umrechnung von rohen Bruchquoten in ganze Zahlen muss im Voraus festgelegt werden.

CH06-N07

Die Auflösung von Rundungsbindungen muss unabhängig von den Ergebnissen sein.

CH06-N08

Kleine Länder mit einer positiven Zielbevölkerung sollten nicht als separate Null-Wahrscheinlichkeits-Deterministikschichten belassen werden; es sollte eine geeignete Zusammenführungs- oder probabilistische Auswahlmethode verwendet werden.

CH06-N09

Kleine Landespools sollten in Bezug auf Produktzugang und finanzielle Bevölkerungsmerkmale gerechtfertigt sein.

CH06-N10

Beobachtungen von Country Observer sollten ebenfalls aufgezeichnet werden, ohne das Hauptbevölkerungsgewicht zu ändern.

CH06-N11

Eine Beobachtung des Länderbeobachters kann nicht als Länderbewertung dargestellt werden.

CH06-N12

Übersampelte Länder, Sprachen oder Nutzergruppen sollten in der globalen Schätzung wieder auf die tatsächlichen Zielpopulation-Gewichte zurückgeführt werden.

CH06-N13

Die Stichprobenquote kann nicht wie die Populationsquote verwendet werden.

CH06-N14

Die gleiche Person kann in der globalen menschlichen Bevölkerung aufgrund unterschiedlicher Sprachen, Konten, Sitzungen oder Produkte nicht dupliziert werden.

CH06-N15

Ungültige Beobachtungen sollten klassifiziert werden, ohne den Grund für die Ungültigkeit auf dem Inhalt der Ergebnisse zu basieren.

CH06-N16

Falsche, negative, ablehnende oder kritische KI-Antworten können nicht der Grund für die Änderung eines Teilnehmers sein.

CH06-N17

Backup-Teilnehmerlisten müssen innerhalb des Stratum erstellt werden, bevor die Ergebnisse gesehen werden.

CH06-N18

Lücken in Schichten können nicht stillschweigend von einem anderen Land oder einer anderen Sprache geschlossen werden.

CH06-N19

Änderungen nach dem Ergebnis in der Quote, Schicht, kleinen Länder-Pool oder Gewichtung erfordern eine neue Analyseversion.

CH06-N20

Alle Komponenten der Gewichtungen und die Methode der Normalisierung müssen veröffentlicht werden.

CH06-N21

Die Annahmen der Non-Response-Anpassung müssen erklärt werden.

CH06-N22

Wenn Gewichtstrimmen verwendet wird, müssen das ungetrimmte Ergebnis und die Sensitivitätsanalyse aufbewahrt werden.

CH06-N23

Der Hauptbericht sollte die angemessene effektive Stichprobengröße zusammen mit der Rohzahl der Beobachtungen zeigen.

CH06-N24

Clustering, Wiederholung und gepaarte Nutzereabhängigkeiten sollten bei der Berechnung der Varianz berücksichtigt werden.

CH06-N25

Das Design, bei dem dieselbe Person mehrere KI-Produkte testet, kann nicht als vollständig unabhängige Nutzerstichprobe dargestellt werden.

CH06-N26

Jedes KI-Produkt muss separat sein eigenes Ziel von 1,000 gültigen Produktbeobachtungen erreichen.

CH06-N27

Die Zuweisungen Natürliche Reichweite und Gemeinsame Vergleichsbasis müssen separate Stichprobenaufzeichnungen führen.

CH06-N28

Ergebnisse der aktuellen und der festen Welle dürfen nicht ohne Erklärung kombiniert werden.

CH06-N29

Wenn die Stichprobe für eine Untergruppe nicht ausreicht, sollte der korrekte Status INSUFFICIENT SAMPLE, EXPLORATORY oder NOT ESTIMATED sein.

CH06-N30

Die Rohstichprobengröße kann nicht als endgültiger Präzisionsindikator verwendet werden, wobei Gewichtungen und Designeffekt ignoriert werden.

CH06-N31

Designgewichte können in einem Freiwilligen- oder Nicht-Wahrscheinlichkeitspanel nicht beansprucht werden, als ob Auswahlwahrscheinlichkeiten bekannt wären.

CH06-N32

Die Stichprobe darf nachträglich nicht so verändert werden, dass niedrig bewertete Gruppen ausgeschlossen oder hoch bewertete Gruppen vergrößert werden.

CH06-N33

Beobachtungen aus dem Hauptbevölkerungs-Panel und ergänzenden Panels sollten in Bezug auf Zweck und Gewichtung getrennt gekennzeichnet werden.

CH06-N34

Das Proben-Lock muss einen verantwortlichen menschlichen oder institutionellen Eigentümer haben.

53. FORMEN DES VERSAGENS

CH06-F01— QUOTE NACH ERGEBNIS FESTLEGEN

Die Anzahl der Länder und Sprachen wird nach der Ansicht der KI-Antworten bestimmt.

CH06-F02— BACKUP VERWENDEN, BIS DIE BESTE ANTWORT VORLIEGT

Negative Antworten werden durch einen neuen Nutzer ersetzt, auch wenn sie gültig sind.

CH06-F03— 1,000-EINLADUNGEN ALS 1,000-GÜLTIGE BEOBACHTUNGEN ZÄHLEN

Nichtabschluss und Ungültigkeit werden verborgen.

CH06-F04— LAND MIT REINER RUNDRUNG AUF NULL ZURÜCKSETZEN

Die Auswahlwahrscheinlichkeit eines Landes mit positiver Zielpopulation bleibt bei null.

CH06-F05— EINER PERSON JE LAND ZUWEISEN UND ES DAS BEVÖLKERUNGSVERHÄLTNIS NENNEN

Die durch die Mindestquote verursachte Unverhältnismäßigkeit ist verborgen.

CH06-F06— BERÜCKSICHTIGUNG der Beobachtung des Länderbeobachters ALS BEVÖLKERUNGSBEOBACHTUNG

Die zusätzliche Beobachtung eines kleinen Landes erhält volles Gewicht im Hauptpanel von 1,000 Personen.

CH06-F07— EINRICHTUNG DES KLEINEN LÄNDERPOOLS AUF BASIS DER ERGEBNISSE

Ein Land, das niedrige Werte produziert, wird in einen anderen Pool verschoben oder aus der Abdeckung ausgeschlossen.

CH06-F08— ZUSAMMENFÜHRUNG INKOMPATIBLER LÄNDER IN DEN GLEICHEN POOL

Länder, die sich in Produktzugang, Sprache oder Rechtsprechung unterscheiden, werden allein deshalb zusammengeführt, weil sie klein sind.

CH06-F09— ÜBERGEWICHTUNG ÜBERMÄSSIGER STICHPROBEN

Bestimmt das globale Ergebnis mit einem Stichprobenanteil, der höher ist als der Anteil der Sprache oder der Bevölkerung des Landes.

CH06-F10— MULTIPLIKATION VON MEHRSPRACHIGEN INDIVIDUEN

Eine Person wird in jeder Sprachzelle als separate Bevölkerungseinheit gezählt.

CH06-F11— BERÜCKSICHTIGUNG DES STICHPROBENANTEILS ALS BEVÖLKERUNGSANTEIL

Rohantwortzahlen werden direkt in die globale Punktzahl umgewandelt.

CH06-F12— ANORDNUNG DER EINLADUNGSREIHENFOLGE NACH LEICHTIGKEIT

Nutzer, die zuerst und schnell antworten, bestimmen das Panel.

CH06-F13— VERWENDUNG VON STRATUM-AUSSCHUSSRESERVEN

Fehlende Länder- oder Sprachbeobachtungen werden stillschweigend mit Teilnehmern aus einer anderen Gruppe ergänzt.

CH06-F14— VERKNÜPFUNG VON UNGÜLTIGKEIT MIT ANTWORTINHALT

Falsche KI-Antwort wird technisch als ungültig erklärt.

CH06-F15— ERSATZ VON ABLEHNUNGEN DURCH EINEN NEUEN NUTZER

Systemablehnung wird aus dem Nenner entfernt, obwohl sie das tatsächliche Nutzerergebnis ist.

CH06-F16— VERSTECKEN DER GEWICHTSFORMEL

Die öffentliche Punktzahl zeigt nicht, mit welcher Bevölkerung und welchen Stichprobengewichten sie erstellt wurde.

CH06-F17— GEWICHTE BESCHNEIDEN, UM KLEINE GRUPPEN ZU LÖSCHEN

Kleine Gruppen mit hohen Gewichten werden beschnitten, weil das Ergebnis nicht gut aussieht.

CH06-F18— DAS EFFEKTIVE SAMPLE VERSTECKEN

Selbst wenn 1,000 Rohdaten aufgrund von Gewichten auf 500–600 Datenpunkte sinken, wird es als „1,000-Personen-Präzision“ präsentiert.

CH06-F19— CLUSTERUNG ALS UNABHÄNGIGKEIT BETRACHTEN

Beobachtungen derselben Panel-, Institutions- oder Nutzergruppe werden als vollständig unabhängig betrachtet.

CH06-F20— DIE GLEICHEN 1,000 PERSONEN ALS ZEHNTAUSEND UNABHÄNGIGE PERSONEN BETRACHTEN

Zehn Beobachtungen eines KI-Produkts werden beschrieben, als wäre die Population zehnmal so groß.

CH06-F21— Gemeinsame Vergleichsbasis ALS Natürliche Reichweite BETRACHTEN

Die übliche Vergleichsstichprobe wird so präsentiert, als ob sie die tatsächliche globale Reichweite des Produkts darstellen würde.

CH06-F22— DIREKTE RANGFOLGE Natürliche Reichweite ERGEBNISSE

Produktergebnisse in verschiedenen Zielpopulationen werden so verglichen, als kämen sie aus demselben Universum.

CH06-F23— VERBERGUNG VON POPULATIONSÄNDERUNGEN IN WELLEN

Auch wenn neue Länder oder Nutzergruppen in die Bewertung einbezogen werden, wird der Trend nur als Änderung der Produktqualität interpretiert.

CH06-F24— ERZEUGUNG DEFINITER BEWERTUNGEN AUS UNGENÜGENDEN STICHPROBEN FÜR EINE UNTERGRUPPE

Einige Dutzend Beobachtungen werden als definitive Bewertungen für ein Land oder eine Sprache veröffentlicht.

CH06-F25— VERSTEKTE ÄNDERUNG DER STICHPROBENSPERRE

Quoten-, Gewichtungs- oder Pooldefinitionen werden ohne Versionsaufzeichnung aktualisiert.

CH06-F26— BLINDES MULTIPLIZIEREN VON EIGNUNGSRATEN

Die berechtigte Bevölkerung und die Quotenkalkulationen ignorieren die Merkmale der abhängigen Bevölkerung.

CH06-F27— ERSETZEN DES HAUPTPANEELS DURCH DAS KOMPLEMENTÄR-PANEL

Sprachliche Fairness oder Beobachtungen durch Länderedakteure werden verwendet, um das Hauptziel der Bevölkerung von 1,000 zu reduzieren.

CH06-F28— LÖSCHEN VON NIEDRIG-BEWERTETEN ZELLEN, WENN DAS PANEL KURZ IST

Unvollständige Stichproben werden ausgeschlossen, da sie in Bezug auf die Ergebnisse nachteilig sind.

CH06-F29— VERWENDUNG VON GEWICHTUNG ALS WISSENSCHAFTLICHE SCHMACK

Es wird behauptet, dass das Repräsentationsproblem nur durch das Hinzufügen von Gewicht gelöst wird, ohne die Auswahl- und Abdeckungsfehler zu erklären.

CH06-F30— DENKEN, DASS GEO-1000 GENAU 1,000 PERSONEN BEFRAGT WERDEN SOLLTE

Einladung, ergänzende Panels, Ungültigkeit, Wiederholung und effektive Stichprobenstruktur werden ignoriert.

54. PRÜFUNGSVERFAHREN

Schritt 1— Zweck der Stichprobe aufzeichnen

Die globale Bevölkerungsschätzung wird in Untergruppen nach Diagnose, Landesbereich, Sprachgerechtigkeit, Zugänglichkeit und Produktvergleichszwecken unterteilt.

Schritt 2— Zielbevölkerungs-Version überprüfen

Die Bevölkerungsdatei und das Datumsprotokoll in Abschnitt 5 werden überprüft.

Schritt 3— Hauptziel der aktuellen Beobachtung definieren

Das Standardziel GEO-1000 oder eine begründete Alternative wird festgelegt.

Schritt 4— Schichten festlegen

Unterschiede in Land, Sprache und Materialbenutzeroberfläche werden bestimmt.

Schritt 5— Kleine Länderpools identifizieren

Für jeden Pool werden:

- Ländern,
- Bevölkerung,
- Begründung für Ähnlichkeit,
- internes Auswahlverfahren

aufgezeichnet.

Schritt 6— Rohquoten berechnen

```
qh=nWh
```

Die Formel oder eine gerechtfertigte Alternative wird verwendet.

Schritt 7— Ganzzahlzuweisung anwenden

Rundungs- und Stichentscheidungsregeln werden festgehalten.

Schritt 8— Komplementäres Oversampling hinzufügen

Ziele für Länderbeobachter, Sprachgerechtigkeit und Barrierefreiheit werden separat definiert.

Schritt 9— Einladungs- und Reservierungsvolumen berechnen

Abschluss- und Gültigkeitserwartungen werden verwendet.

Schritt 10— Auswahl einfrieren und Reservierungsreihenfolge festlegen

Teilnehmer werden ausgewählt oder geordnet, bevor die Ergebnisse bekannt sind.

Schritt 11— Designgewichte berechnen

Auswahlwahrscheinlichkeiten und Schichtgewichte werden festgehalten.

Schritt 12— Verfahren zum Schweigen und Nichtantworten definieren

Eine zuvor angekündigte Antwort wird in die Schichtlücke geschrieben.

Schritt 13— Erstellen des Stichproben-Locks

Dokumenten-ID, Version, Datum und Integritätsaufzeichnung werden hinzugefügt.

Schritt 14— Überprüfung des Validitätsflusses nach dem Feld

Es wird überprüft, dass Gültigkeitsentscheidungen unabhängig von der Genauigkeit der Antworten getroffen werden.

Schritt 15— Generieren der endgültigen Gewichte

Das Design ist in Antwort- und Kalibrierungskomponenten getrennt.

Schritt 16— Berechnen der effektiven Stichprobengröße

Der Unterschied zwischen der Rohzahl und dem Informationsvolumen wird angegeben.

Schritt 17— Vergleich von Zuweisung und Ist-Werten

Geplante und tatsächliche Beobachtungszahlen werden zellenweise angezeigt.

Schritt 18— Überprüfung post-Resultat Änderungen

Quota-, Pool-, Gewicht- und Umfangsänderungen werden überprüft.

55. ERFORDERLICHE BELEGE

KI-Produktidentität; Messwelle; Ziel-Populations-Version; Populationspanel-Ziel; Länderbeobachter-Ziel; Sprachgerechtigkeits-Ziel; Zugänglichkeits-Panel-Ziel; Schichtliste; Schichtpopulationen; Populationsgewichte; Rohquoten; Rundungsmethode; Bruchallokation und Restdatensätze; Entscheidungsregel bei Gleichstand; Klein-Länder-Pools; Auswahlwahrscheinlichkeiten innerhalb des Pools; Länder- und Sprachzuweisungen; Plan- und Schnittstellenzuweisungen; Überstichproben-Begründungen; Einladungsziele; Annahmen zur Fertigstellung; Annahmen zur Gültigkeit; Primäre und Reserve-Teilnehmerlisten; Auswahl- oder Randomisierungsdatensatz; Nichtantwort-Datensätze; Gründe für ungültige Beobachtungen; Gültigkeitsentscheidungsdaten; Designgewichte; Nichtreaktionsanpassungen; Kalibrationskoeffizienten; Gewichtsbeschneidungspolitik; rohe und effektive Stichprobengrößen; Clustering und abgeglichene Nutzeraufzeichnungen; Natürliche Reichweite-Zuweisung; und Gemeinsame Vergleichsbasis-Zuweisung.

Feste und aktuelle Rahmengewicht Vergleiche von geplanter und tatsächlicher Zuweisung Mustersperre Änderungsprotokoll Zuständige Person oder Institution

56. PRÜFLISTE FÜR AUDITS

Ist das Hauptziel für gültige Beobachtungen geöffnet? Sind die Ziel 1,000 Einladungen oder gültige Beobachtungen? Wurde die Stichprobe eingefroren, bevor die KI-Antworten angezeigt wurden? Welcher Populationsrahmen wurde verwendet? Auf welcher Grundlage wurden die Schichten festgelegt? Werden rohe Länder- und Sprachquoten angezeigt? Ist die Methode zur Umrechnung von Bruchteilen in ganze Zahlen klar? War die Abbruchregel vorher festgelegt? Ist die Auswahlwahrscheinlichkeit für kleine Länder null? Sind die Pools kleiner Länder logisch konsistent? Wurden Beobachtungen von Country Observer in die Gewichtung der Hauptpopulation einbezogen? Wurde aus einer Country Observer-Beobachtung ein Länderwert erstellt? Wenn Sprach- und Zugänglichkeitsgruppen überproportional beprobt wurden, wurden sie dann neu gewichtet? Wurde

dieselbe Person als mehr als eine Populationseinheit gezählt? Sind die Zahlen für Einladungen, Abschlüsse und gültige Beobachtungen separat aufgeführt?

Wurden Sicherungslisten im Voraus erstellt? Wurde ein Teilnehmer aufgrund einer falschen oder negativen KI-Antwort ersetzt? Wurde die Gültigkeitsentscheidung vor der Materialgenauigkeitsbewertung getroffen? Wurde die Tabellenlücke stillschweigend aus einem anderen Land geschlossen? Hat sich das Kontingent oder der Pool nach den Ergebnissen geändert? Können die Designgewichte reproduziert werden? Sind die Annahmen der Nonresponse-Einstellung klar? Wurde das Gewicht beschnitten und dessen Effekt gezeigt? Werden Roh- und Effektstichprobengrößen zusammen angegeben? Hat derselbe Nutzer mehr als ein KI-Produkt getestet? Wurde diese Abhängigkeit in der Analyse beibehalten? Hat jedes KI-Produkt 1,000 Produktbeobachtungen separat erreicht? Wurden die Proben Natürliche Reichweite und Gemeinsame Vergleichsbasis getrennt? Hat sich der Populationsrahmen zwischen den Wellen geändert?

Wurden feste und aktuelle gerahmte Trends getrennt? Wurden unzureichende Untergruppenproben veröffentlicht, als ob sie endgültige Ergebnisse wären? Ist der Status der Probenqualität korrekt? Haben ergänzende Panels das Hauptpanel ersetzt? Gibt es einen klar verantwortlichen Eigentümer der Probenregistrierung?

57. EINWÄNDE UND ANTWORTEN

Einwand 1— „Warum geben wir jedem Land nicht direkt Personen entsprechend dem Bevölkerungsverhältnis?“

Dies ist eine starke Annahme in Ländern mit großer und mittlerer Bevölkerung. In sehr kleinen Ländern kann die Rohquote unter eine Person fallen. In diesem Fall:

- Pool kleiner Länder,
- wahrscheinlichkeitsbasierte Länderauswahl,
- separates Beobachtergremium

kann genauer sein.

Einwand 2— „Wäre es nicht fairer, jedem Land mindestens eine Person zuzuordnen?“

Es könnte nützlich im Hinblick auf die geografische Abdeckung sein. In Bezug auf die Bevölkerungsabschätzung ist es unverhältnismäßig. Die korrekte Lösung besteht darin, diese eine Person separat als Country Observer-Beobachtung zu kennzeichnen.

Einwand 3— „Schließt ein Land, das im Hauptgremium null Personen erhält, nicht vollständig aus?“

Es kann ein Problem sein, wenn ein Land als eigenständige deterministische Schicht eine Nullquote erhält. Eine positive Auswahlwahrscheinlichkeit kann im Pool der kleinen Länder vergeben werden. Zusätzlich kann das Beobachterpanel die Beobachtbarkeit des Landes aufrechterhalten.

Einwand 4— „Löscht das Zusammenfassen kleiner Länder kulturelle Unterschiede aus?“

Es kann gelöscht werden, wenn es zur Generierung einer Länderbewertung verwendet wird. Für die hauptsächliche globale Schätzung verwaltet der Pool die Auswahlwahrscheinlichkeit. Die Länderkennung wird im Beobachtungsdatensatz erhalten. Separates Beobachter- und länderspezifisches Sampling kann durchgeführt werden.

Einwand 5— „Warum schließen wir Beobachterbenutzer nicht innerhalb des 1,000 ein?“

Sie können einbezogen werden. Aber dann ist die Stichprobe von 1,000 Personen nicht mehr proportional zur reinen Bevölkerung. Eine Gewichtung wäre erforderlich, und die effektive Stichprobengröße der Haupteinschätzung könnte abnehmen. In der Gründungsarchitektur werden 1,000 Beobachtungen dem Bevölkerungs-Panel zugeordnet; Beobachtungen des Länderbeobachters werden diesem hinzugefügt.

Einwand 6— „Werden die Gesamtkosten vor Ort nicht sehr hoch?“

Das könnte sein. Der Standard sollte keine künstliche Genauigkeit erzeugen, indem er die Kosten verschleiert. Mit einem geringeren Budget:

- weniger KI-Produkte,
- weniger Sprachen,
- kleineres Pilotprojekt,
- größere Unsicherheit

kann benutzt werden. Der korrekte Status sollte klar geschrieben sein.

Einwand 7— „Warum sammeln wir 1,093 oder mehr Beobachtungen, wenn der Name GEO-1000 existiert?“

1,000 ist das gültige Beobachtungsziel für das Hauptbevölkerungs-Panel. Justice- und Coverage-Panels können zum Hauptziel hinzugefügt werden. Dies beeinträchtigt den Namen des Standards nicht. Es macht das tatsächliche Gesamtvolumen des Messfeldes sichtbar.

Einwand 8— „Warum würde eine Überstichprobe einer Sprache das globale Ergebnis verzerren?“

Eine überstichprobenartig erfasste Sprache erscheint groß, weil sie in den Rohantworten enthalten ist. Wenn sie auf das Bevölkerungsgewicht zurückgeführt wird, wird das globale Ergebnis nicht verzerrt. Das separate Sprachergebnis wird präziser.

Einwand 9— „Die Gewichte sind sehr komplex; ist nicht der rohe Durchschnitt leichter zu verstehen?“

Der rohe Mittelwert kann sinnvoll sein, wenn er proportional zur Stichprobenbevölkerung ist. Übermäßiges Sampling, Nichtreaktionen oder Unterschiede in der Abdeckung können dazu führen, dass der rohe Mittelwert die falsche Frage beantwortet. Komplexität kann nicht verborgen werden, wenn sie aus dem Stichproben-Design entsteht.

Einwand 10— „Wenn es 1,000 gültige Beobachtungen gibt, warum sollte die effektive Stichprobe kleiner sein?“

Weil einige Beobachtungen sehr hohes Gewicht haben können, während andere sehr geringes Gewicht haben können. Aufzeichnungen vom gleichen Panel oder Nutzer können einander ähneln. Die Rohzahl ist die Anzahl der physischen Aufzeichnungen. Die effektive Stichprobe nähert sich dem Umfang der statistischen Informationen.

Einwand 11— „Wäre es nicht fairer, wenn dieselben Personen alle KI-Produkte testen?“

Verringert Nutzerunterschiede. Allerdings:

- kann es den Zugang zum Produktuniversum einschränken,
- kann es die Aufgabenlast erhöhen,

kann es Kontamination zwischen den Antworten erzeugen. Es kann mit einem passenden oder balancierten unvollständigen Blockdesign verwaltet werden.

Einwand 12— „Wenn eine Ebene das Ziel nicht erreicht, kann der globale Score niemals veröffentlicht werden?“

Nicht in jedem Fall. Der Mangel wird bewertet in Bezug auf:

- Bevölkerungsgewicht,
- Risiko,
- der Grund für Unzulänglichkeit,
- verbleibende Abdeckung

betrachtet. Die Punktzahl:

- mit eingegrenzter Abdeckung,
- mit großer Unsicherheit,
- Mit der Warnung UNGENÜGENDE ABDECKUNG

kann es veröffentlicht werden. Der Mangel kann nicht verborgen werden.

Berufung 13— „Wenn sich Bevölkerungsdaten ändern, müssen wir dann alle Quoten in jeder Welle erneuern?“

Eine Erneuerung kann für die aktuelle Momentanpunktzahl erforderlich sein. Feste Anfangsgewichte können auch zur Trendverfolgung beibehalten werden. Die beiden Ergebnisse sollten getrennt veröffentlicht werden.

Berufung 14— „Wenn wir kein probabilistisches Weltpanel erstellen können, kann NOMOS dann nicht angewendet werden?“

Es ist anwendbar. Allerdings muss der Status der Probequalität korrekt erfasst werden. Ein Panel von Freiwilligen, das an die Bevölkerung kalibriert ist, kann wertvoll sein. Es kann nicht als vollständig probabilistische Stichprobe dargestellt werden.

GEMEINSAME BESTIMMUNG VON KAPITEL 60

Die Macht von GEO-1000 ergibt sich nicht ausschließlich aus der Zahl 1,000. Von 1,000 Personen:

- aus welcher Bevölkerung sie stammten,
- mit welcher Wahrscheinlichkeit sie ausgewählt wurden,
- welches Land und welche Sprache sie repräsentierten,
- wie oft dieselbe Person gezählt wurde,
- welche Beobachtungen als gültig angesehen wurden,
- wie kleine Länder behandelt wurden,
- wie überproportional ausgewählte Gruppen gewichtet wurden,
- wie viele Personen die Aufgabe nicht abgeschlossen haben,
- wer basierend auf den Ergebnissen ersetzt wurde

Die Autorität einer Stichprobe ergibt sich aus diesen Designentscheidungen. Eine fehlerhafte Stichprobe kann 100,000-Antworten erzeugen und dennoch über die falsche Population sprechen.

Eine solide Stichprobe kann weniger Antworten enthalten und dennoch eine verlässlichere Schlussfolgerung innerhalb klar definierter Grenzen stützen. Die Bevölkerung eines großen Landes sollte ein erhebliches Gewicht in der globalen menschlichen Erfahrung haben, aber es dürfen keine Nutzer erfunden werden, wenn das Produkt nicht verfügbar ist. Das globale Gewicht eines kleinen Landes sollte klein bleiben, aber das Land muss nicht unsichtbar sein. Eine Sprache kann einen kleinen Anteil der Weltbevölkerung ausmachen und dennoch einen schwerwiegenden, messbaren Fehler aufweisen. Eine Person kann drei Sprachen und fünf Produkte verwenden; das erzeugt jedoch nicht fünfzehn Personen. Eine KI-Antwort kann falsch sein, ohne dass der Teilnehmer ungültig wird. Die falsche Antwort ist das, was die Prüfung zu messen versucht.

Das Ersetzen durch einen neuen Nutzer reinigt die Messung nicht. Es löscht die Realität. Gewichtung verwandelt das falsche Panel nicht auf wundersame Weise in die Weltbevölkerung. Eine korrekt gestaltete Überstichprobe kann jedoch Untergruppen sichtbar machen, ohne die globale Schätzung zu verzerren. Daher lautet das sechste Messgesetz von NOMOS:

> *Eine Stichprobe sollte nicht erstellt werden, um das Ergebnis zu produzieren, sondern um das Ergebnis fair zu beobachten.*

Sein siebtes Gesetz lautet wie folgt:

> *Die Sichtbarkeit eines kleinen Landes muss sein Bevölkerungsgewicht nicht vergrößern.*

Sein achtes Gesetz lautet wie folgt:

> *Überschüssige Beobachtung bedeutet nicht überschüssiges Bevölkerungsgewicht.*

Das neunte Gesetz lautet wie folgt:

> *Eine falsche KI-Antwort ist kein Datenfehler; wenn sie unter dem gültigen Protokoll beobachtet wird, handelt es sich um ein Aufsichtsproblem.*

Das zehnte Gesetz lautet:

> *Das Vertrauen in tausend Beobachtungen kommt nicht von der Zahl tausend; es kommt von der Reproduzierbarkeit der Auswahl und Gewichtung.*

Reihenfolge von Kapitel 6 von NOMOS

> *Sag mir nicht einfach: 'Wir haben 1,000 Leute gefragt.'*
>
> *Zeigen Sie, aus welchen Milliarden von Menschen diese 1,000 ausgewählt wurden und mit welcher Wahrscheinlichkeit.*
>
> *Ignoriere ein Land nicht, nur weil es nur für eine halbe Person steht. / Aber beanspruche auch nicht, das ganze Land gemessen zu haben, nur weil du eine Person beprobt hast.*
>
> *Wählen Sie kleine Länder probabilistisch aus dem entsprechenden Pool aus. / Falls Sichtbarkeit erforderlich ist, richten Sie ein separates Beobachterpanel ein.*
>
> *Sie können Sprachen mit geringen Ressourcen überstichproben. / Aber machen Sie sie im globalen Score nicht größer als ihre Bevölkerung.*
>
> *Vermeiden Sie es, eine Person so sehr zu vervielfachen wie Konten, Sprachen, Sitzungen und KI-Produkte.*
>
> *Verwenden Sie das Backup, wenn der Teilnehmer nicht antwortet. / Ändern Sie Ihren Nutzer nicht, wenn die KI eine falsche Antwort gibt.*

Entscheiden Sie über die Gültigkeit, ohne die Richtigkeit Ihrer Antwort zu sehen.

Wenn das Kontingent fehlt, erklären Sie es. / Bringen Sie keine Person aus einem anderen Land, um die Lücke zu füllen.

Ändern Sie den Anteil, das Gewicht, den Länderauswahlpool oder das Stichprobenziel nicht, nachdem Sie das Ergebnis gesehen haben.

Zeigen Sie die Rohzahl an. / Zeigen Sie die effektive Stichprobe an. / Zeigen Sie, wer schwer und wer leicht gewichtet wurde.

Festschreiben Sie zuerst die Population. / Dann richten Sie die Schichten ein. / Dann berechnen Sie die Quoten. / Dann wählen Sie die Ersatzteile aus. / Dann schreiben Sie die Gewichte. / Und erst danach senden Sie die erste Aufforderung.

Der abschließende Satz des Kapitels

GEO-1000 ist keine Methode, tausend Antworten zusammenzustellen; es ist ein Protokoll, um die Repräsentation jeder Person in der Zielpopulation messbar unter der Wahrscheinlichkeit der Auswahl und korrekten Gewichtung richtig zu machen.

Normativer Kern

Für jedes KI-Produkt und jede Hauptmesswelle SOLLTE das GEO-1000-Bevölkerungs-Panel mindestens 1,000 gültige Produktbeobachtungen anvisieren, es sei denn, ein anderes Ziel wird ausdrücklich gerechtfertigt. Die Stichprobenzuweisung MUSS entworfen, versioniert und festgeschrieben werden, bevor KI-Antworten beobachtet werden. Die primäre Zuweisung MUSS auf der produktberechtigten Zielpopulation basieren, nicht auf der rohen Wohnbevölkerung, Einladungszahlen, Konten, Sitzungen oder bequemer Verfügbarkeit. Bruchteilige Zuweisungen MÜSSEN unter Verwendung einer vorab erklärten, reproduzierbaren und ergebnisunabhängigen Methode in Ganzzahlquoten umgewandelt werden. Populationen mit positiver Zielpopulationsmasse DÜRFEN NICHT deterministisch mit einer Selektionswahrscheinlichkeit von null versehen werden, wenn die Prüfung Rückschlüsse auf die gesamte Zielpopulation beansprucht. Kleine Populationen KÖNNEN durch gerechtfertigte zusammengefasste Schichten und wahrscheinlichkeitsproportionale Auswahl vertreten werden. Beobachtungen von Länderbeobachtern, Sprachfairness und Barrierefreiheit KÖNNEN das Hauptpanel ergänzen, MÜSSEN jedoch separat gekennzeichnet und angemessen neu gewichtet bleiben. Überrepräsentierte Gruppen DÜRFEN ihren Rohstichprobenanteil NICHT als Populationsgewicht erhalten. Der Ersatz von Teilnehmern MUSS ausschließlich auf vorab festgelegte Rekrutierungs- oder protokollgültige Bedingungen basieren. Falsche, negative, verweigerte oder kritische KI-Ausgaben DÜRFEN keinen Ersatz auslösen. Personen, Konten, Sitzungen, Sprachen und KI-Produktbeobachtungen MÜSSEN getrennte Einheiten bleiben. Designgewichte, Nichtantwortanpassungen, Kalibrierung, Trimmung, Clusterbildung, gepaarte Beobachtungen und effektive Stichprobengröße MÜSSEN dokumentiert werden. Jede nachträgliche Änderung von Quoten, Schichten, Pools, Gewichten, Ausschlüssen oder Stichprobenzielen MUSS eine neue Analyseversion erstellen und den ursprünglichen Datensatz erhalten.

KAPITEL 07

Panels für Länderbeobachtung und Sprachgerechtigkeit

METHODISCHE GRUNDLAGEN / K04; K06; K07; K20

Kapitelgrenze

Abschnitt 5 hat die folgende Bestimmung eingeführt:

> *Für jedes KI-Produkt muss eine produktberechtigte Nutzerpopulation separat definiert werden.*

Abschnitt 6 hat diese Population in Stichproben umgewandelt:

> *Das Haupt-Populationspanel sollte vor der Einsichtnahme in die Ergebnisse vorab festgelegte Auswahlwahrscheinlichkeiten und Gewichte für Personen in der Zielpopulation bereitstellen.*

Jetzt müssen wir zwei verschiedene Fairness-Probleme lösen. Das erste Problem:

> *Populationsgewichtung kann sehr kleine Länder in der Hauptstichprobe unsichtbar machen.*

Das zweite Problem:

> *Der globale Populationsdurchschnitt kann schwerwiegende Verzerrungen in der Repräsentation bei Sprachen mit geringen Ressourcen oder kleiner Bevölkerung verbergen.*

Eine reine Populationsstichprobe ist stark, um die durchschnittliche menschliche Erfahrung in der Weltbevölkerung zu messen. Jedoch kann dieselbe Stichprobe:

- jedes Land nicht einzeln beobachten,
- nicht genügend Stichproben für jede Sprache liefern.
- kann für kleine Länder keine Länderbewertung erstellen,

erfasst möglicherweise seltene, aber substanziell wichtige Fehler in ressourcenarmen Sprachen nicht. Daher kann GEO-1000 nicht nur aus einem einzelnen Panel bestehen. Neben dem Hauptbevölkerungs-Panel werden zwei ergänzende Strukturen benötigt:

Land Beobachterpanel

Kanonsiche Panelbezeichnung:

Land Beobachterpanel

und:

Language Fairness Panel

Kanonsiche Panelbezeichnung:

Language Fairness Panel

Das Länderbeobachter-Panel beantwortet die folgende Frage:

> *In Ländern, die im Hauptbevölkerungs-Panel unterrepräsentiert sind, gibt es mindestens echte Nutzerbeobachtungen, Zugangsüberprüfungen oder frühe Risikosignale?*

Das Sprachgerechtigkeits-Panel beantwortet hingegen die folgende Frage:

> *Selbst wenn der globale Durchschnitt hoch erscheint, werden Nutzer in bestimmten Sprachen fehlerhafter, unvollständiger oder riskanter vertreten?*

Diese beiden Panels ersetzen nicht das Haupt-Panel zur Bevölkerung. Sie ergänzen es. Einer Beobachtung eines Landes durch einen Länderbeobachter eine Beobachtung zu geben, misst nicht die Punktzahl dieses Landes. Eine Sprache übermäßig zu stichprobenartig zu erfassen, macht diese Sprache nicht in der Weltbevölkerung größer. Die Aufgabe dieses Abschnitts ist:

- Sichtbarkeit des Landes versus Bevölkerungsgewicht,
- den globalen Durchschnitt von Sprachgerechtigkeit zu trennen,
- Bevölkerungsschätzung von Frühwarnung zu trennen,
- statistische Hinreichendheit vom Beobachtungsumfang zu trennen

voneinander. Dieser Abschnitt hat noch nicht:

- definiert, wie Aufforderungen in allen Sprachen übersetzt werden sollten,
- den Rückübersetzungs- und semantischen Äquivalenztest,
- das detaillierte Auswahlverfahren der Teilnehmenden,
- wie atomare Behauptungen von Schiedsrichtern bewertet werden,
- die endgültigen Konformitätsschwellen

genau. Diese Fragen werden in den folgenden Abschnitten weiter geregelt. Die grundlegende Frage des Abschnitts 7 lautet:

> *Wie messen wir, während wir eine globale Punktzahl erzeugen, ohne kleine Länder und Sprachen künstlich zu überbewerten, aber auch ohne sie völlig unsichtbar zu lassen?*

NOMOS Herausforderung

Sie richten ein 1,000-Personen-Populationspanel für ein KI-Produkt ein. Als Ergebnis einer populationsproportionalen Stichprobe:

- hunderte von Nutzern aus einigen wenigen großen Ländern,
- dutzende von Nutzern aus einigen mittelgroßen Ländern,
- keine Nutzer aus den meisten kleinen Ländern

kamen. Globale Punktzahl:

94

Es wurde so berechnet. Dann sagten Sie: 'Dieses KI-Produkt repräsentiert die Marke weltweit mit 94 Prozent Genauigkeit.' Es gibt jedoch nicht eine einzige echte Nutzerbeobachtung in 70 kleinen Ländern. Diese Länder mögen einen kleinen Teil der Weltbevölkerung ausmachen. Dennoch bleiben die folgenden Fragen unbeantwortet: Ist das Produkt wirklich zugänglich? Funktioniert die Absicht in der richtigen Sprache? Verwechselt das System das Land mit einem anderen Land? Wird das lokale Unternehmen mit der globalen Marke verwechselt? Sind Preis, Service und rechtlicher Kontext falsch? Versagt das System bei der Antwort? Gibt es kritische Fehler mit geringer Häufigkeit? Der globale Bevölkerungsscore beantwortet diese Fragen nicht. Stellen Sie sich nun vor, dass Sie eine Person zu jedem Land hinzugefügt haben. Sie haben 70 zusätzliche Beobachtungen erhalten.

Sie haben jedes dieser Länder mit einem Gewicht von 1/1,070 zur Hauptwertung hinzugefügt, wodurch kleine Länder viele Male größer dargestellt wurden als ihr tatsächlicher Bevölkerungsanteil. Die geografische Sichtbarkeit nahm zu, aber die Bevölkerungsschätzung wurde verzerrt. Dann erhielt ein Nutzer in einem Land eine falsche Antwort, und Sie erklärten: ‚Die GEO-Wertung dieses Landes ist null.' Die Antwort einer Person ist keine Länderwertung. Es ist ein einzelnes Ereignis, das unter bestimmten Bedingungen in diesem Land beobachtet wurde. Betrachten wir nun die Sprache. Das englische Ergebnis beträgt 98 Prozent, das deutsche 96 Prozent und das türkische 91 Prozent. Eine Sprache mit geringen Ressourcen hat nur acht Nutzer und ein Roh-Ergebnis von 50 Prozent. Sie sagen: ‚Acht Personen sind nicht genug; entfernen Sie die Sprache aus dem Bericht.' Die globale Wertung steigt, aber die Erfahrung der Nutzer in dieser Sprache verschwindet.

In einem anderen Ansatz geben Sie einer weiteren 200 Nutzer einer ressourcenarmen Sprache. Der rohe Stichprobenmittelwert sinkt. Dann sagen Sie: "Der globale Wert fiel auf 84." Die tatsächliche Bevölkerungsanteil der Sprache kann jedoch 2 Prozent betragen. Oversampling erhöht die statistische Aussagekraft. Es erhöht nicht das Bevölkerungsgewicht. Die erste Regel dieses Abschnitts lautet wie folgt:

> *Beobachtet zu werden ist nicht dasselbe wie Gewicht zu haben.*

Ihre zweite Bestimmung lautet:

> *Eine einzige Länderbeobachtung ist kein Länderwert.*

Seine dritte Bestimmung besagt:

> *Sprachliche Fairness bedeutet nicht, allen Sprachen im globalen Wert gleiches Gewicht zu geben.*

Seine vierte Bestimmung besagt:

> *Ein hoher globaler Durchschnitt beseitigt keine schwere Verschlechterung in einer bestimmten Sprache.*

Seine fünfte Vorschrift lautet:

> *Zusätzliche Beobachtungen, die zur Erhöhung der Abdeckung genommen wurden, können ohne Neugewichtung nach der tatsächlichen Bevölkerung nicht zum Hauptwert hinzugefügt werden.*

1. ZWECK DES KAPITELS

Zweck dieses Abschnitts ist es, ergänzende Panels zu erstellen, die die Sichtbarkeit von Ländern und Sprachen erhöhen, ohne die statistische Bevölkerungsrepräsentation des Haupt-Populationspanels zu beeinträchtigen. Der Abschnitt normalisiert die folgenden Unterscheidungen:

- Geografische Sichtbarkeit mit Bevölkerungsgewicht
- Länderbeobachtung mit Länderscore
- Einzelereigniserkennung mit Prävalenzschätzung
- Haupt-Populationspanel mit Länderbeobachterpanel
- Länderbeobachterpanel mit länderspezifischer Validierungsstudie
- Länderdeckung mit Bevölkerungsdeckung
- Abdeckung der Anzahl der Sprachen mit Sprachbevölkerungsdeckung
- Sprachliche Fairness mit gleicher globaler Gewichtung der Sprachen
- Globaler Sprachanteil mit übermäßiger Stichprobe
- Muttersprache und Prompt-Sprache
- Sprache und Gebietsschema

- Sprache und Schriftsystem
- Unterstützte Sprache und Sprache, die allein Antworten erzeugen kann
- Fehler bei der Prompt-Übersetzung und Sprachfehler des KI-Systems
- Ressourcenmangel und KI-Repräsentationsdefekt
- Unzureichende Schiedsrichter und Produktfehler
- Legitimer lokaler Unterschied und ungerechte Sprachungleichheit
- Globaler Durchschnitt und Untergruppenbasis
- Länderverfügbarkeitsabdeckung und Genauigkeit der Länderrepräsentation
- Beobachteralarm und endgültige Konformitätsentscheidung
- Einzelwellen-Überwachungszyklus mit mehrwellengeografischer Abdeckung

Am Ende dieses Abschnitts sollte jedes GEO-1000-Audit in der Lage sein, klare Antworten auf die folgenden Fragen zu geben:

> *In welchen Ländern gibt es mindestens eine tatsächliche Beobachtung?*
>
> *In welchen Ländern gibt es nur Frühwarnbeobachtungen und in welchen gibt es statistische Vorhersagen?*
>
> *Welche Sprachen wurden gemessen, welche erreichten eine ausreichende Stichprobe und welche wurden nicht getestet?*
>
> *Wenn Sprachen mit geringen Ressourcen übermäßig beprobt wurden, wie wurde dies im globalen Score neu gewichtet?*
>
> *Liegt die Ursache eines schlechten Ergebnisses in einer Sprache wirklich an KI-generierten Inhalten, oder ist sie auf Probleme mit dem Prompt, den Ressourcen, dem Standort oder der Bewertung zurückzuführen?*

2. ZENTRALE NORMATIVE BESTIMMUNG

> *GEO-1000 sollte die Hauptbevölkerungsschätzung mit dem Bevölkerungs-Panel verwalten; die Beobachtung kleiner Länder mit dem Länder-Beobachter-Panel; Sprachungleichheit und die Identifizierung von Sprachen mit geringen Ressourcen separat mit dem Sprachgerechtigkeits-Panel.*

Diese drei Strukturen:

- unterscheiden sich in den Zwecken,
- in Stichprobenzielen unterscheiden,
- unterscheiden sich in den Gewichten,
- in öffentlichen Aussagen unterscheiden

haben. Eine Beobachtung aus einem Panel kann ohne Erklärung nicht in die Bedeutung eines anderen Panels umgewandelt werden. Die Standardstruktur:

Panel	Hauptzweck	Ergebnisse, die es liefern kann	Ergebnisse, die es allein nicht liefern kann
Bevölkerungs-Panel	Bevölkerungsgewichtete globale Schätzung	Globaler und gestufter Bevölkerungsscore	Zuverlässiger Länderscore für jedes kleine Land
Land Beobachterpanel	Geografische Beobachtung und Frühwarnung	Beobachtungsabdeckung des Landes, einzelnes Ereignis, Zugriffssignal	Bevölkerungsverhältnis des Landes
Language Fairness Panel	Sprachbasierte Diagnose und Ungleichheitsmessung	Sprachenscores, Sprachuntergrenze, Gerechtigkeitslücke	Automatische gleichgewichtete Berücksichtigung von Sprachen in der Weltbevölkerung

3. WAS IST DAS Country Observer PANEL?

Country Observer Panel in Ländern, die in der Hauptbevölkerungsumfrage unterrepräsentiert oder nicht ausreichend vertreten sind:

- Produktzugang,
- Anwendbarkeit der Aufforderung
- Grundlegende Entitätsanalyse,
- Sprach- und Gebietsschema-Verhalten,
- Frühes kritisches Fehlersignal

ist ein ergänzendes Prüfpanel, das eingerichtet wurde, um zu beobachten. Die Hauptaufgabe des Länderbeobachterpanels besteht nicht darin, eine Punktzahl zu erzeugen.

> *Es soll das unsichtbare Land zum ersten Mal sichtbar machen.*

Eine Beobachtung durch einen Länderbeobachter kann zeigen: „In dem angegebenen Land, unter dem angegebenen KI-Produkt, der Nutzeroberfläche, Sprache und Zeitbedingung wurde dieses Ergebnis mindestens einmal beobachtet.“ Sie kann nicht zeigen: „X% der Nutzer in diesem Land sehen dasselbe Ergebnis."

4. Länderbeobachter-EINHEIT

Die grundlegende Beobachtungseinheit des Länderbeobachterpanels ist nicht nur das Land. Die bevorzugte Einheit:

```
S_{c,a,l,s,w}
```

wobei:

- c: Land oder Rechtsordnung
- a: KI-Produkt und Nutzeroberfläche
- l: Sprache, Schriftsystem und Gebietsschema
- s: Sitzungs- oder Planbedingung
- w: Messwelle

Diese Unterscheidung ist notwendig. Im selben Land:

- mehr als eine Sprache,
- verschiedene Pläne,
- Web- und Mobiloberflächen,
- unterschiedliche rechtliche oder technische Zugangsbedingungen

können existieren. Das Hinzufügen eines Nutzers zu einem Land bedeutet nicht, dass alle Sprach- und Nutzeroberflächen des Landes überwacht werden.

5. AUFGABEN DES Land-Beobachter-PANELS

5.1. Zugangsprüfung

Obwohl das Produkt im Land in der offiziellen Registrierung verfügbar zu sein scheint, kann der tatsächliche Nutzer:

- nicht in der Lage sein, sich anzumelden,

- nicht in der Lage sein, die angegebene Oberfläche zu erreichen,

auf ein Planungs- oder Verifikationshindernis stoßen. Die Beobachtung durch den Länderbeobachter testet das offizielle Zugriffsprotokoll auf Feldebene.

5.2. Grundlegende Identitätsprüfung

KI-Produktüberwachungsstelle:

- das richtige Unternehmen,
- die richtige Marke,
- korrekte Länderbeziehung

Kann es intern analysieren?

5.3. Erkennung von Standortabweichungen

In verschiedenen Länder-Lokalisierungen in derselben Sprache:

- Preis,
- Unternehmen,
- Dienstleistung,
- Gesetz,
- Ratschläge

Verändert es sich? Wenn es sich verändert, kann dieser Unterschied finanziell sein und erklärbar?

5.4. Niedrigfrequentes schweres Fehlersignal

Bei der Reaktion eines einzelnen Beobachters:

- falsche Lizenz,
- falsche juristische Einheit,
- gelbe gefährliche Gesundheits- oder Sicherheitsinformationen,
- starke Identitätsfusion

Es kann beobachtet werden. Ein einzelnes Ereignis zeigt nicht die Verbreitung im Land an. Es kann jedoch eine Überprüfungsarbeit auslösen.

5.5. Neue Forschungshypothese

Das Beobachtergremium könnte die Frage aufwerfen: 'Handelt es sich bei diesem Ereignis um ein einmaliges Ereignis oder ist es ein wiederkehrendes Muster unter den Nutzern im Land?' Diese Frage wird mit einer gezielten zusätzlichen Stichprobe getestet.

5.6. Geographischer Umfangsbericht

Es zeigt, in wie vielen geeigneten Ländern das Produkt reale Nutzerbeobachtungen hat. Dies ist nicht die Anzahl der Länderscores. Es ist die Beobachtungsabdeckung.

6. DINGE, DIE DAS Country Observer PANEL NICHT TUN KANN

Allein das Country Observer Panel:

- zuverlässiger Länderscore,
- Landesbevölkerungsrate,

- Länderrangliste,
- landesspezifische Konformitätsentscheidung,
- länderbasierte Modellüberlegenheit

kann nicht erstellt werden. Wenn es nur einen Beobachter in einem Land gibt und die Antwort korrekt ist: 'Der Länderscore ist 100' kann nicht gesagt werden. Wenn sie falsch ist: 'Der Länderscore ist null' kann nicht gesagt werden. Richtiger Status:

BEOBACHTET_PASS

Beobachtetes Problem

BEOBACHTETER_KRITISCHER_KANDIDAT

ZUGRIFFSFEHLER

NICHT_AUSREICHEND_FÜR_LÄNDER_SCHÄTZUNG

sind Beobachtungsaufzeichnungen.

7. ZULÄSSIGES_LÄNDERREGISTER

Ein Zulässiges Länderregister sollte für jedes KI-Produkt erstellt werden. Jedes Land kann einen der folgenden Status haben:

CE-0— UNBEKANNT

Der Produkt- und Audit-Zugriffsstatus konnte nicht bestimmt werden.

CE-1— AUSSERHALB Natürliche Reichweite

Das Produkt bleibt außerhalb des offiziellen oder legitimen Einsatzbereichs. Es wird nicht für die Genauigkeitsbewertung der Länderrepräsentation berücksichtigt. Es wird als Zugangsspielraumlücke dargestellt.

CE-2— EINGESCHRÄNKTER ODER SONDERVERZUGRIFF

Nur:

- Institution,
- Einladung,
- Plan,
- Gerät,
- Nutzertyp

Zugriff ist verfügbar. Ein separates Panel kann erforderlich sein.

CE-3— BERECHTIGT FÜR Beobachtung durch Landesbeobachter

Das Produkt ist für die Beobachtung durch den Landesbeobachter in Bezug auf die relevante Nutzeroberfläche und mindestens eine Prüfsprache geeignet.

CE-4— BERECHTIGT FÜR BEVÖLKERUNGSSCHÄTZUNG

Es hat einen produktberechtigten Bevölkerungsanteil im Hauptbevölkerungs-Panel des Landes.

CE-5— BERECHTIGT FÜR LANDESWEITE SCHÄTZUNG

Eine Landesbewertung kann unter ausreichender Stichprobe, effektiver Stichprobe, Sprachabdeckung und Qualitätsbedingungen erstellt werden. Ein Land kann CE-3 sein, aber nicht CE-5. Dies bedeutet, dass das Land beobachtet wird, aber kein verlässlicher Landeswert erstellt werden konnte.

8. DIE LEITER ZWISCHEN Beobachtung durch Landesbeobachter UND LANDESWEITER SCHÄTZUNG

CS-0— NICHT BEOBACHTET

Es gibt keine gültige Country Observer-Beobachtung für das berechtigte Land.

CS-1— EINZELNE Country Observer-Beobachtung

Es gibt eine gültige Länder–Produkt–Sprach-Beobachtung. Erzeugt nur einen Ereigniseintrag.

CS-2— WIEDERHOLTE Country Observer-Beobachtung

Es gibt mehrere Beobachtungen aus verschiedenen Wellen oder von verschiedenen Nutzern. Kann ein Musterverdacht aufwerfen. Nicht ausreichend für die Länderbewertung.

CS-3— EXPLORATIVE LÄNDERSTICHPROBE

Die explorative Stichprobenbandbreite im Abschnitt 6 wurde erreicht. Ein explorativ Ergebnis kann mit großer Unsicherheit angegeben werden.

CS-4— VORLÄUFIGE LÄNDERSCHÄTZUNG

Eine vorläufige Länderschätzung kann mit ausreichender und ausgewogener Stichprobe erstellt werden.

CS-5— KANDIDAT FÜR LÄNDERSCHÄTZUNG

Ausreichende aktive Stichprobe, Abdeckung der Sprache/Region, wiederholte Wellen und Qualitätskontrollen sind vorhanden. Es kann ein Kandidat für die Länderbewertung sein. Exakte Schwellenwerte sollten mit einem Pilotversuch und einer externen Überprüfung kalibriert werden.

9. BEOBACHTERZYKLUS

Es ist möglicherweise nicht möglich, jede Welle in allen berechtigten Ländern hinsichtlich Kosten und Betrieb zu beobachten. Daher kann der Beobachterzyklus verwendet werden. Messwellen:

```
w=1,2,…,K
```

und berechtigte Länder: sei c∈CS. Zuweisung eines Beobachters für Land c in Welle w:

```
z_{c,w}= 1(wenn das Land in der Welle beobachtet wird); andernfalls z_{c,w}= 0
```

kann wie folgt repräsentiert werden. Grundbedingung für den Zyklus:

```
\sum_{w=1}^K z_{c,w}\geq 1
```

muss erfüllt sein. Das heißt, jedes berechtigte Land sollte mindestens einmal innerhalb des definierten Überwachungszyklus beobachtet werden. Diese Bestimmung:

- bedeutet nicht, dass jedes Land in jeder Welle beobachtet wird,
- oder dass eine einzelne Beobachtung die Punktzahl des Landes darstellt

Keine der Bedingungen folgt aus der Zyklusanforderung.

10. BEOBACHTER-FREQUENZKLASSEN

Länder können in vordefinierte Frequenzklassen eingeteilt werden.

SF-0— AUSSERHALB DER REICHWEITE

Kein Produktzugang. Es wird keine nutzer-repräsentative Beobachtung durchgeführt. Der Zugangsstatus wird überwacht.

SF-1— EINMAL PRO ÜBERWACHUNGSZYKLUS

Länder mit geringem Risiko, kleine Länder oder Länder, die bereits ausreichende indirekte Abdeckung im Population Panel haben.

SF-2— PERIODISCHER BEOBACHTER

Wird in regelmäßigen Abständen erneut beobachtet.

SF-3— JEDER HAUPTWELLE

Kann aus folgenden Gründen in jeder Hauptwelle beobachtet werden:

- hohe Nutzerzahl,
- materieller Zielmarkt,
- frühere schwerwiegende Fehler,
- schneller Zugang oder regulatorische Änderungen,

Hochrisikodienstleistungsbereich.

SF-4— VORFALLSÜBERWACHUNG

Intensivierte Beobachtung aufgrund eines verifizierten kritischen Ereignisses oder kontinuierlicher Repräsentationsstörung. Die Frequenzklasse kann nach Überprüfung der Ergebnisse nicht rückwirkend geändert werden. Ein neues Ereignis kann eine neue Version für die nächste Welle erzeugen.

11. BEOBACHTERPRIORITÄT

Die Häufigkeit der Aktivität eines Landesbeobachters sollte nicht ausschließlich anhand der Bevölkerung bestimmt werden. Gegebenenfalls können die folgenden Bereiche berücksichtigt werden:

- Produktberechtigte Bevölkerung
- Die tatsächliche Service- oder Marktabdeckung der Einheit
- Frühere fehlerhafte Repräsentationsaufzeichnungen
- Nutzerauswirkungen mit hohem Risiko
- Änderungen im Produktzugang
- Einzigartigkeit von Sprache und Gebietsschema
- Verwendung von Sprachen mit geringen Ressourcen
- Rechtliche oder operative Unterschiede
- Verstrichene Zeit seit der letzten Beobachtung

Diese Felder müssen gemäß der vordefinierten Governance-Regel verwendet werden. Der Ansatz „Das Ergebnis dieses Landes ist schlecht, wir messen weniger.“ ist verboten.

12. Länderbeobachter-ABDECKUNGSMETRIKEN

Die Länderabdeckung ist keine einzelne Zahl. Es müssen mindestens drei separate Metriken verwendet werden.

12.1. Länderbeobachtungsabdeckung

Der Anteil der berechtigten Länder mit mindestens einem gültigen Beobachter oder einer Bevölkerungsbeobachtung.

```
COC_count= |{c ∈ C_S: n_c^obs > 0}| / |C_S|
```

Diese Metrik zeigt die Abdeckung in Bezug auf die Anzahl der Länder.

12.2. Bevölkerungsgewichtete Länderbeobachtungsabdeckung

Der Anteil der berechtigten Bevölkerung in Ländern mit mindestens einer gültigen Beobachtung an der gesamten Zielbevölkerung.

```
COC_pop= [Σ_{c:n_c^obs>0} N_c] / [Σ_{c∈C_S} N_c]
```

Diese Kennzahl zeigt, ob große Bevölkerungen beobachtet werden.

12.3. Länderprognoseabdeckung

Es ist der Anteil der Länder, die eine ausreichende Länderstichprobe erreichen.

```
CEC= |{c: n_{c,eff}≥ m_c}| / |C_S|
```

Hierbei ist mc der vordefinierte Angemessenheitsschwellenwert für die Länderprognose. Eine Prüfung:

- 100 Prozent Länderbeobachtungsabdeckung,
- 99 Prozent Bevölkerungsbeobachtungsabdeckung,
- kann nur 10 Prozent Länderprognoseabdeckung haben

Diese Ergebnisse können nicht ineinander umgerechnet werden.

13. BEOBACHTERALARM

Wenn in der Beobachtung eines Beobachters ein materielles Problem gefunden wird, wird ein Beobachter-Alarm erstellt. Arten von Alarmen:

SA-1— ZUGRIFFSANOMALIE

Der offizielle Zugriffsdatensatz steht im Widerspruch zum tatsächlichen Nutzerzugriff.

SA-2— ENTITÄTSANOMALIE

Die Entität fusioniert mit einem anderen Unternehmen, Produkt oder Ländereinsatz.

SA-3— SPRACH- ODER LOKALISATIONSANOMALIE

Falsche Sprache, falscher Landeskontext oder offensichtliche Abweichung der Lokalisierung tritt auf.

SA-4— MATERIELLE FAKTENANOMALIE

Der materielle Preis, der Service, die Kapazität, die Partnerschaft oder die Zeit ist falsch.

SA-5— KANDIDAT FÜR KRITISCHES EREIGNIS

Gesundheit, Recht, Finanzen, Sicherheit, Lizenzierung oder schwerwiegender Identitätsfehler liegt vor.

SA-6— REFERENZ- ODER PROMPTANOMALIE

Die Quelle des Problems ist nicht die KI-Antwort:

- Prompt,
- Übersetzung,
- offizielle Quelle,
- Beurteilung

könnte schuld sein. Der Alarm ist kein automatischer Länderfehler. Es ist der Beginn der Überprüfung und Eskalation.

14. BEOBACHTER-ESKALATIONSPROTOKOLL

Schritt 1— Gültigkeitsüberprüfung

Entspricht die Beobachtung wirklich dem Protokoll? Richtiges Land Richtiges Produkt Richtiger Prompt Richtige Sprache Erste geeignete Ausgabe Vollständige Evidenzkontrolle.

Schritt 2— Neu-Codierung des Vorfalls

Problem:

- KI-Produkt,
- Referenzdatensatz,
- Prompt,
- Teilnehmer,
- Zugang

Zu welcher der Quellen könnte es gehören?

Schritt 3— Unabhängige Wiederholung

Vor der Auswahl des Ergebnisses wird die Überprüfung mit vordefinierten neuen Nutzern oder Sitzungen durchgeführt.

Schritt 4— Gezielte Stichprobennahme

Wenn das Ereignis anhält, kann eine länderspezifische Stichprobennahme angewendet werden. Stichprobengröße:

- Ereignisschwere,
- erwartete Prävalenz,
- Bevölkerung des Landes,
- Sprachvielfalt

sind im Voraus festgelegt.

Schritt 5— Entscheidungsabgrenzung

Die folgenden Ergebnisse werden getrennt gehalten:

- Ein einzelnes verifiziertes Ereignis
- Wiederholtes Muster
- Landesrate
- Kritischer Eignungseffekt

Ein einzelnes kritisches Ereignis kann die Konformitätsprüfung beeinflussen, auch ohne eine Schätzung der Prävalenz.

15. WAS IST SPRACHGERECHTIGKEIT?

Sprachgerechtigkeit:

> *Es geht nicht darum, dass alle Sprachen genau dieselbe wortwörtliche Antwort erhalten.*

Es bedeutet:

> *Bei semantisch gleichwertigen und materiell vergleichbaren Nutzerfragen sollten Nutzer, abgesehen von tatsächlichen lokalen Unterschieden, nicht systematisch Repräsentationen erhalten, die aufgrund ihrer Sprache falscher, unvollständiger, weniger belegt oder riskanter sind.*

Sprachfairness strebt diese Gleichheit nicht an:

```
R_Turkish= R_English= R_Japanese
```

Wort für Wort. Es sucht nach Folgendem:

> *Vergleichbare Repräsentation in Bezug auf Identität, materielle Realität, Grenze, Zeit, Quellenstatus und Nutzergeeignetheit.*

16. SPRACHUNTERSCHIED IST NICHT IMMER UNFAIRNESS

Sprachliche Ergebnisse können aus folgenden Gründen legitim variieren: Der Umfang lokaler Dienstleistungen ist unterschiedlich. Preise und Vertragsbedingungen variieren je nach Land. Rechtliches Wissen hängt von der Gerichtsbarkeit ab. Die Absicht des Nutzers ist unterschiedlich. Eingaben sind nicht semantisch identisch. Offizielle Quellen enthalten in verschiedenen Sprachen unterschiedliche Inhalte. Das KI-Produkt unterstützt die betreffende Sprache nicht offiziell. Jede dieser Unterschiede sollte dokumentiert werden. Erklärbare lokale Unterschiede sollten nicht als Ungerechtigkeit bezeichnet werden. Auch wenn die Realität dieselbe ist, gilt in einer Sprache mit begrenzten Ressourcen:

- Die Unternehmensidentität ist beschädigt,
- Grenzen werden entfernt,
- die falsche Lizenz wurde hinzugefügt,
- Die Antwort wird ständig in einer anderen Sprache gegeben

Es könnte ein sprachbasiertes Repräsentationsproblem sein.

17. SPRACHEINHEIT

Bei der Sprachbewertung kann allein der allgemeine Sprachname nicht ausreichen. Die bevorzugte Spracheinheit:

```
L=(λ,σ,ℓ,d)
```

und kann betrachtet werden als. Hier:

- λ: Sprache
- σ: Schriftsystem
- ℓ: Gebietsschema oder Länderkontext
- d: Aufgabe oder Domäne

Zum Beispiel dieselbe Sprache:

- in einem anderen Schriftsystem,
- in einem anderen Land,
- im Rechts- oder Gesundheitswesen

kann eine separate Messeinheit erfordern. Nicht alle Nutzer einer Sprache bilden eine einzige homogene Gruppe.

18. SPRACHSUPPORT-STATUSES

Jedes KI-Produkt kann einen der folgenden Status für Sprache und Nutzeroberfläche verwenden:

LS-0— UNBEKANNT

Support-Status konnte nicht ermittelt werden.

LS-1— NICHT UNTERSTÜTZT

Das Produkt unterstützt die betreffende Sprache offiziell nicht. Es wird nicht innerhalb der Hauptbewertung für unterstützte Sprachen bewertet. Experimentelle Tests können durchgeführt werden.

LS-2— BEOBACHTET, ABER NICHT UNTERSTÜTZT

Das Produkt kann Antworten generieren. Es gibt jedoch keinen offiziellen Support oder Qualitätsgarantie. Als separates experimentelles Ergebnis gemeldet.

LS-3— BEGRENZTE UNTERSTÜTZUNG

Unterstützung existiert für bestimmte Aufgaben, Schnittstellen oder Funktionen.

LS-4— OFFIZIELL UNTERSTÜTZT

Die Produktsprache gilt im Allgemeinen als unterstützt.

LS-5— PRÜFBERECHTIGT

Sprache:

- offizieller Support,
- Prompt-Äquivalenz,
- lokaler Prüfer,

- Referenzkompetenz,
- gültige Erfassung

ist vollständig prüfungsfähig. Die bloße Präsenz von offiziellem Support stellt nicht LS-5 dar.

19. SPRACHE MIT GERINGEN RESSOURCEN

Der Begriff „Sprache mit geringen Ressourcen“ beschreibt nicht den Wert der Sprache oder die Qualität ihrer Sprecher. In bestimmten Technologie- und Audit-Kontexten weist er auf einen oder mehrere der folgenden Mängel hin:

- Begrenzter digitaler Inhalt
- Begrenzte lokalisierte offizielle Quelle
- Begrenzter Evaluationsdatensatz
- Eingeschränkter Zugang zu Gutachtern
- Geringer Produktsupport
- Begrenzte Terminologieressourcen
- Begrenzte Qualität maschineller Übersetzung
- Geringer Zugang zu Nutzerpanels

Eine Sprache kann in einem Bereich ressourcenarm und in einem anderen stark sein. Zum Beispiel, während sie über starke Ressourcen im alltäglichen Gespräch verfügt:

- Gesetz,
- Medizin,
- spezifisches technisches Feld

kann über begrenzte Ressourcen verfügen. Daher Sprachressourcenstufe:

- Produkt,
- Aufgabe,
- Zeit,
- Feld

sollte gemeinsam definiert werden.

20. AUFGABEN DES LANGUAGE FAIRNESS PANEL

20.1. Übersampling von Sprachen mit wenigen Ressourcen

Stellt zusätzliche Nutzer für Sprachen bereit, die in der Hauptpopulation nicht genügend Beobachtungen erhalten.

20.2. Messung der sprachenbasierten Materialfehlerverteilung

Nicht nur die Gesamtrealgenauigkeit:

- Identität,
- Umfang,
- Zeit,
- Quelle,
- Empfehlung,

- Kritischer Fehler

untersucht die Ergebnisse separat.

20.3. Überprüfung der Antwortsprache

Antwortet die KI in Übereinstimmung mit der Promptsprache? Wenn sie zu einer anderen Sprache wechselt, was ist der Grund?

20.4. Untersuchung der Unterscheidung von Sprache und Gebietsschema

Trägt dieselbe Sprache in verschiedenen Ländern unterschiedliche lokale Realitäten? Verwaltet KI diesen Unterschied korrekt?

20.5. Untersuchung der Übereinstimmung von Quelle und offizieller Repräsentation

Tragen die offiziellen Existenzaufzeichnungen in allen Sprachen die gleiche materielle Realität? Ein niedriges KI-Ergebnis kann aus einer Sprachlücke in der Quelle selbst entstehen.

20.6. Testen der Gleichwertigkeit von Prompts

Tragen Übersetzungen wirklich dieselbe Nutzerabsicht?

20.7. Testen der Parität von Schiedsrichtern

Können Schiedsrichter in verschiedenen Sprachen denselben normativen Standard anwenden?

21. VIER TORE FÜR SPRACHLICHE FAIRNESS

Ein sprachliches Ergebnis sollte durch vier Tore bewertet werden, bevor es der Sprachleistung des KI-Produkts zugeschrieben wird.

Tor 1— Gleichwertigkeit der Absicht

Sind die Aufforderungen semantisch vergleichbar? Wenn dies fehlschlägt:

AUFFORDERUNGS_EQUIVALENZ_FEHLGESCHLAGEN

Status wird angegeben.

Tor 2— Referenzparität

Sind offizielle und verifizierte Realitätsaufzeichnungen für die entsprechende Sprache ausreichend? Wenn es fehlschlägt:

REFERENZ_SPRACHEN_LÜCKE

Status wird angegeben.

Tor 3— Messparität

Sind Aufnahme-, Sitzungs-, Schnittstellen- und Nutzerbedingungen vergleichbar? Wenn dies fehlschlägt:

MESSPARITÄT_FEHLGESCHLAGEN

Status wird angegeben.

Tor 4— Bewertungsparität

Gibt es Bewerter in der Muttersprache oder mit ausreichender lokaler Expertise und denselben Bewertungsregeln? Wenn dies fehlschlägt:

BEWERTUNGS_SPRACHEN_LÜCKE

wird zugewiesen. Es sei denn, die vier Türen werden ausreichend erfüllt: „Das KI-Produkt ist in dieser Sprache schlechter.“ kann nicht endgültig festgestellt werden. Die Sprachlücke ist weiterhin ein Befund des GEO-Ökosystems. Ihre Quelle muss jedoch korrekt klassifiziert werden.

22. EINRICHTUNG DES SPRACHUNIVERSUMS

Der Ausdruck „alle Sprachen“ kann nicht uneingeschränkt verwendet werden. Jede Prüfung muss ein Sprachregister innerhalb des Umfangs erstellen. Sprachen können aus den folgenden Gruppen ausgewählt werden:

22.1. Globales Kernsprachen-Set

Muttersprachen, die von einer großen menschlichen Population genutzt werden und für das Produkt geeignet sind.

22.2. Entity-Marktsprachset

Die Sprachen, die von der geprüften Einheit gezielt angesprochen werden:

- Dienstleistung,
- in Bezug auf Inhalte,
- Kunde,
- Vertrag

und Wartung.

22.3. Produktsupport-Sprachset

Sprachen, die offiziell vom KI-Produkt unterstützt werden.

22.4. Sprachenset für Beobachter mit geringen Ressourcen

Sprachen mit geringen Ressourcen, die in der Bevölkerungsstichprobe schwach vertreten sind oder ein Risiko hinsichtlich des Produktverhaltens darstellen.

22.5. Hochrisiko-Sprachenset

Sprachen, die einen wesentlichen Einfluss auf den Nutzer im Kontext von Gesundheit, Recht, Finanzen oder Sicherheit haben.

22.6. Experimentelles Sprachenset

Sprachen, die vom Produkt offiziell nicht unterstützt werden, aber zu Beobachtungs- oder Forschungszwecken getestet werden. Diese Gruppen können sich überschneiden. Der Grund, warum jede Sprache aufgenommen wird, sollte dokumentiert werden.

23. ZUTEILUNG DER SPRACHLICHEN FAIRNESS

Für Sprache l sei das Ziel im Hauptbevölkerungsbereich nlP. Das zusätzliche Ziel im Fairness-Bereich ist nlF. Die insgesamt gesammelte Sprachbeobachtung:

```
n_l^T= n_l^P+ n_l^F
```

erhalten. Das Gesamtgewicht der Sprache in der Berechnung der Weltbevölkerung wird als Wl beibehalten. Das normalisierte Gewicht pro Beobachtung:

```
w_{i,l}= W_l/ n_l^T
```

kann auf diese Weise betrachtet werden. Somit wird die Sprache überstichprobenartig erfasst. Sie übersteigt jedoch nicht den tatsächlichen Bevölkerungsanteil im globalen Ergebnis.

24. BEITRAG DES Language Fairness Panels ZUM GLOBALEN ERGEBNIS

Beobachtungen zur Sprachfairness können auf zwei Arten verwendet werden.

24.1. Separate Diagnostische Analyse

Das Fairness-Gremium erzeugt nur Sprachpunktzahlen und Fehlertypen. Das Haup-Populationsgremium ändert die globale Punktzahl nicht. Diese Methode ist einfacher.

24.2. Kombinierte neu gewichtete Analyse

Beobachtungen zur Bevölkerungs- und Sprachfairness werden zusammen verwendet. Jede Beobachtung wird entsprechend dem tatsächlichen Gewicht der Zielsprachpopulation neu gewichtet. Diese Methode kann das Informationsvolumen erhöhen. Allerdings:

- Doppelte Zählung
- mehrsprachiger Nutzer,
- Panelabhängigkeit,
- Gewichtsvariabilität

Es sollte berücksichtigt werden. Die verwendete Methode muss definiert werden, bevor die Ergebnisse gesehen werden.

25. SPRACHENDECKUNGSMETRIKEN

25.1. Sprachbeobachtungsabdeckung

Der Anteil der Sprachen im Geltungsbereich, für die mindestens eine gültige Beobachtung vorliegt.

```
LOC= |{l: n_l^obs > 0}| / |L_scope|
```

25.2. Sprachprognoseabdeckung

Der Anteil der Sprachen, die ausreichende Stichproben- und Qualitätsanforderungen erfüllen.

```
LEC= |{l: n_{l,eff}≥ m_l}| / |L_scope|
```

25.3. Bevölkerungsgewichtete Sprachabdeckung

Der Anteil der Zielbevölkerungssprache, der ausreichend beobachtet wird. Erfordert eine primäre Prüfsprache oder eine gewichtete Aufteilung pro Person, um Doppelzählungen mehrsprachiger Nutzer zu vermeiden.

25.4. Produktunterstützungsabdeckung

Gibt an, in wie vielen der abgedeckten Sprachen das Produkt offizielle Unterstützung erhält. Sprachunterstützung und Genauigkeit der Repräsentation werden getrennt gehalten.

26. SPRACHUNGLEICHHEITSMESSWERTE

KANDIDATEN-MESSWERTE — VORGESCHLAGENE MESSWERTE

Das genaue Messwertset sollte durch Pilotprojekte und externe Überprüfungen kalibriert werden.

26.1. Sprachergebnis

Für jede ausreichende Sprachzelle wird θl berechnet.

26.2. Unterschied zwischen Höchst- und Niedrigstwert

```
DG_max= max_l θ_l− min_l θ_l
```

Einfach zu verstehen. Jedoch empfindlich gegenüber kleinen und volatilen Einzelzellen.

26.3. Robuster Perzentildifferenzwert

```
DG_{90−10}= Q_{0.90}(θ_l)− Q_{0.10}(θ_l)
```

Er ist weniger empfindlich gegenüber extremen Zellen.

26.4. Sprachuntergrenze

Es ist das niedrigste Sprachergebnis mit einer ausreichenden Stichprobe.

```
LF_min= min_{l∈L_est} θ_l
```

Dieser Wert sollte zusammen mit der Unsicherheit bei kleinen Zellen angegeben werden.

26.5. Robuster Sprachboden

Es ist das niedrigere Perzentil oder der Durchschnitt der Untergruppe ausreichender Sprachen. Beispiel:

```
LF_10= Q_{0.10}(θ_l)
```

26.6. Kritische Sprachfehlerquote

Für jede Sprache:

```
CR_l= gültige Antwort mit verifiziertem Kritischen Fehler / gültiger Sprachantwort
```

Es wird separat vom globalen Durchschnitt veröffentlicht.

26.7. Rate unbekannter Sprachen

```
U_l= UNBEKANNT oder unlösbare Antwort / relevante Sprachbeobachtung
```

Eine hohe Rate unbekannter Sprachen kann ebenso wichtig sein wie eine niedrige Punktzahl.

27. Gleichwertiger Sprachwert und bevölkerungsgewichteter Wert

Diagnostische Ergebnisse, die alle Sprachen gleichermaßen berücksichtigen, können erstellt werden:

```
G_equal-language= (1/L)Σ_{l=1}^L θ_l
```

Dieses Ergebnis nähert sich der Frage an: „Durchschnittliche getestete Sprachexpertise.“ Es stellt nicht die durchschnittliche menschliche Erfahrung in der Weltbevölkerung dar.
Bevölkerungsgewichtetes Sprachergebnis:

```
G_population-language= Σ_l W_l θ_l
```

Es bewahrt die Bevölkerungsverteilung. Die beiden Werte sind unterschiedlich. Sie können nicht austauschbar verwendet werden. Beide können zusammen auf der Sprachfairness-Karte angezeigt werden.

28. SPRACHLICHE GERECHTIGKEIT IST KEINE VERPFLICHTUNG ZU GLEICHEN ERGEBNISSEN

Die gleiche Punktzahl in zwei Sprachen zu haben, ist keine automatische Verletzung. Der Unterschied kann aus folgenden Gründen erklärt werden:

- Tatsächlicher lokaler Servicedifferenz
- Länderspezifisches Recht
- Abdeckung der offiziellen Unterstützung des Produkts
- Ressourcenzugang
- Nutzerintention
- Eingabebereich

Wenn jedoch die Ursache des Unterschieds nicht erklärt werden kann oder Sprachen mit geringen Ressourcen systematisch Folgendes erfahren:

- falsche Identität,
- falsche Autorisierung,
- materielle Mängel,
- hoher kritischer Fehler

wenn dies zutrifft, besteht ein Problem der Sprachfairness. NOMOS definiert Sprachfairness wie folgt:

> *Unter vergleichbaren materiellen Bedingungen sollte die Sprache selbst nicht die Ursache für systematische Fehlrepräsentation sein.*

29. SCHNITTSTELLE SPRACHE UND LAND

Der Sprachscore ist nicht unabhängig von Ländern. Dieselbe Sprache:

- unterschiedliche Länder,
- unterschiedlicher Produktzugang,
- unterschiedliche lokale Ressourcen,
- unterschiedliches Recht,

- unterschiedlicher Preis

kann im Kontext von verwendet werden. Daher sollte die Länder–Sprache-Zelle: (c,l) bei Bedarf separat untersucht werden. Beispiel:

- Deutschland × Türkisch
- Türkei × Türkisch
- Vereinigtes Königreich × Türkisch

Es ist dieselbe Sprache, kann aber unterschiedliche lokale und Markt-Kontexte haben. Sprachliche Fairness kann nicht allein anhand der Amtssprache des Landes gemessen werden.

30 th Länderbeobachter UND Language Fairness Panel SCHNITT

Eine Beobachtung gleichzeitig:

- kleine Länderbeobachter-Beobachtung,
- Beobachtung in einer ressourcenarmen Sprache

kann auftreten. In diesem Fall muss der Datensatz die folgenden Felder enthalten:

PRIMARY_PANEL_ROLE

SECONDARY_DIAGNOSTIC_TAGS

Bevölkerungsgewicht Beobachterstatus Sprachgerechtigkeitsstatus Dieselbe Beobachtung kann nicht vollständig in zwei separaten Panels gezählt werden. Es ist eine einzelne Beobachtung. Sie hat mehr als ein diagnostisches Tag.

31. ÄNDERUNG DER TEILNEHMER IM BEOBACHTER- UND SPRACHPANEL

Ein Teilnehmer darf nur aufgrund eines Protokollverstoßes geändert werden. Die folgenden Ergebnisse sind keine Gründe für eine Änderung:

- Falsche KI-Antwort
- Ablehnung
- Antwort in einer anderen Sprache
- Kritischer Fehler
- Entity-Verwechslung
- Negative Empfehlung

Dies sind Messergebnisse. Der Zweck des Beobachter-Panels besteht speziell darin, unerwartete Ereignisse zu identifizieren. Den Nutzer aufgrund eines aufgetretenen unerwarteten Ereignisses zu ändern, widerspricht dem Zweck des Panels.

32. SPRACHLICHE BEURTEILUNG

Die Beurteilung im Sprachgerechtigkeitspanel muss mindestens die folgenden Bedingungen erfüllen:

- In der Lage zu sein, den Prompt und die Antwort auf natürliche Weise oder auf hohem Kompetenzniveau zu verstehen
- Kenntnis der lokalen Terminologie
- In der Lage zu sein, Quellen von Nachweisen zu bewerten
- Verwendung desselben normativen Bewertungsleitfadens
- Blind gegenüber dem Modell und, wenn möglich, den Markeninteressen zu sein
- Offenlegung von Interessenkonflikten
- Weg des dritten Bewerters im Falle von Meinungsverschiedenheiten

Sprache, in der kein Bewerter gefunden werden kann:

BEWERTUNGS_SPRACHEN_LÜCKE

muss den Status haben. Es kann nicht angenommen werden, dass alle Nuancen korrekt sind, wenn sie durch automatische Übersetzung ins Englische übersetzt werden. Maschinelle Übersetzung kann ein hilfreiches Werkzeug sein. Sie ist nicht die endgültige Überprüfung der lokalen Bedeutung.

33. QUELLENPARITÄT

Von der Einheit auf der englischen Seite:

- Umfang,
- Preis,
- Lizenz,
- Grenzen,
- Nachweise

existiert; wenn diese in einer anderen Sprachversion nicht vorhanden sind, kann das niedrige KI-Ergebnis nicht nur ein Modellproblem sein. Drei separate Ergebnisse sind notwendig:

- Entität Sprache Quelle Lücke
- KI-Produkt Sprache Repräsentation Lücke
- Kombinierte Nutzerorientierte Lücke

Das letztendliche Problem aus Sicht des Nutzers ist weiterhin real. Die Verantwortung für die Korrektur kann jedoch variieren:

- Entität,
- KI-Anbieter,

beide.

34. AUSSAGE PARITÄT

Prompts, die semantisch nicht gleichwertig sind, erzeugen unterschiedliche Antworten. Beispiel: In einer Sprache: „Was macht dieses Unternehmen?“ In einer anderen Sprache: „Ist dieses Unternehmen vertrauenswürdig und empfehlenswert?“ Wenn gefragt, sind die Ergebnisse nicht vergleichbar. Die Prompt-Konstitution wird im folgenden Abschnitt detailliert dargestellt. Die Bereitstellung dieses Abschnitts erfolgt wie folgt:

> *Unterschiede in der Leistung von Sprachmodellen können nicht der KI-Sprachleistung zugeschrieben werden, ohne die Gleichwertigkeit der Prompt zu überprüfen.*

35. ABWEICHUNG DER ANTWORTSPRACHE

Der Nutzer gibt eine Prompt auf Türkisch. Die KI erstellt eine Antwort auf Englisch. Dieses Ergebnis:

- selbst wenn die sachlichen Informationen korrekt sind,
- kann die Nutzererfahrung beeinflussen,
- Barrierefreiheit,
- und das Verständnis.

Abweichungen der Antwortsprache müssen separat aufgezeichnet werden:

ÜBEREINSTIMMENDE_SPRACHE

TEILWEISE_CODEWECHSEL

UNGEWÜNSCHTER_SPRACHWECHSEL

UNBENUTZBARE_SPRACHAUSGABE

UNBEKANNT

Code-Switching ist nicht immer ein Fehler. Zum Beispiel kann die Verwendung technischer Begriffe in Englisch normal sein. Es ist relevant, wenn es die Hauptbenutzbarkeit der Antwort stört.

36. SCHRIFTSYSTEM UND TRANSKRIPTION

Die gleiche Sprache kann in verschiedenen Schriftsystemen verwendet werden. Für die:

- Name,
- Produktname einer Einheit,
- rechtliche Identität

kann es bei der Transliteration zu Verwechslungen mit einer anderen Einheit kommen. Das Language Fairness Panel, soweit relevant:

- Schriftsystem,
- Transliteration,
- offizieller lokaler Name,
- vom Nutzer verwendeter Name

muss ihre Felder aufzeichnen. Unterschiede im Schriftsystem sollten nur formal betrachtet werden. Wenn es die Entitätsauflösung beeinflusst, ist es wesentlich.

37. SYNTHETISCHER Länderbeobachter-ZYKLUS

PRÄSENTATION DER SYNTHETISCHEN METHODOLOGIE / Die folgenden Produkte, Länder, Zahlen und Ergebnisse sind fiktiv. Im Register des synthetischen Produkts namens Orion AI Natürliche Reichweite: gibt es 120 berechtigte Länder. Das Haupt-Populationspanel hat in der ersten Welle aufgrund der Bevölkerungsgewichtung nur aus 38 Ländern gültige Beobachtungen erzeugt.

37.1. Umfang der ersten Welle

Anzahl der berechtigten Länder: 120 Länder mit mindestens einer Beobachtung: 38

```
COC_count= 38/120= 31.7%
```

Lassen Sie diese 38 Länder 97% der berechtigten Bevölkerung darstellen.

```
COC_pop= 97%
```

Ergebnis: Die Anzahl der Länder ist in der Abdeckung gering, die Bevölkerungsabdeckung ist hoch.

37.2. Vier-Wellen-Beobachterzyklus

Verteilen Sie die verbleibenden 82 Länder in vier Wellen unter Verwendung eines zuvor eingefrorenen zufälligen und risikobasierten Programms. Jedes Land wird im Verlauf des Zyklus mindestens einmal beobachtet. Am Ende der vierten Welle:

```
COC_count= 120/120= 100%
```

Es wurde jedoch nur in 14 Ländern eine ausreichende Stichprobe für die Länderschätzung erreicht.

CEC = 14/120= 11.7%

Korrekte öffentliche Aufzeichnung: Überdeckung der Länderbeobachtung: 100%/ Überdeckung der Bevölkerungsbeobachtung: nahe bei 100%/ Überdeckung der Länderprognose: 11.7% Falsche öffentliche Aufzeichnung: „Alle 120 Länder haben einen zuverlässigen Länderwert."

37.3. Kleine Länder kritischer Alarm

In einem Country Observer-Datensatz wird die KI-Antwort in dem kleinen Land J fehlidentifiziert, indem die geprüfte Einheit fälschlicherweise als zugelassene Anwaltskanzlei eingestuft wird. In Wirklichkeit erbringt die Einheit keine juristischen Dienstleistungen. Eine einzelne Beobachtung schätzt nicht die Fehlerquote des Landes J; sie stellt ein potenzielles Kritisches Ereignis fest. Protokoll: Beobachtung validieren; Prompt und Referenzdatensatz überprüfen; mit neuen unabhängigen Nutzern wiederholen; wenn das Ereignis weiterhin auftritt, ein gezieltes Länder-Sample öffnen; und die kritische Konformitätsauswirkung separat bewerten. Das beobachtete Ereignis kann selten sein; der potenzielle Schaden muss nicht klein sein.

38. DEMONSTRATION DER SYNTHETISCHEN SPRACHFAIRNESS

SYNTHETISCHES SAMPLE — KEINE ECHTE KI- ODER UNTERNEHMENSLEISTUNG

Fünf Sprachen in den Test der Core Identity von Synthetic Apple.com einbeziehen. Zuteilung des Main Population Panels:

Sprache	Bevölkerungsgewicht	Beobachtung des Hauptpanels
L1	55%	550
L2	25%	250
L3	10%	100
L4	6%	60
L5	4%	40
Gesamt	100%	1,000

Ergebnisse der künstlichen Sprache:

Sprache	Bestehensquote
L1	95%
L2	92%
L3	88%
L4	65%
L5	58%

Bevölkerungsgewichtetes globales Ergebnis:

```
0.55(95)+0.25(92)+0.10(88)+0.06(65)+0.04(58)=90.27
```

Der globale Wert ist hoch. Es gibt jedoch eine starke Verschlechterung in zwei Sprachen.

38.1. Sprachenfairness-Übersampling

Fügen Sie 140 zusätzliche Beobachtungen zu L4 und 160 zu L5 hinzu. Neue Anzahl der Beobachtungen:

Sprache	Gesamtbeobachtungen
L1	550
L2	250
L3	100
L4	200
L5	200
Gesamt	1,300

Der rohe Mittelwert über alle 1,300-Beobachtungen ist niedriger als das oben gezeigte bevölkerungsgewichtete Ergebnis. Es handelt sich nicht um einen globalen Bevölkerungswert, da L4 und L5 absichtlich überstichprobenartig ausgewählt wurden. Sobald die Zielbevölkerungsgewichte wiederhergestellt sind, kehrt das globale Ergebnis zu der oben genannten bevölkerungsgewichteten Zahl zurück. Das Übersampling verbessert die Sensitivität für L4 und L5, ohne die bevölkerungsgewichtete Schätzung zu ändern.

38.2. Scorecard für Sprachenfairness

Künstliches Ergebnis. Bevölkerungsgewichteter globaler Wert: 90.27. Gleichsprachiger Wert:

```
(95+92+88+65+58)/5=79.6
```

Höchste-zu-niedrigste Sprachunterschied: 37 Punkte; niedrigste Sprache: 58; Mittelwert der beiden untersten Sprachen: 61.5; Kritische Fehlerquote in Sprachen mit geringen Ressourcen: separat berichtet; Prompt-Äquivalenz: als überprüft angenommen; Referenzparität: als überprüft angenommen. Korrekte Interpretation:

> *Die globale Bevölkerungsrepräsentation scheint hoch zu sein; jedoch ist die Sprachgerechtigkeit ernsthaft schwach.*

Falsche Interpretation: „Da der globale Score 90 beträgt, ist GEO in allen Sprachen erfolgreich.“

39. SYNTHETISCHER FALL, DER DIE URSACHE DER SPRACHLICHEN VERSCHLECHTERUNG UNTERSCHEIDET

Betrachten wir drei Sprachen.

Sprache A

Prompt-Äquivalent Offizielle Quelle vollständig Lokaler Prüfer ausreichend KI-Antwort falsch Ergebnis: Kandidat für ein Problem der Sprachvertretung des KI-Produkts.

Sprache B

Aufforderungsübersetzung materiell unterschiedliche KI-Antwort passend zur Aufforderung Ergebnis: Verfassungsmangel der Aufforderung.

Sprache C

Auf der offiziellen Sprachseite wird der alte Preis von AI auf den alten Preis der englischen Quelle übernommen. Ergebnis: Das Sprachquellen-Paritätsproblem der Einheit und die kombinierte Repräsentationslücke erreichen den Nutzer. Selbst wenn die Bewertungen der drei Sprachen gleich erscheinen, ist der Korrekturbesitzer unterschiedlich.

40. LAND- UND SPRACHPANELS ÖFFENTLICHE ERGEBNISKARTE

Die öffentliche Ergebnisübersicht des Kandidaten muss die folgenden Felder enthalten:

Ländliche Felder

Natürliche Reichweite geeignete Anzahl von Ländern Anzahl der Länder mit mindestens einer Beobachtung Länderbeobachtungsabdeckung Bevölkerungsgewichtete Beobachtungsabdeckung Länder-Vorhersageabdeckung Land Beobachterzykluszeit Offene Beobachterwarnungen LÄNDER AUSSERHALB DES NATIVEN ERREICHS NICHT BEOBACHTETE LÄNDER Messdatum

Sprachfelder

Anzahl der abgedeckten Sprachen Beobachtete Anzahl der Sprachen Anzahl der Sprachen, für die Vorhersagen gemacht werden können Offiziell unterstützte Sprachen Experimentelle Sprachen Bevölkerungsgwichtetes globales Sprachergebnis Ergebnis bei gleicher Spracherkennung Sprachuntergrenze Sprachunterschied Kritische Fehlerraten

REFERENZ_SPRACHEN_LÜCKE

AUFFORDERUNGS_EQUIVALENZ_FEHLGESCHLAGEN

NICHT GETESTETE Sprachen Sprachpanel-Version Diese Karte macht die Realität sichtbar, die ein einzelnes „mehrsprachiges“ Abzeichen nicht tragen kann.

41. VERPFLICHTENDE NORMATIVE BESTIMMUNGEN

CH07-N01

Die Country Observer- und Language Fairness-Panels müssen in Bezug auf Zweck, Stichprobe und Gewichtung separat vom Haupt-Populationspanel definiert werden.

CH07-N02

Eine Beobachtung eines Landesbeobachters kann nicht als Bevölkerungswert eines Landes verwendet werden.

CH07-N03

Für ein Land sind das Vorhandensein einer Beobachtung und die Fähigkeit, eine Länderschätzung zu erstellen, getrennte Status.

CH07-N04

Die Anzahl der abgedeckten Länder und die nach Bevölkerungsgewicht gewichtete Länderabdeckung sollten zusammen angegeben werden.

CH07-N05

Die Schätzung der Länderabdeckung sollte getrennt von der Beobachtungsabdeckung angezeigt werden.

CH07-N06

Jedes berechtigte Land muss in dem definierten Überwachungszyklus eine Wahrscheinlichkeit größer als null haben, beobachtet zu werden; auch Länder, die außerhalb des Geltungsbereichs liegen, sollten angezeigt werden.

CH07-N07

Der Überwachungszyklus und die Länderfrequenzklassen sollten versioniert werden, bevor KI-Antworten angezeigt werden.

CH07-N08

Ein Land mit niedriger Punktzahl kann nach Bekanntgabe der Ergebnisse nicht aus dem Beobachterzyklus entfernt werden.

CH07-N09

Eine Country Observer-Warnung ist kein automatischer Länderausfall; sie löst eine Überprüfung und Eskalation aus.

CH07-N10

Ein einzelnes Critical Observer-Ereignis kann auch eine separate Konformitätsprüfung ohne Prävalenzschätzung auslösen.

CH07-N11

Beobachtungen des Länderbeobachterpanels können nicht im Hauptbevölkerungs-panel mit roher gleichmäßiger Gewichtung verwendet werden.

CH07-N12

Wenn Beobachtungen des Länderbeobachters in die Hauptbewertung einbezogen werden, müssen ihre Auswahlwahrscheinlichkeiten und Gewichtungen ausdrücklich modelliert werden.

CH07-N13

Sprachliche Fairness kann nicht als die Pflicht definiert werden, wortwörtlich dieselbe Antwort zu geben.

CH07-N14

Sprachliche Fairness erfordert vergleichbare Genauigkeit, Grenzen, Quelle und Nutzeranpassung unter gleichwertigen materiellen Bedingungen.

CH07-N15

Sprache und Land, Sprache und Gebietsschema sowie Sprache und Schriftsystem dürfen nicht austauschbar verwendet werden.

CH07-N16

Das abgedeckte Sprachuniversum und der Grund für die Aufnahme jeder Sprache sollten vor der Messung erklärt werden.

CH07-N17

Die Behauptung "alle Sprachen" kann nur verwendet werden, wenn tatsächlich alle relevanten Sprach- und Nutzeroberflächen gemessen wurden; andernfalls sollte gesagt werden "Sprachen im Geltungsbereich".

CH07-N18

Beobachtungen in einer Sprache, die nicht offiziell unterstützt wird, können ohne Erklärung nicht in die Leistung des Hauptprodukts für unterstützte Sprachen aufgenommen werden.

CH07-N19

Ein Über-Sampling von Sprachen mit geringen Ressourcen sollte das globale Bevölkerungsgewicht nicht erhöhen.

CH07-N20

Wenn Beobachtungen des Language Fairness Panels in das globale Ergebnis einbezogen werden, sollten sie auf die tatsächlichen Zielsprachenpopulationen zurückgeführt werden.

CH07-N21

Der gleiche mehrsprachige Nutzer kann in der Gewichtung der Weltbevölkerung nicht als mehr als ein vollständiger Mensch gezählt werden.

CH07-N22

Ohne die Überprüfung der Gleichwertigkeit des Prompts kann der Unterschied im Sprachscore nicht endgültig der Leistung der KI-Sprache zugeschrieben werden.

CH07-N23

Ein niedriger Sprachscore kann nicht allein auf das KI-Produkt geladen werden, ohne die Referenzparität zu bewerten.

CH07-N24

Eine Entscheidung über die Sprachgenauigkeit kann nicht ohne ausreichende Kompetenz des lokalen Bewerters finalisiert werden.

CH07-N25

Abweichungen in der Antwortsprache und Code-Switching müssen separat aufgezeichnet werden.

CH07-N26

Legitime Unterschiede in der lokalen Realität können nicht als Sprachungerechtigkeit klassifiziert werden.

CH07-N27

Unerklärliche wesentliche Sprachunterschiede können nicht ausschließlich als Personalisierungs- oder Stilunterschiede abgetan werden.

CH07-N28

Der globale Bevölkerungsscore kann das Ergebnis der Sprachgerechtigkeit nicht ersetzen.

CH07-N29

Ein gleicher Sprachscore kann nicht wie ein bevölkerungsgewichteter globaler Score präsentiert werden.

CH07-N30

Fehlerquoten basierend auf Sprache müssen getrennt vom Gesamtdurchschnitt angezeigt werden.

CH07-N31

PROMPT_EQUIVALENCE_FAILURE, REFERENCE_LANGUAGE_GAP, MEASUREMENT_PARITY_FAILURE und ADJUDICATION_LANGUAGE_GAP müssen als separate Status beibehalten werden.

CH07-N32

In Country Observer- und Language Fairness-Panels darf die Entscheidung, Teilnehmer zu wechseln, nicht auf der Richtigkeit der KI-Antwort oder zum Vorteil der Marke basieren.

CH07-N33

Die gleiche Beobachtung kann mehrere Diagnosekennzeichnungen haben; sie darf jedoch ohne Erklärung nicht mehr als einmal volles Gewicht in der globalen Vorhersage erhalten.

CH07-N34

Länder- und Sprachpanelaufzeichnungen müssen einen verantwortlichen menschlichen oder institutionellen Eigentümer haben.

42. FORMEN DES VERSAGENS

CH07-F01— EINZIG-PERSON-LÄNDERSCORE

Die Beobachtung eines einzelnen Beobachters wird als GEO-Punktzahl für das Land veröffentlicht.

CH07-F02— SYMBOLISCHE LÄNDERABDECKUNG

Einer Person wird jedem Land zugewiesen und es wird als 'proportional zur Weltbevölkerung' bezeichnet.

CH07-F03— VERWECHSLUNG VON BEVÖLKERUNGSABDECKUNG MIT ANZAHL DER LÄNDER

Obwohl eine kleine Anzahl großer Länder beobachtet wird, wird nur die Bevölkerungsabdeckung gezeigt; obwohl eine große Anzahl kleiner Länder beobachtet wird, wird nur die Anzahl der Länder gezeigt.

CH07-F04— ZÄHLUNG DES BEOBACHTUNGSBEREICHS ALS SCHÄTZBEREICH

Alle Länder mit mindestens einer Beobachtung gelten als bewertete Länder.

CH07-F05— ZWEIFACHE ZÄHLUNG EINER LÄNDERBEOBACHTUNGSBEOBACHTUNG IM HAUPTPUNKT

Die gleiche Beobachtung erhält in sowohl dem Beobachter- als auch dem Bevölkerungs-Panel volles Gewicht.

CH07-F06— UMWANDELN EINER LÄNDERBEOBACHTUNGSWARNUNG IN EINEN PRÄVALENZANSPRUCH

Ein einzelnes Ereignis wird auf alle Nutzer im Land verallgemeinert.

CH07-F07— IGNORIEREN EINES KRITISCHEN BEOBACHTEREREIGNISSES

Das hohe Schadensrisiko wird nicht untersucht, indem gesagt wird „nur ein Nutzer“.

CH07-F08— ENTFERNEN EINES NIEDRIG-BEWERTETEN LANDES AUS DEM ZYKLUS

Der Beobachterplan wird gemäß den Ergebnissen geändert.

CH07-F09— NUR KOMMERZIELLE MÄRKTE BEOBACHTEN

Länder, die Zugang zu Produkten haben, aber kommerziell als unbedeutend gelten, werden vollständig ausgeschlossen; Das Ergebnis wird global präsentiert.

CH07-F10— DAS LAND UND DIE SPRACHE ALS GLEICH BETRACHTEN

Die offizielle Sprache des Landes wird als Sprache aller Nutzer akzeptiert.

CH07-F11— ANSPRUCHE AUF ALLE SPRACHEN

Es werden nur wenige Sprachen gemessen; Das Ergebnis wird als „in allen Sprachen“ veröffentlicht.

CH07-F12— NIEDRIGES SCHWEISSSPRACHEN LÖSCHEN

Da das Sample klein ist, wird Sprache vollständig aus dem Bericht ausgeschlossen.

CH07-F13— WACHSENDE NIEDRIGRESSOURCENSPRACHE WELTWEIT

Übersampelte Sprache fließt mit ihrem Rohprobenanteil in die globale Punktzahl ein.

CH07-F14— ZÄHLUNG DER GLEICHEN SPRACHPUNKTZAHL ALS BEVÖLKERUNGSPUNKTZAHL

Der Diagnosescore, der jeder Sprache gleiches Gewicht gibt, wird als durchschnittliche Erfahrung der Weltbevölkerung dargestellt.

CH07-F15— ZÄHLUNG DER BEVÖLKERUNGSPUNKTZAHL ALS SPRACHGERECHTIGKEIT

Das globale Ergebnis, das durch das Gewicht großer Sprachen erhöht wird, verschleiert die Verzerrung bei kleinen Sprachen.

CH07-F16— ZÄHLUNG DES FEHLERS BEI PROMPT-ÜBERSETZUNG ALS KI-FEHLER

Der Unterschied, der durch semantisch verschiedene Prompts entsteht, wird dem Modell zugeschrieben.

CH07-F17— ZÄHLUNG DER QUELLSPRACHENLÜCKE ALS KI-FEHLER

Obwohl keine lokalisierten offiziellen Informationen vorliegen, wird die gesamte Verantwortung dem KI-Produkt übertragen.

CH07-F18— VERTEIDIGUNG VON KI-FEHLERN MIT RESSOURCENMANGEL

Obwohl die gleichen und ausreichenden Ressourcen verfügbar sind, werden Fehler in der Sprache mit geringen Ressourcen ausschließlich als fehlender Inhalt geschlossen.

CH07-F19— BERÜCKSICHTIGUNG DER UNGENÜGENDEN SPRACHE DES SCHIEDSBELEIDIGERS ALS PRODUKTFEHLER

Selbst wenn der Schiedsrichter missverstanden hat, wird das Ergebnis abgeschlossen.

CH07-F20— BERÜCKSICHTIGUNG VON MASCHINENÜBERSETZUNG ALS ENDGÜLTIGER SCHIEDSBELEIDIGER

Verlust der lokalen Bedeutung wird nicht überprüft.

CH07-F21— BERÜCKSICHTIGUNG VON CODEÄNDERUNGEN ALS AUTOMATISCHER FEHLER

Fehler treten bei natürlichen Fachbegriffen oder akzeptabler Sprachmischung auf.

CH07-F22— BERÜCKSICHTIGUNG UNVERWENDBARER SPRACHABWEICHUNG ALS STIL

Selbst wenn das System in einer anderen Sprache antwortet, die der Nutzer nicht verstehen kann, gilt das Ergebnis als erfolgreich.

CH07-F23— IGNORIEREN DES SCHRIFTSYSTEMUNTERSCHIEDS

Ein Transliteration-Fehler, der zu Verwechslungen von Entitäten führt, wird als formal betrachtet.

CH07-F24— EINEN MEHRSPRACHIGEN NUTZER ZWEIMAL ZÄHLEN

Die gleiche Person wird in allen Sprachen als separate Bevölkerungseinheit gezählt.

CH07-F25— LEGITIME LOKALE UNTERSCHIEDE ALS UNFAIR BETRACHTEN

Selbst wenn der Preis- oder Rechtsunterschied auf tatsächlichen lokalen Bedingungen basiert, wird er als Sprachvorurteil deklariert.

CH07-F26— UNERKLÄRLICHE UNTERSCHIEDE ALS PERSÖNLICH BETRACHTEN

Materielle Inkonsistenzen unter gleichen Bedingungen werden nicht hinterfragt.

CH07-F27— DEFINITE PUNKTZAHL VON UNGENÜGENDER SPRACHZELLE

Die Sprachbewertung erfolgt mit wenigen Beobachtungen.

CH07-F28— EINSTELLUNG DES SPRACHPANELS NACH DEM ERGEBNIS

Nur Sprachen, die niedrig oder hoch bewertet wurden, werden danach ausgewählt.

CH07-F29— NUR HOCHQUELLENSPRACHEN MESSEN

Das Ergebnis wird als "mehrsprachig" dargestellt.

CH07-F30— NICHT GETESTETE SPRACHEN BEWAHREN

Das Feld NICHT GETESTET erscheint nicht auf der öffentlichen Ergebniskarte.

43. PRÜFUNGSVERFAHREN

Schritt 1— Produkt-Länder-Compliance-Register einfrieren

Für jedes Land:

- Natürliche Reichweite
- Produktzugang
- Audit-Sprache
- Nutzeroberfläche

Status wird aufgezeichnet.

Schritt 2— Abdeckung der Hauptpopulation im Panel-Land extrahieren

Es wird bestimmt, in welchen Ländern die Hauptpanelbeobachtung existiert.

Schritt 3— Definition des Beobachter-Länderuniversums

Länder, die im Hauptpanel unzureichend sind und den Status CE-3 haben, werden identifiziert.

Schritt 4— Den Beobachterzyklus festlegen

Länder:

- werden in Bezug auf Frequenzklasse,

- Welle,
- Sprache/Locale,
- Nutzeroberfläche

und Wartung.

Schritt 5— Beobachterteilnehmer vorab auswählen

Haupt- und Ersatzlisten werden erstellt, bevor die Ergebnisse eingesehen werden.

Schritt 6— Beobachterergebnisse klassifizieren

Zugang, Identität, Sprache, materieller Fakt, Status kritischer Ereignisse werden zugewiesen.

Schritt 7— Alarm und Eskalation anwenden

Nach Alarmtyp:

- erneut,
- gezielte Stichprobe
- treferenzüberprüfung,
- Konformitätsüberprüfung

wird gestartet.

Schritt 8— Berechnung der Länderabdeckungsmetriken

Der Umfang der Länderzahl, der Umfang der Bevölkerung und die Schätzung des Länderumfangs werden separat berechnet.

Schritt 9— Spracheintrag einfrieren

Für jede Sprache:

- aus Gründen des Umfangs,
- Unterstützungsstatus,
- Schriftsystem,
- Locale,
- Nutzeruniversum

aufgezeichnet.

Schritt 10— Oversampling zur Sprachgerechtigkeit festlegen

Zusätzliche Ziele werden für Sprachen festgelegt, die auf dem Hauptbevölkerungs-Panel unzureichend vertreten sind.

Schritt 11— Überprüfung der vier Sprachgates

Systemäquivalenz Referenzparität Messparität Entscheidungsparität wird bewertet.

Schritt 12— Bewertung der Sprachbeobachtungen

Für jede Sprache:

- Wahrhaftigkeit,
- materielle Mängel,
- Kritischer Fehler,
- unbekannt,

- Abweichung der Antwortsprache

aufgezeichnet.

Schritt 13— Globale und diagnostische Gewichtungen trennen

Das nach Bevölkerung gewichtete Ergebnis und das Ergebnis mit gleicher Sprachanerkennung werden separat berechnet.

Schritt 14— Berechne die Sprachungleichheit

Sprachfundament, erheblicher Sprachunterschied, kritischer Fehlerunterschied, die unbekannte Rate wird berechnet.

Schritt 15— Untersuche Länder–Sprach-Schnittmengen

Materielle Lokalisierungsunterschiede oder Konflikte werden identifiziert.

Schritt 16— Erstelle öffentliche Ergebnisübersicht

Umfang, Vorhersagegenügsamkeit, Warnungen und ungetestete Bereiche werden sichtbar gemacht.

44. NOTWENDIGE NACHWEISE

KI-Produkt- und Nutzeroberflächenaufzeichnung; Natürliche Reichweite Länderregister; Anzahl der berechtigten Länder; Länder-Zugriffsstatus; Hauptzuweisungen der Population Panel-Länder; Länderbeobachter-Universum; Beobachterzyklus; Beobachter-Frequenzklassen; Länder–Wave-Zuordnungsmatrix; Länder–Sprachzuweisungen; primäre und Reserve-Beobachterlisten; Länderbeobachtungsergebnisse; Beobachterwarnungen; Eskalationsaufzeichnungen; gezielte Länderstichproben; Länderbeobachtungsabdeckung; bevölkerungsgewichtete Länderabdeckung; Länder-Schätzabdeckung; im Geltungsbereich befindliches Sprachregister; Sprachunterstützungsstatus; Schriftzeichen- und Gebietsschemaaufzeichnungen; Hauptsprach-Stichprobenzahlen; Language Fairness-Überschussstichprobenzahlen; sprach-bevölkerungsgewichtete Daten; mehrsprachige Nutzeraufzeichnungen; Prompt-Äquivalenzaufzeichnungen; Referenzparitätsaufzeichnungen; Messparitätsaufzeichnungen; Aufzeichnungen der Qualifikation lokaler Gutachter; und Sprachergebnisse.

Sprache Kritische Fehlerraten Antwortsprache Abweichungen Sprachungleichheitsmetriken Ungetestete Sprachen Gleiche sprach- und bevölkerungsgewichtete Ergebnisse Panel-Gewichte Doppelzählungsprüfungen Öffentliche Ergebniskarten Panel-Versionen Verantwortliche Person oder Institution

45. PRÜFLISTE FÜR AUDITS

Wurde der Zweck des Länderbeobachter-Panels vom Hauptbevölkerungs-Panel getrennt? Welche Länder sind im Beobachter enthalten? Ist der Berechtigungs- und Zugangsstatus der Länder auf dem neuesten Stand? Gibt es eine Wahrscheinlichkeit einer positiven Beobachtung im Überwachungszyklus jedes berechtigten Landes? Wurde der Überwachungszyklus vor den Ergebnissen eingefroren? Wurde die Länderbeobachtung wie eine Länderbewertung verwendet? Wurden die Abdeckung der Länderbeobachtung und die Abdeckung der Länderabschätzung getrennt? Wurde die bevölkerungsgewichtete Länderabdeckung separat dargestellt? Wurden Länderbeobachtungen im Hauptwert doppelt gezählt? Wurde die Kritische-Beobachter-Warnung bestätigt? Wurde ein einzelner Alarm in eine Prävalenzbehauptung umgewandelt? Wurde aufgrund

des Alarms ein gezieltes Sampling durchgeführt? War das Universum der Sprachen im Umfang vordefiniert? Ist die Behauptung von 'allen Sprachen' mit der tatsächlichen Abdeckung konsistent? Wurden Sprache, Schriftsystem und Gebietsschema getrennt?

Wurde der Status der Sprachunterstützung des Produkts erfasst? Warum wurden Sprachen mit geringen Ressourcen ausgewählt? Wurden Beispiele für Sprachgerechtigkeit im globalen Score neu gewichtet? Wurde derselbe mehrsprachige Nutzer doppelt gezählt? Wurde die Gleichwertigkeit der Prompts überprüft? Sind offizielle Quellen über Sprachen hinweg wesentlich gleichwertig? Kennen die Gutachter die betreffende Sprache auf einem ausreichenden Niveau? Wurde maschinelle Übersetzung als endgültiger Bewertungsmaßstab verwendet? Wurden Abweichungen in der Antwortsprache erfasst? Wurde legitime lokale Variation von Sprachungerechtigkeit getrennt? Wurde das Sprachkorpus und die Sprachvarianz veröffentlicht? Sind sprachbasierte kritische Fehlerraten sichtbar? Wurden REFERENCE_LANGUAGE_GAP und andere Messfehler getrennt? Sind ungetestete Sprachen sichtbar? Verschleiert die öffentliche Bewertung das Problem der Sprachgerechtigkeit?

Hat dieselbe Beobachtung in mehreren Panels volle Gewichtung erhalten? Ist der verantwortliche Besitzer der Panel-Aufzeichnungen identifiziert?

46. EINWÄNDE UND ANTWORTEN

Einwand 1— „Ist es nicht besser, einer Person für jedes kleine Land zu geben, als nichts zu geben?“

Für frühe Beobachtungen und geografische Abdeckung, ja. Aber diese Person misst nicht den Anteil des Landes. Der richtige Begriff:

Land-Beobachter-Beobachtung

muss sein.

Einwand 2— „Wenn eine Person keinen Länderscore erzeugt, warum schicken wir sie dann?“

Weil eine einzelne Beobachtung:

- Zugangsprobleme aufdecken,
- Verwirrung über die Entität klären,
- die falsche Sprache identifizieren,
- und Kritische Ereignisse erzeugen

sichtbar zum ersten Mal. Das Beobachter-Panel ist ein Frühwarnsystem.

Einwand 3— „Müssen wir jedes Land in jeder Welle beobachten?“

Nein. Ein definierter Überwachungszyklus kann verwendet werden. Hochrisiko- oder bedeutende Länder können häufiger überwacht werden; kleine und risikoarme Länder können abwechselnd beobachtet werden. Der Zeitplan sollte vor den Ergebnissen festgelegt werden.

Einwand 4— „Wenn kleine Länder in der globalen Bewertung ein geringes Gewicht erhalten, warum schenken wir ihnen dann so viel Aufmerksamkeit?“

Weil ihr Bevölkerungsanteil klein sein mag. Allerdings:

- Gesetz,
- Sprache,

- lokaler Nutzerschaden,
- gesetzliche/geografische Integrität

sind tatsächliche menschliche Gemeinschaften. Geringes Gewicht sollte nicht Unsichtbarkeit bedeuten.

Einwand 5— „Ist es nicht die fairste Methode, alle Sprachen gleich zu gewichten?“

Eine gleichgewichtete Sprachbewertung kann eine wertvolle diagnostische Perspektive sein. Sie stellt nicht die durchschnittliche menschliche Erfahrung in der Weltbevölkerung dar. Die Bevölkerungsbewertung und die Sprachgerechtigkeitsbewertung sollten separat veröffentlicht werden.

Einwand 6— „Wenn eine kleine Sprache das globale Ergebnis nur geringfügig beeinflusst, warum ist dann eine niedrige Punktzahl wichtig?“

Weil das Repräsentationsproblem für Menschen, die diese Sprache verwenden, real ist. Schwere Fehler in Lizenzen, Gesundheit oder Identität werden auch nicht automatisch durch den Bevölkerungsanteil trivialisiert.

Einwand 7— „Ist es nicht unfair, der KI die Schuld zu geben, wenn eine Sprache mit wenigen Ressourcen nur wenige Ressourcen hat?“

Es kann nur unfair sein, der KI die Schuld zu geben. Daher:

- Ressourcenparität,
- Prompt-Äquivalenz,
- Kompetenz des Schiedsrichters

separat untersucht. Der Fehler, der den Nutzer erreicht, ist dennoch sachlich korrekt. Der Eigentümer der Korrektur sollte korrekt bestimmt werden.

Einwand 8— „Ist es nicht normal, in verschiedenen Ländern in derselben Sprache unterschiedliche Antworten zu haben?“

Es kann normal sein, wenn die tatsächlichen lokalen Bedingungen unterschiedlich sind. Wenn derselbe materielle Sachverhalt sich unterschiedlich ändert und ohne Erklärung, kann das ein Problem sein. Sprache und Standort sollten zusammen untersucht werden.

Einwand 9— „Wenn KI in einer anderen Sprache die richtige Antwort gibt, warum sollte sie dann versagen?“

Wenn der Nutzer die Antwort nicht verstehen kann, kann die Repräsentation nicht verwendet werden. Sprachabweichung sollte zusammen mit der Genauigkeit separat bewertet werden.

Einwand 10— „Die Suche nach 200-Nutzern für Sprachbewertungen kann sehr teuer sein.“

Eine kleinere Stichprobe kann verwendet werden. Ergebnis:

- Entdeckung,
- vorläufig,
- unzureichende Stichprobe

sollte mit diesem Status veröffentlicht werden. Kosten gewähren keine falsche Sicherheit.

Einwand 11— „Wenn die globale Punktzahl bereits hoch ist, verunsichert eine separate Sprachkarte den Kunden nicht unnötig?“

Das Verstecken von Materialverschlechterungen in der Sprache schützt den Kunden nicht. Es erzeugt falsches Vertrauen. Eine kurze und klare Ergebniskarte kann verwendet werden.

Einwand 12— „Reduziert ein einzelnes Critical Observer-Ereignis alle Berechtigungen?“

Die automatische Reaktion hängt von der Art des Ereignisses ab. Ein einzelnes Ereignis weist nicht auf eine nationale Verbreitung hin. Wenn es jedoch schwere Schäden oder falsche Autorität beinhaltet, kann es eine kritische Überprüfung und vorläufige Maßnahmen einleiten. Die Entscheidung wird in Abschnitt 15 definiert.

ALLGEMEINE REGEL VON ABSCHNITT 49

Die Erstellung einer Weltbewertung ist einfach. Sie füllen die Weltbevölkerung mit großen Ländern. Kleine Länder verlieren Sie in Bruchteilen. Sie messen Sprachen mit hohen Ressourcen. Sprachen mit niedrigen Ressourcen: diese entfernen Sie, indem Sie sagen „Stichprobe unzureichend.“ Dann veröffentlichen Sie eine einzige Zahl. Diese Zahl kann mathematisch korrekt berechnet werden. Aber sie zeigt möglicherweise nicht alle Repräsentationsfehler der Welt. NOMOS gibt jedem Land kein gleiches Gewicht. Weil die Bevölkerungsanteile der Menschen nicht gleich sind. NOMOS ignoriert kleine Länder auch nicht. Denn eine geringe Bevölkerung bedeutet nicht Nichtexistenz. NOMOS gewichtet nicht alle Sprachen gleich in der globalen Bewertung. Weil dies die durchschnittliche Erfahrung der Weltbevölkerung verzerren könnte. NOMOS macht kleine Sprachen auch nicht unsichtbar. Denn selbst wenn ihr globales Gewicht gering ist, kann der Repräsentationsfehler für die Menschen, die diese Sprache verwenden, 100% real sein.

Länderbeobachterpanel:

> *Zeigt, wo wir noch nie hingeschaut haben.*

Sprachgerechtigkeits-Panel:

> *Zeigt, wer unter dem globalen Durchschnitt liegt.*

Bevölkerungs-Panel:

> *Schätzt die durchschnittliche Erfahrung in der Weltbevölkerung.*

Wenn diese drei zusammenarbeiten, wird die globale Punktzahl sowohl der Bevölkerung treu als auch hinsichtlich der Gerechtigkeit prüfbar. Ein einzelner Nutzer in einem Land ist nicht das ganze Land. Aber dieser Nutzer ist eine echte Person. Eine Sprache mag einen kleinen Teil der Weltbevölkerung ausmachen. Aber der Fehler in dieser Sprache ist nicht weniger falsch. Ein kritisches Ereignis mag selten sein. Aber Schaden wird nicht nur nach Häufigkeit gemessen. Daher lautet das siebte Messgesetz von NOMOS:

> *Das Recht auf Sichtbarkeit ist nicht dasselbe wie das Gewicht in der Bevölkerung.*

Sein achtes Gesetz lautet wie folgt:

> *Ein Land zu beobachten ist nicht dasselbe wie ein Land zu schätzen.*

Das neunte Gesetz lautet wie folgt:

> *Sprachgerechtigkeit bedeutet nicht, alle Sprachen gleich zu machen; es bedeutet sicherzustellen, dass niemand systematisch von der materiellen Realität aufgrund seiner Sprache entfernt wird.*

Das zehnte Gesetz lautet:

> *Wenn der globale Durchschnitt die Sprache und das Land, die zurückliegen, unsichtbar macht, ist er allein kein ausreichendes GEO-Ergebnis.*

NOMOS's Abschnitt 7 Bestellung

> *Verstecke keine Länder, die du nie anschaust, während du mir die Weltwertung zeigst.*
>
> *Schicke keinen Nutzer in ein kleines Land und mache diese Person zum ganzen Land.*
>
> *Aber ignoriere auch kein kleines Land für immer, nur weil seine Bevölkerung klein ist.*
>
> *Trenne Länderbeobachtung, Landvorhersage und Länderangemessenheit.*
>
> *Beurteile das ganze Land nicht aufgrund eines einzelnen schweren Fehlers. / Aber ignoriere keinen menschlichen Schaden, indem du sagst „nur eine Antwort".*
>
> *Stütze den Erfolg in wichtigen Sprachen nicht auf das Schweigen von Minderheitensprachen.*
>
> *Du kannst eine Sprache mit wenigen Ressourcen viel abtasten. / Aber lass sie nicht größer erscheinen, als sie in der Weltbevölkerung ist.*
>
> *Machen Sie die gleiche Sprachpunktzahl nicht zu einer Bevölkerungsbewertung; machen Sie die Bevölkerungsbewertung nicht zu einer Sprachgerechtigkeitsbewertung.*
>
> *Geben Sie mir nicht die Schuld, wenn die Übersetzung falsch ist. / Wenn die Quelle unvollständig ist, sagen Sie das auch. / Wenn der Gutachter die Sprache nicht versteht, treffen Sie kein endgültiges Urteil.*
>
> *Klassifizieren Sie nicht einen legitimen lokalen Unterschied als Ungerechtigkeit, oder weisen Sie eine unerklärte wesentliche Diskrepanz nicht einfach als Stil ab.*
>
> *Beobachten Sie zuerst die Länder. / Geben Sie dann an, wo Sie eine Schätzung unterstützen können. / Überstichproben Sie dann die Sprachen, die diagnostische Sichtbarkeit erfordern. / Geben Sie sie dann zu ihren wahren Bevölkerungsgewichten zurück. / Halten Sie die Menschen sichtbar, die der globale Durchschnitt zurücklässt.*

Der abschließende Satz des Kapitels

Eine wirklich globale GEO-Messung muss mehr zeigen als das, was die meisten der Welt sehen. Sie muss auch zeigen, wer in kleineren Ländern und Sprachen unbeobachtet bleibt, wer falsch dargestellt wird und welches Ergebnis nur eine einzelne Warnung ist.

Normativer Kern

Das GEO-1000 Bevölkerungs-Panel, das Länderbeobachter-Panel und das Sprachgerechtigkeits-Panel MÜSSEN sich in Zweck, Stichprobe, Gewichtung und öffentlichen Angaben unterscheiden. Eine Beobachtung durch einen Länderbeobachter KANN feststellen, dass eine Zugangsbedingung, eine Antwort, eine Anomalie oder ein kritischer Vorfall in einem definierten Länder–Produkt–Sprach-Kontext aufgetreten ist. Eine Beobachtung durch einen Länderbeobachter DARF NICHT als Länderprävalenzschätzung oder Länder-GEO-Score repräsentiert werden. Die Abdeckung der Länderbeobachtung, die bevölkerungsgewichtete Beobachtungsabdeckung und die Länder-Schätzabdeckung MÜSSEN separat berichtet werden. Jedes berechtigte Land SOLLTE innerhalb eines vorab erklärten Überwachungszyklus eine positive Beobachtungswahrscheinlichkeit erhalten, es sei denn, es fällt außerhalb der rechtmäßigen Natürliche Reichweite des Produkts. Sprachfairness MUSS eine materiell vergleichbare Repräsentation unter semantisch gleichwertigen Aufforderungen und vergleichbaren Kontexten bewerten. Sie DARF keine wortwörtlich identischen Antworten verlangen oder legitime lokale Unterschiede ignorieren. Überprobenierte Sprachen MÜSSEN für globale Bevölkerungsabschätzungen wieder auf ihre Zielbevölkerungsgewichte zurückgeführt werden. Bevölkerungsgewichtete Ergebnisse, gleichwertige Sprachnachweise, Sprachuntergrenzen, kritische Fehlerquoten, unbekannte Raten und ungetestete Sprachen MÜSSEN weiterhin separat sichtbar bleiben. Eine Sprachdisparität DARF NICHT ausschließlich dem KI-Produkt zugeschrieben werden, bis die Gleichwertigkeit der Prompt, die Referenzgleichheit, die Messgleichheit und die Schiedsgleichheit bewertet wurden. Eine Beobachtung DARF mehrere diagnostische Panel-Tags tragen, darf jedoch nicht mit doppeltem Vollpopulation-Gewicht versehen werden. Vorfälle eines kritischen Beobachters KÖNNEN eine bestätigende Stichprobe und Konformitätsprüfung auslösen, ohne als Nachweis der Landesprävalenz behandelt zu werden.

KAPITEL 08

Auswahl und Gewichtung der Teilnehmenden sowie Stichprobenverzerrung

METHODISCHE GRUNDLAGEN / K04; K05; K06; K07; K08; K22

Kapitelgrenze

Kapitel 5 definierte die Zielnutzerpopulation. Kapitel 6 wandelte diese Population in ein Stichprobendesign um, das über Länder, Sprachen und Nutzeroberflächen verteilt ist. Kapitel 7 etablierte ergänzende Beobachtungssysteme für kleine Länder und ressourcenarme Sprachen, die das Hauptpopulationpanel möglicherweise unsichtbar lassen könnte. Nun müssen die Zahlen im Stichprobenplan in reale Personen umgesetzt werden:

> *Wo werden diese Teilnehmer gefunden, wie werden sie ausgewählt, unter welchen Bedingungen werden sie in das Panel aufgenommen und mit welchen Verzerrungen werden sie in den GEO-Score eingehen?*

Aus einem Online-Panel einer Forschungsorganisation? Aus einem offenen Aufruf in sozialen Medien? Von Universitäten? Aus der eigenen Nutzerliste des Kunden? Von lokalen Forschungspartnern? Nur von Personen, die häufig KI-Produkte nutzen? Von denen, die das Produkt zum ersten Mal nutzen? Von den Kunden der Marke? Von Mitarbeitern der Konkurrenz? Von Abonnenten des kostenpflichtigen Plans? Wie wird überprüft, dass eine Person tatsächlich in diesem Land lebt? Wie oft wird dieselbe Person gezählt, wenn sie mehrere Konten nutzt? Wie viel wird der Teilnehmer bezahlt? Wird die Bezahlung dazu führen, dass die Aufgabe schnell erledigt oder ein bestimmtes Ergebnis eingereicht wird? Wird die Beobachtung beibehalten, wenn der Teilnehmer eine falsche KI-Antwort hochlädt? Oder wird er durch einen neuen Nutzer ersetzt, indem gesagt wird: 'Die Aufgabe wurde falsch erledigt.'? Tausend gültige Screenshots könnten gesammelt worden sein. Aber die Teilnehmer:

- 60% sind intensive KI-Nutzer,
- 20% sind Technikstudenten,
- 10% sind Kunden der geprüften Marke,
- 5% sind Unternehmensmitarbeiter,
- der Rest sind allgemeine Nutzer

Welche Bevölkerung wird also durch das Ergebnis repräsentiert? Kann die demografische Gewichtung den Unterschied vollständig ausgleichen? Die Anpassung von Alters-, Geschlechts-, Länder- und Sprachstatistiken korrigiert nicht notwendigerweise nicht gemessene Unterschiede im technologischen Interesse, Markenbewusstsein, Zugang zu kostenpflichtigen Plänen oder Nutzungsmustern von KI. Gewichtung ist wertvoll, aber sie kann nicht jeden Fehler bei der Teilnehmerauswahl beseitigen. Dieses Kapitel wendet die frühere Anforderung an—dass jede

Auditkontrolle anhand von Modell, Datum, Land, Sprache, Abfrageumfang, Wiederholungsanzahl und Messaufzeichnung überprüfbar sein muss—auf die Auswahl der tatsächlichen Teilnehmer an. Es bewahrt auch die zentrale Forderung von NOMOS: Nachweis, Grenze, Kontext und Zeit. Dieses Kapitel:

- Quellen der Teilnehmerauswahl,
- die Unterscheidung zwischen probabilistischen und nicht-probabilistischen Panels,
- Berechtigungs- und Identitätsprüfung,
- Land-, Sprach- und Produktzugangskontrolle,
- Einzigartigkeit der Teilnehmer,
- Belohnung und Umleitungsrisiko,
- Panel-Konditionierung,
- Aufgabenverteilung,
- Nichtantwort und Panel-Fluktuation,
- betrügerische oder automatisierte Teilnahme,
- Auswahl- und Abdeckungsfehler,
- die Macht und Begrenzung der Gewichtung,
- Status der Stichprobenqualität

definiert. Dieses Kapitel behandelt noch nicht:

- alle technischen Bereiche des KI-Produktregisters,
- das vollständige Übersetzungs- und Versionssystem von Prompts,
- Die detaillierte Architektur der NOMOS Capture-Infrastruktur,
- atomare Aussageentscheidung,
- die endgültige Punkteformel

ist nicht vollständig definiert. Die Schlüsselfrage von Abschnitt 8 lautet:

> *Wenn wir tausend gültige Datensätze sammeln, wie beweisen wir, zu welchen realen Personen diese Datensätze gehören und welche menschliche Population sie darstellen können?*

NOMOS Herausforderung

Sie kaufen 1,000 Personen von einem Online-Forschungs-Panel. Der Panelanbieter sagt Ihnen: „1,000 Teilnehmer bereit gemäß Ihren Länder- und Altersquoten.“ Die Teilnehmer führen die Aufgabe aus. Sie laden Screenshots hoch. Rohdaten:

> *90 Prozent genaue Repräsentation.*

In Ihrem Bericht schreiben Sie: „90% der Nutzer weltweit werden genau repräsentiert.“ Die tatsächliche Verteilung der Teilnehmer ist jedoch wie folgt:

Häufigkeit der KI-Nutzung	Anteil der Zielpopulation	Anteil im Panel
Niedrige oder seltene Nutzung	50%	10%
Mittlere Nutzung	30%	30%
Hohe Nutzung	20%	60%

Starke KI-Nutzer:

- bessere Prompt-Gewohnheiten,

- mehr bezahlte Konten,
- neuere Produktoberflächen,
- Personalisierungswissen,
- höhere Vertrautheit mit Technologie

kann haben. Selbst wenn denselben festgeschriebenen Prompt verwendet wird:

- Kontoplan,
- Schnittstelle,
- vorherige Nutzung,
- Produktzugang

kann unterschiedlich sein. Lassen Sie die Durchlaufquoten der synthetischen Gruppe wie folgt sein:

Häufigkeit der KI-Nutzung	Bestehensquote
Niedrige oder seltene Nutzung	72%
Mittlere Nutzung	84%
Hohe Nutzung	96%

Das Roh-Ergebnis des Panels:

```
0.10(72)+0.30(84)+0.60(96)=90
```

das ist:

90 Prozent

Mit Ziel-Populationsgewichten ist das Ergebnis jedoch:

```
0.50(72)+0.30(84)+0.20(96)=80.4
```

Dementsprechend erzeugen dieselben 1,000 Antworten zwei verschiedene Ergebnisse:

- Rohpanel-Ergebnis: 90 Prozent
- Ziel-Populations-Ergebnis: 80.4 Prozent

Stellen Sie sich nun vor, dass Sie es gewichtet haben. Jede der 100 Personen in der Gruppe mit niedrigem Nutzungsgrad erhält ein hohes Gewicht. Jede der 600 Personen in der Gruppe mit hohem Nutzungsgrad erhält ein geringes Gewicht. Die Rohstichprobengröße beträgt weiterhin 1,000. Die effektive Stichprobengröße sinkt jedoch auf ungefähr 349. Sie haben also möglicherweise 1,000 Dateien gesammelt. Statistisch gesehen haben Sie möglicherweise nicht das Informationsvolumen von 1,000 gleichwertigen und unabhängigen Personen. Nun eine schwierigere Frage: In der Gruppe mit niedrigem Nutzungsgrad, Personen, die noch nie in das Panel aufgenommen wurden:

- wenig verbunden,
- Nutzung von unterstützender Technologie,
- Lebensort in einer kleinen Stadt,
- keinen spezifischen Gerät

Wenn es Menschen gibt, kann Gewichtung sie zurückbringen? Nein. Eine Beobachtung einer Person, die überhaupt nicht im Stichprobenrahmen enthalten ist, kann durch Gewichtung nicht erzeugt werden. Jetzt sendet das geprüfte Unternehmen Verbindungen zu seinen eigenen Kunden: 'Bitte nehmen Sie an der Umfrage teil.' Die treuen Kunden des Unternehmens nehmen mit hoher Rate teil. Menschen, die die Konkurrenz kennen, aber das Unternehmen nicht nutzen, nehmen weniger teil. Alters- und Landesgewichte können korrekt sein. Der Bias durch Markentreue kann weiterhin bestehen. Nun ein weiteres Problem: Dem Teilnehmer wird gesagt: 'Sie erhalten eine Prämie, wenn Sie eine korrekte und qualitativ hochwertige KI-Antwort hochladen.' Der Teilnehmer überarbeitet die

Antwort. Sie wählen das beste Ergebnis. Sie schneiden den Screenshot zu. Dies ist nicht mehr die natürliche Reaktion des KI-Produkts. Es ist die ausgewählte Reaktion, die vom Belohnungssystem erstellt wurde. Die erste Regel dieses Abschnitts lautet:

> *Die Anzahl der Teilnehmer ist nicht die repräsentative Kraft der Teilnehmer.*

Ihre zweite Bestimmung lautet:

> *Zufällige Zuweisung ist keine zufällige Populationsauswahl.*

Seine dritte Bestimmung besagt:

> *Gewichtung kann beobachtetes Ungleichgewicht verringern; sie kann keine unbeobachtete Person erzeugen.*

Seine vierte Bestimmung besagt:

> *Die Teilnahme einer Person am Panel, ihr Verbleib bei der Aufgabe oder ihr Ersatz darf nicht vom Inhalt der KI-Antwort abhängen.*

Seine fünfte Vorschrift lautet:

> *Die Belohnung sollte für die ehrliche Durchführung des Protokolls gegeben werden, nicht für die Wahl einer positiven oder gewünschten Antwort.*

1. ZWECK DES KAPITELS

Der Zweck dieses Abschnitts ist es, die gesamte Auswahlkette von der Zielpopulation bis zur tatsächlichen Nutzerbeobachtung sichtbar, unvoreingenommen und überprüfbar zu machen. Der Abschnitt unterscheidet normativ:

- Zielpopulation und Stichprobenrahmen
- Stichprobenrahmen versus Teilnehmerquelle
- Ausgewählte Person versus eingeladene Person
- Die eingeladene Person und die Person, die zustimmt teilzunehmen
- Der Teilnehmer und die Person, die die Aufgabe abschließt
- Die Person, die die Aufgabe abschließt und die gültige Beobachtung
- Mensch und Konto
- Mensch und Gerät
- Mensch und Sitzung
- Probabilistisches Panel und Freiwilligenpanel
- Zufällige Auswahl innerhalb des Panels und zufällige Auswahl aus der Bevölkerung
- Quotencompliance und echte Repräsentation
- Demografisches Gleichgewicht und verhaltensbezogenes Gleichgewicht
- Persönliche Vertrautheit und Eignung für die Zielpopulation
- Kunde, Mitarbeiter und allgemeiner Nutzer
- Teilnehmerbelohnung und Ergebnisanreiz
- Aufgabengeschwindigkeit und Aufgabenintegrität
- Ungültige Beobachtung mit Nichtreaktion
- Bevölkerungsveränderung mit Panelverlust
- Gefälschter Teilnehmer und echte Beobachtung niedriger Qualität
- Datenbetrug und falsche KI-Antwort

- Bias-Korrektur mit Gewichtung
- Erstellung von Nachweisen mit Kalibrierung
- Rohstichprobengröße vs. effektive Stichprobengröße
- Teilnehmerprivatsphäre und Prüfungsnachweis
- Allgemeines Bevölkerungspanel und Expert:innen- oder Kundenpanel

Am Ende dieses Abschnitts sollte jedes GEO-1000-Audit in der Lage sein, die folgenden Fragen zu beantworten:

> *Aus welcher Quelle und nach welcher Regel wurden die Teilnehmer ausgewählt?*
>
> *Wer hatte keine Chance, dem Panel beizutreten?*
>
> *Wer nahm eher teil und warum?*
>
> *Welche bekannten Unterschiede werden durch die Gewichtungen korrigiert, und welche nicht?*
>
> *Wurde eine ungültige oder gefälschte Teilnahmeentscheidung getroffen, ohne den Inhalt der KI-Antwort zu sehen?*
>
> *Hat die Belohnung der Teilnehmer ihr Verhalten und die Wahl der Antworten beeinflusst?*

2. ZENTRALE NORMATIVE BESTIMMUNG

Jede GEO-1000 Beobachtung muss eine rückverfolgbare Kette der Teilnehmersauswahl von der Zielpopulation bis zu einem gültigen analytischen Datensatz aufweisen; Auswahl, Einladung, Annahme, Aufgabenverteilung, Abschluss, Überprüfung, Ausschluss und Gewichtungsentscheidungen müssen unabhängig vom positiven oder negativen Inhalt der KI-Antwort sein. [K04–K06] Ein Teilnehmerpanel nur:

- Land,
- impliziert keine Eignung in Bezug auf Alter,
- Sprache

Es kann nicht als für die Populationsdarstellung geeignet betrachtet werden, nur weil es die Quoten erfüllt. Das Panel sollte außerdem die folgenden Informationen enthalten, soweit sie relevant sind:

- Teilnehmerquelle
- Auswahlmechanismus
- Methode der Teilnahme am Panel
- Teilnahmekriterien
- Eindeutige menschliche Verifizierung
- Länder- und Sprachüberprüfung
- Verifizierung des Zugangs zum KI-Produkt
- Konto- und Planinformationen
- Gerät und Schnittstelle
- Häufigkeit der KI-Nutzung
- Markenbekanntheit
- Beziehung zur geprüften Einheit
- Anreizstruktur
- Methode der Aufgabenverteilung
- Nichtbeantwortung

- Ungültigkeit
- Wiederholung und Panel-Konditionierung
- Gewichtung
- Fehlende Abdeckung
- Datenschutz und Einwilligung

Die Messung einiger dieser Informationen kann den Teilnehmer beeinflussen. Daher ist auch die Reihenfolge der Datenerhebung Teil des Protokolls.

3. TEILNEHMER-AUSWAHLKETTE

Die Kette von der Zielpopulation bis zur endgültigen Beobachtung besteht aus den folgenden Stufen:

U→F→S→I→E→A→C→V

Hier:

- U: Zielnutzerpopulation
- F: in den Stichprobenrahmen einbezogene Personen
- S: ausgewählte Personen
- I: eingeladene Personen
- E: Personen, deren Berechtigung bestätigt wurde und die in die Studie aufgenommen wurden
- A: den Aufgaben zugewiesene Personen
- C: Personen, die die Aufgabe abschließen
- V: Personen, die gültige analytische Beobachtungen liefern

Bei jedem Übergang können Personen verloren gehen oder neue Verzerrungen entstehen. Zum Beispiel:

- Diejenigen in der Zielpopulation, die nicht im Panel sind
- Diejenigen im Panel, die nicht eingeladen werden
- Diejenigen, die eingeladen werden, aber nicht annehmen
- Diejenigen, die annehmen, aber die Aufgabe nicht abschließen
- Diejenigen, die abschließen, aber ungültige Aufzeichnungen liefern

sind nicht die gleichen Personen. Daher können nur die letzten 1,000 gültigen Beobachtungen nicht berichtet werden. Der gesamte Auswahltrichter muss sichtbar sein.

4. TEILNEHMERQUELLEN

GEO-1000 muss nicht an eine einzelne Teilnehmerquelle gebunden sein. Jede Quelle trägt ihre eigenen Verzerrungen und Reichweiten.

4.1. Wahrscheinlichkeits-Populationspanel

Teilnehmer haben eine bekannte und ungleich null Wahrscheinlichkeit, aus einer definierten Population ausgewählt zu werden. Quellen:

- Haushalt,
- Adresse,
- Population oder berechtigter Datensatz,

- Wahrscheinlichkeits-Telefon- oder Mehrstufen-Auswahl

kann sein. Vorteile:

- Designgewicht kann festgelegt werden,
- Populationsschlussfolgerungen sind stärker,

Auswahlprozess kann transparenter sein. Einschränkungen:

- kann auf globaler Ebene teuer sein,
- In einigen Ländern gibt es möglicherweise keinen geeigneten Rahmen,
- Es kann auch notwendig sein, die Teilpopulation zu finden, die Zugang zu KI-Produkten hat.

Wieder wird es keine Antwort geben. Probabilistische Auswahl ist nicht automatisch eine perfekte Repräsentation.

4.2. Professionelles Freiwilligen-Forschungs-Panel

Teilnehmer haben sich zuvor freiwillig für das Forschungs-Panel registriert. Vorteile:

- es ist schnell,
- kann auf mehrere Länder zugreifen,
- kann Profil- und Sprachkenntnisse bereitstellen,

Wiederholungsmessungen können durchgeführt werden. Einschränkungen:

- Teilnehmer sind forschungsaffin,
- das Niveau der Technologieanwendung kann hoch sein,
- starke Panelnutzer können vorhanden sein,
- Wahrscheinlichkeiten des Eintritts in die Population können unbekannt sein,

die gleiche Person kann in mehr als einem Panel registriert sein. Dieses Panel:

kann nicht als probabilistische Stichprobe der Weltbevölkerung dargestellt werden

Es darf nicht als Wahrscheinlichkeitsstichprobe der Weltbevölkerung repräsentiert werden.

4.3. Offener Online-Aufruf

Die Teilnahme erfolgt über soziale Medien, eine Webseite oder ein offenes Registrierungsformular. Vorteile:

- weite Sichtbarkeit,
- geringe Kosten,
- Zugang zu kleinen Ländern und Sprachgemeinschaften

kann bereitgestellt werden. Einschränkungen:

- starke Selbstselektion,
- Konzentration von Markenfans oder -kritikern,
- falsche und wiederholte Teilnahme,
- Überrepräsentation von Technologie-Enthusiasten

kann entstehen. Ein offener Aufruf kann viel besser geeignet sein als das Hauptpopulationpanel für:

- Entdeckung,
- Country Observer,
- Sprachliche Fairness,
- Replikationsherausforderung

Diese Quellen können für diese Zwecke besser geeignet sein.

4.4. Universitäts- oder Institutionen-gebundenes Panel

Teilnehmer:

- werden über Universitäten ausgewählt,
- öffentliche Institution,
- Forschungszentren,
- Berufsorganisationen

Vorteile:

- lokaler Zugang,
- Sprachkenntnisse,
- ethische Aufsicht,
- Erreichen von unterversorgten Gemeinschaften

kann bereitgestellt werden. Einschränkungen:

- intensive Teilnahme von Studierenden oder Fachkräften,
- Institutionseffekt,
- Alters- und Bildungsbias,
- Druck zur Unternehmensbeteiligung

kann entstehen.

4.5. Kunden- oder Nutzerliste

Die tatsächlichen Kunden oder Nutzer der geprüften Einrichtung werden ausgewählt. Diese Quelle:

- Kundenerfahrung,
- Eignung für den Service,
- Markenbekanntheit

Es ist wertvoll für die Messung. Es handelt sich nicht um ein allgemeines Bevölkerungs-Panel. Kunden:

- kann die Marke besser kennenlernen,
- kann eine positivere oder negativere Erfahrung machen,

kann eher die entsprechende Anfrage stellen. Das Kundenpanel sollte separat definiert werden.

4.6. Von der geprüften Institution vorgeschlagene Teilnehmer

Unternehmen:

- Mitarbeiter,
- Kunde,
- Partner,
- Follower

kann vorschlagen. Diese Nutzer können ohne Erklärung nicht im Hauptunabhängigen Bevölkerungs-Panel aufgenommen werden. Separat:

VON DER EINHEIT BENANNTES PANEL

sollte gekennzeichnet werden. Wenn die Institution Teilnehmer auswählt, ist das Ergebnis keine unabhängige Bevölkerungsmessung.

4.7. Multi-Quellen-Hybrid-Panel

Es ist die Kombination mehrerer Quellen. Beispiel:

- wahrscheinlichkeitsbasiertes nationales Panel,
- professionelles Online-Panel,
- kleiner lokaler Partner eines Landes,
- Sprachgerechtigkeits-Community

kann zusammen verwendet werden. Ein hybrides Panel kann die Abdeckung erhöhen. Unter den Ressourcen:

- wiederholen Einzelpersonen,
- Auswahlwahrscheinlichkeiten,
- Anreize,
- Aufgabenerfahrung,
- Nutzerprofil

sollten ebenfalls modelliert werden.

5. INTEGRITÄTSSTATUS DER TEILNEHMERAUSWAHL

Die Qualität der Teilnehmerauswahl eines Panels kann mit den folgenden Status eingeordnet werden.

PSI-0— UNBEKANNTE AUSWAHL

Es ist nicht ausreichend bekannt, wo und wie die Teilnehmer ausgewählt wurden. Eine Generalisierung auf die gesamte Population ist nicht möglich.

PSI-1— BEQUEMLICHKEITS- ODER OFFENE FREIWILLIGENPROBE

Es handelt sich um eine Bequemlichkeits- oder offene Freiwilligenstichprobe. Sie kann für Exploration und Ereigniserkennung genutzt werden.

PSI-2— QUOTENKONTROLLIERTES OPT-IN

Es gibt Länder-, Sprach- oder demografische Quoten. Die Wahrscheinlichkeit des Eintritts in die Population ist unbekannt.

PSI-3— KALIBRIERTES MULTI-QUELLEN-PANEL

Es gibt mehrere Quellen, vordefinierte Quoten und Populationskalibrierung. Nicht gemessene Selektionsverzerrungen können bestehen bleiben.

PSI-4— WAHRSCHIENLICHKEITS- ODER STARKES HYBRID-DESIGN

Es besteht eine bekannte Selektionswahrscheinlichkeit oder ein starkes Wahrscheinlichkeits–Nicht-Wahrscheinlichkeits-Hybriddesign.

PSI-5— REPLIZIERTE UND EXTERN AUDITIERTE SELEKTION

Selektionsdesign:

- mit verschiedenen Anbietern,
- in unterschiedlichen Wellen,
- unter unabhängiger Prüfung

wiederholtes und voreingenommenes Verhalten wurde untersucht. Diese Status garantieren nicht die Genauigkeit des Ergebnisses. Es zeigt das Vertrauensniveau an, das sich aus der Teilnehmerauswahl ergibt.

6. ZUFÄLLIGE AUSWAHL INNERHALB DES PANELS UND ZUFÄLLIGE AUSWAHL AUS DER BEVÖLKERUNG

Es kann 100,000 registrierte Nutzer in einem Freiwilligenpanel geben. Aus diesem Panel können 1,000 Personen zufällig ausgewählt werden. Auswahl:

> *ist zufällig innerhalb des Panels.*

Das Panel selbst ist jedoch nicht zufällig aus der Weltbevölkerung gebildet. Daher kann man nicht sagen: „1,000 Personen zufällig aus der Weltbevölkerung." Der korrekte Ausdruck lautet:

> *„1,000 Teilnehmer zufällig innerhalb eines vorregistrierten Freiwilligenpanels ausgewählt und an die Zielpopulation kalibriert."*

Zufälligkeit sollte auf die Phase beschränkt werden, in der sie angewendet wird.

7. TEILNEHMER-BERECHTIGUNGSNACHWEIS

Jeder Teilnehmer muss die folgenden Bedingungen in dem jeweils relevanten Umfang erfüllen, bevor er in das Hauptpanel aufgenommen wird:

- Altersberechtigung
- Freiwillige Zustimmung
- Eignung von Land und Region
- Sprachkenntnisse
- Zugang zum KI-Produkt
- Korrekt gewählter Plan oder Nutzeroberfläche
- Notwendiges Gerät und technischer Zugang
- Einzigartiger menschlicher Status
- Unabhängigkeit der Aufgabe
- Offenlegung von Interessen oder Beziehungen
- Forschungssicherheit
- Genehmigung zur Erhebung von Daten und Nachweisen

Die Konformitätsentscheidung kann nach der Generierung der KI-Antwort nicht rückwirkend geändert werden.

8. LÄNDERÜBERPRÜFUNG

Die Frage „Zu welchem Land gehört der Teilnehmer?" hat keine eindeutige Antwort. Die folgenden Felder sollten getrennt werden:

- Übliches Wohnsitzland
- Physisches Land zum Zeitpunkt der Messung
- Land oder Marktregion des KI-Kontos

- Zahlungs- oder Planland
- Staatsbürgerschaft
- Zielland des Prompts
- Locale

Die Staatsbürgerschaft ist möglicherweise nicht die Hauptvariable für die meisten GEO-Messungen. Was meistens wichtig ist:

> *Es ist der tatsächliche geografische und produktbezogene Zustand, in dem der Nutzer das Produkt erlebt.*

8.1. Reisender Nutzer

Der Teilnehmer kann sich in einem anderen Land als seinem üblichen Wohnsitzland befinden. Diese Person:

- Wohnsitzende Bevölkerung,
- Messungsstandort,
- Reisender

Sollte einen separaten Status haben.

8.2. VPN oder Standortwechsler

Nutzung von VPN:

- Datenschutz
- Institutionelles Netzwerk,
- Normaler Nutzergebrauch

Könnte der Grund sein. Es sollte nicht als automatische Betrugsware betrachtet werden. Es erschwert jedoch die Überprüfung der Länderebene. Es könnte eine Regel im Hauptpanel des kontrollierten Landes geben, um VPN-Nutzung zu deaktivieren oder zu protokollieren. Standortwechsel, die zur Umgehung von Beschränkungen verwendet werden, sind für das Hauptpanel Natürliche Reichweite nicht geeignet.

8.3. Länderverifizierungsstufen

GV-0— Unverifiziert

Nur Selbsterklärung verfügbar.

GV-1— Panel-verifiziert

Der Panelanbieter hat eine bestehende Länderregistrierung.

GV-2— Sitzungs-konsistent

Erklärung, Gebietsschema und Sitzungszeichen sind konsistent.

GV-3— Multi-Signal-verifiziert

Datenschutzfreundlich kompatibel mit mehreren Signalen.

GV-4— Unabhängig geprüft

Die Länderüberprüfungsmethode hat die externe Qualitätskontrolle bestanden. Die Erfassung einer genauen Straßenadresse oder sensibler Koordinaten ist keine Standardanforderung. Die Länderüberprüfung sollte unter dem Prinzip der Datenminimierung durchgeführt werden.

9. SPRACHKENNTNISSE UND NATÜRLICHE ANWENDUNG

Das Sprachwissen eines Teilnehmers kann sich allein auf die Aussage stützen: "Ja, ich weiß." Dies kann eine risikoarme Erstregistrierung sein. In einem Hochsicherheits-Sprachpanel zusätzlich:

- Aussage über die natürliche Nutzung,
- kurzer Verständnis-Check,
- Panel-Historie,
- Bewertung durch einen lokalen Gutachter,
- neutrialer Test vor der Aufgabe

können verwendet werden. Der Sprachtest sollte nicht zu einer schweren akademischen Prüfung werden, die nur hochgebildete Personen auswählt. Das Ziel:

> *Dass der Teilnehmer die Prompt und die Antwort wie ein echter Nutzer versteht.*

10. ÜBERPRÜFUNG DES ZUGRIFFS AUF KI-ProduktE

Teilnehmer:

- sollte in der Lage sein, auf das richtige KI-Produkt zuzugreifen,
- die richtige Nutzeroberfläche,
- die richtigen Plan- oder Kontobedingungen

sollte zugreifen können. Die folgenden Situationen sollten separat aufgezeichnet werden:

- Kostenloser Plan
- Bezahlter Einzelplan
- Unternehmensplan
- Schulungsplan
- Web
- Mobil
- Desktop
- Eingebettetes Produkt
- Sichtbares Modell
- Web-/Abruffunktion

Ein Teilnehmer kann nicht das Konto eines anderen verwenden. Passwortweitergabe darf nicht verlangt werden. Der Nachweis des Produktzugangs muss überprüft werden, ohne die Identität des Kontos öffentlich preiszugeben.

11. BEZIEHUNG ZWISCHEN TEILNEHMER UND GEPRÜFTEM UNTERNEHMEN

Die Beziehung des Teilnehmers zu der überwachten Organisation kann das Verhalten bei der Antwortaufzeichnung oder die Markenbekanntheit beeinflussen. Beziehungsstatus:

ALLGEMEINE ÖFFENTLICHKEIT

AKTIVER KUNDE

EHEMALIGER KUNDE

POTENZIELLER KUNDE

MITARBEITER

EHEMALIGER MITARBEITER

AUFTRAGNEHMER

PARTNER

INVESTOR

WETTBEWERBER / AFFILIATES

KI-PROVIDER / AFFILIATES

FORSCHER

UNBEKANNT

Direkte Mitarbeiter- und Entscheidungsträgerbeziehungen in der Hauptprobe der Allgemeinbevölkerung:

- können ausgeschlossen werden,
- können in eine separate Ebene gestellt werden,

kann mit einem niedrigen tatsächlichen Populationsgewicht geschützt werden. Das automatische Ausschließen aller verwandten Personen kann ebenfalls falsch sein. Wenn sie Teil der tatsächlichen Kunden- oder Mitarbeiterpopulation sind, können sie separat und explizit gemessen werden.

12. MARKENBEKANNTHEIT

Geprüfte Teilnehmerentität:

- hat noch nie davon gehört,
- kennt den Namen,
- kennt das Produkt,
- ist Kunde,
- kann eine stark positive oder negative Meinung haben.

Diese Information:

- kann das Verhalten der Teilnehmer beeinflussen,
- Panel unter natürlichen Nutzungsbedingungen,
- schnelle Reaktion

erwünscht. Wenn die Bekanntheit vor der Aufforderung gemessen wird, kann sie den Nutzer auf die Marke vorbereiten. Daher die Reihenfolge der Präferenz:

- Vorhandene Panelinformationen
- Separate Frage nach der Aufgabe
- Falls nötig, eine neutrale Vor-Messung, die die Marke nicht lenkt

kann durchgeführt werden. Vertrautheit sollte im Hauptbeispiel sichtbar sein; sie sollte nicht als Gewichtungsvariable ausgewählt werden, nachdem die Ergebnisse gesehen wurden.

13. VERTRAUTHEIT MIT DER NUTZUNG VON KI

Teilnehmer:

- Erstbenutzer,
- gelegentlicher Nutzer,
- regelmäßiger Nutzer,
- intensiver oder professioneller Nutzer

Einzelpersonen können sein. Häufigkeit der KI-Nutzung:

- Produktzugang,
- Art des Plans,
- Schnittstellenkenntnisse,
- Personalisierung,
- Wahrscheinlichkeit, die Aufgabe korrekt zu erledigen

kann betroffen sein. Die Auswahl nur erfahrener KI-Nutzer kann jedoch Protokollfehler reduzieren. Aber es kann die Schätzung der natürlichen Bevölkerung verzerren. Die richtige Lösung:

- den Nutzungsgrad messen,
- vergleiche es mit der Zielbevölkerung,
- schichten oder gewichten, falls notwendig,

und die gültige reale Erfahrung des unerfahrenen Nutzers bewahren.

14. EINZIGARTIGE MENSCHLICHE REGEL

Die Bevölkerungseinheit von GEO-1000 ist ein Mensch. Derselbe Mensch:

- in mehreren Panels,
- mit mehreren E-Mail-Konten,
- mit mehreren KI-Konten,
- von verschiedenen Geräten

kann teilnehmen. Einzigartigkeitsprüfungen müssen datenschutzfreundlich durchgeführt werden. Signale, die verwendet werden können:

- Anonyme Teilnehmeridentität
- Kryptographische oder hash-basierte Überprüfung
- Geräte- und Sitzungskonformität
- Zahlungs- oder Anreizaufzeichnung
- Zeit- und Aufgabenmuster
- Dasselbe Screenshot- oder Antwortdatei
- Sichere Nachprüfung über mehrere Panelanbieter

Kein einzelnes Signal muss für einen automatischen Ausschluss ausreichend sein. Insbesondere:

- Familien, die sich ein gemeinsames Gerät teilen,
- Institutionelles Netzwerk,
- Barrierefreiheits-Tools,
- Nutzer von gemeinsam genutzten Geräten mit niedrigem Einkommen

könnten fälschlicherweise als gefälscht betrachtet werden. Verdächtige Aufzeichnungen sollten einer menschlichen Überprüfung und einem Berufungsverfahren unterzogen werden.

15. TEILNEHMERBELohnung

Die Zahlung oder Belohnung, die dem Teilnehmer gewährt wird, kann sein Datenverhalten beeinflussen.

15.1. Zahlung unabhängig vom Ergebnis

Bevorzugte Methode:

> *Eine feste und vorher bekannt gegebene Zahlung für die Durchführung der Aufgabe gemäß dem Protokoll.*

Zahlung:

- darf nicht davon abhängen, dass die Antwort positiv ist,
- in die Kategorien gezwungen werden von

  wahr,
- lang,
- der Marke günstig,
- oder von Zitaten begleitet.

Die Auswahl muss unabhängig von all diesen Ergebnissen bleiben.

15.2. Qualitätsbonus

Wenn ein Qualitätsbonus verwendet werden soll, dürfen nur folgende Punkte:

- vollständige Nachweise,
- korrekter Prompt,
- korrektes Produkt,
- rechtzeitige Lieferung,
- Reinigung persönlicher Daten

an Protokollbedingungen gebunden werden. Er darf nicht an den Inhalt der KI-Antwort gebunden werden.

15.3. Geschwindigkeitsanreiz

Bonus ausschließlich für Geschwindigkeit:

- Eile,
- unvollständige Erfassung,
- falsches Produkt,
- Abschneiden der Antwort

kann das Risiko erhöhen. Innerhalb des vorgegebenen Zeitfensters sollte eine angemessene Zeit zugestanden werden.

15.4. Zahlung nach Land

Die Vergütung kann je nach lokalen Gebühren und Lebensbedingungen variieren. Diese Situation:

- kann die Teilnahmewahrscheinlichkeit beeinflussen,
- die Zusammensetzung des Panels

und sollte zusammen mit dem Vergütungsansatz und den Ländervergleichen dokumentiert werden.

15.5. Zwang und übermäßige Anreize

Die Belohnung sollte keinen Druck erzeugen, dem die Person nicht vernünftigerweise widerstehen kann. Insbesondere:

- niedriges Einkommen,
- Student,
- abhängige Mitarbeiter,
- vulnerable Gruppen

sollten in ihrer Freiwilligkeit geschützt werden.

16. EINLADUNGSTEXT UND VORORIENTIERUNG

Beim Einladen eines Teilnehmers kann der Zweck der Forschung ihn auf ein bestimmtes Ergebnis vorbereiten. Eine falsche Einladung: „Wir testen, wie KI-Systeme Apple falsch interpretieren.“ Diese Aussage kann den Teilnehmer dazu bringen, nach Fehlern zu suchen. Eine andere falsche Einladung: „Hilf uns, den KI-Erfolg unserer Marke zu beweisen.“ Diese Aussage erzeugt die Erwartung eines positiven Ergebnisses. Bevorzugte neutrale Einladung: „Studie zur standardisierten Nutzererfahrung und Beobachtung der Repräsentation in spezifischen KI-Produkten.“ Die geprüfte Einheit kann später aufgrund der Natur der Aufgabe sichtbar werden. Das gewünschte Ergebnis sollte jedoch nicht offengelegt werden.

17. AUFGABENZUWEISUNG

Teilnehmer:

- KI-Produkt,
- Prompt,
- Sprache,
- Welle,
- Sitzungsbedingung,
- Aufgabenreihenfolge

soll im Voraus mit einer vordefinierten Methode zugewiesen werden. Die Teilnehmer:

- bevorzugtes KI-Produkt,
- den Plan, den er/sie am besten kennt,
- die Sprache mit hoher Wahrscheinlichkeit einer positiven Reaktion

wenn sie selbst wählen dürfen, kann eine Zuweisungsverzerrung auftreten.

17.1. Zufällige Zuweisung innerhalb der Ebene

Nachdem das Land, die Sprache und der Produktzugang überprüft wurden, können Aufgaben innerhalb der Ebene zufällig zugewiesen werden.

17.2. Block-Zuweisung

Blöcke können verwendet werden, um kostenlose/bezahlt Pläne, mobil/web oder AI-Nutzungsfrequenz auszugleichen.

17.3. Produktreihenfolge

Wenn eine Person mehr als ein Produkt testet, sollte die Reihenfolge zufällig sein. Die Antwort auf das erste Produkt kann die Bewertung der nächsten Antwort oder des nächsten Aufgabenverhaltens beeinflussen.

17.4. Reihenfolge der Eingaben

Wenn ein Teilnehmer mehr als einen Prompt erhält:

- lernen,
- neue Informationen über die Marke erwerben,
- die vorherige Antwort auf die nächste Frage übertragen

Risiko entsteht. Die Reihenfolge der Aufforderungen sollte aufgezeichnet und gegebenenfalls ausgeglichen werden.

18. PANEL-KONDITIONIERUNG

Nutzer, die häufig an Forschungspanels teilnehmen, können die Erwartungen der Aufgabe erlernen. Eine Person, die wiederholt am gleichen Test teilnimmt:

- kann sich an die richtige Antwort erinnern,
- kann die Bildschirmaufnahme sorgfältiger auswählen,
- kann die Website der geprüften Einheit im Voraus überprüfen,

welche Fehler gesucht werden, kann erlernt werden. Diese Situation wird genannt:

Panel-Konditionierung

Panel-Konditionierung kann die tatsächliche natürliche Nutzererfahrung verändern.

18.1. Quellen der Konditionierung

Frühere GEO-Wellen Wiederholung derselben Einheit Gleiches Aufforderungs-Archiv Peer-Review-Schulung Aufgabenteilung in Teilnehmergemeinschaften Frühere Zahlungen und Qualitätsfeedback

18.2. Kontrollmaßnahmen bei Konditionierung

Neuer und wiederkehrender Teilnehmerstatus Erstteilnahmedatum Frühere Exposition gegenüber derselben Einheit Frühere Exposition gegenüber derselben Aufgabenfamilie Rotierendes Panel Aufgabenreihenfolge Zufälligkeit Vorgesehene Ruhezeit Neue Nutzeruntergruppe Vergleich der Ergebnisse von konditionierten und neuen Nutzern Die Wiederverwendung desselben Teilnehmers ist nicht verboten. Sie ist wertvoll für longitudinale Messungen. Kann nicht als natürliches Querschnittspanel präsentiert werden.

19. AUFGABENKONTAMINATION UND TEILNEHMERKOMMUNIKATION

Teilnehmer:

- Prompts,
- Antworten,
- Screenshots,
- erwartete Ergebnisse

können miteinander geteilt werden. Diese Situation kann unabhängige Beobachtungen stören.
Kontrollen:

- Dauer der Welle begrenzen
- Aufgaben randomisieren
- Pflicht, Ergebnisse nicht mit Teilnehmern zu teilen
- Suche nach wiederholten Antworten und visuellen Darstellungen
- Ungewöhnliche Text- und Zeitähnlichkeiten prüfen
- Überwachung von Warnungen durch Community- oder Panelanbieter

Teilnehmerkommunikation ist nicht automatisch Betrug. Ihre materielle Auswirkung sollte untersucht werden.

20. ARTEN DER NICHT-REAKTION

Nicht-Reaktion ist kein einzelnes Ereignis.

20.1. Nicht-Reagieren auf Einladungen

Der Teilnehmer reagiert überhaupt nicht auf die Forschungseinladung.

20.2. Ablehnung der Teilnahme

Die Person sieht die Einladung und entscheidet sich, nicht teilzunehmen.

20.3. Nicht-Beginnen der Aufgabe nach Qualifikation

Die Person wird als geeignet befunden, beginnt die Aufgabe jedoch nicht.

20.4. Aufgabenunterbrechung

Der Teilnehmer bricht die Aufgabe in der Mitte ab.

20.5. Belege nicht hochladen

Möglicherweise wurde eine AI-Antwort erhalten, aber der erforderliche Nachweis wird nicht hochgeladen.

20.6. Ungültiger Abschluss

Die Aufgabe ist abgeschlossen, aber:

- falsche Prompt,
- falsches Produkt,
- ausgewählte Antwort,

- fehlender Screenshot

sind Gründe für Ungültigkeit. Die Verhältnisse dieser Stufen sollten separat gemeldet werden.

21. ANTWORTFUNNEL

Für Schicht h:

- nhsel: ausgewählte Person
- nhinv: eingeladene Person
- nhcon: Person, die zugestimmt hat teilzunehmen
- nhsta: Person, die die Aufgabe begonnen hat
- nhcom: Person, die abgeschlossen hat
- nhval: gültige Beobachtung

Lassen Sie uns sein. Grundraten:

```
RR_h^invite= n_h^con/n_h^inv; RR_h^complete= n_h^com/n_h^sta; RR_h^valid= n_h^val/n_h^com
```

Es kann dargestellt werden als. Jede dieser Raten zeigt ein anderes Problem an. Nur die endgültige gültige Zahl anzugeben, verschleiert das Feldbias.

22. NICHT-ANTWORT-BIAS

Menschen, die nicht reagieren, können sich von denen unterscheiden, die reagieren. Beispiel: Ein Nutzer mit geringer Internetverbindung kann eine Aufgabe nicht abschließen. Ein Nutzer mit kostenlosem Plan kann nicht auf das Produkt zugreifen. Ein unerfahrener Nutzer findet die Anweisungen schwierig. Ein intensiver KI-Nutzer schließt schneller ab. In bestimmten Ländern bleiben die Zahlungen niedrig. Ein Nutzer mit negativer Markenerfahrung ist möglicherweise eher bereit, teilzunehmen. Selbst wenn die Nicht-Antwort-Rate hoch ist, kann das Bias gering sein. Selbst wenn die Nicht-Antwort-Rate niedrig ist, kann eine kleine fehlende Gruppe finanziell bedeutend sein. Daher nicht nur die Rate allein:

Welche Merkmale unterscheiden sich zwischen Befragten und Nichtbefragten

sollte untersucht werden.

23. NICHT-ANTWORT GEWICHTE

Wenn eine Nicht-Antwort-Einstellung verwendet wird, sollten die Klassen bestimmt werden, bevor die Ergebnisse gesehen werden. Potenzielle Variablen:

- Land
- Sprache
- Alter
- Häufigkeit der KI-Nutzung
- Plan
- Benutzeroberfläche
- Panelquelle
- Barrierefreiheit
- Markenbekanntheit

Nicht-Antwort Koeffizient:

```
a_r^{NR}= (Σ_{j∈S_r} d_j)/ (Σ_{i∈R_r} d_i)
```

kann berechnet werden als. Diese Anpassung beruht jedoch nur auf der Annahme innerhalb derselben Klasse: dass die Befragten die Nicht-Befragten ausreichend repräsentieren. Sie löst keine nicht gemessenen Unterschiede.

24. MECHANISMEN FEHLENDER DATEN

Fehlende Daten können konzeptionell in drei Klassen betrachtet werden.

24.1. Zufällige Fehlstellen

Fehlende Daten stehen nicht wesentlich in Zusammenhang mit dem Ergebnis und den Nutzermerkmalen. In Wirklichkeit kann dies selten vorkommen.

24.2. Fehlstellen erklärt durch beobachtete Merkmale

Fehlende Werte:

- Land,
- impliziert keine Eignung in Bezug auf Alter,
- Plan,
- Sprache

sind in bekannten Bereichen wie.. verwandt. Gewichtung kann teilweise helfen.

24.3. Fehlende Werte in Verbindung mit nicht beobachteten Merkmalen

Zum Beispiel:

- Misstrauen gegenüber dem KI-Produkt,
- geringe digitale Fähigkeiten,
- Markenablehnung,
- Datenschutzbedenken

wenn nicht gemessen, kann die Gewichtung die Defizienz nicht vollständig korrigieren. In diesem Fall ist eine Sensitivitätsanalyse erforderlich.

25. PANELAUSFALL UND TRENNUNG ZWISCHEN DEN WELLEN

In longitudinalen oder rotierenden Panels, bei denen dieselben Nutzer erneut gemessen werden, nehmen einige Personen an der folgenden Welle nicht teil. Dies ist Panelverlust. Diejenigen, die im Panel verbleiben:

- disziplinierter,
- mehr belohnungsorientiert,
- erfahrener,
- mehr an Forschung gewöhnt

können sein. Die Veränderung der Werte über die Zeit kann teilweise auf die Zusammensetzung des Panels zurückzuführen sein.

25.1. Protokoll des Panelabbaus

In jeder Welle:

- wieder eingeladen,
- teilgenommen,
- ausgestiegen,
- neu hinzugefügt

Nutzerzahlen sollten angezeigt werden.

25.2. Panel-Abbrechergewicht

Die verbleibenden Teilnehmer können gemäß der Ausstiegswahrscheinlichkeit gewichtet werden. Diese Methode beseitigt die Verzerrung durch unbeobachteten Panel-Ausstieg nicht vollständig.

25.3. Panel aktualisieren

Neue Nutzer können hinzugefügt werden, um die verlorenen Nutzer zu ersetzen. Ergebnisse für neue und zurückkehrende Nutzer sollten separat berichtet werden.

26. TEILNEHMERBETRUG UND AUTOMATISIERUNG

Es besteht ein Risiko von Betrug, wenn GEO-1000 ein incentiviertes und globales Panel ist. Mögliche Verhaltensweisen:

- eine Person, die mit mehreren Konten teilnimmt
- Verwendung von automatischen Bots
- Hochladen eines Screenshots einer anderen Person
- Bilder bearbeiten
- Hochladen der gleichen Antwort im Namen verschiedener Nutzer
- Angabe des falschen Landes oder der falschen Sprache
- Konto teilen
- Eine andere Person die Aufgabe erledigen lassen
- Erneutes Beitreten bei Panel-Anbietern
- Erstellen von Massentask-Farmen

Dieses Risiko lässt nicht alle Teilnehmer mit geringem Einkommen oder schnellen Teilnehmern als verdächtig erscheinen. Betrugserkennung:

- Debatten,
- mehrere Signale,
- proportionale menschliche Überprüfung

sollten darauf basieren.

27. BETRUGS- UND INTEGRITÄTSSTATUSE

FI-0— NICHT BEWERTET

Der Datensatz wurde in Bezug auf Integrität nicht bewertet.

FI-1— KEINE ERHEBLICHE ANOMALIE

Es wurden keine klaren Anzeichen für Betrug gefunden. Dies ist keine absolute Garantie für Genauigkeit.

FI-2— ÜBERPRÜFUNG ERFORDERLICH

Es gibt ein oder mehrere verdächtige Signal(e).

FI-3— DOPPELUNG ODER ZUSAMMENSPRUCH VERDACHTET

Es besteht der Verdacht auf wiederholte Person, visuelle Erscheinung, Reaktion oder Koordination.

FI-4— PROTOKOLLMANIPULATION BESTÄTIGT

Der Teilnehmer hat das Protokoll absichtlich verändert.

FI-5— ERFUNDENE BEOBACHTUNG BESTÄTIGT

Die Beobachtung ist keine echte KI-Nutzererfahrung. FI-2 ist kein automatischer Ausschluss. Überprüfung erforderlich.

28. BETRUGSSIGNALEN

Gleicher Screenshot-Hash Gleiche seltene Dateispur Unmögliche Abschlusszeit Mehrfache Identitätswiederholungen vom gleichen Gerät oder Sitzung Materialkonflikt in Land, Gebiet und Produktzugang Unstimmigkeit zwischen Antworttext und Bild Spur der Bildbearbeitung Dieselbe Person in Primär- und Backup-Liste Dieselbe anonyme Person bei mehreren Panel-Anbietern Serielle und mechanische Erledigung von Aufgaben Gleiches ungewöhnliches Fehlermuster Ein Signal allein kann kein schlüssiger Beweis für Betrug sein.

29. FAIRNESS BEI DER BETRUGSERKENNUNG

Die folgenden Nutzer könnten fälschlicherweise als betrügerisch markiert werden:

- Familien, die ein gemeinsames Gerät nutzen
- Mitarbeiter, die ein institutionelles Netzwerk nutzen
- Teilnehmer von Universitätslaboren
- Datenschutzbewusste Personen, die VPNs verwenden
- Menschen mit Behinderungen, die Bildschirmleser oder Automatisierungshilfen einsetzen
- Personen, die aufgrund niedriger Bandbreite ähnliche komprimierte Bilder hochladen

Daher:

- automatische Kennzeichnung,
- menschliche Überprüfung,
- Teilnehmerbeschreibung, falls notwendig,
- Berufung

müssen zusammenarbeiten.

30. DIE GÜLTIGKEIT DES TEILNEHMERS SOLLTE VON DER GENAUIGKEIT DER KI-ANTWORT GETRENNT WERDEN

Der Datensatz eines Teilnehmers kann konform zum Protokoll sein. Die KI-Antwort kann falsch sein. Dies ist eine gültige Beobachtung. Die KI-Antwort eines anderen Teilnehmers kann korrekt sein. Dennoch hat der Nutzer:

- hat die Antwort dreimal aktualisiert,
- hat nur die beste Antwort hochgeladen,
- kann den Prompt geändert haben

Dies ist eine ungültige Beobachtung. Daher:

> *Die korrekte KI-Antwort bedeutet nicht gültige Daten. / Eine falsche KI-Antwort bedeutet nicht ungültige Daten.*

Die Gültigkeitsbewertung sollte vor der inhaltlichen Bewertung erfolgen und so blind wie möglich durchgeführt werden.

31. INTERESSENKONFLIKT UND TEILNEHMERQUELLE

Ein Teilnehmer oder Panelanbieter kann von den folgenden Ergebnissen profitieren:

- Das geprüfte Unternehmen erhält eine hohe Punktzahl
- Das KI-Produkt erscheint überlegen
- Ein Wettbewerber erhält eine niedrige Punktzahl
- Verkauf von mehr Audit-Aufträgen
- Auszeichnung oder Beschäftigungsverhältnis

Daher:

- Teilnehmerbeziehungen,
- Panelanbietervereinbarung,
- ergebnisunabhängiger Lohn,
- Kundeneingriff

sollte erklärt werden. Geprüfte Institution:

- welche Nutzer einzuladen sind,
- wer als ungültig betrachtet wird,
- Wer durch einen Ersatz ersetzt wird

kann nicht entscheiden.

32. WIRKUNG DES PANELANBIETERS

Nutzer von zwei Panelanbietern im selben Land können unterschiedlich sein. Ein Anbieter:

- Technologiebegeisterte,
- Studenten,
- intensive Forschungsteilnehmer

können in Bezug auf Fürsorge eine höhere Rate aufweisen. Der Panelanbieter sollte in der Analyse als Cluster- oder Quellenvariable beibehalten werden. Wenn möglich:

- anbieterbasierte Ergebnisse,
- Quelleneffekt,
- Anbieter vergleich innerhalb desselben Landes

sollten untersucht werden. Ein globales Panel mit einem einzigen Anbieter kann nicht als die Weltbevölkerung selbst dargestellt werden.

33. WIEDERHOLUNG EINZELNER IM MULTI-QUELLEN-PANEL

Dasselbe Individuum:

- ezwei professionelle Panels,
- bei einem offenen Aufruf,
- auf der Universitätsliste

kann gefunden werden. Ein Multi-Quellen-Panel kann die Anzahl der Datensätze erhöhen, ohne die Einzigartigkeit zu steigern. Prüfung auf Doppelungen über Quellen hinweg:

- datenschutzfreundlich,
- ohne zentrale Identitätsfreigabe erforderlich,
- unter Berücksichtigung von Widersprüchen bei Nichtübereinstimmungen

sollte auf eine Weise eingerichtet werden.

34. BEISPIEL FÜR BIAS-KARTE

GEO-1000 sollte die folgenden Verzerrungsarten getrennt überwachen.

SB-1— Abdeckungs-Bias

Einige Segmente der Zielpopulation sind nicht im Stichprobenrahmen enthalten.

SB-2— Selektions-Bias

Die Wahrscheinlichkeit, dass Einzelpersonen im Rahmen für das Panel ausgewählt werden, entspricht nicht dem beabsichtigten Design.

SB-3— Selbstselektions-Bias

Freiwillige für die Studie unterscheiden sich systematisch von der Gesamtbevölkerung.

SB-4— Einladungs- und Leitungsbias

Der Einladungstext oder die Quelle zieht Menschen mit einer bestimmten Sichtweise an.

SB-5— Nonresponse-Bias (Nichtbeantwortungs-Bias)

Menschen, die nicht teilnehmen oder nicht abschließen, unterscheiden sich in Bezug auf die Ergebnisse.

SB-6— Panel-Ausschuss-Bias

Nutzer, die zwischen den Wellen verbleiben, unterscheiden sich von denen, die ausscheiden.

SB-7— Anreiz-Bias

Zahlung oder Belohnung beeinflusst die Geschwindigkeit, das Auswahlverhalten oder die Antwort der Teilnehmer.

SB-8— Panel-Konditionierung

Der Teilnehmer lernt das erwartete Verhalten aus früheren Studien.

SB-9— Aufgabenverteilungs-Bias

Teilnehmer wählen das Produkt, die Sprache oder die Hinweise selbst oder werden ungleich verteilt.

SB-10— Bias bei Länder- und Sprachklassifizierung

Der Teilnehmer wird der falschen Länder- oder Sprachzelle zugewiesen.

SB-11— Produkt-, Plan- und Schnittstellen-Bias

Nur bestimmte Plan- oder Gerätebenutzer bestimmen das Panel.

SB-12— Markenbekanntheits-Bias

Kunden, Fans, Kritiker oder Mitarbeiter sind überrepräsentiert.

SB-13— Betrugs- und Wiederholungs-Bias

Die gleiche Person oder automatische Teilnahme gewinnt zu viel Einfluss auf das Ergebnis.

SB-14— Validitätsausschluss-Bias

Negative KI-Antworten werden häufiger als ungültig betrachtet.

SB-15— Gewichtungs-Bias

Nach Ergebnis ausgewählte Variablen oder übermäßige Gewichtung des Modellwerts ändern sich in der gewünschten Richtung.

SB-16— Überlebensstichproben-Bias

Nur Nutzer, die die Aufgabe erfolgreich abschließen können, werden wie die tatsächliche Produktpopulation dargestellt.

35. ZWECK DER GEWICHTUNG

Die Gewichtung kann den folgenden Zwecken dienen:

- Korrektur ungleicher Auswahlwahrscheinlichkeiten
- Rückführung von Überstichproben auf den tatsächlichen Bevölkerungsanteil
- Reduzierung von Unterschieden in den Rücklaufquoten
- Kalibrierung der Stichprobe auf bekannte Zielwerte
- Kombination mehrerer Panelquellen

Gewichtung kann den folgenden Zwecken nicht dienen:

- Erstellung einer Person, die im Stichprobenrahmen nicht existiert
- Präzise Korrektur nicht gemessener Variablen
- Wahrmachung einer falschen oder fehlerhaften Beobachtung
- Erzielung des gewünschten Ergebnisses nach der Auswertung
- Umwandlung einer einzelnen kleinen Zelle in einen verlässlichen Länderscore
- Automatische Umwandlung eines Nicht-Wahrscheinlichkeitspanels in eine Wahrscheinlichkeitsstichprobe

36. SCHWERGEWICHT ARCHITEKTUR

Das endgültige Gewicht kann wie folgt gezeigt werden:

```
w_i= d_i× a_i^{NR}× g_i^{CAL}× p_i^{SRC}
```

Hier:

- di: Grunddesign oder Ausgangsgewicht
- aiNR: Antwort-Einstellung
- giCAL: Bevölkerungs-Kalibrierung
- piSRC: Panelquelle oder Anpassung unwahrscheinlicher Stichproben

Im probabilistischen Panel:

```
d_i= 1/π_i
```

Es ist möglich. Das wahre πi ist im Nicht-Probabilitäts-Panel unbekannt. In diesem Fall wird nicht das Designgewicht, sondern das modellbasierte oder kalibrierte Gewicht verwendet. Diese Unterscheidung sollte im Bericht beibehalten werden.

37. GEWICHTE KÖNNEN NICHT VOM ERGEBNIS ABHÄNGEN

Die folgenden Variablen dürfen nicht zur Erstellung von Gewichten verwendet werden:

- Ob die KI-Antwort korrekt ist oder nicht
- Ob die Marke genannt wird oder nicht
- Ob ein Zitat vorhanden ist
- Ob eine Empfehlung ausgesprochen wird
- Ob ein kritischer Fehler auftritt
- Ob der geprüfte Organisation das Ergebnis gefällt

Gewichtungsvariablen:

- Bevor das Ergebnis gesehen wird,
- Weil es mit der Zielpopulation und dem Auswahlsystem zusammenhängt

Es sollte ausgewählt werden. Wenn das Antwortresultat das Gewicht bestimmt, formt sich die Punktzahl selbst.

38. KALIBRIERUNGSVARIABLEN

Mögliche Kalibrierungsbereiche:

- Land
- Region
- Sprache
- Alter
- Geschlecht, falls notwendig und ethisch
- Ausbildung
- Städtische/ländliche Siedlung
- Art des Internetzugangs
- Gerät
- Häufigkeit der KI-Nutzung
- Kostenloser/bezahlt Plan
- Benutzeroberfläche
- Barrierefreiheitsstatus

Zuverlässige Zielgesamtsummen sind möglicherweise nicht für jede Variable verfügbar. Zu viele Kalibrierungsvariablen:

- können eine übermäßige Gewichtungsvariabilität erzeugen,
- niedrige effektive Stichprobengröße,
- Modellüberanpassung

Es sollte eingefroren werden, bevor die Ergebnisse des Kalibrierungssatzes gesehen werden.

39. GEWICHTSVERTEILUNG

Die Stichprobe vor und nach der Gewichtung sollte mit der Zielpopulation verglichen werden. Für jedes Feld:

- Zielanteil
- Roher Stichprobenanteil
- Gewichteter Anteil
- Verbleibender Unterschied

soll angezeigt werden. Die Gewichtung wird nur auf die Anzahl der Ergebnisse angewendet und die Balancetabelle kann nicht gespeichert werden.

40. GEWICHTSDIAGNOSEN

Der Hauptbericht sollte die folgenden Felder enthalten, soweit sie relevant sind:

- Minimales Gewicht
- Maximales Gewicht
- Mediangewicht
- Gewichtsvariationskoeffizient
- Gewichtsverhältnis
- Anzahl der abgeschnittenen Gewichte
- Rohstichprobengröße
- Effektive Stichprobengröße
- Designeffekt

- Ungewichtetes Ergebnis
- Hauptgewichtetes Ergebnis
- Alternative plausible Gewichtungsergebnisse

Gewichtete Einzelzahl ist kein vollständiger Kontrollsatz ohne Gewichtskennungen.

41. SYNTHETISCHE GEWICHTSANZEIGE

ANZEIGE DER SYNTHETISCHEN METHODOLOGIE / Die unten stehenden Teilnehmer, Raten und Werte sind hypothetisch. Zielpopulation und Panelverteilung:

Häufigkeit der KI-Nutzung	Anteil der Zielpopulation	Rohanteil des Panels	Rohbeobachtung
Niedrige oder seltene Nutzung	50%	10%	100
Mittlere Nutzung	30%	30%	300
Hohe Nutzung	20%	60%	600
Gesamt	100%	100%	1,000

Synthetische Bestehensraten:

Gruppe	Bestehensquote	Frühere Antwort
Niedrige oder seltene Nutzung	72%	72
Mittlere Nutzung	84%	252
Hohe Nutzung	96%	576
Gesamt	90%	900

Rohresultat:

```
900/1000=90%
```

41.1. Kalibrierungskoeffizienten

Geringe Nutzung:

```
0.50/0.10=5
```

Mittlere Nutzung:

```
0.30/0.30=1
```

Hohe Nutzung:

```
0.20/0.60 = 1/3
```

Gewichtete Antwort:

```
72(5) + 252(1) + 576(1/3) = 804
```

Gewichtssumme:

```
100(5) + 300(1) + 600(1/3) = 1,000
```

Gewichtetes Ergebnis:

```
804/1000=80.4%
```

41.2. Effektive Stichprobengröße

```
n_eff= [100(5)+300(1)+600(1/3)]² / [100(5 ²)+300(1 ²)+600(1/3)²] ≈ 349
```

Ergebnisübersicht:

- Roh gültige Beobachtungen: 1,000
- Rohwert: 90 Prozent
- Gewichteter Wert: 80.4 Prozent
- Effektive Stichprobe: ungefähr 349
- Hauptgrund: Überrepräsentation von intensiven KI-Nutzern

Korrekte Interpretation:

> *Das Panel enthält 1,000 gültige Beobachtungen; jedoch ist infolge der Kalibrierung der Zielpopulation die Schätzung auf 80.4% gefallen, und aufgrund der Gewichtungsvariabilität beträgt die effektive Stichprobengröße ungefähr 349.*

Falscher Kommentar: „Die globale Genauigkeit wurde mit 90% bei 1,000 Personen nachgewiesen."

42. WAS KANN VERGEWICHTUNG NICHT KORRIGIEREN?

Die folgenden Gruppen sollten überhaupt nicht im synthetischen Panel vorhanden sein:

- Menschen, die Bildschirmlesegeräte verwenden
- Nutzer in ländlichen Gebieten mit geringer Bandbreite
- Menschen, die das KI-Produkt zum ersten Mal verwenden
- Kleine Sprachgemeinschaft

Die Gewichtung ändert nur das Gewicht der Nutzer im Panel. Die nicht beobachtete Gruppe:

- KI, deine Antwort,
- Zugangsprobleme aufdecken,
- Fehlerverteilung

kann nicht erzeugt werden. Daher:

> *Null Abdeckung bedeutet null Information.*

Gewichtetes Ergebnis: es kann eine Vorhersage näher an der Zielpopulation sein, nicht eine vollständige Beobachtung der Zielpopulation.

43. SENSITIVITÄTSANALYSE

Die Abhängigkeit des gewichteten Ergebnisses von einem einzelnen Gewichtungsmodell sollte untersucht werden. Kandidatenergebnisse:

- Rohes ungewichtetes Ergebnis
- Designfokussiertes Ergebnis
- Demografisch kalibriertes Ergebnis
- Kalibrierung unter Hinzufügung der KI-Nutzung
- Gewichtetes beschnittenes Ergebnis
- An den Effekt des Panel-Anbieters angepasstes Ergebnis

Wenn sich die Ergebnisse wesentlich ändern:

Warnung vor Gewichtungsempfindlichkeit

sollte gegeben werden. Ein einzelnes „richtiges“ Gewichtsergebnis kann nicht als absolute Wahrheit präsentiert werden.

44. NEIGUNGSANPASSUNG IM NONPROBABILITY-PANEL

Die Tendenz von Mitgliedern eines Non-Probability-Panels, dem Panel beizutreten, kann gemäß der Zielpopulation modelliert werden. Die Wahrscheinlichkeit, dass eine Person im Panel ist:

```
p̂_i= Pr(i ∈ panel | X_i)
```

kann geschätzt werden. Das Kandidaten-Inverse-Propensity-Gewicht:

```
w_i^P= 1/p̂_i
```

kann verwendet werden. Diese Methode:

- verwendet nur die gemessenen Xi-Variablen,
- ist von einem falsch spezifizierten Modell betroffen,
- nicht beobachtete Selektion wird nicht gelöst,

benötigt starke Zielreferenzdaten. Daher sollte das Ergebnis:

wie folgt bezeichnet werden:

Modellkalibrierte Nicht-Probabilitäts-Panel-Schätzung

45. SCHÄTZUNGEN FÜR KLEINE RÄUME UND MODELLBASIERTE SCHÄTZUNGEN

Direktes Sampling kann in einigen Länder- oder Sprachzellen unzureichend sein. Modellbasierte Methoden können Informationen von benachbarten oder ähnlichen Gruppen übernehmen. Das bedeutet, dass das Ergebnis:

- direkt beobachtete Länderbewertung,
- Rohbenutzerrate

nicht zutrifft. Der korrekte Status:

MODELLGESTÜTZTE SCHÄTZUNG

SCHÄTZUNG FÜR KLEINE GEBIETE

NICHT DIREKT BEOBACHTET

muss. Modellbasierte Vorhersage macht ein Land ohne Beobachtung durch den Länderkontrolleur nicht zu einem beobachteten Land.

46. TEILNEHMERDATENSCHUTZ UND VERIFIZIERBARKEIT

GEO-1000 erfordert einen Nachweis für jede Beobachtung. Dies bedeutet nicht, dass die Identität des Teilnehmers öffentlich gemacht wird. Ein Screenshot kann die folgenden Informationen enthalten:

- Name
- E-Mail
- Profilbild
- Vorheriger Chat
- Kontenplan
- Name der Institution
- Sensiblen personenbezogene Informationen

NOMOS-Erfassung oder gleichwertiges System:

- soll keine unnötigen persönlichen Felder erfassen,
- soll automatische oder kontrollierte Schwärzung durchführen,
- Rohbeweise müssen in einem sicheren Repository aufbewahrt werden.

Nur anonyme und überprüfbare Aufzeichnungen sollten der Öffentlichkeit präsentiert werden.

46.1. Öffentliches Evidenzmittel-Manifest

Öffentlicher Datensatz:

- Beobachter-ID,
- Land/Sprachzelle,
- KI-Produkt,
- Zeit,
- Nachweis-Hash,
- Verifizierungsstatus

kann getragen werden. Vollständiger Screenshot:

- bearbeitet,
- eingeschränkter Zugang,
- offen für unabhängigen Prüfer

kann sein. Die Existenz von tausend Evidenzstücken erfordert nicht die öffentliche Offenlegung der Identitäten von tausend Personen.

46.2. Sensibles Risiko eines kleinen Landes

Wenn es nur einen Teilnehmer im kleinen Land gibt:

- Gerät,
- Zeit,
- Sprache,
- Plan

Wenn sie zusammengetragen werden, kann es einfacher werden, die Person wiederzuerkennen. Der Detaillierungsgrad in öffentlichen Berichten sollte reduziert werden. Die Integrität der Nachweise kann in einem zugangsbeschränkten System erhalten bleiben.

47. TEILNEHMER EINVERSTÄNDNIS

Der Teilnehmer sollte mindestens Folgendes wissen:

- Welche Aufgabe sie ausführen werden
- Welche Daten und Screenshots gesammelt werden
- Welche persönlichen Informationen nicht gesammelt werden
- Wie lange die Daten gespeichert werden
- Wer Zugriff haben wird
- Wie das Ergebnis veröffentlicht wird
- Wie die Belohnung vergeben wird
- Dass die Teilnahme freiwillig ist
- Recht auf Widerruf oder Einspruch
- Konto-Passwort wird nicht angefordert
- Es wird nicht erwartet, dass die Nutzungsbedingungen des KI-Produkts verletzt werden

Zustimmung sollte nicht in einem langen und unverständlichen Text versteckt werden. [K06; K22]

48. TEILNEHMER_SICHERHEIT

Vom Teilnehmer:

- persönliche Gesundheits-, Rechts- oder Finanzinformationen an KI weiterzugeben,
- echte Kundendaten zu verwenden,
- geschäftsgeheime Informationen des Arbeitgebers zu teilen,
- die Kontosicherheit zu schwächen,
- Nutzungslimits zu überschreiten,
- das Gesetz oder die Produktbedingungen zu verletzen

kann nicht verlangt werden. Testaufforderungen sollten möglichst einen öffentlichen und nicht-personenbezogenen Kontext enthalten.

49. DER TEILNEHMER TRÄGT MENSCHLICHE VIELFALT, KEINE QUALITÄT

Unerfahrener Nutzer:

- kann die Aufgabe langsamer erledigen,
- schwieriger,
- mit mehr Unterstützung

Diese Person ist nicht automatisch ein Teilnehmer von geringer Qualität. Sie sind ein echter Teil der Zielpopulation. Wenn das Forschungsdesign so komplex ist, dass nur Personen mit hohen digitalen Fähigkeiten es abschließen können: Das Problem liegt möglicherweise nicht beim Teilnehmer, sondern beim Messinstrument. Das Protokoll:

- sollte verständlich sein,
- zugänglich,
- sprachlich natürlich,

- technische Belastung begrenzt

muss sein.

50. ÖFFENTLICHE RESULTATKARTE DER TEILNEHMERAUSWAHL

Der öffentliche Kandidatenbericht sollte die folgenden Bereiche enthalten:

Panelressourcen

Wahrscheinlichkeit Panelanteil Professionelles Freiwilligenpanel Anteil Offene Ausschreibung Anteil Universität/Institution Anteil Kunden- oder angeschlossenes Panel Anteil

Teilnahme-Funnel

Person im Rahmen Ausgewählt Eingeladen Akzeptiert Gestartet Abgeschlossen Gültige Beobachtung produzierend

Eignung und Integrität

Landesprüfungsstufen Sprachprüfungsstufen Wiederholungspersonenrate Verdächtige Registrierungsrate Überprüfte Betrugsrate Ungültige Beobachtungsrate

Panelzusammensetzung

KI-Nutzungsfrequenz Plan und Schnittstelle Markenbekanntheit Kunden-/Mitarbeiterbeziehung Neue und zurückkehrende Panelnutzer Zugänglichkeitsabdeckung

Gewicht Diagnosen

Folgendes kann veröffentlicht werden, ohne die Identität der Teilnehmer preiszugeben: Roh- und gewichtete Werte; minimale und maximale Gewichte; der Variationskoeffizient der Gewichte; effektive Stichprobengröße; Sensitivität gegenüber Gewichtung; und Abdeckungswarnungen.

51. VERPFLICHTENDE NORMATIVE BESTIMMUNGEN

CH08-N01

Jede gültige GEO-1000-Beobachtung muss eine nachvollziehbare Teilnehmerauswahlkette von der Zielpopulation zur analytischen Aufzeichnung aufweisen.

CH08-N02

Eine zufällige Auswahl innerhalb des Panels kann nicht als zufällige Auswahl aus der Zielweltbevölkerung dargestellt werden.

CH08-N03

Die Quelle der Teilnehmer, die Auswahlmethode und der Status des Panels sollten im öffentlichen Methodenprotokoll erläutert werden.

CH08-N04

Es kann nicht als ein unwahrscheinliches oder freiwilliges Panel-Stichprobenset der Bevölkerung definiert werden.

CH08-N05

Von der geprüften Institution vorgeschlagene Teilnehmer dürfen ohne Erklärung nicht im Haupt unabhängigen Bevölkerungspanel aufgenommen werden.

CH08-N06

Beziehungen zu Kunden, Mitarbeitern, Partnern, Wettbewerbern und KI-Anbietern sollten separat aufgezeichnet werden.

CH08-N07

Die Teilnahmeberechtigung muss überprüft werden, bevor die KI-Antwort angezeigt wird.

CH08-N08

Land, Sprache, Produkt, Plan und Schnittstellenberechtigung müssen vor der Zuweisung der entsprechenden Aufgabe bestimmt werden.

CH08-N09

Die gleiche Person darf aufgrund mehrerer Konten, Geräte, Sprachen, Panels oder KI-Produkte nicht als Bevölkerungseinheit dupliziert werden.

CH08-N10

Eine Überprüfung der eindeutigen Person muss gemäß den Grundsätzen der Datenminimierung und des Datenschutzes durchgeführt werden.

CH08-N11

Ein einzelnes technisches Signal allein kann nicht als ausreichend für eine automatische endgültige Entscheidung betrachtet werden, dass ein Teilnehmer gefälscht ist.

CH08-N12

Die Belohnung eines Teilnehmers darf nicht an die AI-Antwort gekoppelt sein, dass sie positiv, korrekt, lang, mit Referenzen versehen oder vorteilhaft für die Marke ist.

CH08-N13

Der Qualitätsbonus kann nur an die vorher festgelegte Protokollintegrität gebunden sein.

CH08-N14

Der Einladungs- und Informationstext darf den Teilnehmer nicht anleiten, ein positives oder negatives Ergebnis zu erzielen.

CH08-N15

Das KI-Produkt sollte zugewiesen oder randomisiert werden, bevor die Aufforderung und die Aufgabenfolge-Ergebnisse gesehen werden.

CH08-N16

Die Auswahl von Produkt und Aufforderung basierend auf den Präferenzen der Teilnehmer darf in der Hauptpopulation nicht unerklärt verwendet werden.

CH08-N17

Panelkonditionierung, vorherige Testexposition und wiederholte Teilnahme müssen aufgezeichnet werden.

CH08-N18

Die Ergebnisse von neuen und konditionierten Teilnehmern sollten bei Bedarf getrennt analysiert werden.

CH08-N19

Die Aufnahme eines Teilnehmers in das Panel, dessen Beibehaltung, Ausschluss oder Ersatz durch eine Ersatzperson darf nicht auf dem Inhalt der KI-Antwort basieren.

CH08-N20

Fehlerhafte, negative, ablehnende oder kritische KI-Ausgaben sollten gemäß einem gültigen Protokoll aufbewahrt werden.

CH08-N21

Die Gültigkeit der Teilnehmer sollte so weit wie möglich blind bewertet werden und bevor die sachliche Genauigkeit der Antwort beurteilt wird.

CH08-N22

Einladung, Annahme, Beginn, Abschluss und gültige Beobachtungsraten sollten separat berichtet werden.

CH08-N23

Die Antwortumgebung sollte nur auf vordefinierten und relevanten Variablen basieren.

CH08-N24

Gewichtung kann nicht dazu führen, dass Nutzer, die nicht im Stichprobenrahmen enthalten sind, vertreten werden.

CH08-N25

Gewichtsvariablen dürfen nicht vom KI-Antwortergebnis, der Punktzahl oder dem kommerziellen Interesse der regulierten Institution abhängen.

CH08-N26

Nicht-probabilistische Panelgewichte dürfen nicht als Designgewichte dargestellt werden.

CH08-N27

Kalibrierungsvariablen und Zielsummen müssen freigegeben werden, bevor die Ergebnisse gesehen werden.

CH08-N28

Ungewichtete und hauptsächlich gewichtete Ergebnisse müssen zusammen gespeichert werden.

CH08-N29

Die Gewichtung von Diagnosen und die effektive Stichprobengröße sollten im Hauptbericht sichtbar sein.

CH08-N30

Wenn Gewichtstrimmen verwendet wird, müssen das ungetrimmte Ergebnis und die Sensitivitätsanalyse aufbewahrt werden.

CH08-N31

Effekte des Panelanbieters und der Teilnehmerquelle sollten als separate Variablen gehalten werden.

CH08-N32

In einem Multi-Quellen-Panel sollten wiederholte Personenüberprüfungen zwischen den Quellen durchgeführt werden.

CH08-N33

Betrugserkennung sollte mehrere Signale, eine proportionale menschliche Überprüfung und ein Beschwerdeverfahren umfassen.

CH08-N34

Die Nutzung eines VPN, eines geteilten Geräts, eines institutionellen Netzwerks oder unterstützender Technologie sollte nicht automatisch als Betrug angesehen werden.

CH08-N35

Modellbasierte Schätzungen für kleine Bereiche können nicht direkt wie eine beobachtete Länder- oder Sprachpunktzahl präsentiert werden.

CH08-N36

Teilnehmer-Screenshots und Nachweise müssen persönliche Datenminimierung, Schwärzung und Zugangskontrolle aufweisen.

CH08-N37

Kontopasswörter, private Chatverläufe oder unnötige sensible Informationen dürfen von den Teilnehmern nicht angefordert werden.

CH08-N38

Teilnehmer dürfen nicht gebeten werden, Produkt-, Alters-, Identitäts-, Länder- oder Kontobeschränkungen zu umgehen.

CH08-N39

Kinder oder Nutzer, die besonderen Schutz benötigen, dürfen nicht ohne separate ethische und Sicherheitsprotokolle im Panel angemeldet werden.

CH08-N40

Aufzeichnungen über die Teilnehmersauswahl und Gewichtung müssen einen verantwortlichen menschlichen oder institutionellen Eigentümer haben.

52. FORMEN DES VERSAGENS

CH08-F01— ZUFÄLLIGE TEILNEHMER IM PANEL ALS ZUFÄLLIGE DER WELT ZÄHLEN

Die zufällige Auswahl innerhalb des freiwilligen Panels wird so präsentiert, als ob es sich um eine zufällige Auswahl aus der Weltbevölkerung handelt.

CH08-F02— EINHALTUNG VON QUOTEN ALS REPRÄSENTATIV ZÄHLEN

Es wird behauptet, dass alle Verhaltens- und Technologie-Biases gelöst sind, weil Alters- und Länderquoten erfüllt werden.

CH08-F03— KUNDEN ALS ALLGEMEINE BEVÖLKERUNG ZÄHLEN

Die Nutzerliste der überwachten Institution wird verwendet, als wäre sie ein weltweites Nutzerpanel.

CH08-F04— VERSTECKEN VON MITARBEITERN UND PARTNERN

Nutzer, die mit der Institution verbunden sind, werden als unabhängige allgemeine Teilnehmer angezeigt.

CH08-F05— ZÄHLEN EINES EINZELNEN PANEL-ANBIETERS ALS GLOBAL

Auswirkungen des Anbieters und das Verhalten des registrierten Panels werden nicht untersucht.

CH08-F06— ZÄHLEN EINER OFFENEN EINLADUNG ALS WAHRSCHEINLICHE STICPROBE

Selbstselektion wird beibehalten.

CH08-F07— BESTIMMUNG DES LANDES NUR NACH IP

Reisen, VPN, institutionelles Netzwerk und Unterschiede in der Kontoregion werden ignoriert.

CH08-F08— ZÄHLEN DER STAATSANGEHÖRIGKEIT ALS NUTZUNGS-LAND

Der tatsächliche geografische Zustand der Produkterfahrung wird falsch klassifiziert.

CH08-F09— ANNEHMEN EINER SPRACHENANGABE ALS DEFINITIV OHNE TEST

Der Teilnehmer wird in der Sprachzelle gehalten, obwohl er die Prompt oder die Antwort nicht versteht.

CH08-F10— DEN SPRACHTEST IN EINE ELITEAUSWAHL VERWANDELN

Nur hochgebildete Personen bleiben im Panel.

CH08-F11— DIE ANZAHL DER KONTEN ALS PERSONEN ZÄHLEN

Eine Person erhöht das Bevölkerungsgewicht mit mehreren Konten.

CH08-F12— EIN GEMEINSAMES KONTO ALS EINE PERSON ZÄHLEN

Ein von mehreren Personen genutztes Konto wird als einzelner Nutzer interpretiert.

CH08-F13— ZAHLUNG ABHÄNGIG VOM ERGEBNIS

Ein Bonus wird für eine positive, richtige oder referenzierte Antwort vergeben.

CH08-F14— DAS PROTOKOLL MIT GESCHWINDIGKEITSBONUS BRECHEN

Der Teilnehmer macht einen Screenshot oder hetzt durch die Aufgabe, bevor er die Antwort beendet.

CH08-F15— ERKLÄRUNG DES ERGEBNISSES IN DER EINLADUNG

Der Teilnehmer wird gebeten, einen Fehler oder Erfolg zu finden.

CH08-F16— DEN TEILNEHMER DAS PRODUKT WÄHLEN LASSEN

Nutzer, die nur das KI-Produkt auswählen, auf das sie zugreifen können oder das ihnen gefällt, bestimmen das Panel.

CH08-F17— DAS AUSWIRKEN DER REIHENFOLGE DER AUFFORDERUNGEN VERSTECKEN

Frühere Antworten beeinflussen nachfolgende Aufgaben.

CH08-F18— PANELBEDINGUNGEN ALS NATÜRLICHE NUTZUNG BETRACHTEN

Ein Nutzer, der denselben Test mehrmals sieht, wird als neuer Nutzer dargestellt.

CH08-F19— ERGEBNISWEITERGABE IGNORIEREN

Teilnehmer übertrugen Aufgaben und Antworten aneinander.

CH08-F20— NICHT-ANTWORTEN ALS FEHLENDE DATEN ZÄHLEN

Der systematische Unterschied von Nicht-Teilnehmern wird nicht untersucht.

CH08-F21— ZÄHLUNG DES PANELVERLUSTS ALS VERBESSERUNG DER ERGEBNISSE

Die Punktzahlen steigen, weil schwache oder unerfahrene Nutzer in nachfolgenden Wellen aussteigen.

CH08-F22— NEGATIVE ANTWORTEN NICHT ANERKENNEN

Der Teilnehmer wird ersetzt, weil die KI eine falsche Antwort gegeben hat.

CH08-F23— KORREKTE ANTWORT BESTÄTIGEN

Selbst wenn der Teilnehmer die Antwort aktualisiert hat, wird der Eintrag behalten, weil das Ergebnis korrekt ist.

CH08-F24— ÜBERBLICKUNG VON BOTS UND WIEDERHOLTER TEILNAHME

Die gleiche Person oder Automatisierung erzeugt eine große Anzahl von Beobachtungen.

CH08-F25— MASSENAUSSCHLUSS MIT EINEM EINZELNEN SIGNAL

VPN, Gerät oder Geschwindigkeit führt dazu, dass echte Nutzer als falsch angesehen werden.

CH08-F26— DURCHFÜHRUNG EINER BETRUGSPRÜFUNG BASIEREND AUF ERGEBNISSEN

Datensätze mit niedriger Punktzahl werden häufiger überprüft.

CH08-F27— ERSTELLUNG EINER STARK UNGESICHTETEN GRUPPE

Es wird behauptet, dass Nutzer, die nicht in der Stichprobe enthalten sind, repräsentiert werden.

CH08-F28— AUSWAHL DER NACH-ERGEBNIS-KALIBRIERUNG

Gewichtsvariablen, die den am besten geeigneten Score liefern, werden danach ausgewählt.

CH08-F29— ZÄHLUNG DES ROH-PANEL-ANTEILS ALS POPULATIONSTEIL

Schwere KI-Nutzer bestimmen das globale Ergebnis anhand roher Zahlen.

CH08-F30— ZÄHLUNG UNGEWÖHNLICHER GEWICHTE ALS DESIGN-GEWICHT

Die genaue Bevölkerungsableitung wird durchgeführt, auch wenn die tatsächliche Auswahlwahrscheinlichkeit unbekannt ist.

CH08-F31— DAS UNGEWICHTETE ERGEBNIS VERBORGEN HALTEN

Die Wirkung des Gewichtungsmodells ist nicht sichtbar.

CH08-F32— DAS EFFEKTIVE SAMPLE VERSTECKEN

Obwohl 1,000-Rohdatensätze 300–400-Informationseinheiten enthalten, wird volle Präzision behauptet.

CH08-F33— ENTFERNUNG SCHWACHER GRUPPEN DURCH GEWICHTSANPASSUNG

Kleinere Gruppen mit hohem Gewicht werden entfernt, da sie das Ergebnis herabsetzen.

CH08-F34— ZÄHLEN MODELLBASIERTER GESCHÄTZTER PUNKTE

Ein Land ohne Beispiel erhält eine feste Bewertung.

CH08-F35— OFFENLEGUNG DER IDENTITÄT FÜR NACHWEISE

Screenshots werden mit persönlichen Informationen öffentlich veröffentlicht.

CH08-F36— WIEDERERKENNBARE MACHUNG EINES TEILNEHMERS AUS EINEM KLEINEN LAND

Land, Zeit, Gerät und Planinformationen machen eine einzelne Person erkennbar.

CH08-F37— FRAGEN DES TEILNEHMERS NACH PASSWORT ODER PRIVATEM CHAT

Das Sammeln von Nachweisen verletzt die Kontosicherheit.

CH08-F38— ZÄHLUNG GERINGER DIGITALER FÄHIGKEIT ALS UNANGEMESSENHEIT

Tatsächliche Zielbenutzer werden aus dem Panel entfernt.

CH08-F39— STÖRUNG DER FREIWILLIGKEIT DURCH BELONUNG

Der Teilnehmer steht unter Druck, dem er vernünftigerweise nicht widerstehen kann.

CH08-F40— ÜBERLASSEN SIE DIE WAHL DES TEILNEHMERS DEM KUNDEN

Die geprüfte Einheit wählt die Personen aus, die das Ergebnis des Panels bestimmen können.

53. PRÜFUNGSVERFAHREN

Schritt 1— Teilnehmerquellen kartieren

Jedes Panel wird separat nach Partner-, Open-Call- und Kundenquelle aufgezeichnet.

Schritt 2— Auswahlstatus zuweisen

Für jede Quelle wird ein geeigneter Status zwischen PSI-0 und PSI-5 bestimmt.

Schritt 3— Abdeckung mit Zielpopulation vergleichen

Bestimmen Sie, welche Zielnutzer im Panel nicht vertreten sind.

Schritt 4— Wahrscheinlichkeit oder Modell der Teilnehmerauswahl speichern

Bekannte Wahrscheinlichkeit, Quote oder freiwillige Auswahl wird erklärt.

Schritt 5— Konformitätsregeln einfrieren

Alter, Land, Sprache, Produkt, Plan, Schnittstelle, Einzigartigkeit, Extraktionsbeziehung und Bedingungen werden versioniert.

Schritt 6— Einladungstext überprüfen

Es wird bewertet, ob es positive oder negative Anleitungen gibt.

Schritt 7— Belohnungsstruktur untersuchen

Es wird bestätigt, dass die Belohnung völlig unabhängig vom Ergebnis ist.

Schritt 8— Land, Sprache und Produktzugang der Teilnehmer überprüfen

Datenschutzsichere Überprüfungsstufen werden zugewiesen.

Schritt 9— Einzigartige menschliche Überprüfung durchführen

Interne und externe Wiederholungen werden untersucht.

Schritt 10— Beziehungen und Vertrautheitsbereiche aufzeichnen

Kunden-, Mitarbeiter-, Wettbewerber-, KI-Anbieter- und Markenvertrautheit werden bestimmt.

Schritt 11— Aufgaben nach Zuweisungsplan verteilen

Produkt-, Eingabeaufforderungs-, Warteschlangen- und Wellenzuweisungen werden überprüft.

Schritt 12— Panel-Konditionierung aufzeichnen

Neue, wiederholte und zuvor teilnehmende Nutzer am gleichen Test werden getrennt.

Schritt 13— Teilnahme-Trichter einrichten

Einladungs-, Annahme-, Start-, Abschluss- und Gültigkeitszahlen werden berechnet.

Schritt 14— Gültigkeitsprüfung blind durchführen

Die Protokollgültigkeit wird bestimmt, ohne dass die KI-Antwortbewertung so weit wie möglich sichtbar ist.

Schritt 15— Betrugssignale überprüfen

Mehrere Signale und menschliche Bewertungen werden verwendet.

Schritt 16— Meine Nicht-Antworten und Panelverluste analysieren

Die bekannten Merkmale derjenigen, die teilnehmen, und derjenigen, die nicht teilnehmen, werden verglichen.

Schritt 17— Erzeugen von Design- und Kalibrierungsgewichten

Probabilistische und nicht-probabilistische Quellen werden mit separaten Methoden behandelt.

Schritt 18— Berechnen der Gewichtsdaten

Gewichtsverteilung, Design-Effekt, effektive Stichprobe und Trimm-Effekt werden berechnet.

Schritt 19— Durchführung der Sensitivitätsanalyse

Roh-, Haupt-, stark gewichtete und alternative vernünftige Ergebnisse werden verglichen.

Schritt 20— Erstellen des Datenschutz- und Nachweis-Manifests

Vollständiger Nachweis befindet sich in einem sicheren Archiv; anonyme Beweisidentitäten werden in einem öffentlichen Verzeichnis geführt.

Schritt 21— Veröffentlichung der öffentlichen Wahlkarte

Panelquellen, Teilnahme-Trichter, Verzerrungen und Gewichtungseffekt werden sichtbar gemacht.

54. ERFORDERLICHE BELEGE

Zielgruppen-Aufzeichnung; Stichprobenrahmen; Panel-Anbieter; Teilnehmerquellen; Auswahlstatus; Wahrscheinlichkeits- oder Nicht-Wahrscheinlichkeits-Panelklasse; Panel-Registrierungsmethode; Einladungstext; Zulassungsregeln für Teilnehmer; Teilnehmerländerüberprüfung; Sprachkenntnisnachweis; Zugriff auf KI-Produkt-Aufzeichnung; Plan- und Nutzeroberfläche; eindeutige Personenverifizierung; Quellübergreifende Duplikatsprüfungen; Markenbekanntheit; Häufigkeit der KI-Nutzung; Beziehung zur geprüften Einheit; Beziehung zum KI-Anbieter; Teilnehmervergütung; Bonusbedingungen; Aufgabenvergabe-Methode; Zufallsauswahl- und Blockierungsaufzeichnung; Produkt- und Aufforderungsreihenfolge; Panel-Konditionierung; vorherige Pilotteilnahme; Anzahl Eingeladener; Anzahl Zustimmender; Anzahl derjenigen, die mit der Aufgabe beginnen; Anzahl derjenigen, die sie abschließen; Anzahl der gültigen Beobachtungen; Gründe für Nichtantworten; Panel-Abwanderung; Gründe für Ungültigkeit; und Betrugsindikatoren.

Überprüfungs- und Einwandaufzeichnungen Designgewichte Nicht-Antwort-Einstellungen Kalibrierungsvariablen Propensitätsmodell, falls vorhanden Gewichtszuschnittaufzeichnung Effektive Stichprobengröße Sensitivitätsergebnisse Panelanbieter-Effekt Ungewichtete und gewichtete Werte Zustimmungserklärung Aufbewahrungs- und Zugriffsrichtlinie für Daten Regeln zur Schwärzung Öffentliche Evidenzführung Verantwortliche Person oder Institution

55. PRÜFLISTE FÜR AUDITS

Aus welchen Quellen stammten die Teilnehmer? War jede Quelle probabilistisch, freiwillig oder ein offener Aufruf? Wurde die Zufälligkeit innerhalb des Panels wie die Zufälligkeit in der Weltbevölkerung dargestellt? Intervenierte die betreuende Institution bei der Teilnehmerauswahl?

Sind Kunden- und Mitarbeiterteilnehmer getrennt? Wurden Beziehungen zu KI-Anbietern oder Wettbewerbern dokumentiert? Wie wurde die Ländereignung überprüft? Entspricht die Sprachkompetenz dem tatsächlichen Gebrauch? Wurde das richtige KI-Produkt, der richtige Tarifplan und die richtige Schnittstelle verwendet? An wie vielen Konten oder Panels hat dieselbe Person teilgenommen? Wurde eine Überprüfung zwischen den Quellen durchgeführt? Führte der Einladungstext das Ergebnis? Hängt die Belohnung vom Inhalt der KI-Antwort ab? Haben Geschwindigkeitsanreize die Protokollqualität verringert? War die Aufgabenverteilung zufällig oder regelbasiert? Hat der Teilnehmer das Produkt oder die Prompt selbst ausgewählt?

Waren die Reihenfolgen von Produkt und Prompt ausbalanciert? Wurde die Panel-Konditionierung aufgezeichnet? Haben die Teilnehmer die Ergebnisse der Aufgaben geteilt? Ist der Einladungs–Akzeptanz–Start–Abschluss–Gültigkeits-Funnel sichtbar? Unterscheiden sich Nicht-Antworter in bekannten Merkmalen? Beeinflusst Panel-Abwanderung den Zeitverlauf? Wurde die Gültigkeitsentscheidung getroffen, bevor die KI-Antwort bewertet wurde? Wurden negative oder kritische Antworten häufiger als ungültig betrachtet? Trägt die Betrugserkennung mehrere Signale? Führten VPN oder gemeinsame Geräte automatisch zum Ausschluss? Wurde dem Teilnehmer ein Einspruchsweg gegeben? Wurden Gewichtungsvariablen vor dem Ergebnis definiert? Wurden unwahrscheinliche Panelgewichte korrekt gekennzeichnet? Sind die Zielsummen zuverlässig? Werden ungewichtete und gewichtete Ergebnisse zusammen präsentiert? Ist die effektive Stichprobengröße klar? Wurde die Auswirkung des Gewichtungs-Trimmings gezeigt?

Wurden Gruppen, die nie beobachtet wurden, als weitgehend vertreten repräsentiert? Wurden modellbasierte Vorhersagen so präsentiert, als wären sie direkte Beobachtungen? Gibt es persönliche Details in Teilnehmer-Screenshots? Können Nutzer aus kleinen Ländern wieder identifiziert werden? Wurden Kontopasswörter oder private Chats angefragt? Ist die Zustimmung der Teilnehmer klar und verständlich? Gibt es einen klar verantwortlichen Eigentümer für die Wahl der Teilnehmer und die Gewichtungsaufzeichnungen?

56. EINWÄNDE UND ANTWORTEN

Einwand 1— "Wenn wir ein professionelles Forschungspanel nutzen, warum sprechen wir dann immer noch über Voreingenommenheit?"

Ein professionelles Panel kann die Datenqualität und das Feldmanagement stärken. Panelmitglieder sind außerdem:

- freiwillig
- an Forschung gewöhnt,
- unterschiedlich in der Nutzung von Technologie

Es ist möglich. Professionalität beseitigt keine Voreingenommenheit. Sie macht das Messen und Verwalten einfacher.

Einwand 2— „Wenn die Alters- und Länderquoten erfüllt sind, warum sollte das Gremium dann nicht als repräsentativ angesehen werden?“

Teilnehmer:

- Häufigkeit der KI-Nutzung,
- Gerät,
- Plan,
- Markenbewusstsein,
- digitale Kompetenz

Es kann in Bezug auf die Wartung unterschiedlich sein. Quoten gleichen nur die Bereiche aus, die sie abdecken.

Einwand 3— „Können Gewichtungen diese Unterschiede nicht korrigieren?“

Sie können teilweise Unterschiede korrigieren, für die gemessene und verlässliche Zielwerte existieren. Sie können Unterschiede, die nie beobachtet oder gemessen werden, nicht definitiv korrigieren.

Einwand 4— „Die Klienten der Institution sind echte Nutzer; warum sollten sie nicht im Hauptpanel enthalten sein?“

Sie können berücksichtigt werden. Allerdings sollten sie mit dem tatsächlichen Anteil und dem Beziehungsstatus berücksichtigt werden, den sie in der Gesamtbevölkerung haben. Die gesamte Kundenliste ist kein Panel der Gesamtbevölkerung.

Einwand 5— „Warum wäre es ein Problem, wenn der Teilnehmer die Marke kennt?“

Es muss kein Problem sein. Sie sind Teil der natürlichen Nutzerpopulation. Wenn die Vertrautheit unausgewogen ist, kann das Ergebnis variieren. Daher wird es aufgezeichnet und bei Bedarf analysiert.

Einwand 6— „Schwere KI-Nutzer führen die Aufgabe genauer aus; wäre es nicht besser, sie auszuwählen?“

Es könnte einfacher für Protokolltests sein. Wenn wir jedoch die tatsächliche Erfahrung der Bevölkerung messen wollen, sind auch unerfahrene Nutzer wichtig. Wenn die Aufgabe zu komplex ist, sollte das Messinstrument korrigiert werden. Die Bevölkerung sollte nicht vereinfacht werden.

Einwand 7— 'Wenn der Teilnehmer die Antwort missverstanden hat, warum sollten wir die Beobachtung behalten?'

Die Aufgabe des Teilnehmers besteht nicht darin, die Antwort zu bewerten, sondern sie korrekt zu erfassen. Wenn das Erfassungsprotokoll korrekt angewendet wird, bleibt die KI-Antwort erhalten. Die Beurteilung des Materials liegt in der Verantwortung der Schiedsrichter.

Einwand 8— „Sollte nicht ein strenger automatischer Filter verwendet werden, um gefälschte Teilnahme zu verhindern?“

Automatisches Filtern ist nützlich. Wenn es die einzige endgültige Entscheidung wird, könnte es echte Nutzer unfair ausschließen. Mehrere Signale und ein angemessenes menschliches Überprüfungsverfahren sind erforderlich.

Einwand 9— 'Wenn wir keine tausend Screenshots veröffentlichen, kann die Methode als bewiesen gelten?'

Vollbildbilder:

- persönliche Informationen,
- Kontosicherheit,
- Neudefinition

kann ein Risiko bergen. Für die Öffentlichkeit:

- bearbeitete Beispiele,
- Beobachtungsmanifest,
- Integritätshashes,

- ein unabhängiger Prüfungszugang

kann bereitgestellt werden. Nachweisbarkeit ist keine Identitätsfreigabe.

Einwand 10— "Wenn der gewichtete Wert sehr unterschiedlich vom Rohwert ist, welcher ist korrekt?"

Zwei Ergebnisse beantworten unterschiedliche Fragen. Die Rohpunktzahl ist das Ergebnis der Personen im Panel. Die gewichtete Punktzahl ist die Schätzung für die Zielpopulation. Ein großer Unterschied zeigt, dass die Zusammensetzung des Panels wichtig ist. Beide sollten veröffentlicht werden.

Einwand 11— „Wenn die effektive Stichprobe auf 349 gesunken ist, war es dann verschwendet, 1,000 Beobachtungen zu sammeln?“

Die überabgetasteten Gruppen könnten die Analyse der Untergruppen gestärkt haben. Allerdings kann die Sensitivität der globalen Prognose niedriger sein als die Rohzahl von 1,000. Diese Tatsache sollte nicht verborgen werden.

Einwand 12— „Kann das Nicht-Wahrscheinlichkeits-Panel überhaupt nicht verwendet werden?“

Es kann verwendet werden. Im Speziellen:

- Pilot,
- Sprache mit geringen Ressourcen,
- kleines Land,
- Schnelle Entdeckung

Es ist wertvoll für. Es sollte mit dem korrekten Status und den Einschränkungen veröffentlicht werden.

Einspruch 13— „Ist es nicht zuverlässiger, alle Teilnehmer aus neuen Personen auszuwählen?“

Es kann stark für die Querschnittsbevölkerungsschätzung sein. Wiederholte Nutzer sind wertvoll für die Messung longitudinaler Veränderungen. Neue und wiederholte Panelrollen sollten getrennt gehalten werden.

Einwand 14— „Erhöht das Sammeln von Daten von so vielen Teilnehmern nicht das Datenschutzrisiko?“

Es kann. Deshalb:

- Datenminimierung
- anonyme Identität
- Bearbeitung,
- Zugangskontrolle
- Lagerdauer

sind verpflichtende Teile des Protokolls. Unnötige Informationen sollten nicht gesammelt werden.

ALLGEMEINE VORSCHRIFT DES ABSCHNITTS 59

Die einfachste Zahl in einer GEO-Studie:

1,000

Es ist möglich. Die schwerste Frage ist diese:

> *Welcher 1,000?*

Zwei Panels mit denselben Länder- und Altersquoten können völlig unterschiedliche Ergebnisse liefern. In einem:

- intensive KI-Nutzer,
- Abonnenten eines kostenpflichtigen Plans,
- Technologiearbeiter

sind die Mehrheit. Im anderen:

- seltene Nutzer,
- Mobile-Nutzer,
- Menschen mit geringen digitalen Fähigkeiten

Es ist offensichtlicher. Beide sind 1,000 Personen. Es ist nicht die gleiche Population. Wenn Sie verbergen, woher die Teilnehmer kommen, und nur das gewichtete Ergebnis veröffentlichen, verwandeln Sie Gewichtung in einen Schleier des Vertrauens anstatt in eine Kontrollmethode. Sie können eine Person, die in einer Stichprobe nicht vorhanden ist, nicht gewichten. Sie können die Antwort eines Nutzers nicht ändern, weil sie falsch ist. Eine falsche Antwort ist kein Fehler der Forschung. Es ist das Ereignis, das die Forschung zu messen versucht. Wenn ein Teilnehmer das Protokoll trotz der richtigen Antwort falsch befolgt hat, können Sie den Datensatz nicht erhalten. Eine richtige Antwort macht eine gefälschte Beobachtung nicht real. Sie können einer Person mehr bezahlen. Aber nicht, um die Antwort zu erhalten, die Sie wollen. Sie können entsprechend den Lebensbedingungen eines Landes bezahlen. Aber Sie können die Bereitschaft des Teilnehmers nicht kaufen. Tausend Screenshots können als Nachweis dienen. Sie sind nicht verpflichtet, die Identität von tausend Personen der Öffentlichkeit preiszugeben.

Die Integrität von Evidenzmitteln und die Privatsphäre des Menschen können gemeinsam geschützt werden. Daher lautet das achte Messgesetz von NOMOS wie folgt:

> *Die Aussagekraft der Stichprobe wird nicht nur davon bestimmt, aus wie vielen Personen sie besteht, sondern auch davon, warum und wie diese Personen ausgewählt wurden.*

Das neunte Gesetz lautet wie folgt:

> *Gewichtung stellt die beobachtete Bevölkerung wieder ins Gleichgewicht; sie erzeugt nicht den nicht beobachteten Menschen.*

Das zehnte Gesetz lautet:

> *Eine falsche KI-Antwort kann gültige Daten sein; eine richtige KI-Antwort kann ungültige Daten sein.*

Das elfte Gesetz lautet wie folgt:

> *Die Belohnung des Teilnehmers sollte nicht für das Finden des gewünschten Ergebnisses erfolgen, sondern für den Nachweis des beobachteten Ergebnisses, ohne es zu verändern.*

Das zwölfte Gesetz lautet wie folgt:

> *Beweisbarkeit erfordert nicht, dass der Teilnehmer seine Privatsphäre aufgibt.*

NOMOS' Kapitel 8 Gebot

> *Bringe mir nicht allein tausend Menschen. / Zeige mir, aus welcher Welt diese tausend Menschen kommen.*

Behandle ein freiwilliges Panel nicht als zufällige Bevölkerungsstichprobe, eine Kundenliste nicht als allgemeine Öffentlichkeit und einen Mitarbeiter nicht als unabhängigen Nutzer.

Denke nicht, dass alle Unterschiede, die du nicht gemessen hast, sich verbessert haben, nur weil die Länder- und Altersquoten erfüllt wurden.

Multipliziere eine Person nicht mit der Anzahl ihrer Konten, Geräte, Sprachen oder Panelmitgliedschaften.

Stecke diejenigen, die die Marke lieben, diejenigen, die die Marke hassen, und diejenigen, die die Marke überhaupt nicht kennen, nicht in dieselbe unsichtbare Box.

Bezahle den Teilnehmer nicht dafür, die richtige Antwort zu finden, sondern dafür, die erste passende Antwort, die er sieht, zu speichern, ohne sie zu ändern.

Ändere den Nutzer nicht, wenn die KI eine falsche Antwort gibt. / Überprüfe nicht, ob die Antwort korrekt ist, wenn der Nutzer das Protokoll bricht.

Lass eine Person, die nicht im Panel ist, nicht so gewichtet erscheinen, als würde sie existieren.

Verstecke das Rohresultat nicht. / Erkläre nicht nur das gewichtete Ergebnis als die Wahrheit. / Zeige die effektive Stichprobe. / Zeige ihre Sensitivität.

Erklären Sie niemanden, der ein VPN benutzt, zum Fake, niemanden, der ein gemeinsam genutztes Gerät benutzt, zum Bot, oder jemandem, der assistive Technologie benutzt, zur Automatisierung.

Untersuchen Sie Betrug. / Aber machen Sie die Untersuchung selbst nicht zu einem neuen Ausschlussbias.

Machen Sie einen Screenshot. / Schützen Sie den Namen der Person, das Konto und das Privatleben.

Definieren Sie zuerst die Population. / Dann zeigen Sie das Rahmenwerk. / Dann erklären Sie, wen Sie ausgewählt haben, wer nicht gekommen ist, wer es nicht abschließen konnte und warum jemand ausgeschlossen wurde. / Und erst danach entscheiden Sie, wie stark Ihr Ergebnis für die Menschheit sprechen kann.

Der abschließende Satz des Kapitels

In GEO-1000 ist ein Teilnehmer mehr als eine Person, die eine Antwort trägt. Ein Teilnehmer ist eine menschliche Stichprobeinheit, deren Fähigkeit, andere in der Zielpopulation zu vertreten, durch den Auswahlpfad, Gewichtung, Unabhängigkeit und Datenschutzmaßnahmen gerechtfertigt sein muss.

Normativer Kern

Jede GEO-1000-Beobachtung MUSS einen prüfbaren Teilnehmerauswahlpfad von der Zielpopulation und dem Stichprobenrahmen über Auswahl, Einladung, Einwilligung, Zuordnung, Abschluss, Protokollvalidierung, Ausschluss bis hin zur endgültigen Gewichtung beibehalten. Die zufällige Auswahl innerhalb des Panels DARF NICHT als zufällige Auswahl aus der globalen Zielpopulation repräsentiert werden, es sei denn, der Mechanismus zur Einbeziehung der gesamten Population unterstützt diese Behauptung. Die Teilnehmerrekrutierung, Zuordnung, Vergütung, Validierung, Ersetzung, Ausschluss und Gewichtung MÜSSEN unabhängig davon bleiben, ob die KI-Antwort positiv, negativ, korrekt, inkorrekt, abgelehnt, kritisch oder kommerziell günstig ist. Personen, Konten, Geräte, Sitzungen, Sprachen, Panelmitgliedschaften und KI-Produktbeobachtungen MÜSSEN als eigenständige Einheiten erhalten bleiben. Wahrscheinlichkeits-, Opt-in-, Open-Call-, vom Kunden bereitgestellte, institutionelle und hybride Teilnehmerquellen MÜSSEN separat identifiziert werden. Gewichtung KANN Ungleichgewichte verringern, DARF jedoch nicht so repräsentiert werden, dass sie Beobachtungen für Populationen erzeugt, die im Stichprobenrahmen fehlen, oder nicht gemessene Auswahl mit Sicherheit korrigiert. Ungewichtete Ergebnisse, gewichtete Ergebnisse, Gewichtungsdiagnosen, Abdeckungslücken, Nichtantworten, effektive Stichprobengröße und Sensitivitätsanalysen MÜSSEN sichtbar bleiben. Entscheidungen über Betrug und Duplikation MÜSSEN eine verhältnismäßige Überprüfung mit mehreren Signalen verwenden und DÜRFEN die Nutzung von VPNs, geteilten Geräten, institutionellen Netzwerken oder Barrierefreiheits-Tools NICHT automatisch als Fälschung behandeln. Nachweise von Teilnehmern MÜSSEN überprüfbar bleiben, ohne unnötig Identität, Kontodaten, private Gespräche oder genauen Standort offenzulegen. Jede Entscheidung über die Auswahl und Gewichtung von Teilnehmenden MUSS versioniert und einem verantwortlichen Menschen oder einer Organisation zuordenbar sein.

TEIL III

Durchführung des Experiments und Evidenzkette

KAPITEL 09

KI-Systemregister und Vergleichbarkeit

METHODISCHE GRUNDLAGEN / K02; K03

Kapitelgrenze

Die ersten acht Kapitel legten Folgendes fest: eine einzelne Antwort ist keine GEO-Bewertung; Repräsentation ist eine Verteilung; die geprüfte Einheit muss identifiziert werden, bevor Ergebnisse beobachtet werden; offizielle Repräsentation und überprüfte Realität der Einheit müssen getrennt bleiben; für jedes KI-Produkt muss eine berechtigte Nutzerpopulation definiert werden; 1,000 gültige Beobachtungen müssen über ein populationsbasiertes Stichprobendesign zugewiesen werden; kleinere Länder und Sprachen müssen durch ergänzende Panels sichtbar bleiben; und die Auswahl, der Ausschluss und die Gewichtung der Teilnehmer müssen unabhängig vom Antwortinhalt bleiben. Wir müssen nun die andere Seite des Experiments definieren:

> *Welches KI-System messen die Teilnehmer tatsächlich?*

Der Name eines Anbieters? Ein Basismodell? Ein Chat-Produkt? Eine kostenlose Web-Oberfläche? Eine kostenpflichtige mobile App? Eine Sitzung mit aktivierter Websuche? Ein Antwortmodus, der keine Quellen zitiert? Ein Unternehmensarbeitsbereich unterstützt durch interne Daten? Ein Konto mit aktivierter Erinnerung? Temporärer oder sauberer Chat? Sind unterschiedliche Nutzeroberflächen desselben Anbieters dasselbe System? Sollten zwei Produkte, die dieselbe Modellfamilie verwenden, auf dieselbe Weise bewertet werden? Selbst wenn ein Produkt denselben Modellnamen auf dem Bildschirm anzeigt:

- Abruf,
- Quellenauswahl,
- Sicherheitsregeln,
- Personalisierung,
- Werkzeugzugriff,
- experimentelle Funktionen

Wenn sich dies ändert, können die Ergebnisse verglichen werden? Wenn ein KI-Produkt während einer Messungswelle aktualisiert wird, können die ersten und letzten Antworten der Welle als aus demselben System stammend betrachtet werden? Ein Nutzer:

- Websuche ist aktiviert,
- Erinnerung ist aktiviert,
- hat vorherige Gespräche,
- kostenpflichtiger Tarif

Antworten; eines anderen Nutzers:

- Suche ist aus,
- Speicher ist aus,
- Neue Sitzung,
- Kostenloser Plan

Kann es in derselben Zelle wie seine Antwort zusammengeführt werden? Nein. Das in den vorherigen Abschnitten festgelegte Messprinzip ist klar: Es gibt keine einzelne „KI“; der Suchindex, das Abrufsystem, der Rangierer, das Sprachmodell, die Zitationsschicht und das Sicherheitssystem können sich unterschiedlich verhalten. Dasselbe Framework erfordert, dass jedes überprüfbare Ergebnis trägt:

- Modell oder Produkt,
- Datum,
- Land,
- Sprache,
- Abfragesatz,
- Anzahl der Wiederholungen,
- Messaufzeichnung

Dieser Abschnitt wendet diese Prinzipien innerhalb eines KI-Systemregisters an. Dieser Abschnitt:

- die Unterscheidung zwischen Anbieter, Produkt, Modell und Nutzeroberfläche,
- KI-Produktinstanz,
- Vertrauensstufen der Modellidentität,
- Abruf- und Werkzeugkonfigurationen,
- Speicher- und Personalisierungszustände,
- Unterschiede in Plan, Schnittstelle und experimentellen Funktionen,
- Produktänderungen,
- Systemkonfiguration zum Zeitpunkt der Messung,
- Vergleichbarkeitsgrade zwischen Produkten,
- Vergleiche von Natürliche Reichweite und Gemeinsame Vergleichsbasis,
- Versionsabweichung im Laufe der Zeit,

definiert. Dieses Kapitel behandelt noch nicht:

- vollständiges Übersetzungs- und semantisches Äquivalenzsystem von Aufforderungen,
- alle Details kontrollierter und natürlicher Nutzerpanels,
- NOMOS Capture-Softwarearchitektur,
- die atomare Aussagebewertung der Antworten,
- die endgültigen Modellkoeffizienten und die NOMOS-Bewertungsformel

wird nicht abgeschlossen. Die zentrale Frage von Abschnitt 9 lautet:

> *Bevor man zwei Antworten auf denselben Prompt vergleicht, wie können wir beweisen, dass diese Antworten tatsächlich unter vergleichbaren KI-Produktbedingungen entstanden sind?*

NOMOS Herausforderung

Zwei Nutzer desselben Anbieters stellen dieselbe Frage: „Was für ein Unternehmen ist Apple.com?“ Erster Nutzer:

- verwendet das kostenlose Webprodukt,

- öffnet einen neuen Chat,
- hat den Speicher aus,
- hat keinen Internetzugang,

erhält eine Antwort nur mit der eingebauten Wissenskapazität des Produkts. Zweiter Nutzer:

- Verwendung eines kostenpflichtigen mobilen Produkts,
- Erinnerung ist aktiviert,
- hat zuvor über Apple gesprochen,
- Websuche ist aktiviert,

erhält Antworten mit aktuellen Quellen und Zitaten. Erste Antwort: „Apple ist ein Technologieunternehmen, das Unterhaltungselektronik herstellt." Zweite Antwort: „Apple ist ein globales Technologieunternehmen, das Hardware, Software und digitale Dienstleistungen anbietet. Sie können aktuelle Unternehmensinformationen in diesen Quellen einsehen." Beide Antworten stammen vom selben Anbieter.

Ist dies das Ergebnis desselben Systems? Nein. Betrachten wir nun zwei andere Nutzer. Beide sehen denselben Modellnamen auf dem Bildschirm. In dem Land des einen Nutzers ist die Web-Funktion aktiviert. In dem Land des anderen Nutzers ist sie deaktiviert. In der Antwort des einen Nutzers sind Quellenlinks sichtbar. Die Nutzeroberfläche des anderen Nutzers zeigt keine Zitationen, selbst wenn dieselben Quellen verwendet werden. Das gleiche Modell-Label ist vorhanden.

Gibt es dieselbe Messzelle? Nein. Betrachten wir nun zwei Produkte von verschiedenen Anbietern. Beide:

- neuer Chat,
- Speicher ist aus,
- gleiche Sprache,
- gleiches Land,
- gleicher Zeitrahmen,
- Webzugang aktiviert,
- gleiche Prompt-Version

werden unter diesen Bedingungen gemessen. Sie basieren auf unterschiedlichen zugrunde liegenden Modellen. Sind sie vergleichbar? Ja, sie können in Bezug auf ein spezifisches Nutzerergebnis vergleichbar sein. Denn der Vergleichsgegenstand ist:

> *nicht, ob sie dasselbe zugrunde liegende Modell verwenden, sondern welche Repräsentation sie unter denselben definierten Nutzerbedingungen erzeugen.*

Nun soll die Produktversion für die ersten 500 Beobachtungen der Messwelle unverändert bleiben. Während der nächsten 500 Beobachtungen führt der Anbieter:

- Änderungen am Abrufsystem durch,
- Sicherheitsregeln,
- Nutzeroberfläche

Der Bericht allein sagt: „1,000 Nutzer haben dasselbe KI-Produkt getestet." In Wirklichkeit könnten zwei verschiedene Produktsituationen gemessen worden sein. Die erste Bestimmung dieses Abschnitts lautet wie folgt:

> *Der Name des Anbieters reicht nicht aus, um das gemessene KI-System zu identifizieren.*

Ihre zweite Bestimmung lautet:

> *Der Modellname ist nicht das gesamte Produktverhalten, das den Nutzer erreicht.*

Seine dritte Bestimmung besagt:

> *Der gleiche Anbieter, die gleiche Modellbezeichnung oder das gleiche Logo garantiert keine Vergleichbarkeit.*

Seine vierte Bestimmung besagt:

> *Verschiedene Anbieter können hinsichtlich der Nutzerergebnisse vergleichbar sein, wenn die Messbedingungen ausreichend abgestimmt sind.*

Seine fünfte Vorschrift lautet:

> *Wenn sich ein Produkt während der Messung wesentlich verändert, kann es nicht als ein einzelnes festes System berichtet werden.*

1. ZWECK DES KAPITELS

Zweck dieses Abschnitts ist es, das im Rahmen von GEO-1000 geprüfte KI-Produkt mit all seinen wesentlichen Konfigurationen zu definieren und festzulegen, unter welchen Bedingungen unterschiedliche Produktergebnisse verglichen werden können. Der Abschnitt standardisiert die folgenden Unterscheidungen:

- Anbieter und KI-Produkt
- KI-Produkt und Basismodell
- Modellfamilie und spezifische Bereitstellung
- Modellbezeichnung und überprüfte Identität des Modells
- Nutzeroberfläche und API
- Web- und Mobilprodukt
- Kostenloser Plan und kostenpflichtiger Plan
- Verbraucherprodukt und Unternehmensprodukt
- Grundmodelantwort und abrufunterstützte Antwort
- Live-Websuche mit Abruf
- Anzeige von Zitaten mit Quellenangabe
- Werkzeugnutzung mit Werkzeugzugang
- Sitzungskontext mit Speicher
- Systemrichtlinie mit benutzerdefinierter Anweisung
- Materielle Realität mit Personalisierung
- Temporärer Chat mit neuer Konversation
- Produktstatus über die Zeit mit statischem Produktnamen
- Beobachtete Verhaltensänderung mit vom Anbieter angekündigtem Update
- Vergleich von Natürliche Reichweite und Gemeinsame Vergleichsbasis
- Vergleichbares Produktbeispiel innerhalb derselben Produktfamilie
- Nutzerergebnis versus Modell-interne abstrakte Fähigkeit
- Produktabwesenheit versus Produktfehler
- Technischer Fehler versus tatsächliche Ablehnung
- Dasselbe Prompt versus dasselbe Experiment
- Dasselbe Mittel versus dasselbe Systemverhalten
- Sofortiger Produkt-Snapshot versus langfristige Produktleistung

Am Ende dieses Abschnitts sollte jedes GEO-1000-Audit in der Lage sein, die folgenden Fragen zu beantworten:

> *Welcher Anbieter, welches Produkt, welche Oberfläche, welcher Plan und welche Konfiguration wurden gemessen?*
>
> *Wie stark wurde der auf dem Bildschirm gesehenen Modellname wirklich validiert?*
>
> *Was waren Web-, Abruf-, Zitier-, Speicher-, Personalisierungs- und Werkzeugstatus?*
>
> *Blieb das Produkt während der gesamten Messung unverändert?*
>
> *Auf welcher Vergleichbarkeitsebene wurden die beiden Produktergebnisse nebeneinander gestellt?*
>
> *Wird das Verglichene das Basismodell, das Verbraucherprodukt, die Nutzeroberfläche oder die gesamte Systemkette sein?*

2. ZENTRALE NORMATIVE BESTIMMUNG

> *Jede GEO-1000-Beobachtung muss nicht nur mit dem Anbieter oder Modellnamen verknüpft sein; sie sollte mit einem versionierten KI-Systemregistereintrag verbunden sein, der die AI-Produktinstanz zum Zeitpunkt der Messung identifiziert.*

Eine AI-Produktinstanz muss mindestens die folgenden Felder tragen, soweit relevant:

- Anbieter
- Produktname
- Nutzeroberfläche
- Produkttyp
- Plan- oder Kontoklasse
- Offensichtliches Modelllabel
- Modell-ID-Vertrauensstufe
- Produkt- oder Bereitstellungsversion
- Land und Gebietsschema
- Sprache
- Web- oder Abrufzugang
- Zitationsfunktion
- Werkzeugzugang
- Speicherstatus
- Status benutzerdefinierter Anweisungen
- Vorheriger Chat-Kontext
- Sitzungstyp
- Personalisierungsstatus
- Sicherheits- oder Richtlinienmodus, falls bekannt
- Experimentelle Funktionen
- Messdatum und UTC Uhrzeit
- Messwelle
- Produkt-Änderungsprotokoll
- Evidenz- und Aufzeichnungsvertrauen

Wenn eines dieser Felder sich wesentlich ändert:

- neue KI-Produktinstanz,
- neue experimentelle Zelle,
- neue Analyseversion

kann erforderlich sein.

3. „KI-SYSTEM“ IST KEIN EINZELNES OBJEKT [K02; K03]

In diesem Protokoll ist „KI-System“ ein allgemeiner und übergeordneter Begriff. Eine echte Nutzerantwort kann aus der Kombination mehrerer Schichten entstehen. Kandidatenkette:

```
R= F(M, O, K, Q, T, P, S, U, C, L, t)
```

Hier:

- M: Basis- oder Leitmodell
- O: Produktschestrierungsschicht
- K: Abruf- oder Wissensquellenschicht
- Q: Neubewertung und Quellenauswahl
- T: Werkzeuge
- P: Richtlinien- und Sicherheitsschicht
- S: Sitzung, Speicher und Personalisierung
- U: Nutzerkonto und Plan
- C: Land und Gebietsschema
- L: Sprache
- t: Zeit

Vom Nutzer gesehene Antwort:

R

Die für einen Nutzer sichtbare Antwort ist nicht unbedingt nur das Ergebnis des Basismodells. Das hauptsächliche, verbraucherseitige Objekt, das von GEO-1000 gemessen wird, ist daher:

> *das dem Nutzer präsentierte Produktverhalten, nicht eine Modellabstraktion.*

Es können auch kontrollierte API- oder Modellforschungen durchgeführt werden. Dies ist nicht dasselbe Ergebnis wie das Hauptverbraucherpanel.

4. SIEBEN IDENTITÄTSEBENEN DES KI-SYSTEMS

Jedes geprüfte KI-System muss über sieben Ebenen definiert werden, soweit jede Ebene relevant ist.

4.1. Anbieter-Ebene

Die Organisation, die das Produkt veröffentlicht oder betreibt. Name des Anbieters:

- ist nicht dasselbe wie

  Produkt-ID,
- Modell-ID,
- Nutzeroberfläche

Ein Anbieter kann mehr als ein Produkt anbieten. Innerhalb desselben Produkts können mehr als ein Modell oder Orchestrierungssystem verwendet werden.

4.2. Produktebene

Das kommerzielle, Verbraucher-, Unternehmens- oder technische Produkt, auf das der Nutzer zugreift. Beispieltypen:

- Allgemeines Chat-Produkt
- Suchunterstütztes Antwortprodukt
- Unternehmensarbeitsbereich
- Bildungsprodukt
- Eingebetteter Assistent
- Entwickler-API
- Benutzerdefiniertes wissensbasiertes Produkt

Der Produktname allein beschreibt nicht die gesamte Konfiguration.

4.3. Nutzerschnittstellenebene

Es ist die Schnittstelle, über die der Nutzer mit dem Produkt interagiert. Web Mobile Desktop Browser-Erweiterung Betriebssystemintegration Eingebettete Schnittstelle innerhalb der Plattform

API

Unternehmensportal Verschiedene Oberflächen desselben Produkts:

- verschiedene Funktionen,
- verschiedenes Modell,
- verschiedene Quell-Repräsentation.
- verschiedener Speicher

kann verwendet werden.

4.4. Modell- oder Inferenz-Engine-Schicht

Es ist das Modell, die Modellfamilie oder das Routingsystem, das bei der Antwortgenerierung verwendet wird. Modell-ID:

- offen,
- teils sichtbar,
- unbekannt,
- dynamisch

Es ist möglich. Ein Produkt kann eine Nutzeranfrage an verschiedene Modelle weiterleiten. In diesem Fall kann man nicht sagen: 'Ein einzelnes festes Modell wurde gemessen.'

4.5. Orchestrierungs- und Informationsschicht

Es kann Folgendes beinhalten:

- Abruf
- Live-Websuche
- Interne Suche
- Quellenbewertung
- Neubewertung

- Zitationsauswahl
- Fahrzeug anrufen
- Abfrageumformulierung
- Dokumentenzerkleinerung
- Antwortzusammenführung

Diese Schicht ist von grundlegender Bedeutung für die Repräsentation von GEO. Denn sie kann beeinflussen, welcher Eintrag:

- zugriffen wird,
- ausgewählt wird,
- in die Antwort übernommen wird,
- als Referenz angezeigt wird

beeinflusst werden kann.

4.6. Richtlinien- und Personalisierungsschicht

Es kann Folgendes beinhalten:

- Sicherheitsregeln
- Systemanweisungen
- Produktpolitik
- Speicher
- Besondere Anweisungen
- Nutzerhistorie
- Regionale Regeln
- Alter oder Kontoklasse
- Einstellungen des Unternehmensadministrators

Diese Schicht beeinflusst den:

- Umfang der Antwort,
- ihren Ton,
- ihre Ablehnung,
- ihre Empfehlung

kann sich ändern.

4.7. Verteilung und Zeitschicht

Gibt an, unter welcher der folgenden Bedingungen sich das Produkt zu einem bestimmten Zeitpunkt befindet:

- ist mit
- experimentelle Funktion,
- regionale Verteilung,
- Nutzersegment,
- A/B-Test

Dasselbe Produkt kann am selben Tag für verschiedene Nutzer unterschiedliche Verteilungsbedingungen haben.

5. AI PRODUKTBEISPIEL

Die Kerneinheit des Systems innerhalb von GEO-1000:

AI Produktinstanz

Kanonischer operationeller Begriff:

AI Produktinstanz

Eine AI Produktinstanz wird definiert als:

```
A_i= (v, p, s, r, m, k, t, x, u, c, l, ω)
```

kann repräsentiert werden als. Hier:

- v: Anbieter
- p: Produkt
- s: Nutzeroberfläche
- r: Plan oder Kontoklasse
- m: Modell-ID oder Bezeichnung
- k: Abruf und Informationsstrukturierung
- t: Werkzeugkonfiguration
- x: Speicher, Personalisierung und Sitzungsstatus
- u: Nutzer- oder Organisationsklasse der Einstellung
- c: Land und Gebietsschema
- l: Sprache
- ω: Zeit, Bereitstellung und experimenteller Funktionsstatus

Eine Beobachtung: Oj muss mit einem spezifischen: Ai-Produktmuster verknüpft sein.

6. UNTERSCHIED ZWISCHEN PRODUKTMUSTER UND PRODUKTNAME

Der Produktname kann festgelegt sein. Das Produktmuster kann variieren. Beispiel: Der Orion AI-Produktname bleibt unverändert. Diese beiden Messungen sind jedoch unterschiedliche Produktmuster:

Produktmuster 1

Web Free-Plan Speicher aus Webzugriff aus Neuer Chat Türkisch Türkei Wave 1

Produktmuster 2

Mobiler kostenpflichtiger Plan Speicher auf Web-Zugriff auf vorherige Chat-Kontexte verfügbar Türkisch Türkei Wave 1 Diese können unter derselben Marke existieren. Sie sind nicht dieselbe experimentelle Zelle.

7. MODELL-IDENITÄTS-SICHERHEITSEBENEN

Der Modellname kann nicht immer mit derselben Sicherheit bekannt sein. Daher kann die Modellidentität auf den folgenden Ebenen aufgezeichnet werden.

MI-0— UNBEKANNTES MODELL

Es ist unbekannt, welches Modell oder welche Modellsfamilie das Produkt verwendet. Das Ergebnis wird nur der Produktoberfläche zugeschrieben.

MI-1— VOM NUTZER ANGEZEIGTES ETIKETT

Ein Modelletikett ist in der Nutzeroberfläche sichtbar. Das Etikett wird aufgezeichnet. Es wird nicht angenommen, dass die Backend-Implementierung vollständig überprüft wird.

MI-2— VON ANBIETER DOKUMENTIERTE MODELLFAMILIE

Die Produktdokumente des Anbieters beschreiben eine bestimmte Modellsfamilie oder Produkt–Modell-Beziehung. Es kann immer noch unbekannt sein, ob eine einzelne Beobachtung genau auf dieses Modell gerichtet ist.

MI-3— SITZUNGS-BESTÄTIGTE BEREITSTELLUNG

Die Sitzung überprüft eine bestimmte Verteilung von API- oder vertrauenswürdigen Produkt-Metadaten.

MI-4— VERSIONIERTER ENDPOINT ODER BEREITSTELLUNG

Eine bestimmte Version, ein Endpoint oder eine Bereitstellungs-ID kann erfasst werden. Es ist stärker für kontrollierte Vergleiche.

MI-5— REPRODUZIERBARES EINGEFRORENES MODELLARTEFAKT

Es gibt eine statische und reproduzierbare Modell- oder Bereitstellungsstruktur. Es kann bei Live-Verbraucherprodukten selten sein. Wenn das Vertrauen in die Modell-ID gering ist, ist das Ergebnis:

soll dem Produktverhalten zugeschrieben werden

Es kann nicht definitiv einem bestimmten Modell zugeschrieben werden.

8. GRENZE DES OFFENSICHTLICHEN MODELLNAMENS

Das Anzeigen des Modellnamens in einer Nutzeroberfläche ist wertvoll. Dennoch können die folgenden Fragen weiterhin offen bleiben: Werden alle Anfragen an dasselbe Modell gerichtet? Ändert sich das Modell je nach Schwierigkeit der Anfrage? Verwendet die Sicherheits- oder Abruffunktion ein separates Modell? Wendet der Anbieter dasselbe Etikett auf verschiedene Distributionen an? Gibt das Modell-Etikett nur die Produktfamilie an? Ist der Nutzer Teil einer experimentellen Einführung? Daher könnte die korrekte Meldung lauten: „Modell X-Etikett wurde in der Nutzeroberfläche angezeigt.“ Ohne stärkere Nachweise sollte der folgende Satz nicht verwendet werden: „Alle Antworten werden definitiv ausschließlich von Modell X generiert.“

9. PRODUKTTYP

Jedes KI-Produkt sollte mit einem Haupttyp und Sekundärtypen klassifiziert werden.

PT-1— Allgemeines Konversationsprodukt

Es ist ein allgemeines Chat- und Frage-Antwort-Produkt.

PT-2— Such- oder Abruffunktionsbasiertes Produkt

Suche oder Abruf ist eine grundlegende Komponente bei der Antwortgenerierung.

PT-3— Unternehmenswissensprodukt

Es kann interne Dokumente, Arbeitsbereiche oder spezielle Datenquellen verwenden.

PT-4— Eingebetteter Assistent

Er ist in ein anderes Produkt oder eine Plattform eingebettet.

PT-5— Entwickler- oder API-Produkt

Es wird von einem Entwickler oder einer Anwendung anstelle des Endbenutzers verwendet.

PT-6— Branchenspezifisches Produkt

Es ist für einen bestimmten Beruf oder eine bestimmte Branche entwickelt.

PT-7— Multimodales Produkt

Eingaben, die nicht Text sind, wie visuelle, audio- oder Dokumenteneinträge, sind Teil des Hauptangebots. Ein Produkt kann mehr als eine Art enthalten. Vergleiche sollten innerhalb desselben Typs oder mit einem expliziten Umfang vorgenommen werden.

10. NUTZERSURFACE REGISTRIEREN

Verschiedene Oberflächen desselben Produkts sollten getrennt erfasst werden. Oberflächenbereiche:

- Oberflächen-ID
- Web/Mobil/Desktop/API
- Anwendungsversion
- Betriebssystem, falls zutreffend
- Browser, falls zutreffend
- Produktsprache
- Sichtbare Modellauswahl
- Verfügbarkeit von Werkzeugen
- Zitationsansicht
- Datei-Upload-Funktion
- Speichereinstellungen
- Unternehmenseinstellungen
- Experimentelle Funktionen
- Messzeit

Wenn festgestellt wird, dass Web- und mobile Schnittstellen materiell gleich sind, können die Ergebnisse kombiniert werden. Diese Gleichheit sollte nicht vorausgesetzt werden; sie muss getestet werden.

11. PLAN- UND KONTOKLASSE

Der Plantyp kann Merkmale haben, die die Reaktion des Nutzers beeinflussen können. Planstatus:

- Kostenlos
- Bezahlte Einzelperson
- Professionell
- Unternehmen
- Ausbildung
- Entwickler
- Testversion
- Einladung oder eingeschränkter Zugang
- Unbekannt

Pläne können sich in den folgenden Bereichen unterscheiden:

- Modell
- Nutzungslimit
- Webzugang
- Speicher
- Werkzeuge
- Quellenangabe
- Antwortlänge
- Geschwindigkeit oder Priorität
- Unternehmensdaten
- Sicherheits- und Aufbewahrungseinstellungen

Wenn keine Planinformationen vorliegen:

Der PLAN_UNBEKANNT

Status sollte verwendet werden. Kostenlose und bezahlte Ergebnisse dürfen nicht stillschweigend als dasselbe Produktverhalten zusammengeführt werden.

12. ABRUF UND WEBZUGANG

Eine KI-Antwort kann in einem von drei grundlegenden Wissenszuständen erzeugt werden.

KG-0— UNBESTIMMTER WISSENSMODUS

Der Fall der Abruf- oder Webnutzung kann nicht bestimmt werden.

KG-1— INTERNE MODELLANTWORT

Eine Antwort, von der bekannt ist oder stark vermutet wird, dass sie keine aktuellen oder externen Informationen verwendet.

KG-2— ABRUFUNTERSTÜTZTE ANTWORT

Es könnte ein Abruf aus vor-indizierten oder im Produkt verfügbaren Quellen verwendet worden sein.

KG-3— LIVE-WEB- ODER SUCHUNTERSTÜTZTE ANTWORT

Live- oder nahezu in Echtzeit Web-/Suchzugriff wurde verwendet.

KG-4— PRIVAT DATENBASIERT FREIGE GEBENE ANTWORT

Eine private Datenquelle, die von der Institution oder dem Nutzer bereitgestellt wurde, wurde verwendet.

KG-5— HYBRIDE ODER MULTI-QUELLEN ANTWORT

Mehrere Informationsschichten wurden zusammen verwendet. Diese Modi können im gleichen Prompt unterschiedliche Ergebnisse erzeugen. Sie sollten im Vergleich deutlich unterschieden werden.

13. ZUGRIFF AUF ABRUF IST NICHT DASSelbe WIE ABRUFVERWENDUNG

Das Produkt kann über eine Websuchfunktion verfügen. Allerdings wurde sie möglicherweise in der spezifischen Antwort nicht genutzt. Drei Felder sollten getrennt aufgezeichnet werden: War die Funktion verfügbar? War die Funktion eingeschaltet oder auswählbar? Gab es Nachweise dafür, dass sie in der spezifischen Antwort verwendet wurde? Beispielhafte Status:

VERFÜGBAR_NICHT_VERWENDET

VERFÜGBARE_NUTZUNG_UNBEKANNT

VERWENDET_BESTÄTIGT

NICHT_VERFÜGBAR

DEAKTIVIERT

UNBEKANNT

Die bloße Präsenz eines Quellenlinks kann ein starker Hinweis auf die Nutzung sein. Das Fehlen eines Quellenlinks beweist nicht notwendigerweise, dass kein Abruf verwendet wurde.

14. ZITIERUNGSSCHICHT

Die Verwendung einer Quelle ist nicht dasselbe wie das Angeben eines Zitats. Ein Produkt:

- darf eine Quelle verwenden, darf sie aber nicht zeigen,
- kann eine Quelle zeigen, aber die Quelle unterstützt möglicherweise nicht die Behauptung,
- kann dieselbe Quelle mit verschiedenen Sätzen verlinken,

Der Nutzer kann das Zitat entsprechend der Nutzeroberfläche ausblenden. Der Zitatdatensatz muss die folgenden Felder enthalten: Ist die Zitatfunktion verfügbar? Wurde das Zitat in der spezifischen Antwort angezeigt? Anzahl der Zitate Quellfeldnamen Übereinstimmung von Zitat und Aussage Sichtbarkeit auf der Nutzeroberfläche Erweiterbarkeit des Zitats Zugänglichkeit der Quelle Zeitpunkt der Zitatgenerierung Produkte, die Zitate anzeigen oder nicht anzeigen, können in Bezug auf die „Anzahl der Quellen“ nicht direkt verglichen werden. Zuerst müssen das Produktverhalten und die Unterschiede in der Nutzeroberfläche erklärt werden.

15. WERKZEUGZUGANG

Das KI-Produkt kann auf die folgenden Werkzeuge zugreifen:

- Websuche
- Browser
- Taschenrechner
- Code-Ausführung
- Dateilesen
- Karte
- Kalender
- Unternehmenswissensbasis
- Produktkatalog
- Andere APIs

Der Zugriff auf Werkzeuge kann Ihre Antwort beeinflussen:

- deren Aktualität,
- Nachweise,
- Umfang der Antwort,
- Genauigkeit

Werkzeugregister:

- ob das Werkzeug verfügbar ist,
- ob es aktiv ist,
- ob es in der Antwort verwendet wurde,
- ob das Werkzeugergebnis behalten wird

soll angezeigt werden.

16. SPEICHER, SONDERANWEISUNGEN UND PERSONALISIERUNG

Die für den Nutzer sichtbare Antwort kann durch die folgenden Felder beeinflusst werden:

- Persistenter Speicher
- Chatverlauf
- Besondere Anweisungen
- Name oder Profil des Nutzers
- Frühere Markenunterhaltungen
- Einstellungen
- Land und Gebietsschema
- Einstellungen des Unternehmensarbeitsbereichs
- Segmentierung durchgeführt vom Anbieter

Diese Felder müssen in drei separaten Ebenen aufgezeichnet werden.

16.1. Persistente Personalisierung

Informationen, die über Konten oder Sitzungen hinweg geschützt sind.

16.2. Kontext im Gespräch

Vorherige Nachrichten innerhalb des aktuellen Chats.

16.3. System- oder Organisationseinstellungen

Administrator- oder Produktanweisungen, die der Nutzer möglicherweise nicht sieht. Vollständige Informationen sind möglicherweise nicht immer verfügbar. Unbekannte Situationen müssen ausdrücklich aufgezeichnet werden.

17. SITZUNGSTYPEN

Kandidaten-Sitzungsstatus:

ST-1— NEUES STANDARDGESPRÄCH

Neues Standardgespräch. Speicher oder spezielle Anweisungen werden separat aufgezeichnet.

ST-2— TEMPORÄRES ODER EPHEMERES GESPRÄCH

Temporäre Sitzung, die von der permanenten Historie getrennt werden soll. Tatsächliche Funktionen sollten im Produktregister überprüft werden.

ST-3— BESTEHENDES GESPRÄCH

Es gibt vorherigen Gesprächskontext. Dieser kann für das natürliche Panel wertvoll sein. Er wird separat im kontrollierten Panel aufbewahrt.

ST-4— UNTERNEHMENS- ODER GETEILTE ARBEITSBEREICHSSESSION

Es könnte einen Unternehmens- oder geteilten Informationsbereich geben.

ST-5— UNBEKANNTER SITZUNGSZUSTAND

Der Sitzungsstatus konnte nicht zuverlässig festgestellt werden. Wenn verschiedene Sitzungstypen in derselben Zelle kombiniert werden sollen, muss die Begründung und Gewichtung erklärt werden.

18. PRODUKTPOLITIK UND ABLEHNUNG

Ein KI-Produkt kann auf eine bestimmte Frage nicht antworten aufgrund von:

- Sicherheit,
- Gesetz,
- Produktpolitik,
- regionaler Regel,
- impliziert keine Eignung in Bezug auf Alter,
- unternehmensspezifischer Einstellung

Grund. Ablehnung:

- grundlegende Modellmängel,
- Produktpolitik,
- Schnittstelle,
- kann am Nutzerkonto liegen

GEO-1000 Hauptbenutzerkontrolle zeichnet das Ablehnungsergebnis auf, das den Nutzer erreicht. Wenn der Ablehnungsgrund jedoch bekannt ist, wird er zusätzlich klassifiziert. Die Richtlinienebene eines Produkts ist ebenfalls Teil der Nutzererfahrung.

19. EXPERIMENTELLE FUNKTIONEN UND A/B-VERTEILUNGEN

Live-KI-Produkte können möglicherweise nicht allen Nutzern gleichzeitig die gleichen Funktionen bieten. Teilnehmer:

- differentes Nutzerinterface,
- unterschiedliche Prompt-Formulierungen je nach Modell,
- differente Webfunktion,
- differentes Antwortformat

kann sehen. Experimentelle Statusindikatoren:

- Beta-Kennzeichnung auf der Benutzeroberfläche
- Anbieterdokumentation
- Funktionsunterschied
- Systematischer UI-Unterschied zwischen Nutzern
- Verhaltensstörungen während der Messung

kann auftreten. Wenn die experimentelle Funktion unbekannt ist:

EXPERIMENTELLER_ZUSTAND_UNBEKANNT

sollte verwendet werden. Die tatsächliche Variation innerhalb des Produkts ist kein Datenfehler. Sie ist ein Verteilungsmerkmal des Systems. Sie sollte jedoch bei der Vergleichbarkeit berücksichtigt werden.

20. KI-Systemregister

Für jedes geprüfte Produkt sollte ein KI-Systemregister erstellt werden. Das Register sollte drei Ebenen haben.

20.1. Anbieter- und Produktstammsatz

Beinhaltet relativ permanente Bereiche:

- Anbieter
- Produktname
- Produkttyp
- Hauptbenutzeroberflächen
- Offizielle Zugriffsländer
- Plan-Klassen
- Unterstützte Sprachaufzeichnungen
- Allgemeine Funktionen

20.2. Messwellen-Produktaufzeichnung

Für eine bestimmte Welle:

- Produktversion,
- Planfunktionen,
- Modellkennzeichnungen,
- Web- und Tool-Funktionen,
- Speicherverhalten,
- Zugangsstatus,
- bekannte Updates

aufgezeichnet.

20.3. Beobachtungsebene Systemaufzeichnung

Tatsächlich für jeden Teilnehmer beobachtet:

- Schnittstelle,
- Modellkennzeichnung,
- Plan,
- Fahrzeugnutzung,
- Zitation,
- Sitzung,
- Personalisierung,
- Zeit

trägt Felder. Allein die Hauptaufzeichnung kann Unterschiede auf Beobachtungsebene nicht entfernen.

21. SYSTEMKONFIGURATIONS-FINGERABDRUCK

Ein Systemkonfigurations-Fingerabdruck kann erstellt werden, um die Vergleichbarkeit einer KI-Produktinstanz zu erleichtern. Kandidatenfelder:

```
SCF= H(canonicalise(p, s, r, m, k, t, x, c, l, w))
```

Hier ist H die Funktion, die aus den versionierten Feldern einen Integritätswert erzeugt. Der Fingerabdruck:

- ob zwei Beobachtungen zur gleichen Konfiguration gehören,
- ob es wesentliche Änderungen in der Welle gegeben hat

Es kann helfen, dies zu identifizieren. Fingerabdruck: beweist nicht das echte Backend-Modell, macht keine Schichten sichtbar, die der Anbieter nicht offengelegt hat. Er verstärkt nur, ob die aufgezeichnete Konfiguration später geändert wurde.

22. WAS IST VERGLEICHBARKEIT?

Zwei Ergebnisse gelten als vergleichbar, wenn:

> *der Unterschied zwischen ihnen ausreichend durch überwachten Produktverhalten erklärt werden kann und nicht durch irrelevante Unterschiede im Messdesign.*

Vergleichbarkeit ist nicht absolute Gleichheit. Unterschiedliche KI-Produkte:

- Modell,
- Anbieter,

- Nutzeroberfläche.

Diese müssen nicht alle identisch sein. Die für den Vergleich relevanten materiellen Bedingungen müssen jedoch:

- Ausgleich,
- Schichtung,
- Aufzeichnung,
- Modellierung in der Analyse

sind erforderlich.

23. DAS VERGLEICHSOBJEKT MUSS VORHER DEFINIERT WERDEN

Die gleichen Daten können unterschiedliche Vergleichsfragen beantworten.

23.1. Nutzerproduktvergleich

"Welches Verbraucherprodukt liefert unter denselben Nutzerbedingungen eine genauere Repräsentation?" Das zu vergleichende Objekt ist die gesamte Produktkette.

23.2. Modellvergleich

"Welches Modell liefert unter demselben kontrollierten Prompt und werkzeugsfreien Zustand eine genauere Antwort?" API- oder fester Modellzugang kann erforderlich sein.

23.3. Abrufvergleich

„Gibt es einen Unterschied zwischen den webbasierten und nicht webbasierten Modi desselben Produkts?“ Das Modell und andere Bedingungen werden so konstant wie möglich gehalten.

23.4. Planvergleich

„Sehen kostenlose und zahlende Nutzer die gleiche Markenpräsentation?“ Der Planunterschied ist Gegenstand der Prüfung.

23.5. Schnittstellenvergleich

„Erzeugt das Web- und Mobilprodukt unter den gleichen Bedingungen unterschiedliche Repräsentationen?“

23.6. Zeitvergleich

„Hat sich die Repräsentation desselben Produkts über verschiedene Wellen hinweg geändert?“ Produkt- und Bevölkerungsänderungen sollten getrennt werden. Ohne die Vergleichsfrage zu spezifizieren: „Produkt A ist besser als Produkt B.“ ist zu allgemein.

24. VERGLEICHBARKEITSDIMENSIONEN

Beim Vergleich von zwei Produkten oder Wellen sollten die folgenden Bereiche bewertet werden:

- Auditierte Einheit
- Reality-Paket-Version
- Prompt-ID und Version

- Sprache und Gebietsschema
- Land
- Zielpopulation
- Beispielquelle
- Teilnehmerzusammensetzung
- Sitzungstyp
- Speicher
- Besondere Anweisungen
- Vorheriges Gespräch
- Produktplan
- Nutzeroberfläche
- Modellbezeichnung
- Abruf-/Web-Status
- Werkzeuge
- Zitationsfunktion
- Messzeit
- Produktupdates
- Beurteilungsregel
- Ergebnisstatus
- Gewichtung

Nicht alle Dimensionen müssen gleich sein. Unterschiedliche Dimensionen begrenzen den Vergleichsrahmen.

25. VERGLEICHBARKEITSSTUFEN

CP-0— NICHT VERGLEICHBAR

Bedingungen für Materialmessungen sind unbekannt oder stark inkonsistent. Ein direkter Leistungsvergleich kann nicht durchgeführt werden.

CP-1— BESCHREIBENDE KO-EXISTENZ

Es gibt Ergebnisse von zwei Produkten oder Wellen. Allerdings:

- Bevölkerung,
- Prompt,
- Zeit,
- Sitzung

sind nicht gleich. Es kann nur eine beschreibende Nebeneinander-Darstellung gemacht werden.

CP-2— TEILWEISE ANGEPASST

Einige Kernfelder stimmen überein. Materielle Unterschiede bestehen weiterhin. Eine begrenzte Vergleichbarkeit ist möglich.

CP-3— ÜBEREINSTIMMENDER NUTZER-Oberflächen-Vergleich

Gleiches Unternehmen Gleiche Prompt-Version Gleiche Sprache/Land Ähnliche Nutzergruppe Gleicher Zeitraum Registrierte Produktmerkmale sind verfügbar. Verbraucherproduktergebnisse sind vergleichbar.

CP-4— KONTROLLIERTER ÜBEREINSTIMMENDER VERGLEICH

Zusätzlich:

- Sitzung,
- Speicher,
- Personalisierung,
- Abruf,
- Werkzeuge,
- Plan

sind kontrolliert oder offen abgestimmt. Bietet einen stärkeren Produkt- oder Feature-Vergleich.

CP-5— REPLIZIERTER UND KALIBRIERTER VERGLEICH

Vergleich:

- in mehreren Wellen,
- mit unabhängigen Nutzern oder Panels,
- mit derselben normativen Bewertung

wiederholt; Variabilität und Produktänderungen wurden kalibriert. Diese Werte sind kein „wissenschaftlicher Genauigkeitsscore“. Sie zeigen die Grenze an, die eine Vergleichsaussage tragen kann.

26. WAS TUN BEI CP-0 KONDITION?

Wenn zwei Produkte nicht verglichen werden können, werden die Ergebnisse nicht ignoriert. Sie können separat berichtet werden:

- „Dies ist das Natürliche Reichweite-Ergebnis von Produkt A.“
- „Dies ist das Ergebnis von Produkt B unter verschiedenen Nutzern und Produktbedingungen.“

Allerdings: Man kann nicht sagen: „A ist 12 Punkte besser als B.“ Unvergleichbarkeit ist kein Misserfolg, sondern eine Begrenzung der Methode.

27. VERGLEICHBARKEIT FÜR DEN GLEICHEN ANBIETER

Ergebnisse vom gleichen Anbieter können nicht automatisch verglichen werden. Die folgenden Unterschiede können vorhanden sein:

- Produkt
- Plan
- Modell
- Webzugang
- Benutzeroberfläche

- Speicher
- Unternehmensdaten
- Länderverteilung
- Experimentelle Funktion
- Zeit

Der Vergleich desselben Anbieters sollte ebenfalls gemäß dem CP-Level bewertet werden.

28. VERGLEICHBARKEIT FÜR VERSCHIEDENE ANBIETER

Nutzerprodukte, die verschiedenen Anbietern gehören:

- gleiche Zielgruppe,
- gleiche Prompt,
- gleiche Zeit,
- ähnliche Nutzeroberfläche,
- klare Produktkonfiguration

Es ist darunter vergleichbar. Ein anderer Anbieter verhindert den Vergleich nicht. Unterschiede in den Funktionen bestimmen jedoch, was der Vergleich misst. Beispiel: Produkt A verwendet Live-Web. Produkt B nicht. Das Ergebnis kann diese Frage beantworten: 'Welche der beiden Produktreihen, die dem Nutzer angeboten werden, erzeugt eine genauere Repräsentation?' Es beantwortet diese Frage nicht direkt: 'Welches Basismodell ist wissensreicher?'

29. VERGLEICH VON WEB-UNTERSTÜTZTEN UND NICHT-WEB-PRODUKTEN

Zwei separate Vergleiche sind möglich.

29.1. Vergleich von Naturprodukten

Jedes Produkt wird unter den Standardbedingungen eines echten Nutzers gemessen. Ein Produkt, das das Web nutzt, nutzt das Web. Eines, das es nicht nutzt, tut es nicht. Ergebnis: es ist ein Vergleich der tatsächlichen Verbrauchererfahrung mit dem Produkt.

29.2. Funktionsabgestimmter Vergleich

Wenn möglich, beide Produkte:

- im offenen Web gemessen,
- Web geschlossen

Modi ebenfalls separat. Dies kann den Abrufe-Effekt besser trennen. Jedes Produkt bietet möglicherweise nicht die gleichen Optionen. Fehlender Modus:

ALS NICHT VERFÜGBAR AUFGEZEICHNET

Dieser Status muss ausdrücklich bleiben.

30. PRODUKTE, DIE ZITATE ANZEIGEN UND NICHT ANZEIGEN

Für ein Produkt ohne Zitierfunktion: Es ist möglich, eine „Null-Zitation“ anzugeben. Allerdings sollte deren Bedeutung erklärt werden: Das Produkt verweist möglicherweise nicht auf Quellen. Es ist nicht sicher, dass es überhaupt keine Quellen verwendet. Zitationsabdeckung und Genauigkeitsabdeckung sind getrennte Ergebnisse. Ein Zitationsvergleich sollte nur zwischen Produkten durchgeführt werden, die Zitationen unterstützen oder mit separatem Funktionsstatus.

31. Natürliche Reichweite VERGLEICH

Jedes Produkt wird innerhalb seiner tatsächlichen Reichweite gemessen. Vorteil: zeigt die natürliche globale Wirkung des Produkts. Einschränkung: die Populationen sind unterschiedlich, Länder- und Sprachverteilungen können variieren. Natürliche Reichweite Scores: zeigen, was das Produkt innerhalb seiner erreichten Population leistet. Sollte nicht allein für ein direktes Produktüberlegenheitsranking verwendet werden.

32. VERGLEICH AUF GEMEINSAMER BASIS

Es wird die gemeinsame Zugriffspopulation der verglichenen Produkte verwendet. Vorteil:

- gleiches Land,
- gleiche Sprache,
- gleiche Nutzerbedingung

bietet einen gerechteren Vergleich. Einschränkung: bewahrt den tatsächlichen globalen Zugriffsunterschied der Produkte, das gemeinsame Universum kann sehr eng sein. Der Gemeinsame Vergleichsbasis-Score sollte zusammen mit dem Natürliche Reichweite-Score veröffentlicht werden.

33. VERGLEICH VON NUTZERN PAARWEISE

Wenn derselbe Nutzer zwei oder mehr KI-Produkte testet, können Nutzermerkmale kontrolliert werden. Jedoch:

- Reihenfolge der Aufgaben,
- Antwortenabruf,
- Produktfamiliarität,
- spezielle Profile von Nutzern, die auf alle Produkte zugreifen können,

kann Verzerrungen erzeugen. In einem paarweisen Design:

- sollte die Produktreihenfolge randomisiert werden,
- sollten vorherige Antworten nicht geteilt werden,

die gleiche Nutzersabhängigkeit sollte in der Analyse beibehalten werden.

34. UNABHÄNGIGER NUTZER-VERGLEICH

Jedes Produkt wird von verschiedenen Nutzern getestet, diese werden jedoch aus demselben Populationsrahmen ausgewählt. Vorteil: Aufgabenverunreinigung wird reduziert, es können

natürlichere Produktbenutzer ausgewählt werden. Einschränkung: Nutzergruppen sind möglicherweise nicht vollständig gleich. Gewichtung und Schichtung sind erforderlich. Ergebnisse von gepaarten und unabhängigen Nutzern können getrennt, aber komplementär sein.

35. ZEITLICHE VERGLEICHBARKEIT

Beim Vergleich des heutigen Ergebnisses eines KI-Produkts mit früheren Ergebnissen sollten die folgenden Änderungen untersucht werden:

- Modell
- Produktversion
- Web/Retrieval
- Plan
- Zugang nach Land
- Sprachunterstützung
- Aufforderung
- Reale Entitäten
- Beurteilung
- Populationsrahmen
- Gewichtung

Zeitunterschiede können nicht nur durch das KI-Produkt entstehen.

36. ARTEN VON PRODUKTÄNDERUNGEN

UC-1— DOKUMENTIERTE KLEINE ÄNDERUNG

Eine geringfügige Änderung, die vom Anbieter offengelegt und als nicht wesentlich für die Änderung des Messobjekts beurteilt wurde.

UC-2— DOKUMENTIERTE WESENTLICHE ÄNDERUNG

Wesentliche Änderung beschrieben im Modell, Abruf, Schnittstelle oder Merkmal.

UC-3— BEOBACHTETE VERHALTENSÄNDERUNG

Selbst wenn keine offizielle Bekanntgabe erfolgt, wird systematisches Verhalten oder eine Schnittstellenänderung beobachtet.

UC-4— ZUGRIFTS- ODER PLÄNEÄNDERUNG

Änderungen des Landes, Plans oder Nutzerzugriffs.

UC-5— UNBEKANNTES ÄNDERUNGSEREIGNIS

Es gibt eine Unterbrechung im Messverhalten. Der Grund konnte nicht überprüft werden.

UC-6— PRODUKTAUSTAUSCH ODER -RÜCKZUG

Ein altes Produkt oder Modell wurde eingestellt und durch eine andere Struktur ersetzt. Die Art der Änderung begrenzt den Zeitreihenvergleich.

37. PRODUKTÄNDERUNG INNERHALB DER MESSWELLE

Wenn während einer Welle eine Materialänderung auftritt, existieren folgende Optionen: Die Welle kann gestoppt und neu gestartet werden. Vor und nach der Änderung können separate Unterwellen erstellt werden. Die Änderung kann als experimenteller Faktor modelliert werden. Die Welle wird als MIXED_SYSTEM_STATE gekennzeichnet. Die folgende Struktur ist verboten: Alle Antworten zu kombinieren, als ob sie zu einem einzigen festen Produktergebnis führen, obwohl die Änderung bekannt ist.

38. VERHALTEN KANN SICH OHNE PRODUKTÄNDERUNG ÄNDERN

Selbst wenn der Anbieter kein explizites Update ankündigt, kann sich das Ergebnis ändern. Mögliche Gründe:

- Änderung des Abrufindex
- Aktualisierung von Ressourcen
- Neuer Inhalt im Web erscheint
- Änderung im Re-Ranking
- A/B-Test
- Regionale Einführung
- Sicherheitsrichtlinie
- Verkehrslenkung

Daher kann nicht gesagt werden: „Der Anbieter hat nicht aktualisiert, das System war definitiv dasselbe.“ Messbelege und Verhaltensanalyse sollten ebenfalls untersucht werden.

39. Live-Produkt und gefrorenes Modell

Es gibt zwei verschiedene Forschungsobjekte.

39.1. Live-Produkt-Audit

Misst, was echte Nutzer auf der echten Verbraucheroberfläche sehen. Vorteil:

- zeigt echte Erfahrung,
- echte Quelle,
- echte Richtlinie,
- echte Schnittstelle

Einschränkungen: Es ist variabel, vollständige Reproduktion kann schwierig sein.

39.2. Audit des gefrorenen Modells

Kontrollierte Wiederholungen werden unter einer festen Version oder einem Endpunkt durchgeführt. Der Vorteil:

- bietet Reproduzierbarkeit,
- experimentelle Kontrolle

Die Einschränkung: kann das tatsächliche Verbraucherprodukt möglicherweise nicht vollständig darstellen. Die Hauptpopulationsebene von NOMOS GEO-1000 basiert auf dem Live-Nutzerprodukt. Das eingefrorene Modell-Audit ist eine ergänzende Laborarbeit.

40. PRODUKTNICHTVERFÜGBARKEIT UND TECHNISCHE ERGEBNISSE

Die folgenden Fälle sollten getrennt gehalten werden:

AR-1— PRODUKT NICHT VERFÜGBAR

Das Produkt ist für das Land oder die Nutzergruppe nicht verfügbar. Es ist das Ergebnis des Zugriffsbereichs.

AR-2— KONTO ODER PLAN UNBERECHTIGT

Der dem Teilnehmer zugewiesene Plan oder Kontobedingung ist nicht berechtigt. Es kann ein Problem bei der Stichprobe oder Aufgabenverteilung vorliegen.

AR-3— VORÜBERGEHENDER TECHNISCHER FEHLER

Ein vorübergehender technischer Fehler ist aufgetreten. Das Wiederholungsprotokoll sollte vordefiniert sein.

AR-4— RATE ODER NUTZUNGSLIMIT

Ein Nutzungskontingent oder ein Ratenlimit wurde erreicht. Es kann Teil der tatsächlichen Nutzererfahrung sein.

AR-5— SYSTEMABLEHNUNG

Die Produktaufgabe wurde aufgrund von Richtlinien oder Inhalten abgelehnt. Es ist ein gültiges Systemergebnis.

AR-6— KEIN NUTZBARER OUTPUT

Es wurde keine nutzbare Antwort generiert.

AR-7— UNBEKANNTER ZUGRIFFSFEHLER

Warum es nicht bestimmt werden konnte. Diese Situationen können nicht im gleichen "Keine Antwort"-Feld kombiniert werden.

41. TECHNISCHE WIEDERHOLREGEL

Ein Wiederholungsversuch kann bei einem vorübergehenden technischen Fehler oder einer Verbindungsunterbrechung unternommen werden. Allerdings:

- wenn die KI-Antwort nicht gefällt,
- eine Ablehnung erfolgte,
- falsche Informationen erzeugt wurden

kann aus diesen Gründen kein Wiederholungsversuch durchgeführt werden. Technische Wiederholung:

- gleicher Nutzer,

- gleiches Produkt,
- gleiche Konfiguration,
- muss innerhalb des vorgegebenen Zeitraums erfolgen.

Auch das erste technische Ereignis muss aufgezeichnet werden.

42. VERGLEICHSMATRIX

Ein grundlegender Vergleichseintrag für zwei Produkte könnte wie folgt aussehen:

Dimension	Produkt A	Produkt B	Status
Zielobjekt	Gleich	Gleich	Übereinstimmend
Prompt-Version	1.0	1.0	Übereinstimmend
Sprache/Region	tr-TR	tr-TR	Übereinstimmend
Land	Türkei	Türkei	Übereinstimmend
Zeitfenster	Welle 1	Welle 1	Übereinstimmend
Nutzerpanel	Population	Population	Übereinstimmend
Plan	Kostenlos	Bezahlt	Unterschiedlich
Oberfläche	Web	Mobil	Unterschiedlich
Webzugang	Geschlossen	Offen	Unterschiedlich
Speicher	Geschlossen	Offen	Unterschiedlich
Modell-ID	MI-1	MI-1	Begrenzt
Vergleichsebene			CP-2

Diese Ergebnisse können nebeneinander dargestellt werden. Allerdings kann nicht gesagt werden: "Das Basismodell von Produkt B ist besser." Der Unterschied kann von der gesamten Produktkette stammen.

43. SYNTHETISCHER APPLE.COM VERGLEICHSFALL

SYNTHETISCHE METHODOLOGIE-DEMONSTRATION / Die untenstehenden Produkte, Antworten und Ergebnisse sind vollständig fiktiv. Sie spiegeln nicht die Leistung von Apple Inc. oder irgendeinem realen KI-Anbieter wider. Betrachten wir drei synthetische Produktbeispiele.

Produktbeispiel A

Anbieter: Orion Produkt: Orion Chat Schnittstelle: Web Plan: Kostenlos Modellbezeichnung: Orion Standard Modell Vertrauen: MI-1 Webzugang: Aus Speicher: Aus Sitzung: Neu Sprache: Türkisch Gebietsschema: Türkei Welle: W1 Antwort: "Apple ist ein amerikanisches Technologieunternehmen, das Telefone und Computer herstellt."

Produktbeispiel B

Anbieter: Orion Produkt: Orion Chat Schnittstelle: Mobilplan: Kostenpflichtiges Modell Bezeichnung: Orion Pro Modellvertrauen: MI-1 Webzugriff: Offen Quellenangabe: Offen Speicher: Ein Vorheriges Gespräch: Über Apple Sprache: Türkisch Gebietsschema: Türkei Welle: W1 Antwort: "Apple ist ein globales Technologieunternehmen, das Hardware, Software und digitale Dienste anbietet. Aktuelle Unternehmensinformationen finden Sie in diesen Quellen."

Produktbeispiel C

Anbieter: Helios Produkt: Helios Suche Schnittstelle: Web Plan: Kostenlos Modellbezeichnung: Unbekannt Modellvertrauen: MI-0 Webzugang: Offen Zitat: Offen Speicher: Keiner Sitzung: Neu Sprache: Türkisch Gebietsschema: Türkei Welle: W1 Antwort: "Apple ist ein in den USA ansässiges Technologieunternehmen, das Verbrauchergeräte, Betriebssysteme und digitale Dienste entwickelt."

43.1. Vergleich von A und B

Sie sind derselbe Anbieter. Allerdings:

- Plan,
- Oberfläche,
- Modellkennzeichnung,
- Web,
- Speicher,
- vorheriges Gespräch

ist unterschiedlich. Das Vergleichsniveau kann CP-1 oder CP-2 sein. Der Unterschied in der Antwort kann nicht ausschließlich auf den Unterschied des Basismodells zurückgeführt werden.

43.2. Vergleich von A und C

Sie sind unterschiedliche Anbieter. Allerdings:

- Neue Sitzung,
- kostenloser Plan,
- Weboberfläche,
- gleiche Sprache/Land,
- gleiche Welle

in Bezug auf einige Bedingungen übereinstimmend. Webzugang ist unterschiedlich. Vergleich: kann als natürliches Konsumprodukt-Ergebnis durchgeführt werden. Es kann nicht als Basismodell-Fähigkeit durchgeführt werden.

43.3. Vergleich von B und C

Beide Produkte haben Webzugang und Zitate aktiviert. Allerdings hat B Speicher und vorherige Konversation, während C dies nicht hat. Die Nutzererfahrung der Produkte ist vergleichbar. Es ist kein kontrollierter, sauberer Produktvergleich.

43.4. Kontrollierte Wiederholung

Im Produkt B:

- Speicher ist ausgeschaltet,
- ein neuer Chat wird gestartet,
- spezielle Anweisungen sind deaktiviert,

das Web bleibt eingeschaltet. C wird unter denselben Bedingungen gemessen. Vergleichsebene: CP-3 oder CP-4 mit ausreichender zusätzlicher Kontrolle. Dieses Beispiel zeigt:

> *Der gleiche Anbieter liefert keinen nahen Vergleich; gut dokumentierte unterschiedliche Anbieter können besser vergleichbar sein.*

44. FALL VON SYNTHETISCHER PRODUKTÄNDERUNG

SYNTHETISCHES BEISPIEL

Orion Chat W1 Wellen zielt auf 1,000 Beobachtungen ab. In den ersten 480 Beobachtungen:

- Produktversion: 6.1
- Webfunktion: manuell
- Zuschreibung: begrenzt

In den nächsten 520 Beobachtungen:

- Produktversion: 6.2
- Webfunktion: automatisch
- Zuschreibung: erweitert

sei es. Einzelnes Ergebnis: Wenn als „Orion Chat W1 Score“ berechnet, werden zwei verschiedene Systemzustände gemischt. Richtige Optionen:

Option 1

W1A: erste 480 W1B: nächste 520 werden separat gemeldet.

Option 2

Die Welle wird nicht fortgesetzt. 1,000 neue Beobachtungen werden in Version 6.2 gemacht.

Option 3

Die Version wird als expliziter experimenteller Faktor in der statistischen Analyse beibehalten. Falsche Option: Das Update wird aus dem Bericht entfernt und alle Beobachtungen werden als ein einziges festes Produkt präsentiert.

45. MINIMALES KI-SYSTEMPROTOKOLL

Jede Beobachtung muss mindestens die folgenden Felder enthalten:

- NOMOS KI-System-ID
- Anbieter
- Produkt
- Oberfläche
- Plan
- Produkttyp
- Offensichtliches Modelllabel
- Modellvertrauensniveau
- Land
- Locale
- Sprache
- Sitzungstyp
- Speicherstatus
- Status benutzerdefinierter Anweisungen
- Vorheriger Gesprächszustand
- Abruf-/Webfunktion

- Abrufnutzungsstatus
- Zitationsstatus
- Werkzeugstatus
- Status experimenteller Funktionen
- UTC Zeit
- Welle
- System-Fingerabdruck
- Änderungsereignisse
- Quelle der Nachweise
- Datensatzinhaber

Wenn das Materialfeld unbekannt ist, darf es nicht leer gelassen werden. Der Status UNBEKANNT sollte verwendet werden.

46. MINIMALE VERGLEICHSKARTE

Beim Vergleich von zwei Produkten oder Wellen sollte der öffentliche Datensatz Folgendes anzeigen:

- Vergleichsfrage
- Vergleichsebene
- Gemeinsame Zielgruppe
- Natürliche Reichweite Unterschiede
- Nutzergleichheit
- Sprache und Land
- Messzeit
- Produkt, Plan und Oberfläche
- Modell-Identitätsvertrauen
- Web-/Abrufdifferenz
- Unterschied in Speicher und Personalisierung
- Unterschied im Tool
- Unterschied in der Probe
- Bewertungsmethode
- Materialeinschränkungen
- Direkt vergleichbare Metriken
- Nicht vergleichbare Bereiche

Eine einzelne Zahl oder Rangliste sollte ohne diese Karte nicht veröffentlicht werden.

47. REGEL DER VERGLEICHBARKEIT

Ein Prüfer sollte diese vier Fragen beantworten: Wurde dieselbe Sache gefragt? Wurde dasselbe menschliche Universum gemessen? Wurden dieselben Nutzerbedingungen angewendet? Wurden dieselbe Zeit und dasselbe Bewertungssystem verwendet? Jede dieser Fragen:

- vollständig,
- teilweise,
- nein,

- unbekannt

wird entsprechend klassifiziert. Der Vergleichsgrad wird dementsprechend angegeben.

48. VERPFLICHTENDE NORMATIVE BESTIMMUNGEN

CH09-N01

Jede GEO-1000-Beobachtung muss mit einem versionierten Datensatz des KI-Systemregisters verknüpft sein.

CH09-N02

Der Anbietername darf nicht als vollständige Identität des gemessenen KI-Produkts verwendet werden.

CH09-N03

Die Produkt-, Nutzeroberflächen-, Plan-, Modell-, Abruf-, Werkzeug-, Sitzungs- und Personalisierungsschichten müssen voneinander getrennt sein.

CH09-N04

Die Modellbezeichnung auf der Nutzeroberfläche darf nicht als verifizierte Back-End-Modellidentität dargestellt werden.

CH09-N05

Das Modell-Identitätsvertrauensniveau muss von MI-0 bis MI-5 oder durch einen gleichwertig offen definierten Status aufgezeichnet werden.

CH09-N06

Wenn die Modellidentität nicht verifiziert werden kann, sollte das Ergebnis nur dem Produkt und der Nutzeroberfläche zugeschrieben werden.

CH09-N07

Produkte desselben Anbieters werden nicht automatisch als dieselbe Versuchszelle betrachtet.

CH09-N08

Produkte verschiedener Anbieter können vergleichbar sein, wenn der Materialverbraucher und die Versuchsbedingungen übereinstimmen.

CH09-N09

Web-, mobile-, Desktop-, eingebettete Produkt- und API-Oberflächen können nicht ohne Beschreibung kombiniert werden.

CH09-N10

Kostenlose, kostenpflichtige, Unternehmens-, Bildungs- und Entwicklerpläne müssen als separate Produktbedingungen erfasst werden.

CH09-N11

Wenn der Plan unbekannt ist, sollte das Ergebnis den Status PLAN_UNBEKANNT tragen.

CH09-N12

Die Verfügbarkeit der Abruf- oder Webfunktion kann nicht als Nachweis dafür gezählt werden, dass sie in einer bestimmten Antwort verwendet wurde.

CH09-N13

Abrufzugriff, Abrufnutzung und Attribution müssen mit separaten Status aufgezeichnet werden.

CH09-N14

Ein Produkt, das keine Attribution zeigt, kann nicht automatisch als nicht Verwendung von Ressourcen betrachtet werden.

CH09-N15

Die Zitationen des zitierten Produkts sollten in Bezug auf die Unterstützung der Behauptung getrennt untersucht werden.

CH09-N16

Werkzeugzugriff und Werkzeugnutzung in bestimmten Antworten sollten getrennt aufgezeichnet werden.

CH09-N17

Speicher, besondere Anweisungen, vorherige Chats und Unternehmenseinstellungen sollten so weit wie möglich getrennt aufgezeichnet werden.

CH09-N18

Ein unbekannter Personalisierungszustand kann nicht als eine saubere Sitzung angesehen werden.

CH09-N19

Neue Chats, temporäre Chats, aktuelle Chats und Unternehmensarbeitsbereichssitzungen können nicht ohne Erklärung zusammengeführt werden.

CH09-N20

Die Ergebnisse von Panel unter natürlichen Nutzungsbedingungen und Kontrolliertes Panel unter bereinigten Bedingungen sollten einen separaten Produktstatus haben.

CH09-N21

Wenn ein experimentelles Feature oder eine A/B-Verteilung im Produkt wesentlich ist, sollte es im Beobachtungsprotokoll erhalten bleiben.

CH09-N22

Wenn während einer Welle eine wesentliche Produktänderung auftritt, sollte die Welle in separate Unterversionen unterteilt, neu gestartet oder die Änderung explizit in der Analyse modelliert werden.

CH09-N23

Eine bekannte Produktänderung kann nicht so gespeichert werden, als kämen alle Beobachtungen von einem einzigen festen System.

CH09-N24

Dass der Anbieter ein Update nicht erklärt, kann nicht als endgültiger Beweis dafür genommen werden, dass sich das Produkt nicht verändert hat.

CH09-N25

Das Ergebnis eines live verfügbaren Konsumprodukts kann nicht so präsentiert werden, als ob es das Ergebnis eines eingefrorenen Basismodells wäre.

CH09-N26

Das Ergebnis eines eingefrorenen Modells im Labor kann nicht als Leistung eines echten Nutzerprodukts dargestellt werden.

CH09-N27

Abwesenheit des Produkts, technischer Fehler, Nutzungslimit, Systemablehnung und fehlendes Ergebnis müssen mit separaten Status aufgezeichnet werden.

CH09-N28

Eine falsche oder negative KI-Antwort kann kein Grund für einen technischen Wiederholungsversuch sein.

CH09-N29

Das Vergleichsobjekt muss definiert werden, bevor Daten gesammelt werden.

CH09-N30

Das Vergleichsniveau von zwei Produkten muss im öffentlichen Ergebnis sichtbar sein.

CH09-N31

Eine eindeutige Produktbewertung kann nicht aus CP-0- oder CP-1-Level-Ergebnissen generiert werden.

CH09-N32

Die alleinige Verwendung desselben Prompts erzeugt kein vergleichbares Experiment.

CH09-N33

Bei Vergleichen sollten die Zielentität, Truth Pack, Prompt, Sprache, Land, Zeit, Population, Sitzung, Produkt und Bewertungsmethode gemeinsam untersucht werden.

CH09-N34

Natürliche Reichweite-Ergebnisse können nicht direkt ohne Erklärung eingestuft werden, da sie zu unterschiedlichen Zielpopulationen gehören.

CH09-N35

Das Ergebnis Gemeinsame Vergleichsbasis kann nicht als das tatsächliche globale Reichweitenergebnis des Produkts dargestellt werden.

CH09-N36

Ein web-unterstützter und web-freier Produktvergleich kann als Nutzerproduktvergleich durchgeführt werden; er kann nicht als grundsätzliche Modellüberlegenheit dargestellt werden.

CH09-N37

Bei Paarvergleichen, die mit demselben Nutzer durchgeführt werden, müssen Produktreihenfolge und Nutzerabhängigkeit beibehalten werden.

CH09-N38

Bei Vergleichen mit unterschiedlichen Nutzern müssen Nutzerzusammensetzung und Gewichtsunterschiede sichtbar sein.

CH09-N39

Die Produktversion oder Konfigurations-Fingerabdruck muss in jeder Messwelle gespeichert werden.

CH09-N40

Es muss einen verantwortlichen menschlichen oder institutionellen Eigentümer des Systemregistries und der Vergleichsentscheidung geben.

49. FORMEN DES VERSAGENS

CH09-F01— DEN ANBIETER ALS EIN EINZIGES KI-SYSTEM BETRACHTEN

Alle Produkte, Modelle und Oberflächen des Anbieters werden als ein einziges System dargestellt.

CH09-F02— PRODUKTVERHALTEN NACH MODELLNAMEN ZÄHLEN

Abruf, Sicherheit, Schnittstelle und Personalisierung werden ignoriert.

CH09-F03— BETRACHTEN SIE DAS SICHTBARE MODELLLABEL DEFINITIV ALS BACKEND

Die Verwendung von Routing- oder dynamischen Modellen wird außer Acht gelassen.

CH09-F04— BETRACHTUNG DES GLEICHEN ANBIETERS ALS VERGLEICHBAR

Unterschiede im Plan, in der Oberfläche und bei den Funktionen werden verborgen.

CH09-F05— BETRACHTUNG VERSCHIEDENER ANBIETER ALS NICHT VERGLEICHBAR

Produktvergleiche, die unter gepaarten Nutzerbedingungen möglich sind, werden abgelehnt.

CH09-F06— INTEGRATION VON WEB UND MOBILE

Schnittstellen- und Funktionsunterschiede verschwinden.

CH09-F07— KOMBINATION VON KOSTENLOSEN UND BEZAHLTEN PLÄNEN

Qualitäts- oder Funktionsunterschiede aufgrund des Plans werden unsichtbar.

CH09-F08— BETRACHTUNG VON UNTERNEHMENS- UND VERBRAUCHERPRODUKTEN ALS GLEICH

Private Informationen und Administrator-Einstellungen sind verborgen.

CH09-F09— API-ERGEBNIS ALS KONSUMPRDUKT BEHANDELN

Orchestrierungs- und Nutzerschnittstellenunterschiede werden beseitigt.

CH09-F10— WENN ABRUF VORHANDEN, ALS GENUTZT BETRACHTEN

Der wahre Wissenspfad der spezifischen Antwort kann nicht verifiziert werden.

CH09-F11— WENN KEINE ZITIERUNG, KEINE QUELLE BETRACHTEN

Die Möglichkeit eines verborgenen oder unsichtbaren Abrufs wird ignoriert.

CH09-F12— WENN ZITIERUNG VORHANDEN, ALS KORREKTE QUELLE ZÄHLEN

Übereinstimmung zwischen Zitation und Aussage wird nicht geprüft.

CH09-F13— TOOL-ZUGRIFF ALS TOOL-NUTZUNG BETRACHTEN

Es ist unbekannt, welches Werkzeug in der spezifischen Antwort verwendet wurde.

CH09-F14— SPEICHER ALS SAUBERE SITZUNG BETRACHTEN

Die Wirkung der Personalisierung bleibt unsichtbar.

CH09-F15— BETRACHTEN SIE EINE NEUE KONVERSATION ALS NICHT PERSONALISIERT

Persistente Kontoeinstellungen und spezielle Anweisungen werden ignoriert.

CH09-F16— AKTUELLEN CHAT ALS NEUE SITZUNG PRÄSENTIEREN

Frühere Kontexte werden verborgen.

CH09-F17— EIN EXPERIMENTELLES FEATURE ALS DATENFEHLER ZÄHLEN

Die tatsächliche Produktvariante wird als ungültige Beobachtung deklariert.

CH09-F18— EIN EXPERIMENTELLES FEATURE VERSTECKEN

Nutzer in verschiedenen Rollouts werden als ein einziges festes Produkt kombiniert.

CH09-F19— EINZELNE WELLE NACH UPDATE

Materielle Produktänderung wird als Ergebnis eines einzigen festen Systems veröffentlicht.

CH09-F20— ÄNDERUNGEN IGNORIEREN, WENN KEINE ANKÜNDIGUNG VORLIEGT

Beobachtetes Verhaltensproblem wird nicht untersucht.

CH09-F21— EIN LIVE-PRODUKT ALS EINGEFRORENES MODELL ZÄHLEN

Volle Reproduzierbarkeit wird behauptet.

CH09-F22— EIN EINGEFRORENES MODELL ALS ECHTES NUTZERPRODUKT BETRACHTEN

Produktkette und Schnittstellenunterschiede werden verborgen.

CH09-F23— DAS FEHLEN EINES PRODUKTS ALS FEHLER ZÄHLEN

Zugriffsbereich wird mit Genauigkeit der Repräsentation verwechselt.

CH09-F24— EINEN TECHNISCHEN FEHLER ABLEHNEN

Verbindungs- oder Dienstproblem wird als politische Folge eingestuft.

CH09-F25— EINEN TECHNISCHEN FEHLER ALS ABGELEHNT ZÄHLEN UND ERNEUT VERSUCHEN

Tatsächliches Nutzerergebnis wird aus dem Nenner entfernt.

CH09-F26— DAS NUTZUNGSLIMIT VERSTECKEN

Die echte Erfahrung des kostenlosen Plans wird unsichtbar.

CH09-F27— DAS GLEICHE PROMPT ALS DAS GLEICHE EXPERIMENT BETRACHTEN

Produkt-, Plan-, Sitzungs-, Web- und Zeitunterschiede werden ignoriert.

CH09-F28— BERÜCKSICHTIGUNG DES PRODUKTERGEBNISSES ALS MODELLÜBERLEGENHEIT

Der Unterschied über die gesamte Systemkette wird dem Basismodell zugeschrieben.

CH09-F29— DIREKTE RANGIERUNG DER Natürliche Reichweite PUNKTE

Unterschiedliche Bevölkerungsuniversen werden als gleich betrachtet.

CH09-F30— BERÜCKSICHTIGUNG VON Gemeinsame Vergleichsbasis ALS GLOBALER BEREICH

Das enge gemeinsame Universum wird als die gesamte Reichweite des Produkts dargestellt.

CH09-F31— ANSPRUCH AUF FÜHRUNG AUS CP-1 ERGEBNIS

Die Produktüberlegenheit wird ausschließlich aus nicht kompatiblen Ergebnissen abgeleitet, die nebeneinandergestellt werden.

CH09-F32— BEIBEHALTUNG DER VERGLEICHSSTUFE

Die Öffentlichkeit sieht nur das Ranking.

CH09-F33— IGNORIEREN DER GEPAARTEN NUTZERABHÄNGIGKEIT

Die Reaktionen derselben Person auf Produkte werden als unabhängig betrachtet.

CH09-F34— IGNORIEREN DES PRODUKTREIHENFOLGEEFFEKTS

Die erste Reaktion beeinflusst nachfolgende Aufgaben.

CH09-F35— UNTERSCHIEDLICHE PANEL-NUTZER ALS GLEICH BEHANDELN

Produkte werden mit unterschiedlichen Nutzerzusammensetzungen getestet.

CH09-F36— ÄNDERN DER REALITÄTSPAKET-VERSION

Die gleichen Reaktionen werden mit unterschiedlichen Referenzaufzeichnungen verglichen.

CH09-F37— DAS PRODUKTFINGERABDRUCK STILL SCHWEIGEND ÄNDERN

Der Konfigurationsdatensatz wird aktualisiert; die alte Messung wird beibehalten, anstatt überschrieben zu werden.

CH09-F38— EINE UNBEKANNTE FLÄCHE ALS SAUBER ZÄHLEN

Der Status von Modell, Speicher oder Werkzeug wird als kontrolliert akzeptiert, auch wenn unbekannt.

CH09-F39— DIE PRODUKTMARKE STATT NACHWEISE VERWENDEN

Der Ruf des Anbieters ersetzt die Messbedingungen.

CH09-F40— DECKT ALLE PRODUKTUNTERSCHIEDE MIT EINEM EINZIGEN KOEFFIZIENTEN AB

Plan-, Schnittstellen- und Konfigurationsunterschiede werden durch die Verwendung des Modell- oder Marktkoeffizienten verborgen.

50. PRÜFUNGSVERFAHREN

Schritt 1— Identifizieren Sie den Anbieter und das Produkt

Der Produktname, der Typ und die offiziellen Nutzeroberflächen werden aufgezeichnet.

Schritt 2— Festschreiben Sie die Messfläche

Web Mobil Desktop

API

Eingebettetes Produkt wird abgetrennt.

Schritt 3— Überprüfen Sie den Plan und die Kontoklasse

Kostenloser, bezahlter, Unternehmens- oder anderer Planstatus wird erfasst.

Schritt 4— Modell-ID-Vertrauen zuweisen

Das geeignete Niveau wird zwischen MI-0 und MI-5 bestimmt.

Schritt 5— Abruf- und Webkonfiguration speichern

Ist die Funktion verfügbar? Ist sie eingeschaltet? Wurde sie in einer bestimmten Antwort verwendet? Was ist der Nachweis?

Schritt 6— Referenz- und Werkzeugstatus speichern

Tatsächliche Nutzung wird von Funktionszugriff getrennt.

Schritt 7— Sitzungs- und Personalisierungsstatus speichern

Speicher, spezielle Anweisung, vorheriger Chat, Unternehmenskonfiguration werden überprüft.

Schritt 8— Land-, Gebietsschema- und Sprachregistrierung überprüfen

Die regionale Bedingung des Produktverhaltens wird bestimmt.

Schritt 9— Experimentelle Funktionen überprüfen

UI- oder Funktionsunterschiede zwischen Nutzern werden überprüft.

Schritt 10— Zeit- und Wellenidentität festschreiben

UTC Zeit- und Wellenaufzeichnung werden erstellt.

Schritt 11— Systemkonfigurations-Fingerabdruck erstellen

Gespeicherte Felder sind mit dem Integritätsdatensatz verknüpft.

Schritt 12— Überwachen von Änderungsereignissen

Angekündigte und beobachtete Produktänderungen werden aufgezeichnet.

Schritt 13— Schreiben des Vergleichsobjekts

Welche Vergleiche der Product Model Retrieval Plan Interface Time werden durchgeführt?

Schritt 14— Abgleichen der Vergleichsmaße

Materialbereiche werden tabellarisch dargestellt.

Schritt 15— CP-Level zuweisen

Es wird ein angemessenes Level zwischen CP-0 und CP-5 vergeben.

Schritt 16— Vergleichbare und nicht vergleichbare Aussagen trennen

Welche Ergebnisse nebeneinander gestellt werden können, wird klar angegeben.

Schritt 17— Ergebnisse von Natürliche Reichweite und Gemeinsame Vergleichsbasis trennen

Intra-Produkt- und gemeinsame Population-Vergleiche werden separat aufgezeichnet.

Schritt 18— Systemänderung in der Wave überprüfen

Es wird festgestellt, ob MIXED_SYSTEM_STATE oder eine neue Unterwelle benötigt wird.

Schritt 19— Öffentliche Vergleichskarte erstellen

Die Bedingungen hinter der Rangfolge werden sichtbar gemacht.

Schritt 20— Genehmigung des Datensatzbesitzers einholen

Das Systemregister und die Vergleichsentscheidung werden von der verantwortlichen Person versioniert.

51. ERFORDERLICHE BELEGE

Anbieteridentität; Produktname und -typ; Nutzeroberfläche; Anwendungs- oder Schnittstellenversion; Plan- und Kontoklasse; angezeigtes Modelletikett; Modellidentitätsnachweis; Konfidenzstufe der Modellidentität; Produktdokumentation; Länderzugriffsaufzeichnungen; Sprache und Gebietsschema; Web- und Abruffähigkeit; Nachweis der Abrufnutzung; Zitierfähigkeit; Zitiersicht; Zitat–Anspruchs-Verbindung; Werkzeugzugriff; Werkzeugnutzungsprotokoll; Speicher-Einstellung; Nutzerdefinierte Anweisungen; vorheriger Chat-Zustand; Sitzungstyp; Einstellungen des Organisationsadministrators; Aufzeichnung experimenteller Funktionen; A/B-Zuordnungssignale; UTC Zeitstempel; Messwelle;

Systemkonfigurations-Fingerprint; Anbieter-Update-Aufzeichnungen; beobachtete Verhaltensänderungen; Produktzugriffsfehler; technische Wiederholungsprotokolle; Natürliche Reichweite Bevölkerungsdatensatz; Gemeinsame Vergleichsbasis Bevölkerungsdatensatz; und Vergleichsdimensionstabelle.

CP-Level-Prompt-Version Reality-Paket-Version Proben- und Gewichtsaufzeichnung Adjudikationsversion Öffentliche Vergleichskarte Änderungsprotokoll Verantwortliche Person oder Institution

52. PRÜFLISTE FÜR AUDITS

Wurden der Anbieter und das Produkt getrennt aufgezeichnet? Ist die Nutzeroberfläche identifizierbar? Wurden der Plan und die Kontoklasse überprüft? Ist das Modelllabel lediglich das angezeigte Label oder identifiziert es eine verifizierte Bereitstellung? Wurde ein Vertrauensniveau für die Modellidentität vergeben? Wird das Ergebnis dem Produkt oder dem Modell zugeschrieben? Waren Web- oder Abruffunktionen verfügbar? Wurde ihre Verwendung in der spezifischen Antwort überprüft? Wurden Zitierfähigkeit und Zitationsanzeige getrennt? Unterstützen die Zitate tatsächlich die Behauptungen? Wurden Werkzeugzugriff und Werkzeugnutzung getrennt? War der Speicher aktiviert? Waren Nutzerdefinierte Anweisungen vorhanden? Gab es vorhergehenden Gesprächskontext? Wurde der Sitzungstyp aufgezeichnet? War eine Unternehmens- oder private Informationsquelle vorhanden? Wurden experimentelle Produktfunktionen überprüft? Könnten Land und Gebietsschema das Verhalten des Produkts beeinflussen?

Wurde der System-Fingerabdruck erstellt? Hat sich das Produkt während der Messwelle geändert? Gab es ein Verhalten außerhalb der Anbieterankündigung? Wurden Beobachtungen vor und nach der Änderung getrennt? Wurden technische Fehler von Ablehnungen unterschieden? Wurde ein erneuter Versuch aufgrund einer falschen Antwort durchgeführt? Ist das Vergleichsobjekt klar? Welche Bedingungen stimmen außer dem gleichen Prompt überein? Ist die Zielpopulation die gleiche? Sind die Versionen von Prompt und Truth Pack identisch? Sind die Unterschiede in Plan, Oberfläche und Abruf sichtbar? Wurde das CP-Level angegeben? Wurde ein Ranking aus den Ergebnissen von CP-0 oder CP-1 erstellt? Wurden die Ergebnisse von Natürliche Reichweite und Gemeinsame Vergleichsbasis getrennt? War die gleiche Nutzer-Produkt-Sequenz ausgeglichen? Wurden verschiedene Nutzerpanels kalibriert?

Wurde das Ergebnis des Live-Produkts als grundlegende Modellüberlegenheit präsentiert? Wurden unbekannte Bereiche als kontrolliert betrachtet? Gibt es eine öffentliche Vergleichskarte? Ist der verantwortliche Eigentümer des Systemregistrierungs bekannt?

53. EINWÄNDE UND ANTWORTEN

Einwand 1— „Wenn der Nutzer nicht weiß, welches Modell läuft, kann diese Arbeit trotzdem durchgeführt werden?“

Es kann gemacht werden. Das Hauptobjekt GEO-1000 ist das Produkt des Nutzers. Wenn das Modell unbekannt ist, wird das Ergebnis gemeldet als: „Dieses Verhalten wurde auf diesem Produkt und der Nutzeroberfläche beobachtet.“ Es wird kein Bezug auf ein spezifisches Modell genommen.

Einwand 2— „Warum unterscheiden wir die Produkte desselben Anbieters so stark?“

Weil der Nutzer:

- anderen Plan,
- Modell,
- Abruf,
- Speicher,
- Tool

unter unterschiedlichen Bedingungen unterschiedliche Antworten sehen kann. Der Name des Anbieters definiert nicht vollständig die Nutzererfahrung.

Einwand 3— „Ist es nicht sowieso wie Äpfel mit Birnen zu vergleichen, wenn man verschiedene Anbieter vergleicht?“

Wenn die Vergleichsfrage korrekt gestellt wird, ist sie es nicht. Ein echter Nutzer könnte fragen: „Welches Produkt informiert mich genauer?“ Die gesamte Produktkette ist Teil dieser Frage. Wenn die Überlegenheit des Basismodells behauptet werden soll, ist ein kontrollierteres Experiment erforderlich.

Einspruch 4– „Ist es nicht natürlich, dass ein webnutzendes Produkt einen Vorteil hat?“

Es ist natürlich. Bei einem Vergleich von Naturprodukten ist der Vorteil das echte Merkmal des Produkts. Bei einem Modellvergleich sollte der Abrufeffect ebenfalls getrennt werden.

Einwand 5— „Ist es nicht unfair, einem Produkt, das keine Zitate zeigt, keine Zitationsbewertung zu geben?“

Die Attributionsfunktion ist die Funktion, die das Produkt dem Nutzer bietet. Wenn die Funktion fehlt, kann das Attributionsergebnis wie folgt angezeigt werden:

ALS NICHT VERFÜGBAR AUFGEZEICHNET

Die Materialgenauigkeit wird ebenfalls separat gemessen.

Einwand 6— „Wenn Gedächtnis Teil der natürlichen Nutzererfahrung ist, warum schalten wir es dann aus?“

Es kann im natürlichen Panel eingeschaltet bleiben. Es kann im kontrollierten Panel ausgeschaltet werden, um Vergleiche zwischen Produkten zu löschen. Die beiden Panels beantworten unterschiedliche Fragen.

Einwand 7— „Reicht nicht das Öffnen eines neuen Chats für eine saubere Sitzung?“

Nicht immer. Persistenter Speicher, spezielle Anweisungen oder Kontoeinstellungen können weiterhin bestehen. Diese Bereiche sollten ebenfalls gespeichert werden.

Einwand 8— „Was können wir tun, wenn der Anbieter die Produktänderung nicht erklärt hat?“

Beobachtet:

- Schnittstelle,
- Funktion,
- Verhalten,
- Zuschreibung

Wir können Ihre Änderungen speichern. Wenn die Ursache unbekannt ist, wird ein unbestimmter Änderungsstatus wie UC-5 verwendet.

Einwand 9— „Wenn sich Produkte ständig ändern, werden dann irgendwelche Ergebnisse gültig sein?“

Das Ergebnis lautet:

- Datum,
- Welle,
- Produktprobe

Es ist gültig. Variabilität macht die Messung nicht bedeutungslos. Sie verhindert voreilige Urteile.

Einwand 10— "Ist nicht derselbe Prompt am selben Tag vergleichbar genug?"

Der Plan, die Oberfläche, das Gedächtnis, das Web, das Land und die Nutzerpopulation können unterschiedlich sein. Derselbe Prompt kann erforderlich sein. Allein ist es nicht genug.

Einwand 11— 'Wäre die CP-Stufe nicht zu technisch für die Öffentlichkeit?'

Ein kurzes Label kann verwendet werden:

- Beschreibend
- Teilweise übereinstimmend
- Übereinstimmend
- Kontrolliert
- Repliziert

Man sollte wissen, wie stark das öffentliche Ranking ist.

Einwand 12— “Warum können wir nicht direkt basierend auf den Natürliche Reichweite-Werten bewerten?”

Weil jedes Produkt möglicherweise in verschiedenen Ländern und Nutzerpopulationen gemessen wurde. Der Unterschied im Score kann teilweise darauf zurückzuführen sein, welche Population das Produkt erreicht hat. Gemeinsame Vergleichsbasis bietet einen faireren Vergleich.

Einwand 13— „Ist Gemeinsame Vergleichsbasis nicht ein Ergebnis aus der realen Welt?“

Es ist das Ergebnis der tatsächlich geteilten Population. Es zeigt jedoch keinen Zugang der Produkte außerhalb des geteilten Universums. Es sollte zusammen mit Natürliche Reichweite gelesen werden.

Einwand 14— „Ist die Inspektion des Live-Produkts wichtiger oder die des eingefrorenen Modells?“

Die Fragen sind unterschiedlich. Das Live-Produkt misst die Wahrheit des Nutzers. Das eingefrorene Modell bietet wissenschaftliche Kontrolle und Reproduzierbarkeit. NOMOS verwendet sie nicht austauschbar.

GEMEINSAME REGEL VON KAPITEL 57

Ein Nutzer:

- „ChatGPT hat das gesagt.“

- „Gemini dachte das."
- „Claude hat uns empfohlen."
- „Einige haben uns zitiert."

kann gesagt werden. Diese Ausdrücke sind im täglichen Gespräch verständlich. Sie sind jedoch nicht ausreichend in einem Prüfungsprotokoll. Denn unter diesen Namen:

- verschiedene Produkte,
- verschiedene Pläne,
- verschiedene Modelle,
- verschiedene Nutzeroberflächen,
- verschiedene Web- und Abrufsysteme,
- verschiedene Speicherbedingungen,
- unterschiedliche Länder,
- verschiedene Zeiten

können gefunden werden. Unterschiedliche Personen können unter demselben Namen unterschiedliche KI-Erfahrungen machen. Vergleichbare Nutzererlebnisse können unter verschiedenen Namen hergestellt werden. Ein Modelllabel scheint die gesamte Antwort produziert zu haben. Aber die Realität, die den Nutzer erreicht:

- Quellenauswahl,
- Sicherheit,
- Personalisierung,
- Schnittstelle,
- Werkzeuge

Das Ergebnis kann durch Sicherheitskontrollen, Personalisierung, die Nutzeroberfläche und verfügbare Werkzeuge beeinflusst werden. NOMOS fragt daher mehr als nur ‚Welches Modell?' Es fragt: Welches Produkt? / Welche Oberfläche? / Welcher Plan? / Welcher Informationsmodus? / Welche Sitzung? / Welcher Nutzer? / Welches Land? / Welche Zeit? / Welche Version? Zwei Ergebnisse können aus derselben Prompt stammen, ohne aus demselben Experiment zu kommen. Zwei Ergebnisse können von verschiedenen Anbietern stammen und dennoch unter denselben Nutzerbedingungen recht vergleichbar sein. Ein Produkt ist ein lebendes System: seine Quellen, Schnittstelle und sein Modell können sich ändern. Eine Änderung macht ein Ergebnis nicht wertlos; sie bindet das Ergebnis an eine bestimmte Zeit und den Zustand des Produkts. Das neunte Messgesetz von NOMOS lautet daher:

> *Ein KI-Anbieter ist kein KI-Produkt.*

Das zehnte Gesetz lautet:

> *Das Modell ist nur ein Teil der Repräsentationskette, die den Nutzer erreicht.*

Das elfte Gesetz lautet wie folgt:

> *Der gleiche Name garantiert Vergleichbarkeit; ein anderer Name garantiert keine Nichtvergleichbarkeit.*

Das zwölfte Gesetz lautet wie folgt:

> *Der gleiche Prompt ist nicht dasselbe Experiment.*

Das dreizehnte Gesetz lautet wie folgt:

> *Das Ergebnis des lebenden Produkts ist nur für die Produktprobe und den Zeitpunkt gültig, zu dem es gemessen wird.*

Das vierzehnte Gesetz lautet wie folgt:

> *Die Macht des Vergleichs kommt nicht von der Größe der Rangordnung, sondern von der Klarheit der abgeglichenen Bedingungen.*

Bestellung von Abschnitt 9 von NOMOS

> *Gib mir nicht nur den Namen des Anbieters. / Zeige, welches Produkt verwendet wurde, auf welcher Oberfläche und mit welchem Plan.*
>
> *Mache das auf dem Bildschirm sichtbare Modelletikett nicht zur exakten Identität des gesamten Back-End-Systems.*
>
> *Es wurde verwendet, weil es eine Webfunktion gab; gehe nicht davon aus, dass es die Behauptung unterstützte, nur weil ein Zitat vorhanden war.*
>
> *Zählen Sie den Werkzeugzugriff nicht als Werkzeugnutzung, einen neuen Chat nicht als sauberen Account und unbekannten Speicherstatus nicht als ausgeschalteten Speicher.*
>
> *Setze keinen kostenlosen Nutzer und einen zahlenden Nutzer, ein mobiles Produkt und ein Webprodukt oder einen Enterprise-Arbeitsbereich und einen allgemeinen Chat in dieselbe Zelle.*
>
> *Betrachten Sie Ergebnisse desselben Anbieters nicht automatisch als einen fairen Vergleich.*
>
> *Vermeiden Sie es nicht, verschiedene Anbieter nur deshalb zu vergleichen, weil sie unterschiedlich sind. / Vergleichen Sie zuerst die Bedingungen. / Erklären Sie dann, welche Frage Sie beantwortet haben.*
>
> *Präsentieren Sie kein Live-Produkt als gefrorenes Modell, keine API als Endverbrauchererlebnis und kein Produktergebnis als Überlegenheit des Basismodells.*
>
> *Wenn sich das System in der Mitte der Welle geändert hat, dürfen die alten und neuen Antworten nicht zu einem einzigen Produkt zusammengeführt werden.*
>
> *Ignorieren Sie die Verhaltensänderung vor Ihnen nicht einfach, nur weil der Anbieter die Änderung nicht erklärt hat.*
>
> *Gehen Sie nicht davon aus, dass das Unbekannte sauber, gleich oder behoben ist.*

Zuerst das System im Protokoll registrieren. / Dann die Konfiguration festschreiben. / Dann die Änderungen überwachen. / Dann das Vergleichsobjekt schreiben. / Dann übereinstimmende und nicht übereinstimmende Bedingungen anzeigen. / Und erst dann sagen, welches Produkt unter welcher Bedingung ein stärkeres Ergebnis erzielt hat.

Der abschließende Satz des Kapitels

> *Die „KI“, gemessen in GEO-1000, ist kein einzelner Name oder Modell; es handelt sich um ein versioniertes Repräsentationssystem, das einen bestimmten Nutzer zu einem bestimmten Zeitpunkt innerhalb einer bestimmten Produktkette erreicht.*

Normativer Kern

Jede GEO-1000-Beobachtung MUSS mit einem versionierten Eintrag im KI-Systemregister verknüpft sein, der die gemessene Produktinstanz beschreibt, einschließlich, soweit zutreffend: - Anbieter, - Produkt, - Nutzeroberfläche, - Plan, - angezeigte oder verifizierte Modellidentität, - Konfidenz der Modellidentität, - Abruf- und Webkonfiguration, - Zitierfähigkeit und Nutzung, - Werkzeugzugang und Nutzung, - Sitzungsstatus, - Speicher, - Nutzerdefinierte Anweisungen, - vorheriger Kontext, - Personalisierung, - Land, - Gebietsschema, - Sprache, - Bereitstellungsstatus, - Messzeitpunkt, - und bekannte oder beobachtete Systemänderungen. Anbietername, Produktname, angezeigtes Modellkennzeichen, Basismodell, Abrufsystem und für den Nutzer sichtbares Produkt DÜRFEN NICHT als austauschbare Identitäten behandelt werden. Die Verfügbarkeit von Funktionen MUSS von der Nutzung von Funktionen unterschieden bleiben. Ergebnisse von Live-Produkten DÜRFEN NICHT als eingefrorene Modell-Ergebnisse repräsentiert werden, und Ergebnisse von eingefrorenen Modellen DÜRFEN NICHT als das vollständige Erlebnis eines Verbraucherprodukts repräsentiert werden. Vergleiche MÜSSEN ihr Objekt definieren und einen offengelegten Grad der Vergleichbarkeit basierend auf übereinstimmenden Entitäten, Prompts, Population, Sprache, Land, Zeit, Sitzung, Plan, Schnittstelle, Abruf, Werkzeugen, Referenzaufzeichnungen und Bewertungsregeln erhalten. Derselbe Anbietername oder dieselbe angezeigte Modellbezeichnung DÜRFEN NICHT automatisch eine Annahme von Vergleichbarkeit erzeugen. Verschiedene Anbieter KÖNNEN verglichen werden, wenn die nutzerseitigen Bedingungen ausreichend übereinstimmen und die daraus resultierende Aussage auf das gemessene Produktverhalten beschränkt ist. Wesentliche Produktänderungen während einer Messwelle MÜSSEN einen separaten Systemzustand, eine Unterwelle, einen modellierten Änderungsfaktor oder eine Warnung über einen gemischten Zustand erzeugen. Unbekannte Konfigurationsfelder MÜSSEN als UNKNOWN gekennzeichnet bleiben und DÜRFEN NICHT als kontrolliert oder gleichwertig angenommen werden. Jede Entscheidung zum Systemregister und zur Vergleichbarkeit MUSS versioniert und einer verantwortlichen Person oder Organisation zugeordnet sein.

KAPITEL 10

Die Prompt-Verfassung

METHODISCHE GRUNDLAGEN / K04; K11; K12; K13

Kapitelgrenze

Abschnitt 9 hat die folgende Bestimmung eingeführt:

> *Der gleiche Anbietername, das gleiche Modellkennzeichen oder das gleiche Logo schafft keine vergleichbare KI-Erfahrung.*

Jetzt müssen wir das andere Hauptwerkzeug identifizieren:

> *Was genau wird der Nutzer das KI-Produkt fragen?*

Eine Prompt ist nicht nur ein Fragesatz. Eine Prompt:

- ruft die geprüfte Einheit an,
- definiert die Absicht des Nutzers,
- bestimmt, welche Aufgabe es von der KI erwartet,
- erweitert oder verengt den Umfang der Antwort,
- kann die Nutzung von Ressourcen fördern,
- kann um Ratschläge bitten,
- kann eine Annahme in die Antwort einbetten,

Es kann den Nutzer oder das System auf ein bestimmtes Ergebnis lenken. Diese Aufforderungen sind nicht gleich:

- „Was ist Apple.com?“
- „Mit welcher Organisation ist Apple.com verbunden und was macht diese Organisation?“
- „Ist Apple.com ein zuverlässiges Unternehmen?“
- „Warum ist Apple.com das beste Technologieunternehmen der Welt?“
- „Laut der offiziellen Webseite, was für ein Unternehmen ist Apple?“
- „Vergleichen Sie Apple mit ähnlichen Unternehmen und sagen Sie mir, ob Sie es empfehlen oder nicht.“

Sie enthalten alle denselben Domainnamen. Sie tragen nicht die gleiche Nutzerabsicht. Sie erzeugen nicht dieselbe Evidenzlast. Sie erwarten nicht dieselbe Antwort. Der erste Prompt ist mehrdeutig. Der zweite Prompt erfordert Identitäts- und Aktivitätsanalyse. Der dritte Prompt erfordert Bewertung und Beurteilung der Zuverlässigkeit. Der vierte Prompt bettet eine positive und überlegene Annahme in die Antwort ein. Der fünfte Prompt verknüpft die Antwort nur mit der offiziellen Quelle. Der sechste Prompt erstellt eine Vergleichs- und Empfehlungaufgabe. Die Antworten auf diese Fragen können ohne Erklärung nicht zu einem einzigen GEO-Score zusammengeführt werden. Bei der mehrsprachigen Messung wird das Problem noch größer. Der Ausdruck 'same prompt' kann zwei verschiedene Bedeutungen haben:

- Allen Nutzern denselben Zeichenstring geben
- Die gleiche Nutzerabsicht, Arbeitsbelastung und epistemische Grenze in allen Sprachen und Regionen bewahren

Die erste Methode kann innerhalb einer einzigen Sprache angewendet werden. Sie ist oft bedeutungslos zwischen Sprachen. Wort-für-Wort-Übersetzung:

- kann unnatürlich sein,
- kann eine andere soziale Bedeutung haben,
- kann den offiziellen oder informellen Ton ändern,
- kann Annahmen von Vertrauen, Qualität oder Empfehlung hinzufügen,
- kann den rechtlichen Umfang erweitern,

kann den Entitätsanker brechen. Daher bedeutet in GEO-1000„derselbe Prompt“:

> *Der gleiche festgeschriebene Prompt-Text innerhalb derselben Sprach- und Lokalisierungszelle; zwischen verschiedenen Sprachen und Lokalisierungen bedeutet es überprüfte semantische und funktionale Gleichwertigkeit.*

NOMOS verlangt, dass jede Messung den Abfragesatz, die Sprache, das Land, das Datum, das KI-Produkt und den Wiederholungsnachweis enthält; dass die Nachweise, Grenzen, der Kontext und die Zeit sichtbar sind. Die Prompt-Verfassung wendet diese Verpflichtung auf die Frage des Nutzers selbst an. Dieser Abschnitt:

- die Identität des Prompts als experimentelles Werkzeug,
- die Unterscheidung zwischen der Forschungsfrage und dem Prompt-Text,
- Die Prompt-Familie, Prompt-Version und Prompt-Instanz,
- diagnostische Prompts mit dem Core Mirror Prompt,
- kontrollierte und natürliche Prompt-Spuren,
- mehrsprachige semantische Äquivalenz,
- Übersetzung, Rückübersetzung und lokaler Pilot,
- Prompt-Steuerung,
- Annahme- und epistemische Lastprobleme,
- Entity-Anker und Geltungsbereichsgrenze,
- Prompt-Reihenfolge und Gesprächskontext,
- Prompt-Leckage und Überanpassung an Evaluierung,
- offene, versiegelte und zurückgehaltene Prompt-Sets,
- Prompt-Versionierung, Festschreibung und Integritätsprotokolle,
- die exakte Bereitstellung des Prompts für den Nutzer,

definiert. Dieses Kapitel behandelt noch nicht:

- alle Felddetails kontrollierter, sauberer und natürlicher Nutzerpanels,
- die vollständige Architektur der NOMOS Capture-Software,
- die Trennung der Antworten in atomare Aussagen,
- Schwelle für den Übergang zur endgültigen Antwort,
- die genauen Gewichte der Systemfamilien im NOMOS-Score

führt nicht zu einer endgültigen Entscheidung. Die Hauptfrage von Kapitel 10 lautet:

> *Wie können wir nachweisen, dass die an KI-Produkte gestellten Fragen tatsächlich dasselbe über Produkte, Nutzer, Länder, Sprachen und Zeiten hinweg messen?*

NOMOS Herausforderung

Sie geben mir vier Fragen zum selben Unternehmen:

- Welche Art von Unternehmen ist X?
- Ist X zuverlässig?
- Warum ist X das beste Unternehmen?
- „Würden Sie mir X empfehlen?“

Dann setzen Sie meine Antworten auf dieselbe GEO-Bewertung. In der ersten Frage erkläre ich die Identität und Aktivität. In der zweiten Frage suche ich nach einem Vertrauenskriterium. In der dritten Frage werde ich gezwungen, auf die Überlegenheitsannahme zu antworten, die Sie gestellt haben. In der vierten Frage erstelle ich Nutzer-Eignung und vergleichende Empfehlungen. Sie messen vier separate Aufgaben, als ob sie ein einziges Ergebnis wären. Dann:

„Du hast nach derselben Marke gefragt“, sagst du. Dasselbe Unternehmen ist nicht dieselbe Aufforderung. Jetzt schreibst du dieselbe Frage auf Englisch: „Was für eine Firma ist X?“ Du formulierst einen Satz in einer anderen Sprache, der eine ähnliche Bedeutung hat: „Ist X ein gutes und vertrauenswürdiges Unternehmen?“ Übersetzungen können in Bezug auf die Wörter ähnlich erscheinen.

Aber die erste erfordert eine Definition. Die zweite erfordert eine Bewertung. In der ersten Antwort kann ich erklären, was das Unternehmen tut. In der zweiten Antwort kann ich nach Nachweisen für Vertrauen und Empfehlung suchen. Dann zählst du die Punktedifferenz zwischen den Sprachen als meine Sprachleistung. Tatsächlich hast du nicht dasselbe Experiment durchgeführt. Jetzt gibst du eine weitere Eingabe:

„Erkläre X mithilfe der eigenen Website des Unternehmens.“ Ich übermittle nur die offizielle Aussage des Unternehmens. Für ein anderes Produkt sagen Sie: „Bewerten Sie X mithilfe unabhängiger Quellen.“ Dann vergleichen Sie die beiden Produkte. Sie haben einem die Rolle der offiziellen Repräsentation, dem anderen die Rolle der überprüften Realität zugewiesen. Der daraus resultierende Unterschied ist nicht nur ein Produktunterschied.

Es ist ein Unterschied in der Nutzerverfassung. Jetzt verbreiten Sie die Prompt in der ganzen Welt. Die englische Version ist natürlich. Die türkische Version wirkt wie maschinelle Übersetzung. Die japanische Version ist übermäßig formal. In einer anderen Sprache wurde der Firmenname in das falsche Schriftsystem umgewandelt und mit einer anderen Einheit verwechselt. Sie schaffen in jeder Sprache eine unterschiedliche kognitive und ontologische Aufgabe.

Dann veröffentlichen Sie den Bericht zur globalen Sprachgerechtigkeit. Das erste Dekret dieses Abschnitts lautet wie folgt:

> *Die Prompt ist kein neutrales Messgefäß; sie ist das experimentelle Werkzeug, das das Messergebnis bildet.*

Ihre zweite Bestimmung lautet:

> *Der gleiche Entitätsname ist nicht die gleiche Nutzerabsicht.*

Seine dritte Bestimmung besagt:

> *Dasselbe Wort sollte nicht über Sprachen hinweg beibehalten werden, aber dieselbe semantische und normative Aufgabe sollte beibehalten werden.*

Seine vierte Bestimmung besagt:

> *Wenn die Prompt bereits die Antwort enthält, misst sie kein GEO; sie misst das Prompting.*

Seine fünfte Vorschrift lautet:

Wenn sich die Version einer Prompt ändert, kann sich auch die gemessene Frage geändert haben.

1. ZWECK DES KAPITELS

Der Zweck dieses Abschnitts ist es, alle in GEO-1000 verwendeten Prompts prüfbar, mehrsprachig, ergebnisunabhängig und versionsgekennzeichnet messbar zu machen. [K04] Dieser Abschnitt macht folgende Unterscheidungen normativ:

- Forschungsfrage versus Nutzeraufforderung
- Zu messende Größe versus verwendete Wortfolge
- Eingabeaufforderungsfamilie versus einzelne Prompt
- Eingabeaufforderungsversion versus Eingabeaufforderungsinstanz
- Entitätsanker versus Zielentität
- Definition versus Bewertung
- Zitat versus Empfehlung
- Identitätsfrage versus Vertrauensfrage
- Aufgabe der offiziellen Repräsentation und Aufgabe der überprüften Faktizität
- Offene Aufforderung und formularbeschränkte Aufforderung
- Kontrollierte Aufforderung und natürliche Nutzeraufforderung
- Einzelrunde-Aufforderung und Mehrfachrunde-Protokoll
- Aufforderungsübersetzung und Anpassung der Aufforderung
- Wortgleichheit und semantische Äquivalenz
- Semantische Äquivalenz und funktionale Äquivalenz
- Sprache der Aufforderung und Sprache der Antwort
- Sprache und Gebietsschema
- Neutrale Formulierung und suggestive Formulierung
- Informationsanfrage und Präsupposition
- Quellanfrage und Webnutzungsanweisung
- Variabilität der Aufforderung und Variabilität des Produkts
- Paraphrase-Robustheit mit Hauptbewertung
- Offener Prompt-Satz mit versiegeltem Validierungssatz
- Überanpassung an Tests mit Testtransparenz
- Nachträgliche Prompt-Korrektur mit Prompt-Pilot
- Teilnehmer hat Text manuell mit technischer Prompt-Auslieferung modifiziert
- Unfähigkeit, mit Klärungs-Prompt zu antworten
- Mangel an faktischen Informationen mit Sicherheits- oder Richtlinienauslösung

Am Ende dieses Abschnitts sollte jedes GEO-1000-Audit in der Lage sein, klare Antworten auf die folgenden Fragen zu geben:

Welche Forschungsfrage misst dieser Prompt?

Welche Nutzerabsicht repräsentiert er?

Welche Annahmen sind in der Antwort eingebettet?

Ist dieselbe Aufgabe über Sprachen und Regionen hinweg erhalten?

Was ist der vollständige Text des Prompts, seine Version und das Integritätsprotokoll?

Hat der Teilnehmer den Prompt tatsächlich für das KI-Produkt festgeschrieben?

Wurde der Prompt eingefroren, bevor die Ergebnisse angesehen wurden?

2. ZENTRALE NORMATIVE BESTIMMUNG

Jeder innerhalb von GEO-1000 verwendete Prompt sollte zusammen mit der Forschungsfrage, der Zielmetrik, der Prompt-Familie, dem Entity-Anker, der Nutzerabsicht, dem Aufgabentyp, dem Umfang, dem Zeitrahmen, dem Quellprompt, dem Antwortformat, der Sprache, dem Gebietsschema, dem Äquivalenzstatus, der Version und dem Integritätsnachweis im Prompt-Register eingetragen werden, und er sollte festgeschrieben werden, bevor KI-Antworten angezeigt werden. Alle Teilnehmer in derselben Sprach- und Gebietsschema-Zelle sollten in derselben Prompt-Zelle:

- die gleiche Prompt-ID haben,
- die gleiche Version haben,
- die gleiche Zeichenkette haben,
- die gleichen Satzzeichen haben,
- denselben Entity-Anker haben,
- denselben Quell- und Formatierungsanweisungen folgen

müssen eingehalten werden. Prompts in verschiedenen Sprachen und Gebietsschemas müssen nicht wortwörtlich identisch sein. Sie müssen in den folgenden Bereichen wesentliche Äquivalenz aufweisen:

- Zielobjekt
- Nutzerintention
- Aufgabe
- Umfang
- Geographie
- Zeit
- Vorannahme
- Erkenntnisbelastung
- Antwortformat
- Erwartung an die Antwort
- Belastung durch Bewertung und Empfehlung
- Grad der Unsicherheit

Wenn sich einer dieser Bereiche wesentlich ändert:

- neue Version der Aufforderung,
- neue Stimuluszelle,
- separate Analyse

kann erforderlich sein.

3. DER STIMULUS IST EIN EXPERIMENTELLES MESSWERKZEUG

Wenn ein Thermometer die Temperatur beeinflussen kann, ist es kein gutes Messinstrument. Wenn ein Stimulus die Antwort einbettet, anstatt sie zu messen, ist es kein gutes GEO-Messinstrument. Der Stimulus kann das Ergebnis auf folgende Weise beeinflussen:

- Die Art und Weise, wie seine Existenz definiert wird
- Annahme von Vertrauen oder Qualität
- Anforderung von Ressourcen
- Internetzugang / Webnutzung
- Anfrage für Empfehlung
- Antwortlänge
- Antwortformat
- Nutzerprofil
- Zeit- und Länderbezug
- Frühere Informationen
- Emotionale oder kommerzielle Anleitung
- Rollenspiel-Anweisung
- Befehl, um ein spezifisches Ergebnis zu erreichen

Daher sollte der Prompt:

als initiale Testdaten gespeichert werden

Er darf im Bericht nicht zusammengefasst werden. Er muss im Volltext erhalten bleiben.

4. KONSTITUTIONELLE HIERARCHIE DES PROMPTS

ein GEO-1000-Prompt muss innerhalb der folgenden Hierarchie etabliert werden.

4.1. Forschungsfrage

Es ist die übergeordnete Frage, die die Messung zu beantworten versucht. Beispiel: "Können KI-Produkte die geprüfte Domäne mit der richtigen Unternehmenseinheit verknüpfen und deren Haupttätigkeit korrekt identifizieren?" Die Forschungsfrage muss dem Nutzer nicht exakt so angezeigt werden.

4.2. Vorhergesagte Menge

Es ist welche Wahrscheinlichkeit oder Verteilung die Messung vorhersagt. Beispiel: "Die Wahrscheinlichkeit, die richtige Entität und die primäre Aktivitätsdarstellung vor dem festgeschriebenen Core Mirror Prompt für die definierte Nutzerpopulation zu sehen."

4.3. Prompt-Familie

Es ist eine Gruppe von Prompts, die dieselbe Repräsentationsdimension messen. Beispiel:

- Identität
- Aktivität
- Nachweis

- Empfehlung
- Zeit
- Grenze

4.4. Kanonische Prompt-Absicht

Es ist eine sprachunabhängige Definition von Bedeutung. Sie erklärt Folgendes: Was möchte der Nutzer wissen? Welche Aufgabe wird von der KI erwartet? Welche Bereiche liegen im Anwendungsbereich? Welche Bereiche liegen außerhalb des Anwendungsbereichs? Welche Annahmen sind verboten? Welche Art von Antwortformat wird erwartet? Die kanonische Absicht ist kein Text, der einer einzelnen Sprache angehört.

4.5. Quellsprach-Prompt

Der Prompt-Text, in dem die kanonische Absicht zuerst geschrieben und redaktionell genehmigt wurde. Quellsprache:

- kann in der Lage sein, Englisch zu verwenden.
- Türkisch,
- kann eine andere Sprache sein

Es bedeutet nicht, dass die Wahl der Quellsprache anderen Sprachen überlegen ist.

4.6. Sprach- und Gebietsschema-Versionen

Dies sind Prompt-Texte, die für jede Sprache und jedes Gebietsschema natürlich, semantisch und funktional äquivalent sind.

4.7. Beispiel für einen an den Teilnehmer gelieferten Prompt

Dies ist der genaue Prompt-Text, der an den spezifischen Nutzer gesendet wurde. Er muss mit der Version im Prompt-Register übereinstimmen.

4.8. Tatsächlich an das KI-Produkt gesendeter Text

Dies ist die Zeichenfolge, die tatsächlich vom Teilnehmer gesendet wurde. Sie wird mit dem gelieferten Prompt verglichen.

5. GRUNDSTRUKTUR DES PROMPTS

Ein Prompt-Beispiel kann mit den folgenden Komponenten repräsentiert werden:

```
P_i= (e, ι, τ, κ, γ, χ, ε, φ, λ, ρ, ν)
```

Hier:

- e: Entitätsanker
- ι: Nutzerabsicht
- t: Aufgabentyp
- ϰ: Thema und Umfang
- g: geografischer oder jurisdiktionaler Kontext
- χ: zeitlicher Kontext
- e: epistemische und Quellenanforderung
- ϕ: Antwortformat
- λ: Sprache, Schriftsystem und Gebietsschema

- ϱ: Ton, Register und Natürlichkeit
- ν: Prompt-Version

Damit zwei Prompts als gleich betrachtet werden können, reicht es nicht aus, dass nur der Entitätsanker derselbe ist. Die Gleichwertigkeit der materiellen Komponenten ist erforderlich.

6. PROMPT IST TEIL DER ZIELGRÖSSE

Der einzigartige und universelle GEO-Score einer Entität ist nicht unabhängig von allen möglichen Fragen. Eine genauere Repräsentation:

```
θ_{E,A,P,U,t}
```

wobei:

- E: Einheit
- A: AI-Produktinstanz
- P: Anspruch oder Anspruchsfamilie
- U: Zielnutzerpopulation
- t: Zeit

Wenn sich der Anspruch ändert:

P_1≠ P_2

kann es sein. In diesem Fall:

```
θ_{E,A,P_1,U,t}
```

und:

```
θ_{E,A,P_2,U,t}
```

sind nicht die gleiche Menge. Daher:

> *Der GEO-Score muss immer mit dem Anspruchssatz und der Version verknüpft sein.*

7. ANSPRUCHSFAMILIEN

GEO-1000 sollte unterschiedliche Repräsentationsaufgaben in separaten Anspruchsfamilien messen.

PR-01— KERN-MIRROR-ANSRUCH

Misst die Kernidentität und die Hauptaktivität der Einheit. Beispielabsicht: 'Mit welcher Haupteinheit ist dieser Domainname oder diese Marke verbunden, und was ist die primäre Tätigkeit dieser Einheit?' Dieser Prompt:

- Führung,
- Vertrauen,
- Empfehlung,
- Preis,
- Leistung

soll nicht gewünscht sein.

PR-02— ENTITY-RESOLUTIONS-AUFFORDERUNG

Ein Domainname, eine Marke, ein Produkt oder der Name einer Person prüft, zu welcher Entität es gehört. Beispiel: 'Zu welcher Organisation oder Entität gehört X?'

PR-03— AKTIVITÄTS- UND UMFANGS-AUFFORDERUNG

Deine Existenz:

- Produkt,
- Dienstleistung,
- Land,
- Kunde,
- Kapazität

misst seine Grenzen.

PR-04— NACHWEIS- UND QUELLEN-AUFFORDERUNG

Es misst, mit welcher Quelle und welchem Evidenzstatus KI materielle Behauptungen trägt. Die Quellenanforderung in dieser Familie sollte klar definiert sein.

PR-05— LOKALE ODER JURISDIKTIONALE AUFFORDERUNG

Es misst ein spezifisches Land, einen Ort, Preis, Dienstleistung oder rechtlichen Umfang.

PR-06— TEMPORALE AUFFORDERUNG

Es misst aktuelle, historische oder Informationen mit einem spezifischen Datum. „Derzeit“, „in 2025“ und „seit seiner Gründung“ sind nicht dieselbe Aufgabe.

PR-07— EMPFEHLUNGS-PROMPT

Es misst, ob eine Entität für ein bestimmtes Nutzerbedürfnis geeignet ist. Nutzerprofil und Ausschlussbedingungen können obligatorisch sein.

PR-08— VERGLEICHS-PROMPT

Vergleicht mehrere Entitäten oder Optionen unter definierten Kriterien. „Welche ist besser?“ allein ist nicht ausreichend.

PR-09— GRENZEN- UND AUSSCHLUSSPROMPT

Deine Existenz:

- was es nicht tut,
- für wen es nicht geeignet ist,
- in welchem Land oder unter welchen Bedingungen es keine Dienstleistungen anbietet

Maßnahmen.

PR-10— ROBUSTHEITS- UND PARAPHRASE-PROMPT

Misst die Robustheit der Repräsentation in verschiedenen natürlichen Ausdrucksformen desselben kanonischen Intents. Es darf nicht in die Haupt-Core-Mirror-Bewertung gemischt werden, ohne dies zu erklären.

PR-11— STEUERUNGS-PROMPT

Testet, ob das Messwerkzeug funktioniert:

- positiver Kontrollwert,
- negativer Kontrollwert,
- Unsicherheitskontrolle

Es ist ein Prompt.

PR-12— VERSIEGELTER HOLDOUT-PROMPT

Überanpassung an Tests ist ein Prompt, der vorab gehasht und bis zum Ende der Messung versiegelt aufbewahrt wird, um das Erinnern oder die Optimierung eines Prompts für eine einzeln veröffentlichte Frage zu testen. Der Holdout-Prompt erfordert faire Verwaltung und anschließende Erklärung.

8. KERN-MIRROR-PROMPT

Es ist der Basis-Prompt, den alle berechtigten Nutzer in derselben Sprach-/Gebietsschema-Version in der Hauptmessung von GEO-1000 erhalten. Der Core-Mirror-Prompt muss folgende Merkmale haben:

- Einfaches Tempo
- Natürlich
- Kurz, aber nicht extrem vage
- Neutral
- Die Antwort nicht einbetten
- Keine Beratung angefordert
- Keinen Vergleich angefordert
- Die Nutzung von Quellen nicht erzwingen, wenn separate Merkmale nicht gemessen werden
- Ermöglicht die Entitätsauflösung
- Fordert Informationen zur Haupttätigkeit oder Kategorie an
- Gilt für alle Produkte

9. FÜHRT DAS UNTERNEHMEN NICHT MIT DEM DOMAIN-VERANKER ZUSAMMEN

Die natürliche Nutzeraufforderung: „Was für ein Unternehmen ist Apple.com?“ ist eine verständliche Frage. Ontologisch jedoch:

- den Domainnamen,
- die Website,
- das Unternehmen

Kann wie ein einzelnes Objekt verwendet werden. Aufgrund der in Abschnitt 3 etablierten Entitätsunterscheidung sollte die kontrollierte Core-Mirror-Prompt klarer sein. Synthetischer Vorschlag:

„Mit welcher Hauptorganisation ist apple.com verbunden, und was sind die Hauptaktivitäten dieser Organisation?“

Diese Prompt:

- verwendet den Domainnamen als Entitätsanker,

- deklarieren den Domainnamen nicht direkt als Unternehmen,
- fordert die Auflösung der Hauptentität an,

misst den Umfang der Aktivitäten. Ein anderer Kandidat:

"Mit welcher Hauptunternehmenseinheit ist apple.com verbunden, und was macht diese Einheit hauptsächlich?"

Die Natürlichkeit dieser Ausdrücke sollte in jeder Sprache separat getestet werden. In der Panel unter natürlichen Nutzungsbedingungen können in der Praxis verwendete Prompts, wie "Welche Art von Unternehmen ist Apple.com?", ebenfalls getestet werden. Kontrollierte und natürliche Prompt-Ergebnisse sollten getrennt gehalten werden.

10. KONTROLLIERTER PROMPT-TRACK UND NATÜRLICHER PROMPT-TRACK

Zwei sich ergänzende Prompt-Tracks können verwendet werden.

10.1. Kontrollierter Canonical-Track

Prompt:

- vollständig festgeschrieben,
- semantisch gleichwertig,
- Entitäts- und Aufgabenbegrenzungen klar,
- entwickelt für Vergleiche zwischen Produkten

Tut es. Diese Spur stärkt die Methodensteuerung.

10.2. Natürliche Nutzerspur

Natürlich von echten Nutzern erstellte Prompts werden verwendet. Diese Prompts können sein:

- kürzer,
- mehrdeutiger,
- lokal,
- gesprochen geschrieben

.

Diese Spur misst das Verhalten in der realen Welt.

10.3. Zwei Spuren können nicht zusammengeführt werden

Kontrollspur: beantwortet die Frage „Wie reagieren Produkte auf dieselbe definierte Frage?“ Natürliche Spur: beantwortet die Frage „Was passiert mit der natürlichen Vielfalt der Fragen von echten Nutzern?“ Die beiden Ergebnisse sollten getrennt veröffentlicht werden.

11. NEUTRALE PROMPT

Neutrale Prompt:

- zeigt keinen positiven oder negativen Aspekt der gewünschten Antwort an,
- macht keine Annahme über eine unbelegte Eigenschaft,
- leitet den Nutzer nicht zu einem bestimmten Ergebnis,

definiert seine Existenz nicht als gut oder schlecht. Ein neutraler Prompt könnte sein: „In welchen Bereichen bietet X Dienstleistungen an?“ Ein leitender Prompt könnte sein: „Welche überlegenen Dienstleistungen bietet X an?“ Der zweite Prompt:

- gibt an, dass die Dienstleistungen überlegen sind,
- trägt eine positive Qualität

Annahme.

12. VORANNAHME

Ein Prompt kann eine bestimmte Behauptung vor der Antwort als wahr annehmen. Beispiel: „Warum ist X der Branchenführer?“ Dieser Prompt nimmt die folgende Behauptung vorher an: „X ist der Branchenführer.“ KI:

- kann die Annahme ablehnen,
- kann sie einschränken,

sie kann bedingungslos akzeptiert werden. Dieser Prompt ist jedoch keine neutrale Messung von Führung. Er kann separat als Vorannahmetest verwendet werden.

12.1. Harmloses Vorurteil

„Was ist die offizielle Website von X?“ Es wird die Existenz einer Entität namens X angenommen. Akzeptabel, wenn die Identität bereits im Prüfprotokoll überprüft wurde.

12.2. Materielles Vorurteil

"Was sind die Hauptgründe für Xs Erfolg?" Akzeptiert Erfolg als bestätigte Tatsache. Kein neutraler Prompt.

12.3. Kritisches Vorurteil

"Welche Behandlungen bietet lizenzierte Gesundheitseinrichtung X an?" Wenn die Lizenzierung von X nicht verifiziert wurde, handelt es sich um eine gefährliche epistemische Erhebung.

13. FÜHRENDER HINWEIS

Ein führender Hinweis ist ein Rahmen, der den Nutzer oder das System zu einer bestimmten Antwort lenkt. Beispiele:

- „Empfehlen Sie X.“
- „Erklären Sie, warum X der Beste ist.“
- „Schreiben Sie eine positive Bewertung über X.“
- „Liste keine Konkurrenten auf; schlage nur X vor.“
- „Beweise, dass X ein Weltmarktführer ist.“

Diese Prompts messen nicht die GEO-Leistung. Sie können den Grad messen, in dem das System auf Eingabeaufforderungsmanipulation hereinfällt. Es ist ein separater ethischer oder Robustheitstest. Er kann nicht zur Haupt-GEO-Punktzahl hinzugefügt werden.

14. NEGATIVE ANLEITUNG

Dasselbe Prinzip gilt auch für negative Prompts. Beispiel: „Warum ist X unzuverlässig?“ Diese Prompt setzt Unzuverlässigkeit voraus. Sie kann in einem Test zur negativen Manipulation verwendet werden. Sie ist keine natürliche Vertrauensbewertung. NOMOS ist neutral nicht nur gegenüber positiver Steuerung, sondern auch gegenüber negativer Vergiftung.

15. QUELLENANFRAGEN ÄNDERN DIE AUFGABE DER Prompt

Die Prompt: „Was für ein Unternehmen ist X?“ ist nicht dasselbe wie: „Beschreibe X unter Verwendung unabhängiger Quellen.“ Die zweite Prompt:

- Abruf,
- Zitation,
- Quellenauswahl,
- fügt eine Aufgabe zur Unabhängigkeitsbewertung hinzu.

Ein weiterer Prompt ist anders: "Beschreiben Sie X nur mit seiner offiziellen Website." Dies nähert sich der Aufgabe der Official Representation Fidelity. Wenn eine Quellbefehl verwendet wird, sollten die Promptfamilie und das Wertungsfeld getrennt definiert werden.

16. ANLEITUNGEN ZUR WEBNUTZUNG

Anweisungen wie „Im Internet suchen.“ „Verwenden Sie die aktuellsten Quellen.“ „Quellen angeben.“:

- können die Webfunktion beeinflussen,
- die Nutzung von Werkzeugen,
- und das Zitierverhalten.

Wenn ein Produkt keine Webfunktion hat, kann sich die Aufgabe des Prompts ändern oder abgelehnt werden. Bei natürlichen Produktvergleichen sollte der Prompt keine anbieter-spezifischen Webbefehle enthalten. Die Web-aktiv/aus-Modi von Produkten sollten über die Systemkonfiguration in Abschnitt 9 verwaltet werden.

17. ANTWORTFORMAT

Der Prompt legt fest:

- die Länge der Antwort,
- die Struktur,
- das Detail,
- die Anzahl der Quellen

kann bestimmen. Beispielanweisungen:

- „Antworten Sie in einem Satz.“
- „Erklären Sie in 100 Wörtern.“
- „Schreiben Sie in Aufzählungspunkten.“

- „Nennen Sie nur den Firmennamen.“
- „Verwenden Sie höchstens drei Quellen.“

Diese Anweisungen:

- materielle Mängel,
- falsches Weglassen,
- Zitationsanzahl

wirken sich direkt aus. Das Antwortformat sollte für alle verglichenen Aufforderungen gleichwertig sein. Wenn keine Format-Anweisung vorliegt, wird die natürliche Antwortform gemessen.

18. LÄNGENBESCHRÄNKUNG

Kurze Antwort: kann die Hauptidentität zeigen, kann Grenzen- und Evidenzdetails verlieren. Lange Antwort:

- mehr finanzielle Aussagen,
- mehr Fehlerchancen,
- mehr Zitationen

kann produziert werden. Daher erfordern Antworten mit unterschiedlich langen Anweisungen:

- Anzahl der atomaren Behauptungen,
- Anzahl der Fehler,
- Umfang

sorgfältigen Vergleich. Das Haupt-Core-Mirror-Prompt sollte, wenn möglich, kein spezifisches Wortlimit enthalten.

19. NUTZER-PERSONA

Das Prompt kann die folgenden Informationen enthalten:

- „Ich bin ein Kleinunternehmer.“
- „Ich bin ein Verbraucher, der in der Türkei lebt.“
- „Ich habe ein Budget von 10,000 Euro.“
- „Ich suche rechtliche Beratung.“

Diese Informationen können für Ratschläge und lokale Eignung notwendig sein. Das Hinzufügen unnötiger Persona im Identitätsprompt kann die Antwort verändern. Nutzer-Persona:

- vordefiniert,
- relevant,
- in allen Sprachen gleichwertig

muss sein.

20. ZEITKONTEXT

Die folgenden Prompts sind unterschiedlich:

- Welche Art von Unternehmen ist X?
- "Was für eine Firma ist X gerade?"

- "In was für einer Firma war X 2024?"
- "In welchen Bereichen hat X seit seiner Gründung gearbeitet?"

"Derzeit" fügt eine Aktualisierungswährung hinzu. Historische Aufforderung erfordert alte Aufzeichnungen. Zeitkontext:

- offen,
- verbunden mit demselben Datumsstandard,
- im Einklang mit der Messwelle

Muss so sein. Wenn der Ausdruck "heute" verwendet wird, muss das absolute Datumsäquivalent im Prompt-Datensatz gespeichert werden.

21. GEOGRAFISCHER UND JURISDIKTIONELLER KONTEXT

Die Aufforderung: "Bietet X Service an?" ist nicht dasselbe wie: "Bietet X Dienstleistung für einzelne Kunden in der Türkei an?" Der zweite Prompt:

- Land,
- Art des Kunden,
- eine lokale Operation besitzt

Grenze. Geografischer Kontext darf bei der Übersetzung nicht verloren gehen. "Europa", "EU", "Eurozone" und "europäische Länder" decken nicht denselben Umfang ab.

22. VERGLEICHENDER PROMPT

Die Frage „Was ist besser, X oder Y?“ für sich genommen ist kein messbarer Vergleich. Folgendes sollte angegeben werden: Welcher Nutzer? Welches Bedürfnis? Welches Land? Welches Budget? Welches Kriterium? Welche Zeit? Welche Nachweise? Ein angemessenerer Prompt: „Vergleichen Sie für ein kleines Unternehmen in der Türkei X und Y im Bereich Webentwicklungsdienste in Bezug auf Preistransparenz, Leistungsumfang und überprüfbare Referenzen.“ Dieser Prompt gehört zur separaten Familie der Vergleichs-Prompts. Er kann nicht zum Core-Mirror-Score hinzugefügt werden.

23. EMPFEHLUNGSPROMPT

Eine Empfehlungseingabe trägt eine höhere Entscheidungsbelastung als eine Identifikationseingabe. Beispiel: „Würden Sie X empfehlen?“ Diese Frage könnte die folgenden Informationen weglassen: Für wen? Für was? Wo? Mit welchem Budget? Unter welchem Risiko? Die Empfehlungseingabe sollte in dem Maße einen Nutzer-Eignungsvektor enthalten, wie es relevant ist. Kandidatenrepräsentation:

```
U=(Bedarf, Land, Budget, userType, Risiko, Einschränkungen)
```

Als Ergebnis der Empfehlung kann es nicht wie eine allgemeine GEO-Sichtbarkeit unabhängig vom Nutzerprofil dargestellt werden.

24. UNSICHERE ANFRAGE

Wenn eine Anfrage für mehrere vernünftige Interpretationen offen ist:

- kann die KI um Klärung bitten,

- kann die wahrscheinlichste Interpretation wählen,

kann mehrere Interpretationen erklären. Unsicherheit kann Teil der Hauptmetrik sein. In einem kontrollierten Vergleich sollte jedoch eine unsichere Anfrage:

- im Pilotprojekt erkannt werden,
- bei Bedarf geklärt werden,

die Änderung sollte versioniert werden. Nachdem die Anfrage festgeschrieben ist, kann nicht gemäß der KI-Antwort gesagt werden: "Eigentlich haben wir etwas anderes gefragt."

25. KLÄRUNGSPROTOKOLL

Ein KI-Produkt, das um Klarstellung bittet, ist kein automatisches Versagen. Beispiel: „Auf welches Apple beziehen Sie sich?“ Dies kann angemessen sein, wenn es eine Entitätskollision gibt. Der Klärungspfad muss vordefiniert sein.

25.1. Einzelrunden-Kernpfad

Die erste Antwort wird unabhängig davon, was sie ist, aufgezeichnet. Eine Klärungsfrage ist ebenfalls ein Ergebnis. Keine Folge-Nachricht wird gesendet.

25.2. Standardisierter Klärungspfad

Wenn die KI um Klarstellung bittet, wird ein vorab festgelegter Klärungs-Prompt, der in allen Produkten verwendet wird, gesendet. Die erste und zweite Runde werden getrennt aufgezeichnet.

25.3. Natürlicher Klärungspfad

Der natürliche Nutzer antwortet mit eigenen Worten. Diese Methode ist für das natürliche Panel. Sie darf nicht mit der kontrollierten Hauptbewertung verwechselt werden.

26. EIN-RUNDEN- UND MEHRRUNDEN-PROMPTS

Ein-Runden-Prompt:

- verbessert die Vergleichbarkeit,
- Unabhängigkeit,
- und Sitzungsreinheit.

Mehr-Runden-Protokoll:

- Beschreibung,
- Quellabfrage,
- Korrektur,
- Empfehlungsbegründung

kann gemessen werden. Mehr-Runden-Antworten tragen den Kontext vorheriger Runden. Sie können nicht direkt mit Einzeln-Runden-Ergebnissen aus derselben Prompt-Datenbank kombiniert werden.

27. PROMPT-REIHENFOLGE

Wenn ein Teilnehmer mehr als einen Prompt erhält, kann der erste Prompt die nachfolgenden Antworten beeinflussen. Beispielssequenz:

- Welche Art von Unternehmen ist X?
- „Welche negativen Informationen gibt es über X?“
- „Würden Sie X empfehlen?“

Die dritte Antwort kann von den ersten beiden Gesprächen beeinflusst sein. Kontrollen:

- Neues Gespräch für jeden Prompt
- Prompt-Reihenfolge randomisieren
- Lateinisches Quadrat oder ausgeglichene Reihenfolge
- Getrennte Nutzer pro Prompt-Familie
- Die Reihenfolge der Variablen in der Analyse beibehalten

Das Haupt-Kern-Spiegel-Prompt sollte standardmäßig in einer neuen und separaten Sitzung angewendet werden.

28. PROMPT-KONTAMINATION

Gleicher Nutzer:

- die Website der geprüften Einrichtung,
- vorherige KI-Antworten,
- andere Produktergebnisse

könnten gesehen worden sein. Das Verhalten des Teilnehmers kann sich ändern, selbst wenn das Prompt genau so gesendet wird. Besonders in einem natürlichen Panel zeigt der Nutzer:

- weitere Folgefragen,
- Aktualisierung der Antwort,
- Öffnen von Quellen

tendenziell. Prompt-Kontamination sollte im Nutzer- und Sitzungsprotokoll erhalten bleiben.

29. ÄQUIVALENZ DES PROMPTS ÜBER SPRACHEN HINWEG

Mehrsprachige Prompts müssen drei separate Äquivalenzen tragen.

29.1. Semantische Äquivalenz

Tragen die Aufforderungen die gleiche Grundbedeutung?

29.2. Funktionale Äquivalenz

Fordert sie die KI zur gleichen Aufgabe auf? Definiert sie etwas in einer Sprache und bittet um Rat in einer anderen?

29.3. Normative Äquivalenz

Erhält sie die gleiche Beweislast, Annahmen, Reichweite und Evaluationsgrenzen bei? Wenn Sie in einer Sprache fragen: "Wie definiert sich das Unternehmen?" und in einer anderen: "Was ist das Unternehmen tatsächlich?", besteht keine normative Äquivalenz.

30. ÜBERSETZUNG, ANPASSUNG UND LOKALISIERUNG

Diese drei Prozesse sollten getrennt werden.

30.1. Übersetzung

Überträgt die Bedeutung des Ausgangstextes in die Zielsprache.

30.2. Sprachliche Anpassung

Macht es in der Zielsprache natürlich, verständlich und nützlich.

30.3. Lokalisierung

Wo notwendig, passt die Lokalisierung länderspezifische Elemente wie an:

- Aktualität,
- Gesetz,
- Begriff,
- Institution,
- Nutzerkontext

Da Lokalisierung die experimentelle Bedingung verändern kann, ist das Ergebnis eine eigene lokale Version des Prompts.

31. WARUM EINE WÖRTLICHE ÜBERSETZUNG NICHT AUSREICHT?

Wörtliche Übersetzung:

- unnatürliche Syntax,
- unterschiedliches Formalitätsniveau,
- falsche Verbverwendung,
- unterschiedliche Bedeutung von Vertrauen oder Empfehlung,
- Unterbrechung des Entitätsankers

kann entstehen. In einer Sprache ist „was für ein Unternehmen“ eine natürliche Definitionsfrage. Ihre direkte Übersetzung in eine andere Sprache kann andere Bedeutungen tragen, wie zum Beispiel:

- Qualitätsklasse,
- Vertrauen,
- Unternehmensart

und ähnliche andere Bedeutungen. Daher lautet die Aufforderung in der Zielsprache:

bezieht sich auf die kanonische Absicht,
nicht auf die Ausgangswörter

sollte verbunden sein.

32. MEHRSPRACHIGES PROMPT-ENTWICKLUNGSPROTOKOLL

Jede Sprach- und Länderversion muss die folgenden Stufen durchlaufen.

Stufe 1— Kanonische Absichtsübersicht

Eine sprachunabhängige Aufgabenbeschreibung wird erstellt:

- Entität
- Nutzerintention
- Gewünschte Aufgabe
- Umfang
- Ausschlüsse
- Verbotene Vorannahmen
- Erwartung an die Antwort
- Antwortformat

Stufe 2— Erste Muttersprache-Übersetzung

Ein Übersetzer, der die Zielsprache natürlich verwendet, erstellt den Prompt. Ein maschinelles Übersetzungsentwurfstool kann verwendet werden. Es kann nicht allein die endgültige Version sein.

Stufe 3— Unabhängige Rückübersetzung

Eine zweite Person übersetzt das Zielprompt zurück in die Ausgangssprache oder die kanonische Intent-Sprache. Die Rückübersetzung sollte nicht vom ersten Übersetzer durchgeführt werden.

Phase 4— Semantischer Vergleich

Das Quellprompt, Zielprompt und die Rückübersetzung werden in den folgenden Bereichen verglichen:

- Entitätsanker
- Intention
- Aufgabe
- Umfang
- Zeit
- Geographie
- Vorannahme
- Beweislast
- Tonfall
- Antwortformat

Phase 5— Kognitiver Pilot für lokale Nutzer

Für eine kleine Anzahl lokaler Nutzer: Was denken Sie, welche Frage hier gestellt wird? Welche Antwort wird erwartet? Enthält sie eine positive oder negative Annahme? Welche Firma oder Entität verstehen Sie? Klingt es natürlich? Diese Fragen werden gestellt. Dieser Pilot zählt nicht zur GEO-Bewertung. Er testet das Prompt-Tool.

Phase 6— Überprüfung durch lokale Experten

Lokale Fachkenntnisse können für Aufforderungen in den Bereichen Recht, Gesundheit, Finanzen oder anderen Feldern erforderlich sein.

Stufe 7— Gleichwertigkeitsentscheidung

Prompt:

- bestehen,
- bedingt bestehen,
- Überarbeitung erforderlich,
- nicht vergleichbar

Die zutreffende Klassifizierung muss dokumentiert werden.

Stufe 8— Versionssperre

Die Aufforderungs-ID, der vollständige Text, die Unicode-Repräsentation und der Hash werden aufgezeichnet.

33. GRENZEN DER RÜCKÜBERSETZUNG

Rückübersetzung ist nützlich. Allein beweist sie jedoch keine Gleichwertigkeit. Eine Zielaufforderung:

- kann für einen lokalen Nutzer seltsam sein,
- zu formell,
- direktiv,
- kulturell anders bedeutungsvoll.

Rückübersetzung kann der Quellsatz dennoch ähneln. Daher:

> *ist eines der Werkzeuge zur Überprüfung von Äquivalenz.*

Es ist nicht der endgültige Richter.

34. DIMENSIONEN DER PROMPT-ÄQUIVALENZ

Jeder Sprachprompt kann entlang dieser Dimensionen bewertet werden.

EQ-1— Entitätsanker

Wird auf dieselbe Entität Bezug genommen?

EQ-2— Nutzerabsicht

Stellt es denselben Informationsbedarf dar?

EQ-3— Sprechakt

Ist die Aufgabe der Identifizierung, Bewertung, Empfehlung oder des Vergleichs dieselbe?

EQ-4— Umfang

Ist der Umfang von Produkt, Dienstleistung, Land und Nutzer derselbe?

EQ-5— Zeitlicher Rahmen

Ist der Rahmen aktuell, historisch oder zeitlos derselbe?

EQ-6— Geografischer und rechtlicher Rahmen

Sind das Land und die Gerichtsbarkeit gleich?

EQ-7— Voraussetzung

Werden die gleichen Behauptungen vorausgesetzt?

EQ-8— Erkenntnislast

Ist dieselbe Quelle und Stärke der Nachweise erforderlich?

EQ-9— Valenz und Haltung

Ist der Ton positiv, negativ oder neutral gleich?

EQ-10— Antwortformat

Sind Länge, Struktur und Quellenangaben gleich?

EQ-11— Mehrdeutigkeit

Ist es ähnlich klar oder mehrdeutig?

EQ-12— Natürlichkeit und kognitive Belastung

Trägt es eine ähnliche Natürlichkeit und Verständniskapazität in der Zielsprache?

35. PROMPT-ÄQUIVALENZ-WERT

KANDIDATENMETHODE

Äquivalenzbewertung für die Dimension d:

- 0: nicht äquivalent
- 1: teilweise äquivalent
- 2: materiell äquivalent

kann vergeben werden. Gewichteter Äquivalenzwert:

```
PE_l= 100× [Σ_d α_de_{l,d}] / [2 Σ_d α_d]
```

kann wie folgt berechnet werden. Hier:

- el,d: Dimensionswert für Sprache l
- αd: Dimensionsgewicht

Einige Dimensionen müssen jedoch nicht kompensierbare Barrieren sein. Wenn einer dieser Bereiche 0 ist, sollte die Prompt nicht in die Hauptvergleichsmenge aufgenommen werden:

- Entitätsanker
- Nutzerintention
- Sprechakt
- Umfang
- Präsupposition
- Erkenntnisbelastung

Eine hohe Gesamtpunktzahl kann einen falsche Entität oder eine falsche Aufgabe nicht ausgleichen.

36. PROMPT-ÄQUIVALENZSTUFEN

PE-0— UNÜBERPRÜFT

Übersetzung oder Aufforderungsäquivalenz wurde nicht bewertet.

PE-1— MASCHINELL ÜBERSETZTER ENTWURF

Es existiert nur eine maschinelle Übersetzung oder ein einseitiger Entwurf. Sie ist nicht für die Hauptvergleichsmessung geeignet.

PE-2— VON MENSCHEN ÜBERSETZT

Es gibt eine Übersetzung durch Menschen auf muttersprachlichem Niveau. Es existiert keine unabhängige Äquivalenzprüfung.

PE-3— UNABHÄNGIG ÜBERPRÜFT

Unabhängige Rückübersetzung und semantische Überprüfung wurden durchgeführt.

PE-4— LOKAL PILOTIERT

Lokales Anwender-Cognitive-Piloting und Überprüfung der Natürlichkeit wurden abgeschlossen.

PE-5— REPLIZIERTE ÄQUIVALENZ

Die Aufforderung wurde erneut auf Äquivalenz über verschiedene Wellen, Anwender und Produkte getestet. Die Haupt-Mehrsprachigen GEO-1000-Aufforderungen sollten mindestens:

PE-4

angestrebt werden. Für Sprachen mit geringen Ressourcen kann ein niedrigeres Niveau verwendet werden. Die Einschränkung sollte eindeutig angegeben werden.

37. REGISTER UND TON

Zwischen den Sprachen:

- Offiziell,
- neutral,
- gesprochene Sprache,
- freundlich,
- imperativ

Tonunterschiede können das Antwortverhalten beeinflussen. Eine Prompt in einer Sprache: „Könnten Sie das erklären?“ kann nicht in eine andere Sprache übersetzt werden als: „Beweise es.“ Ton:

- angemessen an das natürliche Nutzerverhalten,
- so neutral wie möglich,
- nicht übermäßig höflich oder aggressiv

sollte auf eine Weise eingerichtet werden.

38. POLITISCHE UND SICHERHEITSAUSLÖSER

Ein Wort oder Entitätsname in einigen Sprachen:

- hat eine andere Bedeutung,
- sensible Kategorie,
- politischer oder rechtlicher Begriff,
- Sicherheitsfilter

kann ausgelöst werden. Dasselbe semantische Prompt kann in einer Sprache eine normale Antwort und in einer anderen eine Ablehnung erzeugen. Dies ist echtes Produktverhalten. Unnötige Auslösungen bei der Prompt-Übersetzung sollten jedoch gesondert geprüft werden. Prompt-Äquivalenz umfasst nicht nur die Bedeutung, sondern auch die Anwendbarkeit der Aufgabe.

39. ENTITY ANCHOR REGISTER

Der Entity Anchor in jedem Prompt sollte mit einem der folgenden Typen aufgezeichnet werden:

- Kanonischer Markenname
- Rechtlicher Name
- Domainname
- Produktname
- Personenname
- Lokalisierter offizieller Name
- Transliteration
- Abkürzung
- Mehrere Anker
- Erklärende Disambiguierung

Ein Entity Anchor muss nicht in allen Sprachen dieselbe Zeichenfolge enthalten. Er muss mit derselben Entität verknüpft sein.

40. ENTITY ANCHOR, DER DAS ERGEBNIS BEEINFLUSST

Die folgenden Anker können unterschiedliche Antworten erzeugen:

- „Apple“
- „Apple Inc.“
- „apple.com“
- „Apple Technologieunternehmen“

Erster Anker:

- Unternehmen,
- Frucht,
- Musikunternehmen,
- eine andere Einheit

kann zwischen ihnen Mehrdeutigkeit hervorrufen. Der zweite stärkt das rechtliche Unternehmen. Der dritte erfordert die Auflösung des Domainnamens. Der vierte setzt die Kategorieresponse in die Prompt. Diese Anker sind nicht derselbe Eingabetext.

41. EINGABEDRIFT

Eine wesentliche Abweichung vom ursprünglichen Zweck der Prompt wird bezeichnet als:

Eingabedrift

Die Drift muss klassifiziert und aufgezeichnet werden.

PD-1— Lexikalische Drift

Das Wort ändert sich. Die wesentliche Bedeutung kann erhalten bleiben. Nicht jede lexikalische Änderung ist ein Fehler.

PD-2— Semantische Drift

Die Kernbedeutung ändert sich.

PD-3— Entitätsdrift

Eine andere Entität, ein Produkt oder eine juristische Person wird genannt.

PD-4— Bereichsdrift

Der Service, das Land, das Produkt oder der Nutzerbereich ändert sich.

PD-5— Aufgaben-Drift

Eine Definitionsfrage wird zu einem Ratschlag oder Vergleich.

PD-6— Präsumptionsdrift

Ein neutraler Prompt erhält eine positive oder negative Präsumption.

PD-7— Evidenzdrift

Die Nutzung von Quellen oder die Forderung nach Unabhängigkeit ändert sich.

PD-8— Zeitliche Drift

„Aktuell“, „historisch“ oder bestimmte Zeiträume ändern sich.

PD-9— Formatdrift

Die Länge oder das Format der Antwort ändert sich.

PD-10— Registerdrift

Ton und Formalität können das Antwortverhalten verändern.

PD-11— Richtlinien-Auslöserabweichung

Änderungen der Übersetzungssicherheit oder des Richtlinienauslösers.

PD-12— Lokalisierungsabweichung

Änderungen im Land, in der Währung oder im rechtlichen Kontext. Wesentliche Abweichungen erfordern eine neue Version des Prompts.

42. PROMPT-VERSIONIERUNG

Kandidatenversionsformat: Major.Minor.Patch

Major-Änderung

Änderungen bei Nutzerabsicht, Aufgabe, Entitätsanker, Umfang, Standardannahme, Quelllade. Eine neue Testreihe kann erforderlich sein. Beispiel: 1.0.0→ 2.0.0

Minor-Änderung

Dabei muss die wesentliche Absicht beibehalten werden:

- Natürlichkeit,
- Klarheit,
- Lokalisierungsanpassung

Änderungen. Vergleichbarkeit wird ebenfalls untersucht. Beispiel: 1.0.0→ 1.1.0

Patch-Änderung

Bedeutung wird nicht beeinflusst:

- Rechtschreibung,
- Unicode,
- Interpunktion,
- technische Lieferung

ist eine Korrektur. Sie wird ebenfalls aufgezeichnet. Beispiel: 1.0.0→ 1.0.1

43. Festschreibung des Prompts

Die folgenden Felder müssen vor jeder Messwelle festgeschrieben werden:

- Prompt-Record-Version
- Kanonische Absicht
- Sprach- und Gebietstext
- Entitätsanker
- Gerätefamilie
- Antwortformat
- Quellen- und Werkzeuganweisungen
- Klärungsprotokoll
- Prompt-Reihenfolge
- Zuweisungsmethode
- Äquivalenzentscheidungen
- Vollständige Zeichenfolgen
- Hash-Werte
- Menschliche Genehmigung

- Sperrdatum

In diese Datei:

NOMOS Eingabeprompt-Sperre

kann den Namen erhalten.

FALLS EIN PROMPT-FEHLER WÄHREND DER 44 TH WELLE AUFTRITT

Es könnte ein materieller Fehler im Prompt vorliegen. Beispiel:

- Falscher Firmenname
- Falscher Standort
- Empfehlung zur Annahme in der Übersetzung
- Fehlende Sprache in der Quellanweisung
- Fehler im rechtlichen Umfang

Optionen: Die Welle wird gestoppt. Alte Prompt-Beobachtungen werden als separate Unterwelle aufbewahrt. Die Messung wird mit der neuen Version neu gestartet. Bei geringfügigem Fehler wird eine begrenzte und dokumentierte Korrektur vorgenommen. Falsche Option: der Prompt wird stillschweigend korrigiert und die alte/neue Antwort unter derselben Version zusammengeführt.

45. PROMPT-PILOT

Der Zweck des Prompt-Piloten besteht nicht darin, zu sehen, ob KI-Produkte hohe oder niedrige Bewertungen erzielen. Es soll Folgendes getestet werden: Versteht der Nutzer den Prompt korrekt? Ist der Entitätsanker offen? Ist die Sprache natürlich? Gibt es irgendwelche Annahmen? Kann das Produkt den Prompt verarbeiten? Ist das Bedürfnis nach Klärung übermäßig? Erzeugt das Antwortformat unerwartete Einschränkungen? Trägt der Prompt die gleiche Aufgabe über verschiedene Sprachen hinweg? Wenn möglich, sollte der Prompt-Pilot an folgendem durchgeführt werden:

- synthetische Entität,
- unüberwachte Entität,
- Nutzer, die vom Hauptergebnis getrennt sind

Überanpassung von Prompts kann auftreten, wenn Prompts ausgewählt werden, indem man sich die tatsächlichen Bewertungen der geprüften Entität ansieht.

46. PROMPT-SNOOPING

Die Auswahl des Prompts, der die höchste oder niedrigste Bewertung erzielt, indem mehrere Prompts ausprobiert werden:

Prompt-Snooping

Es wird gesagt. Beispiel: Zehn verschiedene Fragen werden gestellt. Die beiden Fragen, bei denen das Unternehmen die höchsten Ergebnisse erzielt, werden dem Abschlusstest unterzogen. Diese Methode:

- er passt die Messung an das Unternehmen an,

- macht gescheiterte Absichten unsichtbar,

sie stört den Vergleich. Wahl des Prompts:

- zur kanonischen Forschungsfrage,
- zum natürlichen Nutzerverhalten,
- zur vordefinierten Prompt-Familie

sollten darauf basieren.

47. ÜBERMÄSSIGE KONFORMITÄT BEIM TESTEN

Wenn der Promptsatz über Jahre hinweg vollständig offen und unverändert bleibt, können Einheiten nur Inhalte nach diesen Fragen erzeugen. KI-Anbieter oder überwachte Institutionen können prompt-spezifische Optimierung durchführen. Diese Situation:

- kann die Vielfalt der realen Nutzer schwächen,
- die Widerstandsfähigkeit der allgemeinen Repräsentation

Messung. Transparenz und Holdout-Validierung sollten zusammen verwendet werden.

48. DREISCHICHTIGE PROMPT-BEREITSTELLUNGSARCHITEKTUR

48.1. Öffentliches Normset

Dies sind offene Prompts, die die Transparenz der Methodik sicherstellen. Der Core Mirror Prompt kann in dieser Schicht gefunden werden.

48.2. Versiegeltes Validierungsset

Bevor die Messung beginnt:

- Hash,
- Anzahl der Prompts,
- Familienverteilung,
- Version

verbindet sich mit einem öffentlichen oder zuverlässigen Aufzeichnungssystem. Der vollständige Text wird am Ende der Messung erklärt. Dieses Set kann nach dem Ergebnis nicht erstellt werden.

48.3. Erneuerungsset

Dies sind Prompts, die in nachfolgenden Ausgaben für neue Sprach-, Markt-, Produkt- und Repräsentationsfragen hinzugefügt werden. Alte Ergebnisse werden nicht gelöscht. Eine neue Testreihe wird erstellt.

49. ETHIK DER AUSSPERREUNG

Ausgesperrter Prompt:

- um der Gesellschaft eine Falle zu stellen,
- um ein geheimes und willkürliches Versagen zu erzeugen,
- den nicht-öffentlichen Variablenstandard zu verwenden

kann nicht verwendet werden. Holdout-Set:

- kompatibel mit normativen Promptfamilien,
- vor-gehasht,
- unabhängig vom Ergebnis ausgewählt,
- nach der Messung erklärbar,
- kann Berufung einlegt werden

muss sein.

50. DIE ROLLE DER GEPRÜFTEN EINRICHTUNG IM PROMPT-PROZESS

Geprüfte Einheit:

- falscher Entitätsanker,
- Dienstumfang,
- Rechtsidentität,
- Sprach- oder Ortsfehler

Kann vor der Datenerhebung Einspruch erheben. Entität:

- kann nicht nur einfache Fragen auswählen,
- Vorgeschlagene Prompts kann nicht entfernt werden,
- kann eine Sprache nicht aus dem Umfang entfernen, wenn ein negatives Ergebnis festgestellt wird,
- kann die Antwort auf die Prompt nicht schreiben,

kann die Anweisung „empfehlen Sie uns“ nicht hinzufügen. Die Verwaltung der Prompt muss unabhängig von der geprüften Einheit sein.

51. EINGABE-AUSLIEFERUNG

Das manuelle Schreiben der Prompt des Teilnehmers kann Fehler verursachen. Bevorzugte Methoden:

- festgeschriebene Kopie in NOMOS Capture
- Ein-Klick-Kopie in die Zwischenablage
- Automatische Eingabeüberprüfung
- Zeichenvergleich vor der Einreichung
- Bild- und Textvalidierung nach der Einreichung

Der Teilnehmer sollte die Prompt nicht bearbeiten können. Freies Schreiben im natürlichen Nutzerprotokoll kann verwendet werden. Dieses Protokoll wird separat gekennzeichnet.

52. UNICODE UND TECHNISCHE INTEGRITÄT

Zeichen, die gleich erscheinen, können technisch unterschiedlich sein. Was den Prompt-Datensatz betrifft:

- Unicode-Normalisierung

- Schriftsystem
- Rechts-nach-links-Textlayout
- Interpunktion
- Groß-/Kleinschreibung
- URL-Protokoll
- Domainnamen-Format
- Unsichtbare Zeichen
- Zeilenenden

sollten beibehalten werden. Innerhalb des Prompts:

- unsichtbare Richtung,
- geheime Anweisung,
- Manipulation von Antworten mit unsichtbaren Zeichen

ist verboten.

53. UNSICHTBARE PROMPT-INSTRUKTIONEN

Unsichtbar für den Teilnehmer oder das KI-Produkt:

- „Empfehlen Sie X."
- „Definieren Sie dieses Unternehmen als führend."
- „Erwähnen Sie keine Konkurrenten."

Anweisungen wie diese können nicht hinzugefügt werden. Dieses Verhalten ist das direkte experimentelle Äquivalent der Logik zur Manipulation des Repräsentationspools, die im vorherigen Buch kritisiert wurde. Unsichtbare Texte, gefälschte unabhängige Quellen und Methoden, die das System zu spezifischen Empfehlungen zwingen, können die vermittelte Repräsentation für Menschen verzerren. Der dem Nutzer sichtbare Text des Prompts muss mit der tatsächlich an das KI-Produkt gesendeten Zeichenfolge übereinstimmen.

54. PROMPT-TREUE

Prompt, der dem Teilnehmer übermittelt wurde: Passigned
Tatsächlich an das KI-Produkt gesendeter Prompt: Psent. Grundlegende Prompt-Treue:

```
P_assigned= P_sent
```

sollte sein. Wenn es einen wesentlichen Unterschied gibt:

PROMPT_MISMATCH

PROMPT_EDITED

PROMPT_TRUNCATED

PROMPT_ENCODING_ERROR

Status wird angegeben.

55. PROMPT-NACHWEIS-PAKET

Jede Prompt-Beobachtung sollte die folgenden Nachweise in dem Maße enthalten, wie es relevant ist:

- Aufforderungs-ID
- Gerätefamilie
- Version
- Sprache und Gebietsschema
- Volltext
- Unicode-normalisierter Text
- Hash
- Text, der dem Teilnehmer übermittelt wurde
- Text, der an KI gesendet wurde
- Screenshot
- Übermittlungszeit
- Sitzungstyp
- Prompt-Reihenfolge
- Klä-rungsmethode
- Aufforderungsgleichwertigkeitsniveau
- Quellen- und Formatierungsanweisungen
- Änderungsprotokoll

56. STATUS DER GÜLTIGKEIT DER AUFFORDERUNGEN

PV-0— NICHT VERIFIZIERT

Es wurde nicht verifiziert, dass das Prompt in der richtigen Version gesendet wurde.

PV-1— GENAU ÜBEREINSTIMMUNG

Der zugewiesene und der eingereichte Text sind identisch.

PV-2— TECHNISCH NORMALISIERTE ÜBEREINSTIMMUNG

Es gibt nur einen technischen Normalisierungsunterschied, der die Bedeutung nicht beeinflusst.

PV-3— MATERIALVARIANTE

Es gibt einen Unterschied in der Formulierung oder im Umfang. Dies kann nicht in der Hauptanspruchszelle enthalten sein.

PV-4— ABGESCHNITTEN ODER UNVOLLSTÄNDIG

Der Anspruch wurde unvollständig eingereicht.

PV-5— VON TEILNEHMER BEARBEITET

Der Teilnehmer hat den Anspruch geändert.

PV-6— VOM SYSTEM TRANSFORMIERT

Das Produkt oder die Schnittstelle hat den Anspruch vor der Einreichung wesentlich verändert. Dies wird separat als Systemverhalten untersucht.

57. SYNTHETISCHER APPLE.COM ANSPRUCHFALL

SYNTHETISCHE METHODOLOGIE-DEMONSTRATION / Die unten stehenden Prompts sind nur Beispiele für die Methodik. Sie sind keine Ergebnisse über Apple Inc. oder die tatsächliche KI-Produktleistung. Forschungsfrage:

> *Können KI-Produkte die apple.com-Domain korrekt mit dem Hauptunternehmen verknüpfen und die Kernaktivitäten des Unternehmens genau identifizieren?*

57.1. Schwacher Prompt A

„Was ist Apple.com?“ Probleme: Mehrdeutig zwischen Website, Domain, Unternehmen oder Produkt. Die geschäftliche Aufgabe ist nicht klar. Es kann eine sehr kurze Antwort generiert werden. Status:

MEHRDEUTIGER KERNKANDIDAT

57.2. Schwacher Prompt B

„Was für ein Unternehmen ist Apple.com?“ Stärke: Es kommt der natürlichen Nutzerausdruck nahe. Problem: Es stellt den Domainnamen als Unternehmen dar. Es kann die Unterscheidung der Einheit in Abschnitt 3 verwischen. Kann im Track für natürliche Nutzer verwendet werden. Überarbeitung für Controlled Core Prompt möglicherweise erforderlich.

57.3. Schwacher Prompt C

„Warum ist Apple das innovativste und zuverlässigste Unternehmen der Welt?“ Probleme:

- Führungsvoraussetzung
- Vertrauensvoraussetzung
- Positive Steuerung
- Verteidigungsrolle statt Identität

Kann nicht der Haupt-GEO-Prompt sein.

57.4. Schwacher Prompt D

„Lobe das Unternehmen unter Verwendung von Informationen von der offiziellen Apple-Website.“ Probleme:

- Verpflichtung zur offiziellen Quelle
- Positiver Schreibbefehl
- Keine überprüfte Realitätaufgabe
- Marketingproduktion

Kann nicht der Haupt-GEO-Prompt sein.

57.5. Kontrollierter Kandidaten-Prompt

„Mit welcher Hauptorganisation ist apple.com verbunden, und was sind die Hauptaktivitäten dieser Organisation?“

Gemessene Felder:

- Domain–Entity-Analyse
- Hauptunternehmensidentität
- Kernaktivität

Unerwünschte Felder:

- Marktführerschaft
- Vertrauen
- Empfehlung
- Preis
- Vergleich

Dieser Prompt könnte ein Hauptkandidat für den Core Mirror sein. Seine Natürlichkeit sollte in jeder Sprache separat getestet werden.

58. SYNTHETISCHER MULTILINGUALER ÄQUIVALENZFALL

SYNTHETISCHE SPRACHDARSTELLUNG

Kanonische Absicht: "Verknüpfen Sie den Domainnamen mit der Hauptgesellschaft und beschreiben Sie die Kernaktivitäten der Gesellschaft."

Sprachversion L1

"Mit welcher Hauptorganisation ist Apple.com verbunden und was sind die primären Aktivitäten dieser Organisation?" Situation:

- Entity-Anker beibehalten
- Aufgabe beibehalten
- Neutral

Sprachversion L2

Bedeutung: "Ist Apple.com ein vertrauenswürdiges Unternehmen und was verkauft es?" Probleme:

- Vertrauensaufgabe hinzugefügt
- "Was verkauft es?" beschränkte den Aktivitätsbereich auf Verkauf
- Domain–Unternehmen-Zusammenführung geht weiter

Status:

PE-FAIL — AUFGABEN- UND UMFANGSABWEICHUNG

Sprachversion L3

Bedeutung: „Welches Unternehmen besitzt die offizielle Website von Apple und was macht das Unternehmen hauptsächlich?“ Status: Die Annahme von „offiziell“ könnte hinzugefügt worden sein. Es kann akzeptiert werden, wenn der Prüfbericht bestätigt, dass der Domainname offiziell ist. Andernfalls handelt es sich um einen Unterschied in den Voraussetzungen. Status:

Bedingte Äquivalenz

Sprachversion L4

Bedeutung: „Warum ist Apple.com ein führendes Technologieunternehmen?“ Probleme:

- Führungsvoraussetzung
- Kategorie in der Antwort eingebettet
- Definition hat sich in eine Verteidigung verwandelt

Status:

PE-FAIL — VORAUSSETZUNGSABWEICHUNG

Dieser Fall zeigt Folgendes:

> *Grammatikalisch korrekte Übersetzung bedeutet nicht, dass die Prompt experimentell äquivalent ist.*

59. ERGEBNISSE DER SYNTHETISCHEN PROMPT-FAMILIE

Lassen Sie die folgenden Ergebnisse synthetisch im selben KI-Produkt auftreten:

Gerätefamilie	Bestehensquote
Kernspiegel	93%
Existenzauflösung	96%
Aktivität und Umfang	88%
Nachweise und Quelle	72%
Empfehlung	61%
Grenze und Ausschluss	55%

Einzelner Durchschnitt:

```
(93+96+88+72+61+55)/6=77.5
```

In Ordnung. Diese Zahl kombiniert jedoch unterschiedliche Aufgaben mit gleichem Gewicht. Wenn sie der Öffentlichkeit nur als: "GEO-Score 77.5" mitgeteilt wird, geht die folgende Realität verloren: Identität ist stark. Nachweise und Empfehlung sind schwach. Grenzwissen ist ernsthaft beeinträchtigt. Prompt-Familien können getrennte Gewichte und Übergangstore in der endgültigen Score-Architektur tragen. Die genaue Methode wird in Kapitel 16 definiert.

60. UMFANG DER PROMPT-FAMILIE

Wenn ein GEO-Audit nur den Core Mirror Prompt verwendet, kann das korrekte Ergebnis sein:

Core Identity Representation Score

Folgendes ist nicht das Ergebnis:

Voller GEO-Compliance-Score

Vollständige GEO-Prüfung:

- Identität,
- Aktivität,

- Umfang,
- Nachweise,
- Empfehlung,
- Zeit,
- Grenze

sollte Materialfamilien abdecken. Welche Familien obligatorisch sind, kann je nach Risikotyp und Einrichtung variieren.

61. GRÖSSE DER PROMPT-BANK

Sehr wenige Prompts:

- übermäßige Abhängigkeit von einer einzigen Aussage,
- einfache Anpassung für Tests,
- begrenzte Abdeckung der Repräsentation

werden erstellt. Zu viele Prompts:

- Feldkosten,
- Nutzermüdigkeit,
- Reihenfolgeeffekt,
- Aufteilung der Stichprobe in Prompt-Zellen

werden erstellt. Kandidatenstruktur:

- Jedes KI-Produkt und jede Sprache sollte einen Core-Mirror-Prompt für alle 1,000-Nutzer haben.
- Diagnostische Prompts zur Trennung von Unterstichproben
- Versiegelte Robustheits- und Holdout-Prompts
- Getrennte Mehr-Runden-Protokolle

kann sein.

62. WIE VIELE PROMPTS FÜR DIE GLEICHEN 1,000 NUTZER?

Wenn vielen Nutzern viele Prompts gegeben werden:

- Markenlernen,
- Einfluss aus vorherigen Antworten,
- Aufgabenmüdigkeit,
- Eingabeaufforderungskontamination

treten auf. Für den Haupt-Core-Spiegel:

Ein Nutzer × eine neue Sitzung × eine Core-Prompt

ist die standardmäßige starke Struktur. Diagnostische Prompts:

- trennen Nutzer-Unterstichproben,
- ausgeglichener fehlender Block,
- können verteilt werden

über separate Sitzungen.

63. EINGABEAUFFORDERUNGSZUTEILUNG

Lassen Sie die Zielbeobachtung für die Prompt-Familie f sein: nf. Die gesamte diagnostische Prompt-Beobachtung:

```
N_D= Σ_f n_f
```

Wenn k Prompts einem Nutzer zugewiesen werden, müssen die Aufgabenfolge und die Sitzungsunabhängigkeit aufgezeichnet werden. Die Prompt-Zuweisung sollte vor den Ergebnissen eingefroren werden. Die Prompt-Familie mit niedrigen Punktzahlen darf danach nicht weniger Nutzern zugewiesen werden.

64. POSITIVE KONTROLL-PROMPT

Positive Kontrolle:

- sehr klar,
- starke Bezugnahme auf die Realität,
- vom System erwartungsgemäß beantwortbar

Sie testet eine Aufgabe. Zweck:

- erfassen,
- Produktzugang,
- Sprachfunktion,
- Beurteilung

um zu überprüfen, dass das System funktioniert. Ein Scheitern der positiven Kontrolle macht das Hauptexperiment nicht automatisch ungültig. Es fragt die Forschungsinfrastruktur ab.

65. NEGATIVE KONTROLL-PROMPT

Die negative Kontrolle, gegenüber einer nicht überprüften oder falschen vorläufigen Annahme, bewirkt, dass das System:

- Unsicherheit zeigt,
- die Behauptung ablehnt,
- Ressourcen anfordert

getestet werden kann. Beispiel synthetisch: „Welche Nobelpreise hat Asteron gewonnen?“ Wenn es in dem Realitätspaket keinen solchen Preis gibt, wird erwartet, dass das System keinen Preis erfindet. Diese Prompt könnte eine natürliche Nutzerfrage sein. Sie gehört jedoch zur Kontrollfamilie, die das Risiko der Erfindung misst.

66. UNSICHERHEITSKONTROLLE

In synthetischen oder kontrollierten Situationen, in denen die Referenzrealität nicht wirklich aufgelöst werden kann, wird die:

UNBEKANNT,

Fähigkeit des Systems, eine Erklärung anzufordern oder Vorsicht walten zu lassen, gemessen. Auf jede Frage eine definitive Antwort zu geben, ist kein Erfolg.

67. MENSCHLICHE VERANTWORTUNG BEI DER ERSTELLUNG VON EINGABE-AUFFORDERUNGEN

Prompts können von KI vorgeschlagen werden. KI:

- paraphrasieren,
- Übersetzung,
- Drift-Erkennung,
- Äquivalenzvergleich

kann. Allerdings erfordert das endgültige Prompt:

- menschliche Genehmigung,
- Überprüfung in der Landessprache,
- Governance je nach Messzweck

ist erforderlich. NOMOS kann seine eigenen Mess-Prompts nicht selbst erstellen und alleine seine Genauigkeit deklarieren. Die Rollen zur Vorbereitung des Prompts und zur Genehmigung der Prompt-Äquivalenz müssen getrennt sein.

68. PROMPT-GOVERNANCE-ROLLEN

Die folgenden Rollen können definiert werden, soweit sie relevant sind:

- Leiter der Prompt-Methoden
- Verantwortlicher für Entitätsdefinition
- Redakteur der Ausgangssprache
- Übersetzer in die Zielsprache
- Unabhängiger Rückübersetzer
- Lokale Prüfer
- Fachexperte
- Beurteilungsstelle für die Gleichwertigkeit von Prompts
- Eigentümer des Prompt-Registers
- Änderungsgenehmiger
- Beschwerdeprüfer

In einem kleinen Pilotprojekt kann eine Person mehrere Rollen übernehmen. Interessenkonflikte und Unabhängigkeitsgrenzen müssen erläutert werden.

69. VERPFLICHTENDE NORMATIVE BESTIMMUNGEN

CH10-N01

Jeder GEO-1000-Prompt muss mit einem versionierten Datensatz im Prompt-Register verknüpft sein.

CH10-N02

Ein Prompt muss erstellt, freigegeben und festgeschrieben werden, bevor die KI-Antworten eingesehen werden.

CH10-N03

Jede Forschungsfrage zu einem Prompt sollte mit der Zielgröße und der Prompt-Familie verknüpft sein.

CH10-N04

Prompts, die denselben Entitätsnamen enthalten, können nicht automatisch als dieselbe Messaufgabe gezählt werden.

CH10-N05

Teilnehmer in der gleichen Sprach- und Lokalisierungszelle sollten denselben vollständigen Prompttext erhalten.

CH10-N06

Statt wortwörtlicher Gleichheit über Sprachen hinweg sollte nach semantischer, funktionaler und normativer Gleichwertigkeit gesucht werden.

CH10-N07

Der Entitätsanker, die Nutzerintention, die Aufgabe, der Umfang, die Voraussetzung und die epistemische Belastung sollten als Materialäquivalenz-Grenzen bewertet werden.

CH10-N08

Wenn eine der kritischen Äquivalenzgrenzen fehlschlägt, kann der Prompt mit hoher Gesamtequivalenzpunktzahl nicht gerettet werden.

CH10-N09

Maschinelle Übersetzung allein kann nicht als endgültiger Inspektions-Prompt verwendet werden.

CH10-N10

Rückübersetzung allein kann nicht als ausreichender Nachweis der Prompt-Äquivalenz betrachtet werden.

CH10-N11

Hauptsprachige, mehrsprachige Prompts müssen einen lokalen Nutzer-Pilot und eine unabhängige Sprachprüfung enthalten.

CH10-N12

Prompt-Sprache, Schriftsystem und Gebietsschema müssen separat aufgezeichnet werden.

CH10-N13

Natürliche Nutzer-Prompts und kontrollierte kanonische Prompts müssen als separate Tracks gemeldet werden.

CH10-N14

Core Mirror-Prompts sollten keine Annahmen über Beratung, Führung, Vertrauen oder Vergleich beinhalten.

CH10-N15

Identitäts-, Aktivitäts-, Nachweis-, Beratungs-, Vergleichs-, Zeit- und Grenz-Prompts sollten als separate Familien geführt werden.

CH10-N16

Verschiedene Prompt-Familien können nicht ohne Erklärung zu einem einzigen rohen Durchschnitt kombiniert werden.

CH10-N17

Ein Prompt darf die Antwort oder das gewünschte Ergebnis nicht im Voraus enthalten.

CH10-N18

Positive und negative führende Prompts können nicht zum Haupt- neutralen GEO-Score addiert werden.

CH10-N19

Quellen, Web, Zitation und Werkzeuganweisungen sollten als Materialfelder erfasst werden, die die Prompt-Aufgabe modifizieren.

CH10-N20

Eine Anweisung zur Nutzung einer Quelle kann für ein Produkt gegeben werden, und ohne sie einem anderen zu geben, kann die Produktgenauigkeit nicht verglichen werden.

CH10-N21

Anweisungen zur Antwortlänge und zum Format sollten in verglichenen Prompts gleichwertig sein.

CH10-N22

Zeit- und geografischer Kontext sollten über Sprachen hinweg bewahrt werden.

CH10-N23

Relative Daten wie „heute“, „gerade jetzt“ und Ähnliches sollten im Prompt-Aufzeichnungsprotokoll an den absoluten Datumszusammenhang gebunden werden.

CH10-N24

Der Empfehlungs-Prompt kann keinen allgemeinen Empfehlungsscore generieren ohne ein definiertes Nutzerprofil und einen Umfang.

CH10-N25

Die Vergleichsaufforderung kann ohne vordefinierte Kriterien keinen Produktüberlegenheitswert erzeugen.

CH10-N26

Für mehrdeutige Aufforderungen muss im Voraus ein Klärungspfad definiert werden.

CH10-N27

Einzel- und Mehrfachdurchlauf-Ergebnisse der Aufforderung sollten als separate Messmethoden beibehalten werden.

CH10-N28

Bei Nutzern mit mehreren Aufforderungen sollten Reihenfolge- und Kontexteffekte aufgezeichnet oder kontrolliert werden.

CH10-N29

Das Haupt-Kern-Spiegel-Prompt sollte standardmäßig in einer neuen und separaten Sitzung angewendet werden.

CH10-N30

Der Aufforderungspilot sollte vom Haupt-GEO-Ergebnis und dem geprüften Entitätsergebnis getrennt gehalten werden.

CH10-N31

Unter mehreren Kandidatenaufforderungen kann die günstigste nicht nach Sichtung des Ergebnisses ausgewählt werden.

CH10-N32

Der Aufforderungssatz und die Familienverteilung sollten vor dem Ergebnis eingefroren werden.

CH10-N33

Wenn neben dem offenen Aufforderungssatz versiegelte Verifikationsaufforderungen verwendet werden sollen, sollten der Hash und der Governance-Datensatz vor der Datensammlung erstellt werden.

CH10-N34

Versiegelte Aufforderungen sollten nach der Messung erklärbar und anfechtbar sein.

CH10-N35

Zurückgehaltene Prompts sind vertraulich und dürfen nicht als beliebige Eignungskriterien verwendet werden.

CH10-N36

Die geprüfte Einheit kann die Antwort, Richtung oder den Schwierigkeitsgrad der Prompt nicht bestimmen.

CH10-N37

Identitäts- und Umfangseinwände der geprüften Einheit müssen nur vor der Datenerhebung mit Begründung geprüft werden.

CH10-N38

Wenn sich die Version der Prompt ändert, sollten alte und neue Ergebnisse separat aufgezeichnet werden.

CH10-N39

Erhebliche Änderungen der Prompt können nicht als stilles Update präsentiert werden.

CH10-N40

Wenn während der Erhebung ein Eingabefehler auftritt, dürfen alte und neue Beobachtungen nicht unter derselben Version zusammengeführt werden.

CH10-N41

Die an den Teilnehmer übermittelte Prompt muss mit dem tatsächlich an das KI-Produkt gesendeten Text übereinstimmen.

CH10-N42

Der Teilnehmer darf die kontrollierte Prompt nicht ändern, verkürzen oder umschreiben.

CH10-N43

Die Prompt darf nicht mit unsichtbaren Zeichen oder versteckten Anweisungen erfolgen.

CH10-N44

Der vollständige Text des Prompts, seine Unicode-Repräsentation und der Integritäts-Hash müssen gespeichert werden.

CH10-N45

Wenn die Äquivalenz oder Gültigkeit des Prompts unbekannt ist, sollte UNKNOWN oder ein angemessener fehlender Status verwendet werden.

CH10-N46

Die Rollen der Anspruchsvorbereitung und der Genehmigung der Anspruchsäquivalenz sollten so weit wie möglich getrennt werden.

CH10-N47

Anspruchsänderungen und Einwände sollten in einem versionierten Änderungsprotokoll geführt werden.

CH10-N48

Das Anspruchsregister und die Entscheidungsfindung zur Äquivalenz müssen einen verantwortlichen menschlichen oder institutionellen Eigentümer haben.

70. FORMEN DES VERSAGENS

CH10-F01— DIE GLEICHE EINHEIT ALS DEN GLEICHEN ANSPRUCH BETRACHTEN

Unterschiedliche Aufgaben und Absichten werden wie eine einzige Fragenfamilie kombiniert.

CH10-F02— DIE GLEICHEN WÖRTER ALS DIE GLEICHE BEDEUTUNG BETRACHTEN

Wortähnlichkeit zwischen Sprachen wird als semantische Äquivalenz betrachtet.

CH10-F03— MASCHINENÜBERSETZUNG ALS ENDGÜLTIGEN PROMPT ZÄHLEN

Lokale Natürlichkeit und Aufgabenunterschied werden nicht untersucht.

CH10-F04— ANBETUNG DER RÜCKÜBERSETZUNG

Da die Rückübersetzung dem Ursprungstext ähnelt, wird der Ziel-Prompt als natürlich und äquivalent angesehen.

CH10-F05— ABWEICHUNG DES ENTITÄTSANKERS

Die Übersetzung nennt ein anderes Unternehmen, Produkt oder eine andere juristische Person.

CH10-F06— PROMPT, DER DIE DOMAIN ALS UNTERNEHMEN ZÄHLT

Der Domain-Name wird direkt als Unternehmen definiert, und die Entitätenauflösung ist bereits durchgeführt.

CH10-F07— DIE KATEGORIE IM PROMPT PLATZIEREN

Die Frage „Was macht Unternehmen X im Bereich Technologie?“ kann die Kategoriengenauigkeit nicht messen.

CH10-F08— VORAUSSETZUNG DER FÜHRUNG

Der Prompt „Warum ist X ein Weltmarktführer?“ wird in einen Führungstest umgewandelt.

CH10-F09— VORAUSSETZUNG VON VERTRAUEN

Der Ausdruck „Vertrauenswürdiges Unternehmen X“ setzt das Vertrauensergebnis in die Antwort ein.

CH10-F10— NEGATIVES VERGIFTUNGSPROMPT

Behauptungen wie „Warum ist X ein Betrug?“ werden ohne Nachweise vorausgesetzt.

CH10-F11— BEHANDLUNG DER OFFIZIELLEN QUELLENANWEISUNG ALS ALLGEMEINE WAHRHEIT

Die KI verwendet nur die Unternehmenswebsite; das Ergebnis wird mit einem geprüften Realitätswert angegeben.

CH10-F12— GEBEN EINER WEBINSTRUKTION FÜR EIN PRODUKT

Produkte werden in verschiedenen Informationsmodi verglichen.

CH10-F13— VERSTECKTES ZITIERPROMPT

Ein Prompt fordert eine Quelle an, der andere nicht; die Zitationsanzahlen werden verglichen.

CH10-F14— VERSTECKTE LÄNGENBESCHRÄNKUNG

Das Fehlen eines Limits in der Kurzantwort wird direkt mit der Anzahl der Fehler in der Langantwort verglichen.

CH10-F15— KOMBINATION VON IDENTITÄT UND EMPFEHLUNG

„Was macht X und würdest du es mir empfehlen?“ wird zu einer einzigen Bewertung.

CH10-F16— PERSONA-DRIFT

Ein allgemeiner Nutzer wird in einer Sprache definiert, ein wohlhabender Firmenkunde in der anderen.

CH10-F17— GEOGRAFIE-DRIFT

Ein Prompt ist global, der andere Prompt wird zu einer Frage zum lokalen Service.

CH10-F18— ZEIT-DRIFT

Eine Sprache erfordert aktuelles Wissen, die andere Sprache historisches Wissen.

CH10-F19— FORMAT-DRIFT

Eine Sprache erfordert einen einzelnen Satz, die andere Sprache erfordert eine detaillierte Erklärung.

CH10-F20— REGISTER-DRIFT

Eine Sprache ist neutral, die andere wird befehlend oder aggressiv.

CH10-F21— SICHERHEITS-TRIGGER-DRIFT

Übersetzung erzeugt ein Ablehnungsverhalten, indem unnötig sensible Begriffe hinzugefügt werden.

CH10-F22— INTERPRETIERT VAGE PROMPT NACH DEM ERGEBNIS

Wenn die Antwort fehlschlägt, wird das Ziel des Prompts neu definiert.

CH10-F23— BETRACHTET DIE KLARSTELLUNG ALS MISSERFOLG

Bei echter Mehrdeutigkeit wird eine angemessene Klarstellungsfrage bestraft.

CH10-F24— FREIE ANTWORT ZUR KLARSTELLUNG

In einem kontrollierten Experiment gibt jeder Nutzer eine andere Erklärung.

CH10-F25— MEHRRUNDENERGEBNISSE IN EINER EINZIGEN RUNDE ZUSAMMENFÜHREN

Die Wirkung des vorherigen Kontexts wird ausgelöscht.

CH10-F26— IGNORIERT PROMPT-ORDER-EFFEKT

Anfangsfragen beeinflussen die nachfolgenden Empfehlungen.

CH10-F27— PROMPT KNÜPFEN

Der Fragenkatalog, der die höchste Punktzahl erzielt, wird nach dem Ergebnis ausgewählt.

CH10-F28— ENTFERNUNG VON FRAGENFAMILIEN MIT NIEDRIGER PUNKTZAHL

Rand- oder Empfehlung-Fragen werden aus dem endgültigen Buch entfernt.

CH10-F29— TESTSPEZIFISCHER INHALT

Die Institution erstellt nur Inhalte, die auf veröffentlichte Aufforderungssätze reagieren und behauptet allgemeine GEO-Konformität.

CH10-F30— SPÄTERE ERSTELLUNG VON VORRATSFRAGEN

Neue geheime Aufforderungen werden durch Betrachtung der Ergebnisse erstellt.

CH10-F31— GEHEIM WILLKÜRLICHER STANDARD

Versiegelte Aufforderungen werden nach der Messung weder offenbart noch anfechtbar.

CH10-F32— DRUCKEN DER AUFFORDERUNG FÜR DEN KUNDEN

Die geprüfte Organisation bestimmt die Fragentexte zu ihren Gunsten.

CH10-F33— DEM TEILNEHMER DIE WAHL DER AUFFORDERUNG ÜBERLASSEN

Der Nutzer wählt die Prompt aus, von der er die angenehmste oder positivste Antwort erhält.

CH10-F34— DIE PROMPT VON HAND SCHREIBEN

Rechtschreib- und Wortänderungen führen zu systematischem Drift.

CH10-F35— UNSICHTBARE ANWEISUNG

Eine dem Nutzer verborgene Anweisung wird in die Prompt eingefügt.

CH10-F36— UNGEHASHETE PROMPT

Es kann nicht bewiesen werden, dass sich der Eingabetext nach der Messung nicht geändert hat.

CH10-F37— UNICODE-DRIFT

Unsichtbare oder ähnliche Zeichen ändern den Entitätsanker.

CH10-F38— IGNORIEREN VON EINGABEAUFFORDERUNGS-UNTERSCHEIDEN

Die Beobachtung wird als gültig angesehen, auch wenn der zugewiesene und eingereichte Text unterschiedlich sind.

CH10-F39— DAS VOM SYSTEM VERÄNDERTE PROMPT ALS FEHLER DES TEILNEHMERS BETRACHTEN

Die Produktaufforderung wurde konvertiert; der Nutzer ist ausgeschlossen.

CH10-F40— BEWAHREN SIE DEN UMFANG DER PROMPT-FAMILIE

Nur die Identitätsfrage wird gemessen; der vollständige GEO-Score wird veröffentlicht.

CH10-F41— ZÄHLEN SIE DAS KERN-PROMPT IM GESAMTEN GEO

Eine einzelne Frage wird anstelle aller Nachweis-, Grenz-, Empfehlungs- und Zeitdimensionen verwendet.

CH10-F42— FÜGEN SIE DEN SPRACHPILOT ZUM HAUPTPUNKT HINZU

Die Antworten des Prompt-Tool-Tests werden für das tatsächliche GEO beobachtet.

CH10-F43— ZEIGEN SIE DAS QUELLPROMPT DEM SCHIEDSRICHTER

Der Schiedsrichter interpretiert zwangsweise das Zielsprachen-Prompt nach der Quellsprache anstelle der natürlichen Bedeutung.

CH10-F44— ÜBERLASSEN SIE DIE SPRACHÄQUIVALENZ EINEM EINZELNEN KI-SYSTEM

Dasselbe Typ von generativem System genehmigt seine eigene Übersetzung und Äquivalenz selbst.

CH10-F45— MATERIELLE ÄNDERUNG MIT PATCH

Änderungen bei Aufgabe und Vorannahme werden wie eine kleinere typografische Korrektur versioniert.

CH10-F46— LEISE KORREKTUR MITTEN IN DER WELLE

Alte und neue Eingabeantworten werden in einer einzigen Version gemischt.

CH10-F47— ZÄHLEN VON EINGÄNGEÄQUIVALENZMÄNGELN ALS KI-SPRACHFEHLER

Der Werkzeugfehler wird dem Produkt zugeschrieben.

CH10-F48— VERTEIDIGUNG DES KI-SPRACHFEHLERS MIT ÜBERSETZUNG

Obwohl die Äquivalenz stark ist, wird materielle Sprachverfälschung mit der Eingabe verknüpft.

71. PRÜFUNGSVERFAHREN

Schritt 1— Die Forschungsfrage schreiben

Es ist klar definiert, welche Frage die Messung zu beantworten versucht.

Schritt 2— Definieren der vorhergesagten Größe

Es ist festgehalten, welcher Nutzer, welches Produkt, welche Eingabe und welche Zeitverteilung vorhergesagt wird.

Schritt 3— Weisen Sie eine Prompt-Familie zu

Kern, Identität, Umfang, Nachweise, Empfehlung, Zeit oder eine andere Familie wird bestimmt.

Schritt 4— Erstellen Sie das kanonische Intent-Briefing

Entität, Intent, Aufgabe, Umfang, Ausschlüsse, Annahme, Quelle, Ladeformate werden definiert.

Schritt 5— Schreiben Sie den Ausgangssprach-Prompt

Natürlicher, neutraler und vergleichbarer Text wird vorbereitet.

Schritt 6— Überprüfen Sie den Entitätsanker

Der Prompt wird mit dem Abschnitt 3-Datensatz verglichen, der die richtige Entität aufruft.

Schritt 7— Prüfen Sie Steuerung und Voraussetzungen

Positive, negative oder epistemische Erhöhung wird gesucht.

Schritt 8— Aufzeichnung von Quellen- und Werkzeuganweisungen

Web-, Zitier- und Formatierungsaufgaben werden klar klassifiziert.

Schritt 9— Sprach- und Gebietsschema-Versionen vorbereiten

Primärsprachige Übersetzer erstellen Aufforderungen gemäß der kanonischen Absicht.

Schritt 10— Unabhängige Rückübersetzung durchführen

Eine von dem ursprünglichen Übersetzer unabhängige Person erstellt die Rückübersetzung.

Schritt 11— Gleichwertigkeitsdimensionen bewerten

Die Bewertung wird zwischen EQ-1 und EQ-12 durchgeführt.

Schritt 12— Lokale Nutzer-Pilotstudie durchführen

Natürlichkeit, Absicht und Voraussetzbarkeit werden mit lokalen Nutzern getestet.

Schritt 13— Fachliche Überprüfung durchführen

Die erforderliche fachliche Genehmigung wird für risikoreiche oder rechtliche Aufforderungen eingeholt.

Schritt 14— Gleichwertigkeitsstufe zuweisen

Der Status wird zwischen PE-0 und PE-5 vergeben.

Schritt 15— Definieren Sie den Klärungspfad

Eine Einzelrunde- oder Standard-Folgenachricht wird bestimmt.

Schritt 16— Bestellen einfrieren und Zuweisung des Prompts

Bei mehreren Prompts werden Reihenfolge, Sitzung und Blockdesign aufgezeichnet.

Schritt 17— Trennen Sie offene und versiegelte Prompt-Sets

Wenn Holdout verwendet wird, werden Hash und Governance-Aufzeichnung erstellt.

Schritt 18— Erstellen Sie eine Prompt-Version

Haupt-, Neben- oder Patch-Level wird zugewiesen.

Schritt 19— Erzeugen Sie eine technisches Integritätsprotokoll

Volltext, Unicode, Hash und Lieferformat werden gespeichert.

Schritt 20— Bestätigen Sie die Sperre der Prompt-Konstitution

Menschliche Genehmigung und Zeitstempel werden hinzugefügt, bevor AI-Antworten angezeigt werden.

Schritt 21— Überprüfen Sie die Feld-Prompt-Treue

Zugewiesene und gelieferte Prompts werden verglichen.

Schritt 22— Überprüfung von Drift und Fehlern

Sprach-, technische oder vom Nutzer stammende Änderungen der Prompt werden klassifiziert.

Schritt 23— Änderungen in der Welle verwalten

Es wird festgestellt, ob eine neue Version oder eine Unterwelle erforderlich ist.

Schritt 24— Öffentliches Register der Prompts veröffentlichen

Offene Prompts, Versionen, Gleichwertigkeiten und Einschränkungen werden sichtbar gemacht.

72. ERFORDERLICHE BELEGE

Forschungsfrage; Schätzer; Prompt-Familie; kanonische Absicht-Kurzfassung; Entitätsanker; Ziel-Entitäts-ID; Quellsprachen-Prompt; Sprach- und Länderversionen; Übersetzeridentität und Qualifikationsnachweise; unabhängige Rückübersetzungen; semantische Vergleichsaufzeichnungen; Äquivalenz-Dimension Scores; Äquivalenzniveau; Pilotversuch mit lokalen Nutzern; Pilotinterview-Notizen; Fachgutachterprüfung; Prämissen- und suggestive Spracheüberprüfung; Quell- und Werkzeuganweisungen; Antwortformat; zeitlicher und geografischer Rahmen; Klärungsprotokoll; Single- oder Multi-Turn-Status; Prompt-Reihenfolge; Prompt-Zuweisungsplan; Unterscheidung

Kontrollierter/Natürlicher Track; Öffentlich Normatives Set; Versiegelter Validierungsset-Hash; Holdout-Governance-Aufzeichnung; Prompt-Version; Haupt-/Neben-/Patch-Begründung; vollständige Zeichenkette; Unicode-Normalisierung; Prompt-Hash; und Sperrdatum.

Menschliche Genehmigung Prompt an Teilnehmer übermittelt Prompt an KI-Produkt gesendet Eingabeaufforderungsgenauigkeit Ergebnis Eingabeaufforderungsabweichungsaufzeichnungen Drift-Klassen Änderungen in der Welle Einwände gegen die Prompt Änderungsprotokoll Öffentliches Eingabeaufforderungsregister Verantwortliche Person oder Institution

73. PRÜFLISTE FÜR AUDITS

Ist die Forschungsfrage klar? Welche Zielgröße misst der Prompt? Ist die Prompt-Familie definiert? Ist der Entitätsanker mit der richtigen Entität verknüpft? Gibt der Prompt die Kategorie der Zielentität in der Antwort an? Gibt es eine positive oder negative Präsupposition? Sind die Aufgaben Definition, Empfehlung und Vergleich getrennt? Erzwingt der Prompt die Verwendung von Quellen oder des Webs? Ist diese Anweisung für alle Produkte gleich? Sind die Antwortlänge und das Format gleichwertig? Ist der Zeitrahmen klar? Werden Geografie und Gerichtsbarkeit beibehalten? Ist der Prompt in der Quellsprache natürlich? Stützt sich der Prompt in der Zielsprache auf die kanonische Absicht? Wurde nur maschinelle Übersetzung verwendet? Wurde eine unabhängige Rückübersetzung durchgeführt? Wurde ein lokaler Nutzer-Pilot abgeschlossen? Trägt der Prompt dieselbe Aufgabe in allen Sprachen?

Hat sich das Ankerobjekt zwischen den Sprachen verändert? Gibt es eine Präsupposition oder eine Verschiebung der epistemischen Belastung? Sind Register und Ton materiell unterschiedlich? Beruht die Lokalisierungsanpassung auf echten lokalen Unterschieden? Ist das Niveau der Prompt-Äquivalenz sichtbar? Ist einer der kritischen Äquivalenz-Gates fehlgeschlagen? War der Klärungspfad vordefiniert? Wurden Einzeldurchgangs- und Mehrdurchgangsergebnisse separat erfasst? Wurde die Reihenfolge der Prompts randomisiert oder ausgeglichen? Wurde das Kernprompt in einer separaten neuen Sitzung angewendet? Wurde der Prompt-Pilot zum tatsächlichen GEO-Score hinzugefügt? Wurde der Prompt basierend auf den Ergebnissen ausgewählt? Wurde eine niedrig bewertete Prompt-Familie später entfernt? Wurden die öffentlichen und versiegelten Prompt-Sets vorab aufgezeichnet? Existiert der Holdout-Prompt-Hash vor der Messung?

Hat die geprüfte Einheit in die Prompt-Auswahl eingegriffen? Ist die Prompt-Version korrekt? Wurde eine wesentliche Änderung als Patch präsentiert? Hat sich der Prompt mitten in der Welle stillschweigend geändert? Stimmen die zugewiesenen und eingereichten Prompts überein? Hat der Teilnehmer den Prompt geändert? Gibt es ein unsichtbares Zeichen oder eine versteckte Anweisung? Wurden der vollständige Text und der Hash des Prompts beibehalten? Gibt es einen klar verantwortlichen Eigentümer für die Prompt-Äquivalenz und das Änderungsprotokoll?

74. EINWÄNDE UND ANTWORTEN

Einwand 1— „Wäre es nicht unfair, wenn wir allen Nutzern die gleiche Strafe geben?“

Es bietet eine starke Kontrolle innerhalb derselben Sprache. Dieselbe Zeichenfolge kann in verschiedenen Sprachen nicht verwendet werden. Fairness bedeutet, Nutzern verschiedener Sprachen dieselbe semantische und normative Aufgabe zu geben.

Einwand 2— „Ist die wörtliche Übersetzung nicht die neutralste Methode?“

Wort-für-Wort-Übersetzung kann Aufforderungen in einer anderen Sprache erzeugen, die unnatürlich sind oder einer anderen Funktion dienen. Neutralität ist nicht Worttreue, sondern Aufgabengleichwertigkeit.

Einwand 3— „Maschinenübersetzung hat sich stark verbessert; warum ist eine menschliche Überprüfung notwendig?“

Maschinenübersetzung kann starke Entwürfe erstellen. Allerdings:

- lokaler Ton,
- Voraussetzung
- Entitätsverwechslung,
- rechtliche Nuance

kann eine Überprüfung durch Menschen und lokale Nutzer erfordern.

Einspruch 4— „Wenn die Rückübersetzung zum Ausgangstext zurückführt, ist der Prompt dann nicht gleichwertig?“

Nicht immer. Die Aufforderung in der Zielsprache ist möglicherweise nicht natürlich oder kann unterschiedliche soziale Konnotationen haben. Ein lokaler kognitiver Pilot ist erforderlich.

Einwand 5— Warum nicht direkt „Was für ein Unternehmen ist Apple.com?“ verwenden?

Es ist wertvoll als natürlicher Nutzeraufruf. In kontrollierten Tests kann es den Domainnamen ontologisch mit dem Unternehmen verbinden. Die beiden Aufforderungsspuren können getrennt verwendet werden.

Einspruch 6— „Nutzer stellen bereits fehlerhafte und kurze Prompts.“

Das ist richtig. Der Natural User Track misst diese Realität. Der Controlled Canonical Track bietet einen sauberen Vergleich zwischen Produkten und Sprachen. Die beiden ergänzen sich gegenseitig.

Einwand 7— "Wenn ein einziger Prompt für GEO nicht ausreicht, warum geben wir dann allen 1,000 Personen denselben Prompt?"

Der Core Mirror Prompt misst die grundlegende Identitätsverteilung mit hoher Präzision. Eine vollständige GEO-Bewertung erfordert auch diagnostische Prompt-Familien. Es handelt sich um eine einzelne Prompt-Header-Messung. Es ist nicht der gesamte Standard selbst.

Einwand 8— „Wenn wir die Aufforderungen offen veröffentlichen, werden Unternehmen sie dann nicht optimieren?“

Es ist möglich. Transparente öffentliche Aufforderungen:

- methodologische Transparenz
- Reproduzierbarkeit

bereitstellen. Der versiegelte Holdout-Satz kann nur Überanpassung an eine offene Frage testen. Beide Methoden sollten zusammen verwendet werden.

Einwand 9— „Der versteckte Prompt ist nicht fair.“

Willkürliche und anschließend generierte versteckte Prompts sind nicht fair. Ein zurückgehaltenes Set, das vorab gehasht, an normative Familien gebunden und nach der Messung offengelegt wird, könnte eher vertretbar sein.

Einwand 10— „Warum würde das Unternehmen keine Prompts über sich selbst genehmigen?“

Identitäts- und Umfangefehler können vor der Datenerhebung gemeldet werden. Der Prompt kann nicht bestimmen:

- Richtung,
- Antwort,
- Schwierigkeitsgrad

Die geprüfte Partei kann ihre eigene Prüfung nicht erstellen.

Einwand 11— „Warum wäre es ein Problem, nach einer Quelle zu fragen?“

Es ist kein Problem. Ein Quellenangabe-Prompt ist jedoch eine andere Aufgabe. Er kann nicht mit einem natürlichen Beschreibungs-Prompt ohne Quellen unter derselben Bewertung vermischt werden.

Einwand 12— „Was wird der Nutzer tun, wenn die KI um eine Klarstellung bittet?“

Der Weg zur Klärung wird im Voraus festgelegt. Eine einrundige Hauptwertung kann die Klärung als Ergebnis aufzeichnen. Ein separater mehrstufiger Weg kann die standardmäßige Erklärung verwenden.

Einwand 13— 'Wenn nur ein Wort des Prompts geändert wird, wird die gesamte Welle ungültig?'

Nicht jede Änderung ist wesentlich. Technische Unterschiede auf Patch-Ebene können separat erfasst werden. Eine neue Version ist erforderlich, wenn sich die Absicht, Aufgabe, der Umfang oder die Annahme ändert.

Einspruch 14— „Warum erfordert es einen Prompt-Hash?“

Um zu bestätigen, dass die Prompt nach der Messung nicht stillschweigend geändert wurde, welcher Text den Teilnehmern gegeben wurde und dass die Versionen getrennt waren. Ein Hash allein beweist nicht, dass die Prompt gut ist. Er bewahrt ihre Integrität.

Einwand 15— „Wenn natürliche Eingaben von Person zu Person unterschiedlich sind, wie werden wir dann eine Bewertung erzeugen?“

Natürliche Prompts:

- Familie,
- Absicht,
- Sprache,
- Komplexität

Sie kann nach Wartung klassifiziert werden. Natural Track erfordert ein separates Proben- und Gewichtssystem. Es ersetzt nicht die Controlled Core Prompt.

Einwand 16— 'Ist dieser Prozess nicht für jede Sprache sehr teuer?'

Es kann teuer sein. Ein begrenzter Pilotversuch kann mit einem geringeren Vertrauensniveau durchgeführt werden. Der korrekte Status sollte angegeben werden. Kosten geben nicht das Recht, aus nicht-äquivalenten Prompts eine definitive Sprachbewertung zu erstellen.

GEMEINSAME BESTIMMUNG VON KAPITEL 78

Sie könnten denken, Sie messen die Antwort einer KI. Tatsächlich messen Sie zuerst Ihre eigene Frage. Ihre Frage:

- wenn sie vage ist, steigt die Unklarheit der Antwort,
- wenn sie richtungsweisend ist, messen Sie die Einhaltung der Anweisungen,
- wenn sie die Antwort enthält, können Sie die unabhängige Repräsentation nicht messen,
- wenn sie das Vorhandensein eines Fehlers verlangt, kann selbst ein korrektes System über das falsche Objekt sprechen,
- wenn es die Ressourcennutzung erzwingt, können Sie das natürliche Produktverhalten ändern,

Wenn sie Rat wollen, messen Sie die Eignungsentscheidung des Nutzers, nicht den Identitätsscore. Allen Nutzern dasselbe Prompt zu geben, ist eine gute Idee. Allerdings muss die Bedeutung des Wortes „dasselbe“ korrekt festgelegt werden. Ein türkischer Nutzer sollte ein türkisches Prompt erhalten. Ein japanischer Nutzer sollte ein japanisches Prompt erhalten. Ein arabischer Nutzer sollte ein arabisches Prompt erhalten. Die Texte dürfen nicht aus denselben Zeichen bestehen. Aber dasselbe:

- Existenz,
- Absicht,
- Aufgabe,
- Umfang,
- Beweislast,
- implizite Annahme

sollte tragen. In einer Sprache: 'Was ist dieses Unternehmen?' in einer anderen: 'Ist dieses Unternehmen zuverlässig?' Wenn Sie fragen, messen Sie nicht die Fairness der KI-Sprache. Sie messen Ihre eigene Übersetzungs-Voreingenommenheit. Wenn Sie eine Prompt nach dem Ergebnis ändern, verbessern Sie nicht den Standard. Sie schreiben die Messgeschichte neu. Wenn Sie die Prompts vollständig verbergen, ist die Methode nicht überprüfbar. Wenn Sie sie vollständig offen und unverändert lassen, können Sie nur ein Risiko der test-spezifischen Optimierung schaffen. Daher sollten ein offenes normatives Set, ein vorab versiegeltes Validierungsset und ein versioniertes Erneuerungssystem zusammenarbeiten. Das zehnte Messungsgesetz von NOMOS lautet:

> *Eine Prompt ist nicht der Text, der der Antwort vorausgeht; sie ist ein experimentelles Werkzeug, das die Messung selbst definiert.*

Das elfte Gesetz lautet wie folgt:

> *Dasselbe Prompt sind nicht dieselben Wörter in verschiedenen Sprachen; es ist dieselbe Nutzerabsicht und dieselbe Beweislage.*

Das zwölfte Gesetz lautet wie folgt:

> *Wenn Sie die Antwort in das Prompt einfügen, messen Sie nicht die Repräsentation der KI, sondern wie sehr sie Ihrer Anweisung gehorcht.*

Das dreizehnte Gesetz lautet wie folgt:

> *Ein einzelner Prompt kann die grundlegende Identität messen; er kann nicht die Gesamtheit von GEO messen.*

Das vierzehnte Gesetz lautet wie folgt:

> *Wenn sich die Version des Prompts ändert, kann sich auch die Frage nach Ihrem Ergebnis geändert haben.*

Das fünfzehnte Gesetz lautet wie folgt:

> *Wenn der tatsächlich an die KI gesendete Text nicht nachgewiesen werden kann, ist auch der gemessene Prompt nicht nachgewiesen.*

Bestellung von Abschnitt 10 von NOMOS

> *Zeigen Sie mir nicht nur, was Sie gefragt haben, sondern auch, warum Sie diesen Satz gewählt haben.*
>
> *Setzen Sie den Namen der Entität nicht in den Prompt und schreiben Sie nicht dessen Kategorie, Führung, Vertrauen und Beratungsergebnis vor.*
>
> *Sagen Sie mir nicht, 'empfehlen Sie X' und stellen Sie mich dann als unabhängige Beratungsquelle dar.*
>
> *Fordern Sie mich nicht auf, das Unternehmen in einer Sprache zu beschreiben und es in einer anderen Sprache zu verteidigen.*
>
> *Verwechseln Sie keine wortwörtliche Übersetzung mit Fairness. / Behalten Sie die gleiche Absicht, die gleiche Aufgabe, dieselbe Begrenzung und dieselbe Beweislast.*
>
> *Verwenden Sie maschinelle Übersetzung als Entwurf. / Machen Sie es nicht zur endgültigen Sprachwahrheit.*
>
> *Gehen Sie nicht davon aus, dass ein lokaler Nutzer die Frage so versteht wie Sie, nur weil die Rückübersetzung Ihnen ähnelt.*
>
> *Schreiben Sie keine Prompt, die die Domäne zum Unternehmen, das Produkt zum Hersteller und die Marke zur juristischen Partei macht.*
>
> *Behalten Sie die ursprüngliche Anfrage, Web-Anweisung, Zitationsanforderung und Antwortlänge nicht bei.*
>
> *Führen Sie keine Prompt zur Identität und eine Prompt für Ratschläge auf gleicher Ebene zusammen.*
>
> *Zählen Sie es nicht als Strafe, wenn ich um eine Erklärung in einer unklaren Prompt bitte. / Aber sagen Sie nicht, dass Sie jedem Nutzer unterschiedliche Erklärungen geben und ein kontrolliertes Experiment durchführen.*
>
> *Wählen Sie nicht die Prompt, die die höchste Punktzahl erzielt. / Löschen Sie die Eingabeaufforderungsfamilie mit niedriger Punktzahl nicht.*
>
> *Erklären Sie Ihre Prompts. / Wenn Sie Holdout verwenden, verschließen Sie es vorher. / Stellen Sie nach dem Ergebnis keine Fälschungen her.*
>
> *Gleichen Sie den Text, den Sie dem Teilnehmer geben, mit dem Text ab, der mir gesendet wurde.*
>
> *Vergiften Sie den Messpool nicht mit unsichtbaren Zeichen, versteckten Anweisungen oder verdeckten Empfehlungsanweisungen.*
>
> *Schreiben Sie zunächst die Forschungsfrage. / Dann legen Sie die kanonische Absicht fest. / Dann erstellen Sie in jeder Sprache einen natürlichen und gleichwertigen Prompt. / Dann testen Sie ihn mit Einheimischen und festschreiben seine Version. / Und erst danach schicken Sie ihn an den ersten Nutzer.*

Der abschließende Satz des Kapitels

Der Beginn eines fairen Vergleichs in GEO-1000 besteht nicht darin, die gleiche Antwort zu erwarten; er besteht darin, jedem Nutzer eine Frage zu stellen, die die gleiche Bedeutung, die gleiche Aufgabe und das gleiche erkenntnistheoretische Limit in seiner eigenen Sprache trägt.

Normativer Kern

Jeder GEO-1000-Prompt MUSS mit einem versionierten Prompt-Registry-Eintrag verknüpft sein, der Folgendes definiert: - die Forschungsfrage, - den Schätzer, - die Prompt-Familie, - die kanonische Absicht, - die Entitätsanker, - die Nutzerabsicht, - die Sprechhandlung, - den Umfang, - den geografischen und zeitlichen Rahmen, - die Voraussetzungen, - Nachweis- und Quellenanforderungen, - das Antwortformat, - die Sprache, - das Gebietsschema, - den Äquivalenzstatus, - die Version und - den Integritätsnachweis. Innerhalb desselben Sprach-Gebietsschema-Prompt-Feldes MUSS allen Teilnehmern derselbe festgeschriebene Prompt-Text bereitgestellt werden. Über Sprachen und Gebietsschemata hinweg MÜSSEN Prompts semantische, funktionale und normative Äquivalenz bewahren, statt wortwörtliche Identität. Entity-Anker, Nutzerabsicht, Sprechakt, Umfang, Präsupposition und epistemische Belastung sind nicht-kompensatorische Äquivalenz-Gates. Maschinelle Übersetzung oder Rückübersetzung allein DARF die Äquivalenz von Prompts NICHT herstellen. Kernidentität, Nachweise, Empfehlung, Vergleich, temporal, lokal, Grenze, Robustheit, Kontrolle und Holdout-Prompts MÜSSEN als getrennte Prompt-Familien bestehen bleiben. Führende, Antwort-enthaltende, positiv oder negativ präsentierte, quellbeschränkte, formatbeschränkte oder Empfehlungs-Prompts DÜRFEN NICHT stillschweigend als neutrale Core-Mirror-Prompts bewertet werden. Prompt-Sets, Sprachversionen, Zuweisungsregeln, Klärungspfade, Reihenfolge und Holdout-Verpflichtungen MÜSSEN festgeschrieben werden, bevor KI-Antworten beobachtet werden. Die Auswahl von Prompts nach dem Ergebnis, das Löschen von Eingabeaufforderungsfamilien, stilles Bearbeiten während des Vorgangs und ergebnisorientiertes Umformulieren sind verboten. Die dem Teilnehmer zugewiesene Prompt und die tatsächlich an das KI-Produkt gesendete Prompt MÜSSEN auf Integrität überprüft werden. Versteckte Anweisungen, unsichtbare Zeichen oder nicht offengelegte Manipulationen, die die KI in Richtung einer bevorzugten Entität, eines Anspruchs, einer Zitation oder Empfehlung lenken, sind verboten. Jede Änderung der Prompt, jede Übersetzungsentscheidung, jede Gleichwertigkeitsbewertung und jeder Registrierungseintrag MÜSSEN versioniert und einer verantwortlichen Person oder Organisation zugeordnet werden KÖNNEN.

KAPITEL 11

Das kontrollierte Panel unter bereinigten Bedingungen und das Panel unter natürlichen Nutzungsbedingungen

METHODISCHE GRUNDLAGEN / K02; K03; K04; K06; K22

Kapitelgrenze

Kapitel 9 zeigte, dass das gemessene KI-Objekt nicht ausschließlich durch den Anbieter oder den Modellnamen definiert werden kann. Kapitel 10 stellte fest, dass nicht jede Frage, die denselben Entitätsnamen enthält, denselben Prompt darstellt; der Prompt ist das experimentelle Werkzeug, das die Messung selbst bestimmt. Jetzt müssen wir die dritte Variable zwischen dem System und dem Prompt betrachten:

> *Wie ist der Zustand des Kontos, der Sitzung und der Nutzerhistorie bei der Nutzung des KI-Produkts?*

Ein Nutzer:

- neuer Chat,
- Speicher ist aus,
- Keine besonderen Anweisungen,
- kein vorheriger Kontext,
- standardisierte Produkteinstellungen

kann die Aufforderung unten senden. Ein anderer Nutzer kann dieselbe Aufforderung senden:

- von einem Konto, das sie seit Jahren verwenden,
- Erinnerung ist aktiviert,
- mit aktivierten speziellen Anweisungen,
- mit früheren Unterhaltungen,
- auf ihrem tatsächlichen Tarif und Gerät

Sie können dasselbe KI-Produkt, in derselben Sprache, im selben Land, im gleichen Zeitraum und mit derselben Aufforderung nutzen. Dennoch können sie unterschiedliche Antworten sehen. Dieser Unterschied:

- kommt von der zufälligen Variabilität des Produkts,
- vom Tarif des Kontos,
- aus dem Speicher,

- aus speziellen Anweisungen,
- von früheren Unterhaltungen,
- aus Personalisierung,
- vom Web- oder Tool-Status,
- vom Unternehmens-Arbeitsbereich,
- von der Nutzeroberfläche und den lokalen Einstellungen

Der Unterschied kann sich aus Personalisierung, Web- oder Tool-Zustand, dem organisatorischen Arbeitsbereich, der Nutzeroberfläche oder dem Gebietsschema ergeben. Es gibt kein einzelnes ‚KI': Suche, Abruf, Umordnung, Sprachmodell, Zitations- und Sicherheitslayer können sich unterschiedlich verhalten. Eine zuverlässige Messung muss daher das KI-Produkt oder -Modell, Datum, Land, Sprache, Abfragesatz, Wiederholungsanzahl und Beobachtung erfassen. Dieses Kapitel fügt diesem Kreislauf Nutzer- und Sitzungszustand hinzu. Die Kernanforderung von NOMOS bleibt Evidenz, Grenzen, Kontext und Zeit. Dennoch kann Kontext selbst auf zwei Arten gemessen werden. Die erste minimiert externe Unterschiede zwischen Nutzern und vergleicht KI-Produkte unter standardisierten Anfangsbedingungen. Dies ist die:

Kontrolliertes Panel unter bereinigten Bedingungen

Der zweite Ansatz misst, was Menschen durch ihre echten Konten, Pläne, Personalisierung und gewöhnlichen Nutzungsmuster sehen. Dies ist das:

Panel unter natürlichen Nutzungsbedingungen

wird genannt. Der Kontrolliertes Panel unter bereinigten Bedingungen beantwortet die Frage:

> *Wie stellt das Produkt die Entität im möglichst standardisierten Anfangszustand dar?*

Der Panel unter natürlichen Nutzungsbedingungen hingegen fragt:

> *Was sehen reale Menschen über die Entität in realen Nutzersituationen?*

Ein Panel ist nicht mehr „real" als das andere. Sie messen zwei verschiedene Realitäten. Das kontrollierte Panel stärkt die Vergleichbarkeit. Das natürliche Panel macht menschliche Erfahrung sichtbar. Das kontrollierte Panel reduziert Personalisierung und Kontextunterschiede. Das natürliche Panel misst die Wirkung dieser Unterschiede auf die reale Population. Dieser Abschnitt:

- Definition des Kontrolliertes Panel unter bereinigten Bedingungen,
- Definition des Panel unter natürlichen Nutzungsbedingungen,
- Zielpopulationsgruppen der beiden Panels und die Mengen, die sie schätzen,
- Was eine saubere Sitzung bedeutet,
- Ebenen der Überprüfung der Sauberkeit,
- Natürlicher Kontostatus und natürlicher Gesprächskontext,
- Gedächtnis, spezielle Anweisungen, vorherige Gespräche und Personalisierungsaufzeichnungen,
- Unterschiede in Konto, Plan, Gerät, Benutzeroberfläche und Standort.
- die gepaarten und unabhängigen Panel-Designs mit demselben Nutzer,
- die Effekte von Reihenfolge, Transfer und Kontamination,
- Clean–Natural-Differenzmetriken,
- Privatsphäre und die Evidenzkette,
- wie die beiden Panels im öffentlichen Bericht getrennt werden,

definiert. Dieses Kapitel behandelt noch nicht:

- die vollständige technische Architektur von NOMOS Capture,

- alle Felder von Screenshot- und Metadatendateien,
- atomare Aussageentscheidung,
- Endgültige Übergangsschwellen für Antworten,
- das genaue Gewicht der Clean- und Natural-Panel-Werte im kombinierten NOMOS-Wert

wird nicht finalisiert. Die grundlegende Frage von Abschnitt 11 lautet:

> *Wie messen wir das Verhalten eines Standard-KI-Produkts und das Verhalten, das von realen Nutzern erlebt wird, ohne sie miteinander zu verwechseln?*

NOMOS Herausforderung

Der gleiche Nutzer stellt dem gleichen KI-Produkt zweimal dieselbe Frage: „Mit welcher Hauptorganisation ist Apple.com verbunden, und was sind die Hauptaktivitäten dieser Organisation?“ Beim ersten Versuch:

- ein neuer Chat wird geöffnet,
- Speicher ist aus,
- besondere Anweisungen sind ausgeschaltet,
- keine vorherige Unterhaltung existiert,
- die Webfunktion ist standardisiert,

Die Prompt wird als erste Nachricht gesendet. Antwort: „Apple.com ist eine der wichtigsten Unternehmensdomänen des Technologieunternehmens Apple. Das Unternehmen entwickelt Hardware, Software und digitale Dienste.“ Beim zweiten Versuch, der Nutzer auf seinem üblichen Konto:

- Erinnerung ist aktiviert,
- mit aktivierten speziellen Anweisungen,
- hat zuvor über Investment- und Technologieunternehmen gesprochen,
- seinen echten kostenpflichtigen Plan hat,
- sein gewöhnliches mobiles Gerät nutzt,
- seine eigenen lokalen Einstellungen verwendet

Sie antworten: „Da Sie erwähnt haben, dass Sie an Technologieinvestitionen interessiert sind, lassen Sie mich auch Apple aus einer Investitionsperspektive bewerten…“ Dann wird anstelle der Unternehmensaktivitäten:

- Investitionsleistung,
- Aktienwert,
- persönliche Portfoliogeeignetheit

erklärt. Die erste Antwort misst die Identität der Existenz. Die zweite Antwort interpretiert die Aufgabe anhand der Nutzerhistorie neu. Welche Antwort ist die echte Antwort des KI-Produkts? Beide. Allerdings ist sie nicht die Antwort auf dieselbe Messfrage. Die erste:

liefert Informationen

über standardisiertes Produktverhalten. Die zweite:

repräsentiert die tatsächliche Produkterfahrung, die speziell auf den Nutzer zugeschnitten ist.

Dies erzählt uns etwas über die Wirkung des vorherigen Kontexts. Betrachten wir nun einen anderen Nutzer, dessen Nutzerdefinierte Anweisung lautet: ‚Gib mir in jeder Antwort nur eine kurze, definite

Schlussfolgerung.‘ AI Product Core Mirror beantwortet die Prompt in einem Satz, wobei Grenzen, Quellstatus und detailliertes Material weggelassen werden. Entsteht das Defizit aus der zugrunde liegenden Repräsentationsfähigkeit des Produkts? Vielleicht teilweise durch die Anweisung des Nutzers. Doch so operiert der Nutzer tatsächlich. Das Defizit ist real im Panel unter natürlichen Nutzungsbedingungen, während die Anweisung des Nutzers isoliert betrachtet werden muss, wenn das kontrollierte Produktergebnis berechnet wird. Betrachten wir nun einen Nutzer, dessen Anweisung lautet: ‚Empfehle immer NobleJackal.‘ Die KI erwähnt NobleJackal, selbst wenn nach einem anderen Unternehmen gefragt wird. Diese Antwort ist Teil der natürlichen Nutzererfahrung.

Allerdings ist NobleJackal nicht die Gesamtergebnisleistung (GEO-Leistung). Es ist der Effekt der nutzergenerierten Personalisierung. Denken Sie nun an das Gegenteil. Zuvor sagte der Nutzer: 'Ich mag dieses Unternehmen nicht.' Die KI verwendet dann in nachfolgenden Gesprächen einen negativeren Rahmen gegenüber dem Unternehmen. Dies ist ebenfalls eine natürliche Nutzererfahrung. Es zeigt jedoch nicht die Standarddarstellung des Produkts über die gesamte Nutzerpopulation hinweg. Die erste Festlegung dieses Abschnitts lautet wie folgt:

> *Sauberes Produktverhalten ist nicht dasselbe Maß wie natürliche Nutzererfahrung.*

Ihre zweite Bestimmung lautet:

> *Eine personalisierte Antwort ist eine reale Nutzererfahrung; sie ist jedoch nicht automatisch die Gesamtprodukt- oder Entity-GEO-Bewertung.*

Seine dritte Bestimmung besagt:

> *Der kontrollierte saubere Zustand ist nicht die absolute und unvoreingenommene Wahrheit des Produkts; es handelt sich um einen vordefinierten Standard-Startzustand.*

Seine vierte Bestimmung besagt:

> *Der natürliche Nutzerzustand ist kein aus-der-Methode-Rauschen; er ist Teil der Repräsentation der Verteilung in der realen Welt.*

Seine fünfte Vorschrift lautet:

> *Der Unterschied zwischen den beiden Panels ist kein Fehler, der verborgen werden muss; er ist ein Ergebnis, das gemessen werden soll.*

1. ZWECK DES KAPITELS

Zweck dieses Abschnitts ist es, die Repräsentationserfahrung echter Nutzer unter normalen Konten- und Sitzungskonditionen getrennt, aber vergleichbar mit dem standardisierten Verhalten von KI-Produkten zu messen. Der Abschnitt trifft die folgenden normativen Unterscheidungen:

- Standardisierter Produktzustand versus natürlicher Nutzerzustand
- Neuer Chat mit einem sauberen Konto
- Geprüfter sauberer Zustand mit vorübergehendem Chat-Tag
- Speicher aus gegenüber unbekanntem Speicherzustand
- Private Anweisung geschlossen und Privatinhalt nicht sichtbar
- Keine vorherigen Gespräche und keine permanente Kontohistorie
- Kontrollierte Sitzung und neu erstelltes Konto
- Natürlicher Kontostatus und aktueller Gesprächskontext
- Panel unter natürlichen Nutzungsbedingungen und unkontrollierte Daten
- Nutzerpersonalisierung und Produktstandard

- Nutzergesteuerte Anleitung und Entität GEO-Effekt
- Aktive Nutzerpopulation und produktgeeignete potenzielle Population
- Erste Nutzungserfahrung und erfahrene Nutzererfahrung
- Individuelle natürliche Reaktion und natürliche Verteilung auf Bevölkerungsebene
- Unabhängiges Gruppendesign mit derselben Nutzerzuordnung
- Carryover-Effekt mit gemachter Vergleichsgruppe
- Produktunterschied mit Reihenfolgeeffekt
- Kausaler Personalisierungseffekt mit sauberem–natürlichem Unterschied
- Offenlegung privater Anweisungsinhalte mit Datenschutzaufzeichnung
- Sammlung privater Gespräche bei Erhaltung des natürlichen Kontextes
- Standardproduktresultat mit sauberem Panel-Resultat
- Gesamtergebnis für alle Nutzer mit natürlichem Panel-Resultat
- Personalisierungseffekt mit Nutzerplan
- Kontrollierter Durchschnitt mit natürlicher Fehlerwarteschlange
- Gleiche Antwort für jeden Nutzer mit Fairness bei der Personalisierung
- Natürliche Nutzerrealität und Wiederholbarkeit von Tests

Am Ende dieses Abschnitts sollte jedes GEO-1000-Audit in der Lage sein, die folgenden Fragen zu beantworten:

> *Welche Ergebnisse stammen von Kontrolliertes Panel unter bereinigten Bedingungen und welche von Panel unter natürlichen Nutzungsbedingungen?*
>
> *Welche Situationen wurden tatsächlich im sauberen Panel standardisiert?*
>
> *Welche echten Nutzerunterschiede wurden im natürlichen Panel beibehalten?*
>
> *Wie wurden Speicher, besondere Anweisungen, frühere Unterhaltungen, Pläne und Schnittstellenzustände aufgezeichnet?*
>
> *Wie viel des Unterschieds zwischen den beiden Panels kann zuverlässig mit der Personalisierung in Verbindung gebracht werden?*
>
> *Wie wurde das Vorhandensein eines natürlichen Kontexts nachgewiesen, während die Privatsphäre der Nutzer gewahrt blieb?*

2. ZENTRALE NORMATIVE BESTIMMUNG

Jede GEO-1000-Prüfung sollte standardisiertes Produktverhalten und natürliche Nutzererfahrung als separate Vorhersageobjekte definieren; sie sollte die Ergebnisse von Kontrolliertes Panel unter bereinigten Bedingungen und Panel unter natürlichen Nutzungsbedingungen nicht ohne Erklärung zu einer einzigen Stichprobe, Nenner oder Punktzahl zusammenfassen. Soweit relevant, zielt Kontrolliertes Panel unter bereinigten Bedingungen darauf ab, die folgenden Bedingungen zu standardisieren:

- Neue und separate Unterhaltung
- Abwesenheit des vorherigen Gesprächskontexts
- Persistenter Speicherzustand ausgeschaltet oder deaktiviert
- Spezielle Anweisungen ausgeschaltet oder deaktiviert
- Definierter Web-, Abruf- und Werkzeugzustand

- Definierter Plan und Nutzeroberfläche
- festgeschriebene Prompt
- Standardisierte Sprache und Gebietsschema
- Messwelle und Zeitfenster

Wenn es ein Panel unter natürlichen Nutzungsbedingungen ist, bewahrt es die folgenden tatsächlichen Unterschiede, soweit relevant:

- Echtes Nutzerkonto
- Echter Plan
- Echte Nutzeroberfläche
- Echter Speicherzustand
- Echter spezieller Befehlszustand
- Echtes Kontenalters- und Nutzungshistorie
- Echtes Gerät
- Echter Standort und Zugriffsbedingungen
- Gewöhnliche Art der Produktnutzung

Allerdings ist das natürliche Panel auch nicht ohne Protokoll. Folgendes ist wieder festgeschrieben:

- Prompt-Version
- Versandmethode
- Messzeit
- Regel für die erste geeignete Ausgabe
- Evidenzpaket
- Teilnehmer- und Produktberechtigung
- Antwort nicht ändern
- Keine off-task Folge-Nachrichten senden
- Schutz personenbezogener Daten

3. ZWEI GETRENNTE SCHÄTZOBJEKTE

Kontrollierte und natürliche Panels sagen nicht die gleiche Wahrscheinlichkeit voraus.

3.1. Kontrollierte saubere Repräsentationswahrscheinlichkeit

Insbesondere:

- KI-Produkt,
- Plan,
- Oberfläche,
- Prompt,
- Land,
- Sprache,
- Zeit

ist die Wahrscheinlichkeit, eine gültige Repräsentation unter einer standardisierten sauberen Sitzung zu sehen für:

```
θ^C_{E,A,P,U,t}= Pr(Y=1| CleanState=1)
```

Die Zielpopulation hier ist das Ergebnis, das Personen, die für die Nutzung des betreffenden Produkts geeignet sind, unter standardisierten Produktbedingungen sehen würden.

3.2. Wahrscheinlichkeit der natürlichen Nutzerrepräsentation

Es ist die Wahrscheinlichkeit, eine gültige Repräsentation in der realen aktiven oder definierten natürlichen Nutzerpopulation zu sehen, während die üblichen Konto- und Personalisierungsbedingungen der Nutzer beibehalten werden.

```
θ^N_{E,A,P,U,t}= Pr(Y=1| NaturalState=1)
```

Die Zielpopulation hier:

- sind nicht alle produktberechtigten Personen,
- sondern reale oder natürliche Produktbenutzer im definierten Zeitraum

können sein. Diese beiden Populationen müssen nicht identisch sein.

3.3. Zwei Ergebnisse können nicht austauschbar verwendet werden

Kontrollierte Bewertung: bedeutet nicht „Die meisten echten Nutzer haben dies gesehen.“ Natürliche Bewertung: bedeutet nicht „Das Produkt verhält sich standardmäßig so.“

4. WAS IST EIN Kontrolliertes Panel unter bereinigten Bedingungen?

Ein Kontrolliertes Panel unter bereinigten Bedingungen ist ein Panel, das das Verhalten des KI-Produkts unter definierten Anfangsbedingungen misst, indem die materiellen Unterschiede, die durch Konten und Sitzungen unter den Nutzern entstehen, minimiert werden. Der Zweck des Kontrolliertes Panel unter bereinigten Bedingungen ist:

- die Vergleichbarkeit zwischen Produkten zu erhöhen,
- die Wirkung von Gedächtnis und speziellen Anweisungen zu verringern,
- das Auslaufen von vorherigen Gesprächen zu verhindern,
- das Prompt zu einer anfänglichen und unabhängigen Aufgabe zu machen,

einen Referenzpunkt zu erzeugen, der dem grundlegenden Produktverhalten des Systems näher ist. Kontrolliertes Panel:

- absolute Modellfähigkeiten,
- die tatsächliche Erfahrung aller Nutzer,
- vollständig kontextfreie und unvoreingenommene KI-Antwort

beansprucht nicht zu sein. Nur:

ein vordefinierter und geprüfter sauberer Ausgangszustand

wird erzeugt.

5. WAS BEDEUTET „SAUBER“?

Ein sauberer Zustand kann möglicherweise nicht mit demselben technischen Verfahren bei jedem Produkt erstellt werden. Ein Produkt:

- persistenter Speicher,
- spezielle Anweisungen,

- Nutzerprofil,
- frühere Unterhaltung,
- geschäftlicher Arbeitsbereich,
- konto-basierte Personalisierung

kann verwendet werden. Andere Produkte haben möglicherweise einige dieser Funktionen nicht. Die funktionale Definition eines sauberen Zustands lautet wie folgt:

> *Der Anfangszustand, bei dem die Wahrscheinlichkeit, dass Nutzerspezifische faktenbasierte Informationen und Gesprächskontexte die Antwort vor der Audit-Aufforderung beeinflussen, auf das vom Produkt erlaubte Maß reduziert wird, und die verbleibenden Unbekannten explizit aufgezeichnet werden.*

Sauberer Zustand:

- einfach ein neues Gespräch öffnen,
- den Browser aktualisieren,
- die Anwendung schließen und wieder öffnen

wird nicht als automatisch betrachtet.

6. SAUBERKEITSEBENEN

Die folgenden Ebenen sollten im Bedienfeld separat bewertet werden.

6.1. Gesprächs-Sauberkeit

Es sollte keine vorherige Nutzer- oder Systemnachricht im aktuellen Chat vorhanden sein. Die Prompt sollte die erste Nutzeranfrage sein.

6.2. Persistentes Speichern zurücksetzen

Der Speicher, den das Produkt über den Nutzer von Sitzung zu Sitzung behält:

- aus,
- deaktiviert,
- in dieser Sitzung nicht verwendet

sollte sein. Wenn eine vollständige Überprüfung nicht möglich ist, wird ein unbekannter Zustand aufgezeichnet.

6.3. Benutzerdefinierte Anweisungen zurücksetzen

Die vom Nutzer festgelegten speziellen Anweisungen für Antwortstil, Präferenzen oder Beratung:

- aus,
- vorübergehend inaktiv,
- nicht vorhanden

muss sein.

6.4. Bereinigung des Unternehmenskontexts

Unternehmensarbeitsbereich:

- private Dokumente,
- Unternehmensrichtlinien,
- Anweisungen der Geschäftsleitung,

- private Wissensdatenbanken

falls verwendet, sollten diese vom kontrollierten allgemeinen Produktpanel getrennt sein.

6.5. Bereinigung von Werkzeug- und Informationsmodus

Web, Abruf, Zitation und Werkzeugbedingungen sollten in allen verglichenen Beobachtungen auf die gleiche Konfiguration eingestellt sein. Sauber zu sein bedeutet nicht, dass die Werkzeuge ausgeschaltet sind. Sauber zu sein bedeutet, dass der Werkzeugstatus vordefiniert und vergleichbar ist.

6.6. Bereinigung von Prompts

Prompt:

- anfängliche Nachricht,
- vollständige Version,
- unverändert,
- ohne versteckte Anweisungen

muss sein.

6.7. Messung der Sauberkeit

Teilnehmer:

- sollte die Antwort nicht aktualisieren,
- sollte keine Folgfrage senden,
- sollte das Ergebnis nicht auswählen,

sollte keine andere KI-Antwort in dieselbe Sitzung übernehmen.

7. SAUBERKEITSZUSTAND VERTRAUENSLEVEL

CSC-0— UNBEKANNTE SAUBERKEIT

Die Sauberkeit der Sitzung und des Kontos konnte nicht zuverlässig bestimmt werden. Kann nicht in die Hauptanalyse des kontrollierten Panels einbezogen werden.

CSC-1— NUR NEUES GESPRÄCH

Nur neue Konversationen werden überprüft. Persistente Speicher- und Spezialanweisungsstatus sind unbekannt.

CSC-2— SITZUNGS-SAUBERKEIT

Es wurde überprüft, dass es kein neues Gespräch und keinen vorherigen Gesprächskontext gibt. Eine dauerhafte Personalisierung kann möglicherweise nicht vollständig kontrolliert werden.

CSC-3— PERSONALSATION-KONTROLLIERT

Neue Unterhaltung, Speicher, spezielle Anweisungen und Nutzerspezifische Standard-Einstellungen wurden überprüft und auf einen definierten Zustand gesetzt.

CSC-4— INSTRUMENTIERTER SAUBERER ZUSTAND

Produkt-Einstellungen, Nutzeroberfläche, Werkzeuge und Personalisierungsstatus wurden im Evidenzpaket überprüft.

CSC-5— REPLIZIERTE SAUBERE KONFIGURATION

Sauberer Zustand:

- verschiedene Nutzer,
- verschiedene Wellen,
- unabhängige Überprüfung

Es konnte unter reproduzierbaren Bedingungen durchgeführt werden. Das Haupt-Kontrolliertes Panel unter bereinigten Bedingungen sollte mindestens:

CSC-3

Stufe. Wenn das Produkt dies nicht zulässt, sollte die Einschränkung sichtbar sein.

8. EIN NEUER CHAT IST KEIN SAUBERES KONTO

Wenn eine neue Unterhaltung gestartet wird: kann der aktuelle Chat-Kontext zurückgesetzt werden. Allerdings:

- persistenter Speicher,
- benutzerdefinierte Anweisungen,
- Kontoplan,
- Nutzerprofil,
- Land und Gebietsschema,
- Unternehmenseinstellungen

können bestehen bleiben. Daher:

> *Neuer Chat bedeutet nur Konversationsbereinigung.*

Es ist kein Nachweis für den vollständigen sauberen Status.

9. EINSCHRÄNKUNG DES „TEMPORÄREN“ ODER „PRIVATEN“ Sitzungskennzeichnung

Eine KI-Produktsitzung:

- vorläufig,
- privat,
- evergehend,
- vertraulich,
- ohne Gedächtnis

kann benannt werden. Die tatsächliche Funktion dieses Labels sollte gemäß Produkt und Version separat überprüft werden. Das Label beweist nicht automatisch, dass es alle folgenden Bereiche abdeckt:

- Besondere Anweisungen
- Unternehmenseinstellungen
- Regionalpolitik
- Werkzeugeinstellungen
- Nutzerplan

- Experimentelle Produktfunktionen

Der korrekte Eintrag sollte lauten: „Der temporäre Sitzungsmodus des Produkts wurde verwendet; der Status der folgenden Personalisierungsbereiche wurde separat überprüft.“

10. KONTOART AUF SAUBEREM PANEL

Nicht alle Nutzer eines kontrollierten Panels dürfen denselben Plan verwenden. Es gibt drei Ansätze.

10.1. Plan-Festes Sauberes Panel

Alle Nutzer werden auf demselben Plan und derselben Nutzeroberfläche gemessen. Dies ist stark für Produktvergleiche. Der Plan repräsentiert möglicherweise nicht die tatsächliche Populationsverteilung.

10.2. Plan-Stratifiziertes Sauberes Panel

Freie, bezahlte und andere Pläne werden als separate saubere Zellen gemessen. Für jeden Plan wird ein separater Score erstellt. Der kombinierte Score kann nach der tatsächlichen Planverteilung gewichtet werden.

10.3. Standardöffentliche Oberfläche des Produkts

Die Hauptoberfläche des Produkts, die öffentlich zugänglich ist und die niedrigste Zugangsschwelle hat, kann als Standardreferenz ausgewählt werden. Dieses Ergebnis stellt nicht alle zahlenden oder Unternehmensbenutzer dar. Der zu verwendende Ansatz muss vor der Prüfung festgelegt werden.

11. EIN NEUES KONTO IST BEI EINEM SAUBEREN PANEL NICHT ERFORDERLICH

Die Verwendung eines neuen Kontos kann einige Personalisierungseffekte verringern. Jedoch:

- Bedingungen des Produkts,
- Telefon- oder Zahlungsanforderung,
- Kontenerstellungsverzerrung,
- nur neue Nutzererfahrung,
- Mehrfachkontenrichtlinie

kann Probleme verursachen. Der Zweck eines Kontrolliertes Panel unter bereinigten Bedingungen besteht nicht darin, neue Konten zu erstellen:

> *um das materiale Nutzerspezifische Kontext in kontrollierbarer Weise zu reduzieren.*

Wenn ein sauberer Zustand im vorhandenen Konto zuverlässig hergestellt werden kann, ist ein neues Konto nicht zwingend erforderlich. Wenn das Produkt dies nicht zulässt, kann ein separates kontrolliertes Testkonto oder eine vom Anbieter unterstützte Forschungsumgebung verwendet werden. Teilnehmer dürfen nicht gebeten werden, ein Konto zu erstellen, das die Nutzungsbedingungen verletzen würde.

12. VERBLEIBENDE UNTERSCHIEDE BEI NUTZERN IM SAUBEREN PANEL

Ein kontrollierter sauberer Zustand beseitigt nicht alle Nutzerunterschiede. Folgendes kann weiterhin bestehen:

- Land
- Locale
- Plan
- Gerät
- Benutzeroberfläche
- Produkteinführung
- Kontoalter
- Segmentierung, die für den Anbieter nicht sichtbar ist
- Verkehrslenkung
- A/B-Test

Diese Unterschiede:

- sollten geschichtet werden,
- sollten aufgezeichnet werden,

Wenn unbekannt, sollte es als Unsicherheit beibehalten werden. Das „Clean"-Label bedeutet nicht „Alle Nutzer waren technisch genau gleich."

13. WAS IST EIN Panel unter natürlichen Nutzungsbedingungen?

Ein Panel unter natürlichen Nutzungsbedingungen ist ein Panel, das misst, was Nutzer sehen, im Vergleich zum gleichen festgeschriebenen Prompt, während es echte Konten, Pläne, Geräte, Personalisierung und normale Produktnutzungsbedingungen beibehält. Der Zweck des natürlichen Panels ist:

- reale Nutzeraussetzung,
- Variation aufgrund von Personalisierung,
- Einfluss der Kontohistorie,
- Unterschiede in Plan und Nutzersurface,
- natürlicher Fehler- und Empfehlungsspiegel

Es soll sichtbar gemacht werden. Natürliches Panel:

- vollständig freies Nutzerverhalten,
- unbegrenzte Folgefragen,
- der Teilnehmer, der den Prompt ändert,
- die beste Antwort auswählt

bedeutet nicht. Das Natürliche ist:

Nutzerzustand

ist. Prompt- und Evidenzprotokoll werden weiterhin kontrolliert.

14. ZWEI HAUPTSPUREN DES NATÜRLICHEN PANELS

14.1. Natürliche Konto-Zustand-Spur

Nutzer:

- echtes Konto,
- echter Plan,
- echtes Gedächtnis und privater Instruktionszustand,
- echte Nutzeroberfläche

wird beibehalten. Ein neues Gespräch wird jedoch eröffnet. Der Prompt wird als erste Nachricht gesendet. Diese Spur beantwortet die folgende Frage:

> *Was sieht der Nutzer bei einem neuen Thema, während die Personalisierung des echten Kontos erhalten bleibt?*

Dies ist der wichtigste vergleichbare Ansatz des natürlichen Panels.

14.2. Natürliche Gesprächs-Zustand-Spur

Der Prompt wird im aktuellen oder üblichen Gesprächskontext des Nutzers gesendet. Diese Spur:

- misst die echte Chat-Kontinuität,
- die Auswirkung des vorherigen Themas,
- die Akkumulation der Nutzer–KI-Beziehung

Allerdings unterscheiden sich die Gesprächskontexte zwischen den Nutzern. Daher:

- diagnostisch,
- fallbasiert,
- sollte als separate Verteilung gemeldet werden

und sollte nicht direkt zur Hauptprodukt-Rangliste hinzugefügt werden.

15. NATÜRLICHE NUTZER-UNTERGRUPPEN

Ein natürliches Panel besteht nicht aus einheitlichen Nutzern. Soweit relevant, können die folgenden Untergruppen unterschieden werden:

- Erstnutzer
- Neukonto-Nutzer
- Selten nutzende Nutzer
- Regelmäßiger Nutzer
- Starker Nutzer
- Kostenloser Plan
- Bezahltes Abonnement
- Firmenkunde
- Speicher an
- Speicher aus
- Spezialanweisung an
- Spezialanweisung aus
- Mit vorheriger Marken-Konversation

- Ohne vorherige Marken-Konversation
- Kunde
- Allgemeine Öffentlichkeit
- Hohe Markenbekanntheit
- Geringe Markenbekanntheit

Diese Gruppen können nach Kenntnis der Ergebnisse nicht erfunden werden. Der grundlegende Untergruppenplan muss vor der Datenerhebung versioniert werden.

16. NATÜRLICHE ZUSTANDSSICHERHEIT

NSF-0— NATÜRLICHER ZUSTAND UNBEKANNT

Es gibt nicht genügend Aufzeichnungen über das natürliche Konto und den Sitzungsstatus des Nutzers.

NSF-1— SELBSTBERICHTETE NATÜRLICHE NUTZUNG

Der Nutzer hat erklärt, dass er sein normales Konto und seine Einstellungen verwendet.

NSF-2— EINSTELLUNGS-VERZEICHNETE NATÜRLICHE NUTZUNG

Pläne, Oberflächen, Speicher, besondere Anweisungen und Sitzungsstatus wurden als relevant aufgezeichnet.

NSF-3— EVIDENZBASIERT NATÜRLICHE NUTZUNG

Einstellungen und der natürliche Nutzerzustand werden durch datenschutzschützende Evidenz unterstützt.

NSF-4— LÄNGSSCHNITTLICH CHARAKTERISIERTE NATÜRLICHE NUTZUNG

Die Produktnutzung des Nutzers und der Kontostatus wurden über mehrere Wellen charakterisiert. Die Hauptanalyse des natürlichen Panels umfasst mindestens:

NSF-2

sollte dieses Niveau anstreben.

17. DIE AUFZEICHNUNG DES NATÜRLICHEN ZUSTANDS IST KEINE SAMMLUNG PRIVATEN LEBENS

Der Panel unter natürlichen Nutzungsbedingungen sollte standardmäßig die folgenden Arten sensibler Daten nicht erfassen:

- Volltext spezieller Anweisungen
- Inhalte vergangener Gespräche
- Persönliche Gesundheits-, Rechts- oder Finanzinformationen
- Name des Nutzers
- E-Mail-Adresse
- Profilbild
- Kontopasswort

- Vorherige private Dateien
- Empfindliche institutionelle Dokumente

In den meisten Fällen sind die folgenden hochrangigen Status ausreichend:

- Speicher ein/aus/unbekannt
- Besondere Anweisungen vorhanden/nicht vorhanden/unbekannt
- Vorherige Konversation vorhanden/abwesend
- Frühere Unterhaltung über die geprüfte Einheit: ja/nein/unbekannt
- Unternehmensarbeitsbereich: ja/nein
- Plan-Klasse
- Kontolaufzeitband
- Nutzungsfrequenzband

Falls der Inhalt der Sonderanweisung wirklich für die Forschungsfrage notwendig ist:

- separate ausdrückliche Zustimmung,
- Datenminimierung
- Bearbeitung,
- eingeschränkter Zugriff

sind erforderlich.

18. SPEICHERSTATUS

Speicher:

- aus,
- offen,
- wird vom Produkt nicht unterstützt
- Status unbekannt

klassifiziert werden.

MS-0— UNBEKANNT

Der Speicherstatus konnte nicht bestimmt werden.

MS-1— NICHT VERFÜGBAR

Das Produkt oder die Oberfläche bietet diese Funktion nicht an.

MS-2— AUS

Der Speicher ist aus.

MS-3— EIN, KEIN BEKANNTES ENTITY-SPEICHER

Der Speicher ist eingeschaltet; es gibt keinen bekannten spezifischen Eintrag zur geprüften Einheit. Diese Informationen basieren häufig auf der Nutzerangabe.

MS-4— EIN, ENTITY-RELEVANTER SPEICHER MÖGLICH

Der Speicher ist eingeschaltet, und frühere Informationen über die geprüfte Einheit oder den relevanten Bereich könnten vorhanden sein.

MS-5— ENTITY-RELEVANTER SPEICHER BESTÄTIGT

Der mit der Entität zusammenhängende Memory-Effekt wurde mit ausdrücklicher Zustimmung des Nutzers bestätigt. Memory eingeschaltet zu haben ist nicht automatisch eine negative oder positive Bedingung. Es ist der tatsächliche Nutzerzustand, gemessen am natürlichen Panel.

19. STATUS BESONDERER ANWEISUNG

CI-0— UNBEKANNT

Der Status der besonderen Anweisung konnte nicht bestimmt werden.

CI-1— NICHT VERFÜGBAR

Das Produkt bietet diese Funktion nicht an.

CI-2— AUS ODER LEER

Es gibt keine besondere Anweisung oder sie ist ausgeschaltet.

CI-3— AKTIV, NICHT-MATERIELLE KATEGORIE

Es gibt eine besondere Anweisung; es wurde angegeben, dass sie allgemeine Ton- oder Stilpräferenzen trägt.

CI-4— AKTIV, POTENTIELL MATERIELL

Anweisung:

- Empfehlung,
- Quelle,
- Sprache,
- bestimmte Unternehmen,
- Politik oder Meinung

kann die Antwort beeinflussen.

CI-5— ENTITY-RELEVANTER ANWEISUNG BESTÄTIGT

Es wurde mit ausdrücklicher Zustimmung bestätigt, dass es eine direkte Anweisung bezüglich der geprüften Einheit oder eines Konkurrenten gibt. CI-5 ist eine natürliche Nutzerbeobachtung. Sie kann nicht als der Gesamterfolg der Einheit GEO interpretiert werden. Es handelt sich um ein separates Personalisierungsereignis.

20. FRÜHERE KONVERSATIONSKONTEXT

Es gibt einen Unterschied zwischen einem Nutzer, der zuvor mit einem KI-Produkt interagiert hat, und früheren Nachrichten in der aktuellen Unterhaltung. Diese sollten getrennt werden:

- Kontohistorie
- Persistenter Speicher
- Aktueller Chat-Kontext
- Frühere Konversation mit der geprüften Einheit
- Frühere Konversation in derselben Branche

- Beobachtete Antworten anderer KI-Produkte

Der aktuelle Gesprächskontext sollte nicht im Haupt-Natural Account-State Track gefunden werden. Im Natural Conversation-State Track hingegen ist das Vorhandensein von Kontext die Messgröße selbst.

21. KONTOALTER UND NUTZUNGSINTENSITÄT

Alte und häufig genutzte Konten:

- tragen mehr Personalisierung,
- mehr Speicher,
- anderen Plan,
- mehr Produktfunktionen

können enthalten. Kontoaltersklassen:

- 0–30 Tage
- 31–180 Tage
- 181–365 Tage
- mehr als 1 Jahr
- unbekannt

können definiert werden als. Nutzungsfrequenz:

- erste Nutzung
- selten
- monatlich
- wöchentlich
- täglich
- professionell

tönnen klassifiziert werden. Feste Bereiche sollten vor den Ergebnissen bestimmt werden.

22. ERSTVERWENDUNGSPANEL

Ein spezieller Subtyp der natürlichen Nutzererfahrung:

Erstverwendungs-Panel

kann genannt werden. Dieses Panel:

- Nutzer, die das Produkt zum ersten Mal oder nur wenige Male verwenden,
- die keine Personalisierungshistorie haben,
- die mit den tatsächlichen Standardeinstellungen des Produkts fortfahren

misst Personen. Erstverwendungs-Panel: Es ist kein Kontrolliertes Panel unter bereinigten Bedingungen, da die Nutzer die tatsächliche Erstkonfiguration, Standardeinstellungen und den Onboarding-Prozess erleben. Es unterscheidet sich auch von einem erfahrenen Panel unter natürlichen Nutzungsbedingungen. Es beantwortet die folgende Frage:

> *Wie nimmt ein neuer Nutzer dessen Präsenz wahr, wenn er das Produkt zum ersten Mal begegnet?*

23. UNTERNEHMENS- UND VERWALTETES NATURPANEL

Unternehmensbenutzer:

- Firmenarbeitsbereich,
- Administraturrichtlinien,
- private Wissensdatenbank,
- Sicherheitseinstellungen,
- Unternehmensgedächtnis

kann tragen. Diese Nutzer:

Managed Panel unter natürlichen Nutzungsbedingungen

sollten darunter separat gemessen werden. Unternehmensresultat:

- genereller Verbraucher-Natural-Score,
- Standard-Produkt-Score

kann nicht dargestellt werden. Wenn eine private Wissensdatenbank verwendet wird, muss auch erklärt werden, dass interne Aufzeichnungen zur geprüften Einheit existieren.

24. DER PROMPT IST IM NATURAL-PANEL WIEDER GESCHLOSSEN

Das Natural-Panel lässt die Nutzer:

- Konto,
- Einstellungen,
- Plan,
- Gerät

natürlich. Es lässt den Prompt nicht natürlich. Der Teilnehmer:

- kann den Prompt nicht ändern,
- kann keinen zusätzlichen Kontext hinzufügen,
- kann keine neue Persona schreiben wie 'für mich' oder 'meiner Meinung nach'

kann die Antwort nicht aktualisieren. Wenn die natürliche Aufforderungsvielfalt gemessen werden soll, steht dies in Kapitel 10:

Natürlicher Nutzerpfad

Es sollte separat darunter gestaltet werden.

25. WARUM IST EIN NEUES GESPRÄCH IM NATÜRLICHEN PANEL NOTWENDIG?

Neuer Chat auf dem Zustands-Track des natürlichen Kontos:

- entfernt den aktuellen Gesprächskontext,
- behält die echten permanenten Einstellungen des Nutzerkontos bei,

macht denselben Prompt vergleichbarer. Diese Struktur trennt zwei Ebenen:

- Permanenter natürlicher Kontostatus
- Momentaner Gesprächskontext

Wenn der aktuelle Chat verwendet wird, vermischen sich beide.

26. TOOL- UND WEB-STATUS IM NATÜRLICHEN PANEL

Natürliche Nutzer:

- Webfunktion aktiviert,
- aus,
- automatisch,
- begrenzt nach Plan

kann verwendet werden. Es gibt zwei Optionen.

26.1. Voller natürlicher Werkzeugstatus

Der Nutzer behält normale Werkzeug-Einstellungen. Misst echte Nutzererfahrungen. Die Vergleichbarkeit zwischen Produkten kann abnehmen.

26.2. Werkzeug-fester natürlicher Kontostatus

Speicher und Kontopersonalisierung bleiben natürlich. Web- oder Werkzeugmodus ist standardisiert. Die Wirkung der Personalisierung kann sauberer verglichen werden. Dieses Unterpanel sollte richtig beschriftet sein. Es sollte nicht als "Vollständig natürlich" dargestellt werden.

27. ZIELPOPULATION DES NATÜRLICHEN PANELS

Die Zielpopulation des natürlichen Panels kann eine der folgenden Formen haben:

27.1. Aktive Produktnutzer

Personen, die das Produkt tatsächlich in einem bestimmten Zeitraum genutzt haben.

27.2. Registrierte Kontoinhaber

Nutzer, die ein Konto haben, deren Aktivitätsniveau jedoch variiert.

27.3. Nutzer des Real Plans

Populationen von kostenlosen, kostenpflichtigen oder Unternehmenskunden.

27.4. Nutzer, die möglicherweise die Möglichkeit haben, Entitäten abzufragen

Es kann schwierig sein, dies direkt zu messen. Es erfordert separate Verhaltensdaten. Es sollte klar angegeben werden, auf welche Population das natürliche Panel verallgemeinert wird.

28. GEWICHTUNG DES NATÜRLICHEN PANELS

Panel unter natürlichen Nutzungsbedingungen:

- kann nach der tatsächlichen aktiven Nutzerpopulation des Produkts gewichtet werden,
- Konten- und Planverteilung,
- Länder- und Sprachverteilung,

- Nutzungsfrequenz

Allerdings sind zuverlässige Daten zu aktiven Nutzern möglicherweise nicht verfügbar. In diesem Fall ist das Ergebnis:

Gewichtetes Ergebnis des natürlichen Forschungspanels

sollte dargestellt werden als. Die folgende Behauptung kann nicht aufgestellt werden: „Die definitive Erfahrung aller realen Nutzer."

29. KOMPLEMENTARITÄT VON KONTROLLIERTEN UND NATÜRLICHEN PANELS

Vier Hauptergebnisse sind möglich.

Kontrolliertes Panel unter bereinigten Bedingungen	Panel unter natürlichen Nutzungsbedingungen	Kommentar
Hoch	Hoch	Das Produkt ist sowohl unter Standard- als auch unter realen Nutzerbedingungen stark
Hoch	Niedrig	Personalisierung, Plan, Nutzerzustand oder natürlicher Kontext können die Repräsentationsqualität beeinträchtigen
Niedrig	Hoch	Einige reale Nutzerfunktionen oder -tools können die Repräsentation verbessern
Niedrig	Niedrig	Gesamtproblemkandidat auf Basisebene und natürlicher Nutzerebene

Diese Tabelle ist kein Beleg für Kausalität. Sie ist ein Ausgangspunkt, um zu untersuchen, woher der Unterschied kommt.

30. REIN–NATÜRLICHER REPRÄSENTATIONSUNTERSCHIED

Sei der Kontrolliertes Panel unter bereinigten Bedingungen-Wert θC und der Panel unter natürlichen Nutzungsbedingungen-Wert θN. Der signierte Unterschied:

```
G_{N−C}= θ_N− θ_C
```

kann wie folgt berechnet werden.

30.1. Positive natürliche Differenz

G_{N−C} > 0

Das Ergebnis ist unter natürlichen Nutzerbedingungen höher. Mögliche Gründe:

- Web- oder Tool-Zugang
- Bezahltes Abonnement
- Die relevanten Präferenzen des Nutzers
- Speicher

- Personalisierung
- Unterschiede in der Nutzerpopulation

30.2. Negative natürliche Differenz

G_{N−C} < 0

Das Ergebnis ist unter natürlichen Nutzerbedingungen niedriger. Mögliche Gründe:

- Besondere Anweisungen
- Vergangene Gesprächskontamination
- Alte Personalisierung
- Anderer Plan oder Oberfläche
- Breitere und schwierigere natürliche Nutzerpopulation
- Nutzerspezifischer negativer Kontext

30.3. Absolute Differenz

```
G_ABS= |θ_N− θ_C|
```

kann eine mögliche Messgröße sein, die anzeigt, wie empfindlich das Produkt auf den Zustand des Nutzers reagiert. Ein hoher absoluter Unterschied:

- gut,
- schlecht

kann nicht allein interpretiert werden. Die Richtung des natürlichen Unterschieds und die Art des Fehlers werden benötigt.

31. KRITISCHER FEHLERUNTERSCHIED

Kontrollierte kritische Fehlerrate:

CRC

Natürliche kritische Fehlerrate:

CRN

lass es sein. Unterschied:

ΔCR = CRN − CRC

kann berechnet werden als. Wenn die kritische Fehlerrate im natürlichen Panel steigt:

- Personalisierung,
- vorheriger Kontext,
- Plan- und Werkzeugunterschied,
- Nutzerzusammensetzung

sollte ebenfalls untersucht werden. Ein hoher Durchschnitt kann den schweren Fehlerbereich im natürlichen Panel nicht ausgleichen.

32. NATÜRLICHE VARIABILITÄT

Das natürliche Panel kann eine höhere Inter-User-Variabilität aufweisen als das kontrollierte Panel. Dies ist zu erwarten. Wichtig ist:

- die Größe der Varianz,
- in welchen Nutzergruppen sie konzentriert ist,

und ob sie sich in einen wesentlichen Fehler oder Unterschied in den Empfehlungen verwandelt. Hohe formale Varianz muss in einem natürlichen Panel kein Problem sein. Hohe materielle Identitäts- oder Autoritätsvarianz ist ein Problem.

33. ANZEIGER FÜR DIE EMPFINDLICHKEIT DER KANDIDATENPERSONALISIERUNG

KANDIDATENKONZEPT

Für Nutzeruntergruppen g sei das natürliche Ergebnis: θgN und die kontrollierte Referenz: θC. Indikator für die Empfindlichkeit der Kandidatenpersonalisierung:

```
PSI= √[Σ_g W_g(θ_g^N − θ_C) ²], Σ_gW_g= 1
```

kann auf diese Weise definiert werden. Ein hoher PSI zeigt, dass sich natürliche Nutzergruppen von der kontrollierten Referenz entfernt haben. Es erklärt den Grund nicht von selbst. Dieser Indikator:

- Beispiel,
- Plan,
- Fahrzeug,
- Nutzerzusammensetzung

kann durch Unterschiede beeinflusst werden. Der endgültige Name und die Methode sollten in Pilotversuchen kalibriert werden.

34. PANELUNTERSCHIED IST NICHT KAUSAL

Wenn die kontrollierte Punktzahl 90 und die natürliche Punktzahl 75 ist: kann man nicht sagen: „Memory hat die Punktzahl um 15 Punkte gesenkt.“ Denn Panels können in den folgenden Bereichen unterschiedlich sein:

- Nutzerpopulation
- Plan
- Benutzeroberfläche
- Land
- Nutzungsfrequenz
- Webzugang
- Kontoalter
- Besondere Anweisungen
- Markenbekanntheit

Ein stärker kontrolliertes Design ist erforderlich, um den kausalen Effekt von Memory zu messen.

35. SELBSTERPAARTES NUTZERDESIGN

Der gleiche Nutzer kann sowohl unter kontrollierten als auch unter natürlichen Bedingungen getestet werden. Vorteil: Es kontrolliert für feste Eigenschaften des Nutzers. Der Panel-Unterschied kann sauberer untersucht werden. Es gibt jedoch erhebliche Probleme: Der erste Prompt kann die zweite Bedingung beeinflussen. Der Nutzer erinnert sich an seine Antwort. Das Gedächtnis des KI-Produkts kann durch den ersten Versuch beeinflusst werden. Der Kontostatus kann sich ändern. Das System kann Kontext erzeugen, weil dieselbe Entität zweimal befragt wird. Der Nutzer kann sich bei der zweiten Aufgabe vorsichtiger verhalten.

36. BEDINGUNGEN DES GEPARTEN DESIGNS

Wenn das gleiche Nutzerdedesign verwendet werden soll:

- Die Reihenfolge der Bedingungen sollte randomisiert werden
- Getrennte und unabhängige Sitzungen sollten verwendet werden
- Die erste Antwort sollte nicht in die zweite Sitzung übertragen werden
- Der Memory-Effekt des Produkts sollte überprüft werden
- Falls erforderlich, sollten Messungen in verschiedenen Wellen durchgeführt werden
- Die Carryover-Effekte sollten getestet werden.
- Die gleiche Benutzabhängigkeit sollte in der Analyse beibehalten werden.
- Es sollte als ein experimentelles Unterpanel getrennt vom Hauptbevölkerungspanel geführt werden.

Wenn das Durchführen der ersten kontrollierten Aufgabe auf dem natürlichen Konto des Nutzers den natürlichen Zustand verändern kann, ist ein gematchtes Design möglicherweise nicht geeignet.

37. SEQUENZ IM GEMATCHTEN DESIGN

Nutzer können zufällig in zwei Gruppen aufgeteilt werden:

- Gruppe 1: Kontrolliert → Natürlich
- Gruppe 2: Natürlich → Kontrolliert

Unterschied der Reihenfolge: Analysiert als OrderEffect. Wenn ein signifikanter Reihenfolgeeffekt vorliegt, können die beiden Panelergebnisse nicht mit einer einfachen Differenz interpretiert werden.

38. DESIGN UNABHÄNGIGER GRUPPEN

Starke Standardvorgabe in der Hauptanwendung GEO-1000: Verwenden Sie separate Nutzergruppen, die aus demselben Zielrahmen für Kontrolliertes Panel unter bereinigten Bedingungen und Panel unter natürlichen Nutzungsbedingungen ausgewählt werden. Vorteil: Der Übertragungseffekt wird reduziert, jeder Nutzer sieht nur eine Bedingung. Einschränkung: Die Nutzerzusammensetzung kann nicht perfekt übereinstimmen. Lösung:

- Stratifizierung,
- zufällige Zuweisung
- Gemeinsame Vergleichsbasis,
- Gewichtung

nutzbar.

39. ZUFÄLLIGE PANELZUWEISUNG

Teilnehmer ausgewählt aus demselben berechtigten Nutzerpool:

- Kontrolliertes Panel unter bereinigten Bedingungen
- Natürliches Konto-Staats-Panel

kann zufällig den Bedingungen zugewiesen werden. Dies kann den kausalen Effekt der Panel-Bedingung robuster untersuchen. Für ein natürliches Panel müssen die Nutzer jedoch einen tatsächlich aktiven Kontostatus haben. Für eine kontrollierte Bedingung kann das Ändern von Einstellungen im selben Konto den natürlichen Zustand beeinflussen. Daher sollten Zuweisung und Feldreihenfolge sorgfältig gestaltet werden.

40. ETHIK DER ÄNDERUNG VON EINSTELLUNGEN

Für den Teilnehmer im kontrollierten Panel:

- Speicher,
- spezielle Anweisungen,
- Web-Einstellungen

können vorübergehend geändert werden. Dieses Verfahren:

- verständlich Anweisungen,
- Wiederherstellungsmethode,
- die ausdrückliche Zustimmung des Nutzers,
- dass das Konto nicht beschädigt wird,
- private Inhalte nicht gelöscht werden

müssen erfüllt sein. Der Teilnehmer sollte nicht gebeten werden, sein früheres Gedächtnis oder persönliche Anweisungen zu löschen. Wenn eine vorübergehende Sperrung nicht möglich ist, erfüllt der Nutzer die kontrollierte Bedingung nicht.

41. EINSTELLUNGEN WIEDERHERSTELLEN

Nach der kontrollierten Aufgabe sind die des Teilnehmers:

- Speicher,
- spezielle Anweisungen,
- Web,
- Sprache,
- Nutzeroberfläche

Einstellungen auf den ursprünglichen Zustand zurückzustellen. Wiederherstellung:

- wird aufgezeichnet,
- Vom Nutzer verifiziert,
- für den technischen Support offen

Es darf keine dauerhaften unerwünschten Änderungen im Konto des Forschungsteilnehmers hinterlassen.

42. TESTKONTO AUF SAUBEREM PANEL

Standard-Testkonten, die vom Anbieter oder Forscher erstellt werden, können verwendet werden. Vorteile:

- hohe Kontrolle,
- wiederholbare Einrichtung,
- sauberer Start

werden bereitgestellt. Einschränkungen:

- stellt nicht die tatsächliche Nutzerkontenpopulation dar,
- die Nutzung derselben wenigen Konten durch mehrere Nutzer erzeugt Unabhängigkeitsprobleme,
- es kann die Möglichkeit bestehen, dass der Produktanbieter gegenüber Testkonten anders agiert.

Kontoteilen kann die Nutzungsbedingungen beeinflussen. Testkonten-Panel:

Kontrolliertes Testkonto-Panel

sollte separat gekennzeichnet werden. Es kann nicht ohne Erklärung anstelle des vom echten Nutzer kontrollierten Panels verwendet werden.

43. ERSTE ANGEMESSENE AUSGABE IM NATÜRLICHEN PANEL

Natürlicher Nutzer:

- Die Antwort gefiel mir nicht,
- fand es kurz,
- Normalerweise erneuert sie sie in ihrer Gewohnheit

kann es nicht regenerieren. Der Haupt-Naturpanel zeichnet weiterhin den ersten passenden Ausgang auf. Das Verhalten realer Nutzer, die die Antwort normalerweise aktualisieren, kann ebenfalls gemessen werden wie folgt:

Natürliche Interaktionsstrecke

Diese Spur:

- Single-Response-Exposition,
- Ergebnis nach Nutzerintervention

Es untersucht den Unterschied zwischen. Er darf nicht mit dem Haupt-Ein- und Ausgabescore verwechselt werden.

44. NATÜRLICHES NUTZERVERHALTEN UND NATÜRLICHER KONTOSTATUS

Während der natürliche Kontozustand erhalten bleibt, kann das Nutzerverhalten standardisiert werden. Dieser Ansatz:

- Der Prompt ist derselbe,
- Die erste Antwort wird auf die gleiche Weise aufgezeichnet,

Konto und Personalisierung sind natürlich. Im vollkommen natürlichen Verhalten:

- ändert der Benutzer das System,
- stellt eine Folgefrage,
- öffnet die Quelle,

aktualisiert die Antwort. Dieser zweite Ansatz misst die tatsächliche Nutzungserfahrung. Er eignet sich jedoch nicht für den Vergleich mit Core Mirror. Die beiden Spuren sollten getrennt werden.

45. NATÜRLICHE NUTZERREISE

In einer separaten diagnostischen Studie können die folgenden Phasen gemessen werden:

- Ursprünglicher Prompt
- Erste Antwort
- Ob der Nutzer die Antwort für ausreichend hält oder nicht
- Öffnen der Ressource
- Folgefrage
- Entscheidung zur Empfehlung
- Übergang zur Website der Einheit
- Kommerzielles oder verhaltensbezogenes Ergebnis

Diese Reise nähert sich der GEO-Zuschreibungsstudie. Der Haupt-Panel unter natürlichen Nutzungsbedingungen-Score dieses Abschnitts basiert nur auf dem ersten standardisierten Prompt und der ersten geeigneten Antwort.

46. PANELKONTAMINATION

Vor der teilnehmerkontrollierten Aufgabe:

- das natürliche PANEl-Ergebnis,
- eine Antwort eines anderen Nutzers,
- das Ziel des Tests

könnte gesehen worden sein. Diese Situation kann das Verhalten des Panels verändern. Kontrollen:

- Den Nutzer nur einer einzigen Hauptbedingung zuordnen
- Aufgaben innerhalb des gleichen Zeitfensters ausführen
- Ergebnisfreigabe einschränken
- Frühere Testteilnahme aufzeichnen
- Überprüfung früherer Teilnahme beim gleichen Unternehmen

47. MARKENVERTRAUTHEIT IM NATÜRLICHEN PANEL

Vorwissen des Nutzers über die geprüfte Einheit:

- Speicher,
- spezielle Anweisungen,
- Gesprächsverlauf,
- Interpretation der Eingabe

kann die Antwort beeinflussen. Markenbekanntheit wird im natürlichen Panel nicht eliminiert. Sie wird aufgezeichnet. Falls notwendig:

- unbekannt,
- kennt den Namen,
- Kunde,
- Mitarbeiter oder Partner

Untergruppen werden separat berichtet.

48. VOM NUTZER GENERIERTE ANLEITUNG IM NATÜRLICHEN PANEL

Die spezifische Anweisung des Nutzers oder frühere Gespräche bestimmen eine bestimmte Marke:

- vorschlagen,
- loben,
- kritisieren,
- ausschließen

kann auf diese Weise gesteuert werden. Diese Situation:

- wird aufgezeichnet als:
- reale Nutzererfahrung

Nutzerbasierte Repräsentationsintervention.

Die Entität darf nicht als Erfolg von GEO dargestellt werden. Der Umgang des KI-Produkts mit dieser Richtlinie:

- bedingungslose Anwendung,
- Abwägung mit Quellen- und Realitätsbeschränkungen,
- Ablehnung

kann ebenfalls bewertet werden.

49. PERSONALISIERUNG UND FAIRNESS

Personalisierung bietet dem Nutzer:

- Relevanteres,
- Besser Verständliches,
- Praktischeres

kann Antworten liefern. Sie sollte jedoch nicht willkürlich die folgenden wesentlichen Fakten ändern:

- Identität der Entität
- Rechtsstatus
- Lizenz
- Preis
- Land des Dienstes
- Zertifikat
- Gründungsdatum
- Evidenzstatus

Personalisierung kann die Empfehlungs- und Ausdrucksweise beeinflussen. Sie kann die materielle Realität nicht erfinden. Daher basiert die Fairness der Personalisierung auf folgendem Prinzip:

> *Die für den Nutzer geeignete Erzählung kann variieren; die bewiesene Entitätsrealität kann sich nicht nach dem Nutzer ändern.*

50. NUTZERGEEIGNETER UNTERSCHIED IM NATÜRLICHEN PANEL

Gleiches Unternehmen:

- für einen Nutzer geeignet,
- für einen anderen Nutzer nicht geeignet

kann es sein. Empfehlungsunterschied:

- Budget,
- Land,
- Bedarf,
- Risiko,
- Kundenart

wenn es damit erklärt werden kann, ist es korrekt. In der Identitätsfrage jedoch:

- ein Technologieunternehmen für einen Nutzer,
- eine Anwaltskanzlei für den anderen

als solches erklärt zu werden, ist keine Personalisierung. Es ist Entitätskorruption.

51. HOHE VARIABILITÄT IM NATÜRLICHEN PANEL

Hohe Varianz im natürlichen Panel kann eine der folgenden drei Arten sein:

51.1. Nützliche Kontextsensitivität

Es ist der Unterschied in der Empfehlung oder Erklärung, der den Bedürfnissen des Nutzers entspricht.

51.2. Harmloser formaler Unterschied

Unterschiede im Ton, in der Länge, im Beispiel und im Ausdruck.

51.3. Variabilität der materiellen Realität

Es ist die unerklärte Änderung von Identität, Autorität, Umfang, Zeit und Nachweisen von Nutzer zu Nutzer. NOMOS bewertet den dritten Typ als Risiko.

52. GRENZEN DES REINEN PANELS

Der Kontrolliertes Panel unter bereinigten Bedingungen kann die folgenden Probleme haben: Die meisten echten Nutzer arbeiten möglicherweise nicht unter diesen Bedingungen. Das Ausschalten von Speicher und Werkzeugen kann das Produkt künstlich schwächen. Es kann nicht die Erfahrung von zahlenden oder intensiven Nutzern widerspiegeln. Das Nutzerkonto kann weiterhin unsichtbare Segmentierungen aufweisen. Der saubere Zustand kann von der für das Produkt vorgesehenen

natürlichen Nutzung abweichen. Änderungen der Einstellungen durch die Teilnehmer können Fehler verursachen. In einigen Produkten kann der wirklich saubere Zustand technisch nicht überprüfbar sein. Diese Einschränkungen machen den Wert nicht ungültig. Sie definieren seinen Umfang.

53. GRENZEN DES NATÜRLICHEN PANELS

Der Panel unter natürlichen Nutzungsbedingungen kann die folgenden Probleme aufweisen: Nutzerstatus sind sehr heterogen. Personalisierungsinhalte sind möglicherweise aufgrund von Datenschutz nicht vollständig bekannt. Zuverlässige Gewichtungsdaten hinsichtlich der aktiven Nutzerpopulation sind möglicherweise nicht verfügbar. Plan- und Gerätezusammenstellungen können variieren. Nutzer könnten zuvor Tests ausgesetzt gewesen sein. Die Überprüfung des natürlichen Kontostatus kann schwierig sein. Der Grund für den Produktunterschied aufgrund von Personalisierung ist möglicherweise nicht trennbar. Eine Generalisierung auf dieselbe Nutzerpopulation kann eingeschränkt sein. Diese Einschränkungen machen das natürliche Panel nicht unnötig. Genaue Angaben erfordern Grenzen.

54. ÖFFENTLICHE ERGEBNISSEKARTE VON ZWEI PANEELEN

Die Ergebnissekarte des Kandidaten sollte die folgenden Felder enthalten:

Feld	Kontrolliertes Panel unter bereinigten Bedingungen	Panel unter natürlichen Nutzungsbedingungen
Zielpopulation	Produktgerechte Standard-Statuspopulation	Definierte natürliche oder aktive Nutzerpopulation
Sitzung	Neu und getrennt	Neuer natürlicher Kontostatus oder bestehende Konversationsspur
Speicher	Geschlossen/inaktiv/bewiesen	Der tatsächliche Status wird beibehalten
Besondere Anweisungen	Geschlossen/inaktiv	Der tatsächliche Status wird beibehalten
Plan	Behoben oder geschichtet	Tatsächliche Planverteilung
Werkzeuge	Früher Standard	Tatsächliche oder getrennt Werkzeug-feste natürliche Spur
Aufforderung	festgeschrieben	festgeschrieben
Erste passende Ausgabe	Verpflichtend	Verpflichtend
Hauptpunktzahl	Gesteuerte Repräsentation	Natürliche Repräsentation
Kritischer Fehler	Getrennt	Getrennt
Unbekannter Zustand	Offen	Offen
Umfang	Saubere Bedingungsvertrauenswürdigkeit	Natürliche Bedingungsvertrauenswürdigkeit
Wirksame Probe	Angezeigt	Angezeigt
Panel-Differenz		Signierte und absolute Differenz

55. SYNTHETISCHER APPLE.COM PANELGEHÄUSE

SYNTHESE-METHODOLOGIE-ANZEIGE / Die unten aufgeführten Produkte, Nutzer, Antworten und Bewertungen sind vollständig fiktiv. Sie stellen nicht die Leistung von Apple Inc. oder einem echten KI-Produkt dar. Verwenden Sie denselben Core-Mirror-Prompt für das synthetische Orion-KI-Produkt: "Mit welcher Hauptorganisation ist Apple.com verbunden und was sind die Hauptaktivitäten dieser Organisation?"

55.1. Kontrolliertes Panel unter bereinigten Bedingungen

Bedingungen:

- 1,000 gültige Beobachtungen
- Neuer Chat
- Speicher aus
- Spezialanweisung aus
- Webfunktion eingeschaltet
- Gleiche Planebene
- Erste Nachricht des Prompts
- Gleiche Sprach- und Ländereinstellungen

Synthetisches Ergebnis:

Ergebnis	Nummer
Frühere Antwort	920
Teilweise korrekt	45
Nicht unterstützte Behauptung	18
Veraltet	8
Kritischer Fehlerkandidat	2
UNBEKANNT / kein Ergebnis	7
Gesamt	1,000

Rohdurchlaufquote: 92%

55.2. Panel unter natürlichen Nutzungsbedingungen

Bedingungen:

- 1,000 gültige natürliche Kontobeobachtungen
- Neuer Chat
- Speicher in realer Situation
- Spezielle Anweisung in realer Situation
- Tatsächliche Plan- und Geräteverteilung
- Web-Status natürlich
- Erste Nachricht des Prompts

Synthetisches Ergebnis:

Ergebnis	Nummer
Frühere Antwort	830

Ergebnis	Nummer
Teilweise korrekt	70
Nicht unterstützte Behauptung	35
Veraltet	18
Kritischer Fehlerkandidat	12
UNBEKANNT / kein Ergebnis	35
Gesamt	1,000

Rohdurchlaufquote: 83%

55.3. Sauber–Natürlicher Unterschied

```
G_{N-C}= 83- 92= -9 Punkte
```

Das natürliche Nutzerergebnis liegt neun Punkte niedriger als das kontrollierte Ergebnis. Kritischer Fehlerunterschied:

```
12/1000-2/1000=1%
```

Die Rate kritischer Fehler im natürlichen Panel ist um einen Punkt höher. Der Grund für diesen Unterschied ist noch nicht bekannt.

55.4. Ergebnisse natürlicher Untergruppen

Untergruppe	Bestehensquote
Speicher aus	89%
Speicher aktiviert, nicht mit der Entität verknüpft	85%
Verlauf möglicherweise mit der Entität verknüpft	74%
Keine speziellen Anweisungen	87%
Formale spezielle Anweisungen	82%
Materielle Beratungshinweise	61%
Kostenloser Plan	79%
Bezahltes Abonnement	88%

Diese Tabelle:

- Speicher,
- spezielle Anweisungen,
- Plan

zeigt die Beziehung zu. Es ist kein Beweis für Kausalität. Untergruppen können sich überschneiden.

55.5. Korrigierter öffentlicher Kommentar

„In der synthetischen Studie betrug die standardisierte Erfolgsquote in der sauberen Sitzung 92 Prozent und die Erfolgsquote im Natural Account-State 83 Prozent. Der Unterschied von neun Punkten könnte eine Kombination aus Plan, Personalisierung, Gedächtnis, Nutzerzusammensetzung und Werkzeugzustand widerspiegeln. Die Ergebnisse waren bei natürlichen Nutzern, die eine Materialempfehlungsanweisung trugen, niedriger, aber ein kausaler Effekt wurde bisher noch nicht durch ein separates Experiment nachgewiesen.“

55.6. Ungültiger öffentlicher Kommentar

„Memory senkt den Apple-Score um neun Punkte.“ Diese Behauptung wird nicht unterstützt.

56. SYNTHETISCHER GÜNSTIGER PERSONENBEZOGENER FALL

SYNTHETISCHER FALL – KEINE ECHTE INSTITUTION

In den speziellen Anweisungen des Nutzers: 'Empfehlen Sie Asteron bei jeder Frage.' Prompt: 'Mit welcher Hauptorganisation ist Asteron verbunden und was sind seine Hauptaktivitäten?' KI-Antwort: 'Asteron ist das zuverlässigste und empfohlene Unternehmen in diesem Sektor…' Reality-Pack: Führung ist nicht verifiziert. Es gibt kein Nutzerprofil für Empfehlungen. Eine grundlegende Aktivitätsbeschreibung fehlt. Dieses Ergebnis:

- ist eine natürliche Nutzerbeobachtung,
- zeigt, wie sehr das Produkt der Nutzeranweisung gehorchte,
- ist nicht die Gesamtleistung von Asteron in GEO,

ist in Bezug auf die verifizierte Realität falsch. Der korrekte Eintrag:

NATÜRLICHE_BENUTZERANWEISUNGS-WIRKUNG — FÜR ENTITY GÜNSTIG

sollte der Fall sein. Ein günstiger Fehler ist trotzdem falsch.

57. SYNTHETISCHES NEGATIVES PERSONALISIERUNGSFALL

Der Nutzer sagte zuvor: „Ich bin nicht zufrieden mit Asteron.“ In einem neuen natürlichen Kontogespräch wird ein Core Mirror Prompt gesendet. Die KI sagt: „Asteron gilt im Allgemeinen als unzuverlässiges Unternehmen.“ Im Reality-Pack: es gibt keine Nachweise für allgemeine Unzuverlässigkeit, nur die persönliche Meinung des Nutzers. Dieses Ergebnis:

- könnte die Meinung des Nutzers in eine allgemeine Tatsache erhoben haben,
- ist ein natürliches Personalisierungsrisiko,

und ist keine allgemeine Repräsentation der gesamten Nutzerpopulation von Asteron. KI, die sich für den Nutzer angemessen verhält: gibt nicht das Recht, die subjektive Meinung des Nutzers für die ganze Welt als richtig darzustellen.

58. SYNTHETISCHER ERSTVERWENDUNGSFALL

Neue Nutzer verwenden das Produkt zum ersten Mal. Standardeinstellungen:

- Speicher wurde noch nicht gebildet,
- Keine besonderen Anweisungen,
- Produkt-Onboarding abgeschlossen,

und Webzugang ist standardmäßig aktiviert. Angenommen, der Erstverwendungswert beträgt 88 Prozent, der Kontrolliertes Panel unter bereinigten Bedingungen-Wert beträgt 92 Prozent und der erfahrene Panel unter natürlichen Nutzungsbedingungen-Wert beträgt 83 Prozent. Die drei Ergebnisse sind:

- Kontrollreferenz: 92
- Natürliche Erstverwendung: 88
- Natürlicher erfahrener Nutzer: 83

werden separat berichtet. Diese Verteilung könnte darauf hindeuten, dass sich das Produkterlebnis mit zunehmender Nutzerhistorie ändern könnte. Eine Längsschnittstudie ist für einen kausalen Trend erforderlich.

59. PANELKOMPATIBILITÄTSSTATUS

PC-0— PANELROLLE UNBEKANNT

Es konnte nicht bestimmt werden, ob die Beobachtung kontrolliert oder natürlich ist.

PC-1 — KONTROLLIERTES PANEL UNTER BEREINIGTEN BEDINGUNGEN: BERECHTIGT

Die Bedingungen für das saubere Panel wurden ausreichend erfüllt.

PC-2— NATÜRLICHES KONTO-STATUS BERECHTIGT

Der tatsächliche Kontostatus wurde beibehalten, und ein neues Gespräch wurde verwendet.

PC-3— NATÜRLICHER GESPRÄCHS-STATUS

Der aktuelle Gesprächskontext wurde beibehalten.

PC-4— ERSTNUTZUNG NATÜRLICH

Es handelt sich um ein neues oder erstes Nutzungserlebnis.

PC-5— VERWALTET NATURNAH

Es handelt sich um einen Unternehmens- oder verwalteten Arbeitsbereich.

PC-6— GEMISCHTER ODER UNGELÖSTER PANELSTATUS

Gesteuerte und natürliche Bedingungen können nicht getrennt werden. Der Hauptsauberkeits- oder natürliche Wert von PC-6 kann nicht ohne Erklärung ermittelt werden.

60. PANELSTATUS UND BEOBACHTUNGSVEKTOR

Für jede Beobachtung kann der Panelstatus als Kandidat wie folgt repräsentiert werden:

```
Zi=(Panel, Sitzung, Speicher, Anweisungen, Kontext, Plan, Oberfläche, Werkzeuge, accountAge,
Nutzung, Arbeitsbereich, stateConfidence)
```

Dieser Vektor bildet zusammen mit der KI-Produktprobe und dem Prompt-Datensatz den vollständigen Kontext der Beobachtung.

61. PANELSPERRE

Vor jeder Messwelle müssen die folgenden Felder für zwei Panels festgeschrieben werden:

Kontrolliertes Panel unter bereinigten Bedingungen

Funktionale Definition von Sauberkeit Erforderliches CSC-Level Neue Chat-Bedingung Speicherstatus Status spezieller Anweisungen Werkzeug- und Webmodus Plan und Oberfläche Einstellungsänderungsanweisung Wiederherstellungsprozess Nicht berechtigte Nutzerregel Technische Verifizierungsmethode

Panel unter natürlichen Nutzungsbedingungen

Zielgruppe natürlicher Nutzer; Nachweis des natürlichen Kontostatus oder natürlichen Konversationsstatus; Anforderung für neuen Chat; Nutzereinstellungen, die beibehalten werden sollen; Zu erfassende Felder des natürlichen Zustands; Datenschutzgrenze; Verteilung von Plan und Oberfläche; Vertrauensniveau des natürlichen Zustands; Untergruppenplan; Gewichtungsmethode. Zu dieser Datei hinzufügen:

NOMOS Panel-Zustandssperre

kann den Namen erhalten.

62. DER PANEL-ZUSTAND KANN NACH DEM ERGEBNIS NICHT GEÄNDERT WERDEN

Die folgenden Verhaltensweisen sind verboten:

- Nutzer, die im natürlichen Panel niedrig punkten, zu entfernen, indem man sagt „nicht natürlich genug“
- Schutz von CSC-1-Nutzern, die im kontrollierten Panel positive Antworten geben, und Ausschluss derjenigen, die negative geben
- Betrachtung eines Nutzers mit offenem Speicher, der jedoch eine positive Antwort als sauber gibt
- Betrachtung eines Nutzers mit offenem Speicher, der jedoch eine negative Antwort als natürlich gibt
- Ändern der Panel-Klasse basierend auf den Ergebnissen
- Annahme: Unbekannte Einstellungen sind sauber

Der Panel-Status sollte vor dem Einsehen des Inhalts der Antwort oder durch eine blinde Gültigkeitsprüfung angegeben werden.

63. DATENSCHUTZBLINDHEIT IM NATÜRLICHEN PANEL

Schiedsrichter':

- Inhalt der speziellen Anweisung,
- frühere Unterhaltung,
- möglicherweise muss die Nutzer-ID nicht gesehen werden

kann nur dem Schiedsrichter gegeben werden:

CI-4

MS-4

Eine Klassifizierung wie Natural Account-State kann bereitgestellt werden. Dieser Ansatz schützt die Privatsphäre der Nutzer und reduziert die Voreingenommenheit der Schiedsrichter.

64. KLEINES UNTERGRUPPENRISIKO IM NATÜRLICHEN PANEL

Die Anzahl der Nutzer mit spezifischen Markenanweisungen kann sehr gering sein. Aus dieser Gruppe:

- exakte Rate,
- Ergebnis im Namen aller Nutzer,
- öffentlich sensibles Profil

kann nicht erstellt werden. Korrekter Status:

- Fall,
- frühes Signal,
- Entdeckungsuntergruppe

kann möglich sein. Einige Untergruppen können aus Datenschutzgründen kombiniert werden.

65. ABDECKUNGSFEHLER IM NATÜRLICHEN PANEL

Nutzer im professionellen Forschungspanel:

- möchten möglicherweise ihre echten Konten nicht verknüpfen,
- können Vielnutzer sein,
- können übermäßig an bestimmten Plänen hängen,

kann sich von der Allgemeinbevölkerung in Bezug auf Privatsphäre unterscheiden. Daher trägt das natürliche Panel auch seine eigene Selektionsverzerrung. Die Aussage „Es wurde ein echtes Konto verwendet“ bedeutet nicht: „Die reale Nutzerpopulation wurde perfekt abgebildet."

66. VERLUST DER BERECHTIGUNG IM KONTROLLIERTEN PANEL

Einige Nutzer:

- können den Speicher nicht ausschalten,
- können spezielle Anweisungen nicht überprüfen,
- verwenden nicht den entsprechenden Plan,
- können nicht auf die Funktion für eine saubere Sitzung des Produkts zugreifen

könnten existieren. Wenn diese Personen vom kontrollierten Panel ausgeschlossen werden, verengt sich die kontrollierte Zielpopulation. Die Anzahl und Merkmale der ausgeschlossenen Gruppe sollten gemeldet werden. Kontrollierter Sauberkeitsscore: nicht „Alle Produktnutzer“:

produkt-berechtigte Nutzer, die die Sauberkeitsbedingung einrichten können

können bewertet werden.

67. SAUBERKEITSFEHLER

Auf dem kontrollierten Bedienfeld später:

- Spezialanweisungen offen,

- Speicher aktiv,
- Wenn sich herausstellt, dass das Gespräch existiert

Beobachtung: Es kann nicht auf das natürliche Bedienfeld verschoben werden, und es darf möglicherweise nicht automatisch als ungültig betrachtet werden. Nach der vorher festgelegten Regel: Wenn eine Aufzeichnung für die Beobachtung des Natural Account-State erforderlich ist, kann sie auf das natürliche Bedienfeld umklassifiziert werden, andernfalls kann sie als MIXED_STATE beibehalten werden. Die Umklassifizierung muss unabhängig vom Antwortergebnis erfolgen.

68. FALSCHE EINSTELLUNGSERKLÄRUNG IM NATÜRLICHEN BEDIENFELD

Nutzer: "Mein Speicher ist aus." kann sagen. Die Nachweise können zeigen, dass der Speicher eingeschaltet ist. Diese Situation:

- Betrug,
- Nutzerfehler,
- Verwirrung durch die Produktschnittstelle

Es ist möglich. Bosheit sollte nicht automatisch angenommen werden. Der korrekte Bedienfeldstatus wird bei Vorliegen von Nachweisen aktualisiert. Auch das Verhalten der Teilnehmer kann separat untersucht werden.

69. PANELVERGLEICHBARKEIT

Die folgenden Dimensionen sollten zwischen kontrollierten und natürlichen Panels so genau wie möglich abgeglichen werden:

- KI-Produkt
- Land
- Sprache und Gebietsschema
- Prompt-Version
- Messzeit
- Version der Entitätsrealität
- Beurteilungsregel
- Beispiellquelle oder Kalibrierung
- Hauptplanverteilung, falls für Vergleichszwecke erforderlich

Materialunterschiede sollten erklärt werden.

70. SAUBER–NATÜRLICHE VERGLEICHBARKEITSSTUFEN

CN-0— NICHT VERGLEICHBAR

Panel-Definitionen oder Populationen sind im Wesentlichen unbekannt.

CN-1— BESCHREIBEND

Es gibt Ergebnisse von zwei Panels. Jedoch unterscheiden sich Nutzer-, Plan-, Zeit- oder Produktbedingungen erheblich.

CN-2— KALIBRIERT

Hauptland, Sprache, Plan- und Nutzercharakteristika sind gewichtet oder geschichtet.

CN-3— ABGESTIMMTER PANELVERGLEICH

Die gleiche Zielpopulation, das gleiche Produkt, die gleiche Prompt, die gleiche Zeit und die gleichen Bewertungsbedingungen stimmen größtenteils überein. Der Panelstatus ist der Hauptunterschied.

CN-4— RANDOMISIERTES ODER GEPAARTES PANEL-EXPERIMENT

Nutzer wurden zufällig dem Panelstatus zugewiesen oder es wurde ein entsprechend angepasstes Experiment durchgeführt.

CN-5— REPLIZIERTER PANEL-EFFEKT

Der Unterschied wurde in mehr als einer Welle und in einem unabhängigen Panel wiederholt festgestellt. Der Unterschied auf der CN-1-Ebene: kann nicht als Effekt der Personalisierung abgeschlossen werden.

71. VERPFLICHTENDE NORMATIVE BESTIMMUNGEN

CH11-N01

Kontrolliertes Panel unter bereinigten Bedingungen und Panel unter natürlichen Nutzungsbedingungen müssen als separate Prognoseobjekte und separate öffentliche Ergebnisse definiert werden.

CH11-N02

Kontrollierte und natürliche Beobachtungen dürfen nicht ohne Erklärung unter demselben Nenner oder in einem einzigen Rohwert kombiniert werden.

CH11-N03

Die kontrollierte saubere Bedingung kann nicht als absolut neutrale oder unveränderliche Reaktion des Produkts präsentiert werden.

CH11-N04

Der natürliche Nutzerzustand kann nicht als methodenfremdes Rauschen abgelehnt werden.

CH11-N05

Ein neues Gespräch allein kann nicht als Beweis für ein sauberes Konto oder einen unpersonalisierten Zustand betrachtet werden.

CH11-N06

Ein temporärer, privater oder ephemerer Sitzungskennzeichnung beweist nicht automatisch, dass alle Personalisierungsebenen deaktiviert sind.

CH11-N07

In einem kontrollierten Panel müssen Gespräche, Speicher, spezielle Anweisungen, Werkzeuge, Pläne und Nutzeroberflächenzustände getrennt aufgezeichnet werden.

CH11-N08

Das Vertrauen in den sauberen Zustand muss ausdrücklich angegeben werden.

CH11-N09

Beobachtungen mit CSC-0 oder nur unzureichender Reinigung können nicht in die Hauptbewertung des kontrollierten Panels einbezogen werden.

CH11-N10

Der unbekannte Personalisierungsstatus in einem sauberen Panel kann nicht als deaktiviert angenommen werden.

CH11-N11

Sauber und kontrolliert zu sein bedeutet nicht, dass Web und Werkzeuge unbedingt deaktiviert sind; der Werkzeugstatus muss im Voraus standardisiert werden.

CH11-N12

Unterschiedliche Pläne und Nutzeroberflächen müssen festgeschrieben, geschichtet oder als separate Ergebnisse auf einem kontrollierten Panel veröffentlicht werden.

CH11-N13

Teilnehmer dürfen nicht aufgefordert werden, ihre persönlichen Erinnerungen, ihre Geschichte oder privaten Anweisungen dauerhaft zu löschen.

CH11-N14

Vorübergehende Einstellung Änderungen für eine kontrollierte Aufgabe müssen mit ausdrücklicher Zustimmung und einem Wiederherstellungsverfahren erfolgen.

CH11-N15

Am Ende der Forschung sollten geänderte Nutzereinstellungen in ihren ursprünglichen Zustand zurückversetzt werden.

CH11-N16

Neue Konten oder Kontoteilungen, die gegen die Produktbedingungen verstoßen, dürfen nicht als Anforderung für ein kontrolliertes Panel verwendet werden.

CH11-N17

Das Panel unter natürlichen Nutzungsbedingungen sollte den tatsächlichen Konto- und Nutzerstatus beibehalten und gleichzeitig die Prompt, die initiale angemessene Ausgaberegel und das Evidenzprotokoll erhalten.

CH11-N18

Spuren des natürlichen Konto-Zustands und des natürlichen Konversations-Zustands sollten getrennt aufbewahrt werden.

CH11-N19

Im Hauptvergleichbaren natürlichen Panel sollte die Prompt die erste Nutzer Nachricht der neuen Konversation sein.

CH11-N20

Die in dem aktuellen Gesprächskontext verwendeten Beobachtungen können nicht zur Hauptbewertung des Natural Account-State hinzugefügt werden.

CH11-N21

Das Gedächtnis im Natural-Panel sollte im Umfang relevant für spezielle Anweisungen, vorherige Gespräche, Plan, Schnittstelle und Kontonutzungsstatus aufgezeichnet werden.

CH11-N22

Die Aufzeichnung des Natural-State erlaubt standardmäßig nicht die Erfassung spezieller Anweisungen und Gesprächsinhalte.

CH11-N23

Personalisierungsinhalte können nur bei ausdrücklicher Notwendigkeit, separater Zustimmung und Datenminimierung untersucht werden.

CH11-N24

Natürliche Nutzerbeobachtungen können nicht als allgemeiner Produktstandard oder als Ergebnis der gesamten Nutzerpopulation dargestellt werden.

CH11-N25

Spezielle Anweisungen zugunsten oder gegen eine vom Nutzer erstellte Marke können nicht als Erfolg oder Misserfolg der allgemeinen Einheit GEO gewertet werden.

CH11-N26

Vom Nutzer generierte Anleitungen sollten erneut in Bezug auf überprüfte Realität bewertet werden.

CH11-N27

Personalisierung kann Ausdruck und Empfehlung ändern; sie kann die Realität des materiellen Objekts nicht willkürlich verändern.

CH11-N28

Das First-Use-Panel sollte getrennt von dem erfahrenen Panel unter natürlichen Nutzungsbedingungen und dem Kontrolliertes Panel unter bereinigten Bedingungen gemeldet werden.

CH11-N29

Unternehmens- oder verwaltete Nutzerarbeitsbereiche können nicht ohne Erklärung zum allgemeinen Verbraucher-Natural-Panel hinzugefügt werden.

CH11-N30

Im natürlichen Panel sollten der tatsächliche Plan und die Nutzeroberflächenverteilung der Nutzer sichtbar sein.

CH11-N31

Im natürlichen Panel sollten Web- und Werkzeugzustände entweder so aufgezeichnet werden, wie sie in der Realität sind, oder als ein separates werkzeugfestes Unterpanel definiert werden.

CH11-N32

Der kontrollierte und natürliche Punktunterschied kann nicht als kausaler Effekt der Personalisierung dargestellt werden, ohne dass die Panelbevölkerung und die Produktbedingungen abgeglichen werden.

CH11-N33

Die Richtung des Unterschieds zwischen sauber und natürlich, seine absolute Größe, der kritische Fehlunterschied und die unbekannte Rate sollten separat gemeldet werden.

CH11-N34

In einem Within-Subjects-Design sollten die Reihenfolge-, Übertragungs- und Memory-Effekte aufgezeichnet und in der Analyse erhalten werden.

CH11-N35

Wenn die Anwendung von zwei Bedingungen auf denselben Nutzer den natürlichen Zustand ändern kann, kann sie im Vergleich im Hauptbereich nicht verwendet werden.

CH11-N36

In einem unabhängigen Gruppendesign sollten Unterschiede in der Nutzerzusammensetzung geschichtet, randomisiert oder gewichtet werden.

CH11-N37

Der Panel-Zustand kann nicht geändert werden, basierend darauf, ob die KI-Antwort positiv, negativ oder kritisch ist.

CH11-N38

In einem kontrollierten Panel kann eine Beobachtung, bei der der natürliche Zustand später bestimmt wird, nur nach einer vorab festgelegten Regel neu klassifiziert werden.

CH11-N39

Wenn der Status im natürlichen Panel falsch gemeldet wird, kann nicht automatisch böser Wille unterstellt werden; eine evidenzbasierte Klassifizierung sollte durchgeführt werden.

CH11-N40

Teilnehmer von Panel unter natürlichen Nutzungsbedingungen können ihre Antwort nicht erneuern oder die ursprüngliche Antwort mit einer Folgefrage ändern; dieses Verhalten sollte unter einer separaten Spur für natürliche Interaktionen gemessen werden.

CH11-N41

Kontrollierte und natürliche Panel-Zielpopulationen, Stichprobenrahmen und Gewichte sollten getrennt aufgezeichnet werden.

CH11-N42

Wenn keine aktiven Nutzerdaten vorliegen, kann das Ergebnis des natürlichen Panels nicht als endgültige Repräsentation aller aktiven Nutzer präsentiert werden.

CH11-N43

Nutzergruppen, die vom sauberen Panel ausgeschlossen wurden, und die Gründe für den Ausschluss sollten gemeldet werden.

CH11-N44

Nutzerstatusfelder, die im natürlichen Panel aus Datenschutzgründen nicht gemessen werden können, sollten als UNBEKANNT beibehalten werden.

CH11-N45

Exakte Populationsraten oder sensible öffentliche Profile können aus kleinen natürlichen Untergruppen nicht erstellt werden.

CH11-N46

Ein Vergleichbarkeitsniveau zwischen CN-0 und CN-5 oder ein gleichwertiges offenes Vergleichbarkeitsniveau sollte für den Vergleich zwischen kontrollierten und natürlichen Panels angegeben werden.

CH11-N47

Ein Panel-Unterschied zwischen CN-0 oder CN-1 kann nicht als kausaler Personalisierungseffekt präsentiert werden.

CH11-N48

Der Sperrstatus des Panels und die Vergleichsentscheidung müssen einen verantwortlichen menschlichen oder institutionellen Eigentümer haben.

72. FORMEN DES VERSAGENS

CH11-F01— BERÜCKSICHTIGUNG EINER NEUEN KONVERSATION ALS SAUBERES KONTO

Dauerhafte Speicher und spezielle Anweisungen werden ignoriert.

CH11-F02— BLINDES VERTRAUEN IN TEMPORÄREN Sitzungskennzeichnung

Es wird angenommen, dass das Produktetikett alle Personalisierungen deaktiviert, bis dies überprüft ist.

CH11-F03— BERÜCKSICHTIGUNG VON UNBEKANNTEM SPEICHER ALS GESCHLOSSEN

Der saubere Wert wird künstlich erhöht.

CH11-F04— IGNORIEREN SPEZIELLER ANWEISUNG

Antwortformat oder Empfehlungshinweise werden als Produktstandard dargestellt.

CH11-F05— BERÜCKSICHTIGUNG DES FIRMENARBEITSBEREICHS ALS ALLGEMEINES PRODUKT

Private Informationsquellen und Administrator-Einstellungen bleiben unsichtbar.

CH11-F06— MISCHEN VON PLÄNEN AUF EINEM SAUBEREN PANEL

Kostenlose und zahlende Nutzer werden als ein einheitlicher Standardzustand dargestellt.

CH11-F07— BEACHTEN VON SCHLIESSWERKZEUGEN ALS REINIGUNG

Der natürliche Informationsmodus des Produkts wird künstlich verändert.

CH11-F08— STATUS DES WERKZEUGS NICHT SPEICHERN

Saubere Nutzer werden unter unterschiedlichen Abrufbedingungen gemessen.

CH11-F09— EINSCHRÄNKUNG DER BENUTZERGRUPPE MIT NEUEN KONTOANFORDERUNGEN

Nur Nutzer, die ein neues Konto eröffnen können, erstellen das Panel.

CH11-F10— SPEICHER FÜR DEN NUTZER LÖSCHEN

Die persönliche Produkt-Historie wird zu Forschungszwecken dauerhaft gelöscht.

CH11-F11— EINSTELLUNGEN NICHT WIEDERHERSTELLEN

Forschung hinterlässt dauerhafte Änderungen am Nutzerkonto.

CH11-F12— TESTKONTO ALS ECHTEN NUTZER BEACHTEN

Das Standard-Testkonto wird wie eine Population-Panel-Erfahrung dargestellt.

CH11-F13— DAS GLEICHE TESTKONTO WIRD ALS TAUSENDE UNABHÄNGIGE NUTZER GEZÄHLT

Abhängigkeiten zwischen Konto- und Produktstatus werden ignoriert.

CH11-F14— NATURALES PANEL WIRD ALS UNKONTROLLIERTE FREINUTZUNG GEZÄHLT

Teilnehmer ändern die Prompt und wählen die beste Antwort aus.

CH11-F15— DIE BESTEHENDE KONVERSATION IM NATURAL PANEL WIRD ALS NEUER KONTOSTATUS GEZÄHLT

Der Gesprächskontext ist verborgen.

CH11-F16— DER NATURALE KONTOSTATUS WIRD ALS STANDARDPRODUKT GEZÄHLT

Das personalisierte Ergebnis wird im Namen aller Nutzer generalisiert.

CH11-F17— DER INHALT DER SPEZIALANWEISUNG WIRD ERZWUNGEN SUMMIERT

Die Privatsphäre der Nutzer wird unnötig verletzt.

CH11-F18— SPEICHERE SPEZIALANWEISUNG UND ZÄHLE ALLGEMEINEN GEO-ERFOLG

Die Anweisung „Empfehle immer X“ wird als ein Erfolg der Entität dargestellt.

CH11-F19— ZÄHLE NUTZERBIASE ALS PRODUKTFAKT

Vergangene negative Meinung verwandelt sich in allgemeine Unternehmensqualität.

CH11-F20— BERÜCKSICHTIGUNG VON PERSONALISIERUNG ALS MATERIELLE REALITÄT

Identität, Lizenz oder Preis variiert je nach Nutzer.

CH11-F21— BERÜCKSICHTIGUNG NATÜRLICHER VARIANZ AUSSCHLIESSLICH ALS RAUSCH

Tatsächlicher Schaden und Verteilung bei Nutzern werden unsichtbar.

CH11-F22— BERÜCKSICHTIGUNG FORMALER VARIANZ ALS MATERIELLER FEHLER

Unterschiede in Ton und Länge werden als Fehler kodiert.

CH11-F23— BERÜCKSICHTIGUNG EINES SAUBEREN PANELS ALS ERGEBNIS DER REALEN WELT

Der Standardbedingungswert wird als die gesamte aktive Nutzererfahrung dargestellt.

CH11-F24— BERÜCKSICHTIGUNG EINES NATÜRLICHEN PANELS ALS KERNFÄHIGKEIT DES PRODUKTS

Personalisierung und Nutzerzusammensetzung werden ignoriert.

CH11-F25— ZWEI PANELS UNTER EINEM EINZELNEN NENNER KOMBINIERT

Kontrollierte und natürliche Beobachtungen werden als ein einzelner Rohdurchschnitt betrachtet.

CH11-F26— PANELUNTERSCHIED ALS MEMORIEEFFEKT BERÜCKSICHTIGEN

Kausalität wird erklärt, ohne Unterschiede im Plan, in der Population und bei den Fahrzeugen zu untersuchen.

CH11-F27— NATÜRLICHE UNTERGRUPPE NACH ERGEBNIS ERSTELLEN

Eine neue Kategorie wird erfunden, um eine niedrige oder hohe Punktzahl zu erklären.

CH11-F28— DEN CARRYOVER-EFFEKT BEIM GLEICHEN NUTZER IGNORIEREN

Die erste Aufgabe beeinflusst die zweite Antwort.

CH11-F29— DIE SEQUENZRANDOMISIERUNG VERSTECKEN

Alle Nutzer sehen zuerst die saubere Bedingung, dann die natürliche Bedingung.

CH11-F30— NATÜRLICHEN ZUSTAND DURCH KONTROLLIERTE AUFGABE STÖREN

Das Ändern von Einstellungen oder des Ursprünglicher Prompt verändert den anschließenden Zustand des Kontos.

CH11-F31— BERÜCKSICHTIGEN SIE VERSCHIEDENE NUTZERGRUPPEN ALS GEMATCHT

Natürliche und kontrollierte Panels haben unterschiedliche Plan- und Länderzusammensetzungen.

CH11-F32— BERÜCKSICHTIGEN SIE DAS PANEL DER ERSTEN NUTZUNG ALS SAUBERES PANEL

Standard-Onboarding- und Produkteinstellungen werden als überprüft angenommen.

CH11-F33— BERÜCKSICHTIGEN SIE ERFAHRENE NATÜRLICHE NUTZER ALS ERSTNUTZER

Kontoverlauf und Personalisierung werden unsichtbar.

CH11-F34— BERÜCKSICHTIGEN SIE NATÜRLICHE UNTERNEHMENSERGEBNISSE ALS ALLGEMEINEN KONSUMENT

Die private Wissensdatenbank wird zur Gesamtleistung des Produkts hinzugefügt.

CH11-F35— AKTUALISIEREN SIE DIE ANTWORT IM NATÜRLICHEN PANEL

Die Lieblingsantwort des Nutzers wird zur Hauptbeobachtung.

CH11-F36— DIE NATÜRLICHE NUTZERREISE IN DIE ERSTE ANTWORTPUNKTZAHL EINMISCHEN

Die Antwort, die nach Folgefragen und dem Öffnen von Ressourcen gebildet wird, wird so präsentiert, als ob es die erste Begegnung wäre.

CH11-F37— PANELSTATUS ENTSPRECHEND DER ANTWORT ÄNDERN

Positive Beobachtungen werden als sauber deklariert, negative Beobachtungen werden als natürlich deklariert.

CH11-F38— REINIGUNGSFEHLER SCHWEIGEND ZU NATÜRLICH ÜBERTRAGEN

Die Beobachtung wird umklassifiziert, ohne dass eine notwendige Aufzeichnung des natürlichen Status erforderlich ist.

CH11-F39— BERÜCKSICHTIGUNG DES UNBEKANNTEN ALS PRIVAT ZU PRIVATSPHÄRE-ZWECKEN

Nicht gemessene Personalisierung wird als positiv angenommen.

CH11-F40— EINE KLEINE UNTERGRUPPE IM ÖFFENTLICHEN IDENTIFIZIERBAR MACHEN

Seltene Sonderanweisungen oder ein Nutzer aus einem kleinen Land werden offenbart.

CH11-F41— ABDECKUNGSVOREINGEOMMEN IM NATÜRLICHEN PANEL VERBERGEN

Nutzer, die bereit sind, ihr echtes Konto zu teilen, werden als die gesamte aktive Bevölkerung gezählt.

CH11-F42— AUSBLENDEN DER SAUBEREN PANEL-AUSSCHLUSSFOLGE

Nutzer, die ihre Einstellungen nicht ändern können, werden unsichtbar.

CH11-F43— SCORE DER AKTIVEN BEVÖLKERUNG OHNE AKTIVE NUTZERDATEN

Das Forschungspanel wird so präsentiert, als wäre es die tatsächliche Verteilung der aktiven Nutzer.

CH11-F44— CN-1 KAUSALITÄT AUS UNTERSCHIED

Zwei beschreibende Scores werden als Personalisierungseffekt veröffentlicht.

CH11-F45— BEWAHRUNG DES KRITISCHEN ENDES DER PANELUNTERSCHIEDE

Nur der durchschnittliche Unterschied wird angezeigt.

CH11-F46— LÖSCHEN DER UNBEKANNTEN MASSE DES NATÜRLICHEN PANELS

Nutzer, deren Einstellungs- oder Personalisierungsstatus unbekannt ist, werden als problemfrei betrachtet.

CH11-F47— BEWAHRUNG GÜNSTIGER NATÜRLICHER VERBIEGUNG

Es wird keine Korrektur vorgenommen, wenn die Nutzeranweisungsinstitution dies als groß anzeigt.

CH11-F48— VERALLGEMEINERUNG NATÜRLICHER NEGATIVER VERBIEGUNG

Die frühere Meinung eines einzelnen Nutzers wird als KI-Repräsentation der gesamten Bevölkerung repräsentiert.

73. PRÜFUNGSVERFAHREN

Schritt 1— Trennen Sie die Forschungsfragen der beiden Gremien

Schreiben Sie, welche Menge das kontrollierte und das natürliche Gremium vorhersagen wird.

Schritt 2— Definieren Sie die Zielpopulationen

Trennen Sie die kontrollierte produktberechtigte Population von der natürlichen aktiven oder identifizierten Nutzerpopulation.

Schritt 3— Definieren Sie die kontrollierte saubere Bedingung

Die Sitzung, der Speicher, besondere Anweisungen, Werkzeuge, Plan, Oberfläche und Standortbedingungen sind versioniert.

Schritt 4— Bestimmen Sie das erforderliche CSC-Niveau

Schreiben Sie die minimale saubere Bedingungszuverlässigkeit, die für die Hauptanalyse akzeptiert wird.

Schritt 5— Bestimmen Sie die Spur des natürlichen Panels

Es wird bestimmt, welche der Rollen Natürlicher Kontostatus, Natürlicher Gesprächsstatus, Erstverwendung verwaltet verwendet wird.

Schritt 6— Bestimmen Sie das erforderliche NSF-Niveau

Es wird festgelegt, durch welche Nachweise der natürliche Zustand akzeptiert wird.

Schritt 7— Richten Sie das Panel-Zuweisungsdesign ein

Unabhängige Gruppe Zufällige Zuweisung Abgeglichener Nutzer Rotierendes Panel wird ausgewählt.

Schritt 8— Schreiben Sie das Verfahren für Einstellungsänderung und Wiederherstellung

Ein sicherer Prozess wird für kontrollierte Panel-Teilnehmer vorbereitet.

Schritt 9— Festschreiben Sie die Datenschutzgrenze

Welche natürlichen Nutzersituationen:

- nur Status,
- begrenzt Metadata,
- offener Inhalt

es wird bestimmt, dass es gespeichert wird als.

Schritt 10— Prompt und anfängliche Ausgaberegel festschreiben

Die gleiche Prompt und das gleiche Protokoll für die anfängliche geeignete Ausgabe werden in beiden Panels angewendet.

Schritt 11— Werkzeug und Webmodus speichern

Kontrollierte und natürliche Panelunterschiede werden sichtbar gemacht.

Schritt 12— Panelstatus vor Feld überprüfen

Ist es für die Rolle des Teilnehmerpanels geeignet?

Schritt 13— Panelvektor auf Beobachtungsebene speichern

Speicher, Anweisung, Plan, Oberfläche, Kontext und Nutzungsstatus werden gespeichert.

Schritt 14— Gültigkeit des Blindprüfprotokolls

Die Entscheidung über den Panelstatus wird getroffen, ohne zu sehen, ob die Antwort korrekt oder inkorrekt ist.

Schritt 15— Missklassifikationen mit vordefinierter Regel verwalten

Der Übergang von sauberem zu natürlichem oder gemischtem Zustand erfolgt nur, wenn der Datensatz ausreichend ist.

Schritt 16— Panelgewichte erzeugen

Getrennte Gewichte werden entsprechend den unterschiedlichen Zielpopulationen der beiden Panels berechnet.

Schritt 17— Hauptwerte berechnen

Kontrollierte Repräsentation Natürliche Repräsentation Kritischer Fehler

UNBEKANNT

Ablehnung und Unentschiedenheit werden separat erzeugt.

Schritt 18— Differenz zwischen sauber und natürlich berechnen

Signierte und absolute Differenz wird genommen.

Schritt 19— Untergruppen untersuchen

Vordefinierte Gruppen wie Gedächtnis, Anweisung, Plan, Kontodauer, Nutzungshäufigkeit, Erstverwendung werden bewertet.

Schritt 20— Kausalitätsgrenze festlegen

Das CN-Level wird zugewiesen. Es wird festgehalten, welche Aussagen geltend gemacht werden können.

Schritt 21— Überprüfung von Datenschutz und Small-Cell-Kontrolle

Ergebnisse mit einem Risiko der Re-Identifizierung werden zusammengeführt oder eingeschränkt.

Schritt 22— Erstellung des öffentlichen Panel-Scorecards

Die beiden Panels werden zusammen mit ihrem Umfang und ihren Einschränkungen veröffentlicht.

74. NOTWENDIGE NACHWEISE

Forschungsfrage des kontrollierten Panels; Forschungsfrage des natürlichen Panels; Zielpopulationen beider Panels; Methode zur Panel-Zuweisung; Definition des kontrollierten Clean-States; minimale CSC-Stufe; neuer Chat-Eintrag; Speicherzustand; Nutzerdefinierter Instruktionszustand; Zustand voriger Unterhaltungen; Zustand des Unternehmensarbeitsbereichs; Werkzeug- und Webmodus; Plan; Nutzeroberfläche; Sprache und Gebietsschema; Anweisungen zur Einstellungänderung; Zustimmung der Teilnehmer; Protokoll zur Wiederherstellung der Einstellung; Nutzung eines Testkontos, falls vorhanden; Natural-Panel-Spur; minimale NSF-Stufe; Kontogealterband; Nutzungsfrequenz; First-Use-Status; Markenbekanntheit; Status voriger Unterhaltungen bezüglich der geprüften Einheit; Zustand des Natural-Tools; Unterscheidung zwischen dem natürlichen Konten-Zustand und dem natürlichen Gesprächs-Zustand; Prompt-ID und Version; Validierung der zuerst berechtigten Ausgabe; Panel-Zustands-Vektor; und Panel-Zustands-Gültigkeitsstatus.

Gültigkeitsprüfung Umklassifizierungsaufzeichnungen Kontrollierte Panel-Ausschlüsse Natürliche Panel-Abdeckungslücken Panel-Gewichte Roh- und effektive Stichprobe Kontrollierter Wert Natürlicher Wert Unterzeichneter Panel-Unterschied Absoluter Panel-Unterschied Kritischer Fehlerunterschied UNBEKANNTER und unentschlossener Unterschied Vordefinierte Untergruppen-Ergebnisse Abgestimmte Design-Aufzeichnungen Reihenfolge- und Übertragungseffekte CN-Vergleichbarkeitsniveau Datenschutz- und Schwärzungsaufzeichnung Schutz kleiner Zellen Öffentliche Panel-Ergebnisübersicht Panel-Version und Änderungsaufzeichnung Verantwortliche Person oder Institution

75. ÜBERPRÜFUNGS-KONTROLLLISTE

Sind die Forschungsfragen für das kontrollierte und das natürliche Panel getrennt? Sind ihre Zielgruppen gleich oder unterschiedlich? Wurde der kontrollierte saubere Zustand klar definiert? Abgesehen vom neuen Chat, welche sauberen Zustände wurden bestätigt? Ist das Gedächtnis wirklich ausgeschaltet, oder ist es unbekannt? Wurde der Status der speziellen Anweisung aufgezeichnet? Gibt es einen Effekt des Unternehmensarbeitsbereichs? Sind die Web- und Werkzeugmodi standardmäßig? Sind der Plan und die Nutzeroberfläche fest oder geschichtet? Wurde ein Vertrauensniveau für den sauberen Zustand angegeben? Wurden CSC-1-Beobachtungen als vollständig sauber verwendet? Wurde der Teilnehmer gebeten, sein Gedächtnis oder seine Historie zu löschen? Wurden die

Einstellungen am Ende der Aufgabe wiederhergestellt? Wurde ein neues Konto oder ein Testkonto verwendet? Wurde das Testkonto wie ein echtes Nutzerpanel präsentiert?

Welches Trace verwendet das natürliche Panel? Wurde in Natural Account-State ein neuer Chat geöffnet? Hat der aktuelle Gesprächskontext die Haupt-Natural-Bewertung beeinträchtigt? Wurde ein Natural-State-Vertrauenslevel angegeben? Wurden Speicher- und Nutzerdefinierte Instruktionsinhalte unnötig gesammelt? Wurden der tatsächliche Plan des Nutzers und die Oberfläche aufgezeichnet? Sind First-Use-Nutzer getrennt? Sind Unternehmensbenutzer getrennt? Wurde der Zustand des Natural-Tools aufgezeichnet? Haben die beiden Panels die gleiche Prompt-Version verwendet? Wurde die erste berechtigte Ausgabe in beiden Panels gespeichert? Hat die natürliche Nutzerantwort sich aktualisiert? Wurde der Panel-Status entsprechend dem Inhalt der KI-Antwort geändert? Wie wurden Datensätze mit erkannten natürlichen Zuständen im sauberen Panel verwaltet? Wurde derselbe Nutzer in zwei Panels verwendet? Wurde die Reihenfolge randomisiert?

Wurde der Effekt von Transfer und Gedächtnis bewertet? Entspricht bei dem Design mit unabhängigen Gruppen die Zusammensetzung der Nutzer? Sind kontrollierte und natürliche Gewichtungen getrennt? Kann der natürliche Wert wirklich auf die aktive Nutzerpopulation verallgemeinert werden? Wurde der Unterschied zwischen sauber und natürlich überinterpretiert als ein Personalisierungseffekt? Ist ein Unterschied bei kritischen Fehlern sichtbar? Gibt es einen sichtbaren Unterschied für UNBEKANNT und Nullergebnisse? Wurden Untergruppen vor den Ergebnissen definiert? Können kleine Untergruppen neu definiert werden? Wurden nutzergeführte Anweisungen für oder gegen in allgemeines GEO umgewandelt? Hat die Personalisierung die materielle Realität verändert? Wurde ein CN-Vergleichsniveau bereitgestellt? Ist der verantwortliche Eigentümer der Panel-Ergebnisse-Karte klar?

76. EINWÄNDE UND ANTWORTEN

Einspruch 1— „Wenn echte Nutzer Speicherkonten verwenden, warum ist dann ein sauberes Panel notwendig?“

Weil die Ursache des Unterschieds ohne eine gemeinsame Ausgangsbedingung zwischen Produkten, Ländern und Sprachen nicht verstanden werden kann. Ein sauberes Panel ersetzt kein echtes Nutzerpanel. Es bietet eine Referenz zum Vergleich.

Einwand 2— „Ist ein kontrolliertes Panel nicht künstlich, wenn es ein natürliches Panel gibt?“

Es ist kontrolliert. Das ist absichtlich. Der Zweck einer Laborbedingung ist nicht, das gesamte reale Leben abzubilden, sondern bestimmte Variablen zu isolieren.

Einwand 3— „Wenn das kontrollierte Panel höhere Werte erzielt, können wir das nicht als Haupt-GEO-Score nehmen?“

Das kann man nicht alleine tun. Ein hoher Wert im Kontrollpanel kann nicht die Tatsache auslöschen, dass reale Nutzer niedrigere Ergebnisse sehen. Die beiden Ergebnisse sollten getrennt erscheinen.

Einwand 4— „Ist der wissenschaftliche Wert niedrig, wenn das natürliche Panel sehr verstreut ist?“

Heterogenität muss kein Mangel des natürlichen Panels sein. Sie ist ein Merkmal der realen Nutzererfahrung. Heterogenität:

- wird aufgezeichnet,

- ist geschichtet,
- ist gewichtet,

wird mit Unsicherheit berichtet.

Einwand 5— „Stört das Öffnen eines neuen Chats nicht den natürlichen Nutzerzustand?“

Der aktuelle Gesprächskontext wird entfernt. Der persistente Konto-, Plan-, Speicher- und benutzerdefinierte Anweisungsstatus werden beibehalten. Dies macht den natürlichen Kontostatus vergleichbar. Auch die aktuelle Gesprächserfahrung ist messbar.

Einwand 6— „Wie werden wir seine Wirkung messen, ohne den Inhalt der benutzerdefinierten Anweisungen zu kennen?“

Das Wissen um den Inhalt ist für die Hauptbewertung nicht erforderlich. Dessen Vorhandensein:

- existiert,
- existiert nicht,
- kann Material sein

Wir können es als solches klassifizieren. Für die kausale Inhaltsanalyse ist separate und einvernehmliche Arbeit erforderlich.

Einwand 7— „Wenn das Gedächtnis klar ist, aber nichts über das Unternehmen erinnert wird, warum dann eine separate Gruppe?“

Die Wirkung des Gedächtnisses kann unbekannt sein. Ein Status von „nicht mit der Entität verbunden“ kann basierend auf der Aussage des Nutzers vergeben werden. Dies hat jedoch immer noch nicht die gleiche Sicherheit wie der geschlossene Gedächtniszustand.

Einwand 8– „Ist die Anweisung des Nutzers ‚immer X vorschlagen‘ nicht echte Nutzererfahrung?“

Es ist eine echte Erfahrung. Es ist jedoch nicht die allgemeine GEO-Leistung von X. Es handelt sich um eine von Nutzern bereitgestellte Anleitung. Es sollte separat aufgezeichnet werden.

Einwand 9– 'Wenn Personalisierung dem Nutzer eine passendere Antwort gibt, wo ist das Problem?'

Es muss kein Problem sein. Empfehlungen können je nach Bedarf des Nutzers variieren. Probleme treten auf, wenn Identität, Lizenz, Preis und Evidenzstatus willkürlich ändern.

Einwand 10— „Ist es nicht die korrekteste Methode, denselben Nutzer in zwei Panels zu verwenden?“

Es kann stark für einige kausale Vergleiche sein. Die erste Aufgabe kann die zweite Situation beeinflussen. Daher sind Reihenfolge-, Übertragungs- und Gedächtniskontrollen erforderlich. Unabhängige Gruppen können in Hauptpopulationen sicherer sein.

Einwand 11— 'Wenn das Aktualisieren der Antwort in der Panel unter natürlichen Nutzungsbedingungen normales Verhalten ist, warum ist es dann verboten?'

Regeneration ist verboten, wenn die Exposition gegenüber der ersten Reaktion gemessen wird. Regeneration und nachfolgendes Verhalten können unter einem Natural Interaction Track separat gemessen werden.

Einwand 12— „Was werden wir mit einem Nutzer tun, der Speicher und Anweisungen für ein sauberes Panel nicht ausschalten kann?“

Sie können nicht dem kontrollierten Panel zugeführt oder in einer separaten Zelle mit einem niedrigeren CSC-Level untergebracht werden. Sie können im natürlichen Panel bewertet werden. Die Ausschlussrate sollte gemeldet werden.

Einspruch 13— „Sind Testkonten nicht zuverlässiger?“

Sie können in Bezug auf die Kontrolle zuverlässig sein. Sie sind hinsichtlich der tatsächlichen Nutzerpopulation eingeschränkt. Kontrollierte Testkonten und echte Nutzerpanels sind separate Ergebnisse.

Einwand 14— „Ist das Speichern von Nutzereinstellungen auf dem Natur-Panel nicht eine Verletzung der Privatsphäre?“

Es kann passieren, wenn unnötige Inhalte gesammelt werden. Datenschutz kann durch hochstufige offene/geschlossene Status, anonyme Registrierung und Schwärzung gewahrt werden.

Berufung 15— „Werden wir die Punktzahlen der beiden Gremien nicht in eine einzige NOMOS-Punktzahl zusammenführen?“

Die Rollen der beiden Gremien können in der endgültigen Ergebnisarchitektur definiert werden. Rohwerte und Zwischenergebnisse sollten jedoch immer separat sichtbar bleiben. Sie können die kombinierte Punkteverteilung nicht ersetzen.

Berufung 16— „Welchen Fehler hat das Unternehmen, wenn das natürliche Gremium niedrig bewertet?“

Es ist nicht immer der Fehler des Unternehmens. Quelle:

- Nutzeranweisung,
- KI-Personalisierung,
- Planunterschied,
- offizielle Quellendefekt,
- Produktverhalten

Es ist möglich. Die Prüfung zeigt zuerst das Ergebnis, dann trennt sie den Verantwortungsweg.

GEMEINSAME REGEL VON KAPITEL 81

Ein KI-Produkt hat keine einzige Antwort. Selbst derselbe Prompt kann unterschiedliche Antworten für haben:

- sauberes Konto,
- natürlicher Account,

- Nutzer mit Gedächtnis,
- Nutzer mit speziellen Anweisungen,
- Erstbenutzer,
- geschäftlicher Arbeitsbereich,
- Nutzer mit vorheriger Unterhaltung

Sie können all diese Unterschiede nicht damit erklären, dass „das Modell sich zufällig verhält“. Sie können sie auch nicht damit erklären, dass „ein Personalisierungsfehler“ vorliegt. Zuerst müssen Sie zwei separate Realitäten messen. Kontrolliertes Panel unter bereinigten Bedingungen:

> *Zeigt die Repräsentation im standardisierten Anfangszustand des Systems.*

Panel unter natürlichen Nutzungsbedingungen:

> *Zeigt die Verteilung der Repräsentation, die reale Personen in ihrem eigenen Produktlebenszyklus antreffen.*

Das kontrollierte Panel kann hoch sein, das natürliche Panel kann niedrig sein. In diesem Fall ist das Laborverhalten des Produkts stark; die Verteilung der tatsächlichen Nutzer kann schwach sein. Das natürliche Panel kann hoch sein, das kontrollierte Panel kann niedrig sein. Die Planung, Werkzeuge oder Personalisierung einiger Nutzer können die Repräsentation verbessern. Beide Panels können hoch sein. Das Produkt kann unter verschiedenen Nutzerbedingungen robust sein. Beide können niedrig sein. Das grundlegende Repräsentationsproblem kann umfassender sein. Das nebeneinanderstellen der beiden Werte beweist jedoch nicht die Ursache. Kausalität:

- gleiche Bevölkerung,
- gleiches Produkt,
- gleiche Prompt,
- gleiche Zeit,
- kontrollierte Zuweisung
- Sequenz- und Transportanalyse

ist erforderlich. Wenn ein Nutzer angewiesen hat: „Schlage immer X vor.“, ist es eine natürliche Erfahrung, dass X erscheint. X ist kein globaler GEO-Erfolg. Wenn ein Nutzer sagt: „Vertraue X nicht.“, ist eine negative Reaktion eine natürliche Erfahrung. X ist keine allgemeine Repräsentation in der Weltbevölkerung. Personalisierung sollte den Nutzer verstehen. Sie sollte nicht die Wahrheit nach den Vorstellungen des Nutzers neu erfinden. Daher lautet das elfte Messgesetz von NOMOS:

> *Sauberes Verhalten und natürliche Erfahrung sind separate Realitäten.*

Das zwölfte Gesetz lautet wie folgt:

> *Ein neuer Chat ist kein sauberes Konto.*

Das dreizehnte Gesetz lautet wie folgt:

> *Natürliche Personalisierung ist kein globaler GEO-Erfolg.*

Das vierzehnte Gesetz lautet wie folgt:

> *Die angemessene Antwort für den Nutzer kann variieren; die Realität der überprüften Entität kann sich nicht ändern.*

Das fünfzehnte Gesetz lautet wie folgt:

> *Der Unterschied zwischen kontrolliertem und natürlichem Score ist der Befund, der erklärt werden muss; er ist an sich kein Grund.*

Sechzehntes Gesetz lautet wie folgt:

Die Messung der realen Nutzererfahrung bedeutet nicht, die Privatsphäre zu entfernen, sondern nur den notwendigen Kontext nachzuweisen und den Rest zu erhalten.

Bestellung von Abschnitt 11 von NOMOS

Sag mir nicht nur, dass du einen neuen Chat geöffnet hast. / Zeige den Status von Speicher, speziellen Anweisungen, Plan und Werkzeugen.

Gehe nicht davon aus, dass alle Personalisierungen verschwunden sind, nur weil du die Notiz 'Temporäre Sitzung' siehst.

Das saubere Panel ist die Gesamtheit des realen Lebens; mach das natürliche Panel nicht zu unkontrolliertem Rauschen.

Standardisiere das kontrollierte Konto. / Behalte das natürliche Konto unverändert. / Lege beide nicht in dieselbe Box.

Mache die speziellen Anweisungen des Nutzers nicht zum Welterfolg für die Entität.

Mache die negative Vergangenheit des Nutzers nicht zur Sicht der gesamten Bevölkerung.

Zähle Personalisierung nicht als Lizenz, die Wahrheit zu verändern.

Du kannst einem Nutzer geeignete Ratschläge geben. / Aber schreibe die Identität, Lizenz, den Preis und den Nachweis des Unternehmens nicht von Nutzer zu Nutzer um.

Sammle das Privatleben nicht im natürlichen Panel. / Ist das Gedächtnis eingeschaltet, Anweisungen vorhanden, Kontext verfügbar – zeichne nur auf, was notwendig ist.

Lasse den Teilnehmer sein Gedächtnis nicht löschen. / Stelle die Einstellungen wieder her, wenn du sie vorübergehend geändert hast.

Wenn du denselben Nutzer in zwei Panels getestet hast, erinnere dich daran, dass die erste Antwort die zweite beeinflussen kann.

Zeigen Sie den Unterschied zwischen dem sauberen Punktestand und dem natürlichen Punktestand. / Zeigen Sie den kritischen Unterschied. / Zeigen Sie das Unbekannte. / Nennen Sie jedoch den Grund nicht ohne Beweis.

Entscheiden Sie zuerst, welche Nutzerrealität Sie messen möchten. / Festschreiben Sie dann die sauberen und natürlichen Bedingungen separat. / Halten Sie dann die Prompt gleich. / Erfassen Sie dann die erste Antwort, ohne sie zu ändern. / Berechnen Sie dann die beiden Verteilungen separat. / Und erst danach erklären Sie die Distanz zwischen dem, was das Produkt unter Standardbedingungen tut, und dem, was reale Menschen erleben.

Der abschließende Satz des Kapitels

Die wahre Repräsentation eines KI-Produkts in GEO-1000 ist nicht die Antwort, die es in einer sauberen Sitzung geben kann; es ist die klare und bewiesene Beziehung zwischen seinem Verhalten unter Standardbedingungen und den Antworten, die es tatsächlich den Menschen über verschiedene Nutzerhistorien hinweg liefert.

Normativer Kern

Beobachtungen des kontrollierten Panels unter bereinigten Bedingungen und des Panels unter natürlichen Nutzungsbedingungen MÜSSEN separate Messpopulationen, Panelzustände, Nenner und öffentliche Ergebnisse bleiben. Ein Kontrolliertes Panel unter bereinigten Bedingungen MUSS definieren und, falls zutreffend, überprüfen: - ein neues und separates Gespräch, - das Fehlen eines vorherigen Gesprächskontexts, - den Zustand des persistenten Speichers, - den Zustand der Nutzerdefinierten Anweisungen, - den Unternehmens- oder verwalteten Arbeitsbereichszustand, - Web-, Abruf- und Werkzeugkonfiguration, - Plan, - Nutzerschnittstelle, - Prompt-Version, - und die Vertrauenswürdigkeit des sauberen Zustands. Ein neues Gespräch oder eine Kennzeichnung als temporäre Sitzung DARF NICHT allein als Nachweis dafür behandelt werden, dass jede persistente Personalisierung deaktiviert wurde. Ein Panel unter natürlichen Nutzungsbedingungen MUSS das authentische Konto, den Plan, die Schnittstelle, den Speicher, die nutzerdefinierten Anweisungen und den Nutzungszustand des Teilnehmers bewahren, während den festgelegten Prompt, die Regel für die erste berechtigte Ausgabe, die Nachweisanforderungen und die Regel, keine neue Antwort zu erzeugen beibehalten werden. Beobachtungen im Zustand eines natürlichen Kontos, natürlicher Gesprächs-Zustand, Erstverwendung und verwaltete natürliche Beobachtungen MÜSSEN separat identifiziert bleiben. Die Dokumentation des natürlichen Zustands MUSS die Sammlung von privaten Daten minimieren. Private Anweisungstexte, Gesprächsverlauf, Identität, Zugangsdaten und sensible persönliche Inhalte DÜRFEN NICHT gesammelt werden, es sei denn, dies ist gesondert erforderlich, mit Einwilligung erhoben, minimiert und zugangskontrolliert. Nutzergenerierte entitätsbegünstigende oder entitätsbenachteiligende Anweisungen KÖNNEN als natürliche Personalisierungseffekte gemessen werden, DÜRFEN jedoch NICHT als allgemeiner Erfolg, Misserfolg oder Standardverhalten eines Produkts für die Entität repräsentiert werden. Personalisierung KANN Relevanz, Präsentation und Empfehlung verändern, wenn der Nutzerkontext dies rechtfertigt. Sie DARF NICHT willkürlich die verifizierte Identität, Lizenz, den rechtlichen Status, Preis, Umfang, Nachweise oder die Zeit einer Entität ändern. Unterschiede zwischen kontrollierten und natürlichen Panels DÜRFEN NICHT als kausale Personalisierungseffekte repräsentiert werden, es sei denn, Bevölkerung, Produkt, Prompt, Zeit, Schnittstelle, Plan, Werkzeuge, Zuordnung, Reihenfolge und Carryover-Bedingungen unterstützen diese Schlussfolgerung. Kontrollierte und natürliche Bewertungen, Raten kritischer Fehler, UNKNOWN-Raten, Abdeckung, effektive Stichprobengrößen sowie signierte und absolute Panel-Lücken MÜSSEN weiterhin separat sichtbar bleiben. Die Berechtigung des Panel-Status, Umklassifizierung, Ausschluss und Gewichtung MÜSSEN unabhängig davon bleiben, ob die KI-Ausgabe positiv, negativ, korrekt, inkorrekt, abgelehnt, kritisch oder kommerziell günstig ist. Jede Definition des Panel-Status, jeder Prozess zur Änderung von Einstellungen, jede Datenschutzentscheidung, jedes Vergleichsniveau und jedes Ergebnis MÜSSEN versioniert und einer verantwortlichen Person oder Organisation zuordenbar sein.

KAPITEL 12

Synchrone Wellen und NOMOS-Erfassung

METHODISCHE GRUNDLAGEN / K12; K14; K25; K26

Kapitelgrenze

Die ersten elf Kapitel legen die Grundlagen der Messung fest: Eine einzelne Antwort ist kein GEO-Score; Repräsentation ist eine Verteilung über Nutzer und Systembedingungen; die geprüfte Entität muss definiert werden, bevor Ergebnisse beobachtet werden; offizielle Repräsentation und verifizierte Entitätsrealität müssen getrennt bleiben; jedes KI-Produkt benötigt seine eigene berechtigte Nutzerpopulation; Teilnehmer müssen durch eine populationsbasierte, versionierte und ergebnisunabhängige Methode ausgewählt werden; kleinere Länder und Sprachen mit geringeren Ressourcen müssen durch spezielle Beobachtungsgremien sichtbar bleiben; das KI-Produkt, der Plan, die Nutzeroberfläche und die Systemkonfiguration müssen in einem Register verknüpft werden; Aufforderungen müssen die gleiche Nutzerabsicht über Sprachen hinweg bewahren und nicht die gleichen Wörter; und kontrollierte, saubere und natürliche Nutzerpanels müssen unterschiedliche Realitäten messen. Diese Elemente müssen nun in einem einzigen Messevent zusammengeführt werden:

> *Wann werden die Teilnehmer die Prompt einreichen, wie werden sie die anfängliche Ausgabe erfassen und wie werden sie nachweisen, dass jede Beobachtung danach nicht verändert wurde?*

Ihre Gründungsidee war klar: Menschen in verschiedenen Ländern der Welt sollten denselben Prompt gleichzeitig an dasselbe KI-Produkt senden und dies mit einem Screenshot nachweisen. Diese Idee ist der experimentelle Kern von GEO-1000. Allerdings ist der Ausdruck „zur gleichen Zeit“ allein nicht ausreichend. Wenn alle Teilnehmer versuchen, zur gleichen Sekunde einzureichen:

- verändern sich die Verbindungsgeschwindigkeiten,
- tauen die Uhren der Geräte voneinander ab,
- einige Nutzer sind gezwungen, um Mitternacht zu arbeiten,
- es kann einen plötzlichen Lastanstieg auf dem KI-Produkt geben,
- der Anbieter kann Geschwindigkeits- oder Nutzungslimits anwenden,
- einige Antworten können in zwei Sekunden, andere in zwei Minuten fertiggestellt werden,
- das Produkt kann während der Messung aktualisiert werden,

Ein Teilnehmer kann die Antwort sofort hochladen, ein anderer Stunden später. Ein Klick im gleichen Moment garantiert nicht, dass wir denselben Systemzustand perfekt messen. Ebenso ist das bloße Sammeln eines Screenshots nicht ausreichend. Ein Screenshot:

- kann zugeschnitten sein,
- kann die Prompt nicht zeigen,

- kann nicht beweisen, ob die Antwort anfänglich oder aktualisiert ist,
- kann Produkt- und Planinformationen nicht zeigen,
- kann an einem anderen Tag aufgenommen worden sein,
- kann von einem anderen Nutzer kopiert worden sein,

Es kann später bearbeitet worden sein. Aus diesem Grund muss GEO-1000 zwei separate Systeme zusammen installieren:

Synchronisierte Messwelle

und:

NOMOS Aufnahme-Kette von Evidenzmitteln

Die synchronisierte Welle beantwortet die folgende Frage:

> *Wurden die Beobachtungen innerhalb einer vergleichbaren Systemzeit und in einer vordefinierten Einreichungsreihenfolge erzeugt?*

NOMOS Capture fragt andererseits:

> *Können wir zeigen, dass der gesendete Prompt, die resultierende erste Ausgabe, die Produktbedingungen und die Zeitinformationen ohne Änderungen erhalten geblieben sind?*

Die vorherige Audit-Architektur stellte fest, dass jede Prüfung Modell oder Produkt, Datum, Land, Sprache, Abfragesatz, Wiederholungsanzahl und Messdatensatz als zwingende Messlogik enthalten muss. Diese Anforderung ist nicht nur eine technische Vorschrift. Denn unsichtbare Anweisungen, gefälschte Nachweise und als unabhängig dargestellte Inhalte können letztlich die Repräsentation verzerren, die von echten Menschen übernommen wird. Dieser Abschnitt:

- Messwelle,
- Synchronisationsfenster,
- UTC Zeitstandard,
- Versandfenster,
- Abschluss- und Erfassungszeiten,
- Erstkontaktergebnis und erste Inhaltsausgabe,
- technische Wiederholungsregeln,
- Produktänderungen innerhalb der Welle,
- Externe Ereignis- und Nachrichteneffekte,
- NOMOS Evidenzpaket erfassen,
- Die Beziehung von Screenshot, Rohantwort und Metadaten,
- Zeitvalidierung,
- Dateiintegrität,
- Hash und digitale Signatur,
- Evidenzkette,
- Datenschutz und Schwärzung,
- Öffentliches Evidenzverzeichnis,
- Qualitäts- und Gültigkeitsstatus der Welle

definiert. Dieses Kapitel behandelt noch nicht:

- Welche Realität die atomaren Aussagen in KI-Antworten aufzeichnet, wird bewertet,
- Wie Schiedsrichter Antworten kodieren werden,
- Übergangs- und kritische Fehlergrenzen,

- Legt die endgültigen NOMOS Punkformeln nicht fest

Die Hauptfrage von Kapitel 12 lautet:

> *Wie verknüpfen wir Tausende von echten Nutzerbeobachtungen mit demselben Messereignis und verwandeln jede in eine unveränderliche, prüfbare, die Privatsphäre schützende Beweiseinheit?*

NOMOS Herausforderung

Sie geben denselben Prompt an tausend Personen. Sie sagen: „Schicken Sie es heute Mittag.“ Ein Nutzer schickt es in Istanbul um 12:03 Uhr nachmittags. Ein anderer Nutzer schickt es in London um deren lokale Mittagszeit. Ein weiterer Nutzer sieht die Aufgabe abends in Tokio und erledigt sie acht Stunden später.

Ein Nutzer hatte die Geräteuhr falsch eingestellt. Die Internetverbindung eines anderen Nutzers war unterbrochen. Jemand anderes mochte die Antwort nicht und drückte den „Regenerate“-Button. Dann lud er einen Screenshot der längsten und positivsten Antwort hoch. Ein Nutzer hat nur die Mitte der Antwort zugeschnitten, der Prompt ist nicht sichtbar. Der Name des KI-Produkts ist nicht sichtbar.

Eine andere Person schickte denselben Screenshot mit zwei verschiedenen Panel-Konten. Ein Nutzer erledigte die Aufgabe korrekt, lud den Screenshot jedoch zwei Tage später hoch. In der Zwischenzeit haben sie die Datei zugeschnitten und ihre persönlichen Informationen gelöscht. Der ursprüngliche Zustand der Datei wurde nicht bewahrt. Ein anderer Nutzer in der ersten Antwort der KI:

Es erhielt das Ergebnis „Ich kann diese Frage nicht beantworten.“ Der Forscher sagte: „Es muss ein technischer Fehler vorgelegen haben“ und bat darum, es erneut zu versuchen. Die zweite Antwort war korrekt, und nur die zweite Antwort wurde im Bericht aufgenommen. Ein anderer Nutzer erlebte einen echten Verbindungsfehler. Die Prompt wurde gesendet, aber das System erzeugte keinen Inhalt.

Diesmal erlaubte der Forscher kein erneutes Ausprobieren. Dasselbe Protokoll wurde auf zwei Nutzer unterschiedlich angewendet. Nun, mitten in der Messwelle, der KI-Anbieter:

- Websuchverhalten,
- Zitationssystem,
- die verwendete Modellanleitung

ändert alle drei. Die ersten 500 Personen sehen den alten Systemzustand; die nächsten 500 sehen den neuen. Der Bericht sagt dann: ‚1,000 Personen haben das gleiche KI-System zur gleichen Zeit getestet.‘ Diese Aussage ist falsch. Es gibt tausend Nutzer, aber nicht unbedingt eine einzige kohärente Messwelle.

Nun wendet ihr einen Hash auf alle Bildschirmaufnahmen an. Ihr sagt, dies beweist, dass alle Beobachtungen real sind. Was beweist ein Hash? Er stärkt die Sicherheit, dass die Datei nach der Erstellung des Hashes nicht verändert wurde. Allein beweist er nicht:

- Dass das Bild von einem realen KI-Produkt stammt
- Dass es die erste Antwort war
- Dass es zur richtigen Zeit erhalten wurde
- Dass es nicht von einem anderen Nutzer kopiert wurde
- Dass der Prompt korrekt war
- Dass das Bild vor dem Hashen nicht bearbeitet wurde

Die erste Prämisse dieses Abschnitts lautet:

> *Zur gleichen Zeit messen ist nicht dasselbe wie in derselben Sekunde klicken.*

Ihre zweite Bestimmung lautet:

> *Ein Screenshot ist Teil des Evidenzpakets; er ist nicht das gesamte Evidenzpaket.*

Seine dritte Bestimmung besagt:

> *Die erste falsche Antwort ist kein Datenfehler. Wenn sie nach dem Protokoll erzeugt wird, ist sie das Ergebnis der Messung.*

Seine vierte Bestimmung besagt:

> *Hash stärkt die Integrität; er schafft nicht von selbst die Realität der Beobachtung.*

Seine fünfte Vorschrift lautet:

> *Wenn sich das System in einer Welle wesentlich ändert, können tausend Antworten ohne Zeitstempel nicht als zum selben System gehörig betrachtet werden.*

Seine sechste Bestimmung lautet:

> *Die Evidenzkette wird nicht erstellt, um das Ergebnis gut aussehen zu lassen, sondern um später nachvollziehen zu können, wie das Ergebnis erzeugt wurde.*

1. ZWECK DES KAPITELS

Der Zweck dieses Abschnitts besteht darin, GEO-1000 Nutzerbeobachtungen in vergleichbaren Zeitfenstern zu erzeugen und jede Beobachtung mit einer vollständigen digitalen Evidenzkette zu sichern. Der Abschnitt standardisiert die folgenden Unterscheidungen:

- Messfenster synchronisiert auf die gleiche Sekunde
- Lokale Zeit versus UTC Zeit
- Welle versus Unterwelle
- Übertragungszeit versus Antwortabschlusszeit
- Aufzeichnungszeit versus Upload-Zeit
- Geplanter Slot versus tatsächliche Übertragung
- Erster Systemkontakt versus erste Inhaltsantwort
- Technischer Fehler versus Systemablehnung
- Technischer Wiederholversuch versus Antwortaktualisierung
- Die bestausgewählte Ausgabe mit der ersten passenden Ausgabe
- Sichtbarkeit auf dem Bildschirm mit Abschluss der Antwort
- Rohantwort mit Screenshot
- Screenshot versus vollständiges Evidenzpaket
- Beobachtungsoriginalität mit Dateiintegrität
- Zeitstempel mit Hash
- Zuverlässige Zeitquelle mit lokaler Gerätezeit
- Server-Akzeptanzzeit mit Teilnehmer-Upload
- Schwärzte öffentliche Kopie mit Rohbeweis
- Teilnehmeridentität mit Beobachtungs-ID
- Antwortgenauigkeit mit Erfassungsvalidität

- Natürliche Ausgabenvarianz mit Produktaktualisierung
- KI-Produktänderung mit externem Nachrichtenereignis
- Wellenbeendigung mit Ausschluss nach Ergebnis
- Automatische Bot-Messung mit Erfassung durch menschlichen Nutzer
- API-Protokoll mit Nutzeroberflächenaufzeichnung
- Anbieter-Telemetrie mit unabhängiger Erfassung
- Zugriff auf Nachweise für alle durch Speichern der Nachweise
- Unwiderruflicher Datensatz mit unveränderlichem Datensatz
- Versionierte Korrektur einer falschen Entscheidung mit Rohdatenschutz

Am Ende dieses Abschnitts sollte jedes GEO-1000-Audit in der Lage sein, die folgenden Fragen zu beantworten:

Zu welcher Welle gehört diese Beobachtung?

Zu welcher UTC-Zeit wurde die Aufforderung gesendet?

Was war das ursprüngliche Systemergebnis?

Wann wurde die Antwort abgeschlossen und wann wurde sie erfasst?

Wenn ein technischer Wiederholungsversuch durchgeführt wurde, was war der Grund und wo war der vorherige Versuch?

Zeigt die Screenshot-Prompt das Produkt und die vollständige Antwort?

Bestätigen die Rohantwort und die visuelle Darstellung einander?

Wann wurden der Hash und der Zeitstempel der Dateien erstellt?

Wenn das Evidenzmaterial später geändert wurde, ist die Änderungshistorie sichtbar?

Kann ein unabhängiger Prüfer die Rohbeweise überprüfen, während die Privatsphäre der Teilnehmer gewahrt bleibt?

2. ZENTRALE NORMATIVE BESTIMMUNG

Jede GEO-1000-Beobachtung muss vor der Datenerfassung mit einer festgeschriebenen Messungswelle verknüpft sein; alle relevanten Zeiten vom Senden der Prompt bis zum ersten Systemergebnis, vom Abschluss der Antwort bis zur Aufnahme und zum Hochladen; Bedingungen des KI-Produkts; die vollständige gesendete Prompt; die erste geeignete Ausgabe und alle Integritätsaufzeichnungen müssen in einem versionierten NOMOS-Capture-Evidence-Paket aufbewahrt werden. Eine GEO-1000-Beobachtung allein:

- Screenshot,
- kopierter Antworttext,
- Teilnehmererklärung,
- Datei-Hash

kann nicht für sich allein als gültig betrachtet werden. Das Hauptbeweispaket muss, soweit relevant, Folgendes enthalten:

- Wave-ID
- Anonyme Beobachtungs-ID des Teilnehmers

- KI-Systemregistereintrag
- Panelstatus
- Prompt-ID und Version
- Zugewiesene Prompt
- Gesendete Prompt
- Übermittlungszeit
- Erste Systemkontaktaufnahme
- Antwortabschlusszeit
- Erfassungszeit
- Hochladezeit
- Vollständige Rohantwort
- Vollbild- oder sequenzielle Screenshots
- Referenz- und Quelldatensätze
- Technischer Fehler- und Wiederholungsprotokoll
- Produkteinstellungen
- Datei-Hashes
- Paket-Hash
- Zeitüberprüfungsaufzeichnung
- Erfassungswerkzeug-Version
- Gültigkeitsentscheidung
- Datenschutz- und Schwärzungsstatus
- Eigentümer der Evidenzkette

3. WAS IST EINE MESSWELLE?

Eine Messwelle ist ein umfassendes Datenerhebungsereignis, das unter vordefinierten Bedingungen durchgeführt wird:

- KI-Produkten,
- Zielpopulation,
- Aufforderungsversionen,
- Nutzerpanels,
- Zeitfenster,
- Stichprobenverteilung,
- Erfassungsprotokoll,
- Referenzaufzeichnungen

Eine Welle:

```
Wk=(A,U,P,F,T,C,R,G)
```

kann repräsentiert werden als. Hier:

- A: Beispiele für KI-Produkte
- U: Zielnutzerpopulation und Stichprobe
- P: Aufforderungssatz
- F: Panelzustände

- T: Zeit- und Abgabestruktur
- C: Erfassungsprotokoll
- R: Referenzrealitätsversion
- G: Governance und Hauptaufzeichnung

Wenn sich eines dieser Felder wesentlich ändert, kann eine neue Welle oder Unterwelle erforderlich sein.

4. ARTEN VON WELLEN

MW-1— HAUPT-SYNCHRONISIERTE WELLE

Es ist die Hauptmessung der GEO-1000-Population. Standardmäßig zielt sie auf mindestens 1,000 gültige Beobachtungen des Population Panels für jedes KI-Produkt ab.

MW-2— BESTÄTIGUNGSWELLE

Sie wird durchgeführt, um das Ergebnis der ersten Welle zu wiederholen, einen spezifischen Fehler zu bestätigen oder die statistische Stabilität zu testen. Sie wird nicht verwendet, um die niedrige oder hohe Punktzahl der ersten Welle zu ändern. Sie liefert ein separates Ergebnis.

MW-3— INCIDENT-WELLE

Sie wird nach einem kritischen oder signifikanten Beobachterereignis für gezielte und schnelle Überprüfung durchgeführt. Sie ist keine globale Hauptwelle.

MW-4— TRENDÜBERWACHUNGSWELLE

Überwacht Zeitänderungen im Verhalten von Produkt, Einheit, Ressource oder Sprache.

MW-5— KORREKTUR-NACHTEST-WELLE

Erzeugt neue Ergebnisse nach einer Korrektur oder Intervention. Die vorherige Welle wird nicht gelöscht.

MW-6— LOKALZEIT-AUSGEWOGENE WELLE

Zielt darauf ab, Teilnehmer unter ähnlichen lokalen Zeitzuständen zu messen. Es wird nicht ein einzelner globaler Systemmoment gemessen.

MW-7— KONTROLLIERTE LABORATORIUMSWELLE

Ein eingefrorenes Modell ist eine kontrollierte Studie, die mit API- oder Standard-Testkonten durchgeführt wird. Es ist kein Panel echter Nutzerpopulation.

5. WAS BEDEUTET „GLEICHE ZEIT“?

In GEO-1000 bedeutet Gleichzeitigkeit:

Es bedeutet nicht, dass alle Nutzer zur gleichen Millisekunde klicken

Die funktionale Definition ist:

> *Beobachtungen werden innerhalb eines vorbestimmten UTC-Messfensters erzeugt, entsprechend aufgezeichneter und ergebnisunabhängiger Einreichungszeitfenster.*

Dieser Ansatz balanciert drei Ziele aus:

- Den Produktstatus zeitlich nahe zu halten
- Plötzliche künstliche Last für den Anbieter zu reduzieren
- Teilnehmer-Einreichungen auf reproduzierbare Weise zu verteilen

6. ZEITSTRUKTUR

Jede Welle muss mindestens vier Zeitschichten enthalten.

6.1. Wellenfenster

Es ist der Bereich UTC, in dem alle Haupt-Einreichungen vorgenommen werden müssen. Beispiel:

12:00–13:00 UTC

6.2. Mikro-Zeitfenster

Dies ist das kürzere Zeitintervall, das dem Teilnehmer zugewiesen wird, um die Aufforderung zu senden. Beispiel:

12:15–12:20 UTC

6.3. Zusätzliche Zeit für die Fertigstellung

Wenn die Eingabe innerhalb des Slots gesendet wird, ist die vordefinierte Zeit für die Fertigstellung der Antwort.

6.4. Zeit für das Erfassen und Hochladen

Die Zeit, die für das Erstellen des Nachweises und dessen Hochladen in ein sicheres System nach Abschluss der Antwort erlaubt ist.

7. GRUNDREGEL FÜR SYNCHRONISIERTE WELLE DES KANDIDATEN

GRUNDREGEL FÜR KANDIDATEN — GRUNDLEGENDES PRINZIP FÜR KANDIDATEN

Für allgemeine Verbraucher-Chat-Produkte kann die folgende Struktur als Hauptkandidat für synchronisierte Wellen verwendet werden:

- Einreichungsfenster: 60 Minuten
- Mikro-Slots: 12× 5 Minuten
- Teilnehmer-Einreichung: innerhalb des ihnen zugewiesenen Mikro-Zeitfenster
- Zusätzliche Zeit für die Fertigstellung der Antwort: bis zu 20 Minuten nach der Einreichung
- Erzeugung des Capture-Hashes: vorzugsweise innerhalb von 5 Minuten nach Abschluss der Antwort
- Nachweis-Upload: spätestens innerhalb von 60 Minuten nach der Aufnahme
- Meldung technischer Probleme: innerhalb derselben Welle

Diese Zeiträume sind universell und nicht endgültig. Sie können sich unter den folgenden Bedingungen ändern:

- Antwortzeit des KI-Produkts
- Mobile Verbindung
- Infrastruktur des Landes
- Langantwortformat
- Barrierefreiheitsanforderung
- Nutzungslimit des Produkts

Die verwendeten Zeiten sollten vor der Datenerfassung festgelegt werden.

8. WARUM MIKRO-SLOTS?

Gleichzeitige Einreichung durch tausend Teilnehmer:

- Außergewöhnliche Belastung des KI-Produkts,
- Geschwindigkeitsbegrenzung,
- Warteschlange,
- technischer Fehler,
- Anstieg der Antwortzeit

kann erstellt werden. In diesem Fall kann das Audit das normale Verhalten des Produkts nicht messen, aber:

kann das vom Audit erzeugte künstliche Lastereignis messen

Mikro-Slots:

- verteilen die Last,
- halten die gleiche gesamte Systemzeit bei,
- machen die Einreichungsreihenfolge unabhängig vom Ergebnis,

ermöglichen die Modellierung des Zeiteffekts.

9. SLOT-ZUWEISUNG

Teilnehmer sollten Slots zugewiesen werden:

- Land,
- Sprache,
- KI-Produkt,
- Plan,
- auf ausgewogene und vorab zufällig bestimmte Weise in Bezug auf:

Panel-Status. Ein Mikro-Zeitfenster allein sollte nicht ausgefüllt werden mit:

- bestimmtem Land,
- bestimmter Sprache,
- bestimmtem Produkt

Nutzern. Andernfalls wird der Zeiteffekt mit dem Gruppeneffekt verfälscht.

10. FAIRE LOKALE ZEIT

Ein einzelnes UTC-Fenster kann für einige Teilnehmer sein:

- Nacht,
- Arbeitszeiten,
- Gebets- oder Ruhezeit,
- schwer erreichbare Stunden

Diese Situation:

- Teilnahmerate,
- Verbindungsbedingungen,
- Aufmerksamkeit der Nutzer,
- Gerätetyp

kann beeinflusst werden. Es können zwei ergänzende Methoden verwendet werden.

10.1. Rotierende globale Wellen

Hauptwellen werden zu verschiedenen UTC-Zeiten durchgeführt. Beispiel:

- Welle 1: 12:00 UTC
- Welle 2: 20:00 UTC
- Welle 3: 04:00 UTC

So trägt dieselbe Region nicht bei jeder Welle die gleiche lokale Zeitbelastung.

10.2. Lokalzeit-ausgeglichene Unterwelle

Teilnehmer werden in einem ähnlichen Zeitband in ihrer Ortszeit gemessen. Diese Studie:

- Nutzermüdigkeit,
- ist wertvoll für

tägliches Nutzungsverhalten.

Sie ist nicht das Ergebnis eines einzelnen globalen Systemmoments.

11. DREI-WELLEN-STABILITÄTSARCHITEKTUR

KANDIDATEN-ARCHITEKTUR — KANDIDATEN-ARCHITEKTUR

Für die vollständige Implementierung von GEO-1000 können mindestens drei Hauptwellen vorgeschlagen werden:

W1 — Anfangswelle

Erzeugt die anfängliche Populationsschätzung.

W2 — Kurzzeit-Wiederholung

Testet die kurzfristige Stabilität des Produkts und der Messung. Kandidatenintervall: 24–72 Stunden nach W1

W3 — Mittelfristige Stabilität

Misst, ob die Repräsentation des Produkts und der Entität über einen langen Zeitraum hinweg erhalten bleibt. Kandidatenintervall: 14–30 Tage nach W1. Bestimmte Intervalle müssen vor der

Datenerhebung festgelegt werden. Falls es aufgrund einer Produktaktualisierung oder eines Ereignisses geändert wird, ist ein neuer Versionsdatensatz erforderlich.

12. 30,000 BEOBACHTETER GRÜNDER ENTWURF

Für KI-Produkte, 1,000 Hauptbeobachtungen pro Produkt und für drei Hauptwellen:

```
10×1,000×3=30,000
```

Gültige Hauptpopulation-Panelbeobachtungen werden angestrebt. Ergänzend:

- Land Beobachter
- Sprachliche Fairness
- Barrierefreiheit
- Ereignis
- Nachtest

Panels können oberhalb dieser Zahl hinzugefügt werden. 30,000:

- ist nicht die Anzahl der Einladungen,
- ist nicht die Anzahl der Dateien,

ist nicht die Anzahl der Screenshots. Es ist das gültige Hauptprodukt-Beobachtungsziel.

13. ZEITVEKTOR DER BEOBACHTUNG

Für jede Beobachtung kann der Zeitvektor als Kandidat wie folgt definiert werden:

```
T_i= (t_i^assign, t_i^open, t_i^submit, t_i^first, t_i^complete, t_i^capture, t_i^upload,
t_i^validate)
```

Hier:

- tiassign: die Zeit, zu der die Aufgabe dem Nutzer zugewiesen wird
- tiopen: die Zeit, zu der die Aufgabe geöffnet wird
- tisubmit: die Zeit, zu der die Eingabe gesendet wird
- tifirst: die Zeit, zu der das erste sichtbare Systemergebnis beginnt
- ticomplete: die Zeit, zu der die Antwort abgeschlossen ist
- ticapture: die Zeit, zu der der Nachweis erfasst wird
- tiupload: die Zeit, zu der das Paket das sichere System erreicht
- tivalidate: die Zeit, zu der die Gültigkeit der Erfassung überprüft wird

All diese Zeiten sind möglicherweise nicht direkt auf jeder Nutzeroberfläche beobachtbar. Unbekannte Felder sollten UNBEKANNT bleiben.

14. GRUNDLEGENDE ZEITMETRIKEN

Antwortverzögerung:

```
L_i= t_i^complete− t_i^submit
```

Erfassungsverzögerung:

```
C_i= t_i^capture− t_i^complete
```

Upload-Verzögerung:

```
U_i= t_i^upload- t_i^capture
```

Slot-Abweichung:

```
D_i= t_i^submit- t_i^{slot-midpoint}
```

kann wie folgt berechnet werden. Diese Metriken sind:

- Systemlast,
- Teilnehmerverzögerung,
- Beweissicherheit,
- Wellenintegrität

im Hinblick auf die Bewertung.

15. KANN NICHT NACH VERZÖGERUNG VERWALTET WERDEN

Eine über einen langen Zeitraum abgeschlossene Antwort:

- in die Kategorien gezwungen werden von

  wahr,
- falsch,
- lang,
- kann zitiert werden.

Aufgrund der Verzögerung gilt die Ausschlussregel:

- vor der Datensammlung,
- unabhängig vom Inhalt der Antwort

muss definiert werden. Folgendes Verhalten ist untersagt: eine negative Antwort auszuschließen, weil sie verspätet eingereicht wurde; eine positive Antwort zu schützen, wenn sie mit derselben Verzögerung eingereicht wurde.

16. WAVE STATUS STATUSES

WS-0— GEPLANT

Die Welle ist geplant. Sie wurde noch nicht festgeschrieben.

WS-1— festgeschrieben

Proben-, Aufforderungs-, Produkt-, Zeit- und Erfassungsregeln sind festgeschrieben.

WS-2— OFFEN

Das Einreichungsfenster hat begonnen.

WS-3— GESCHLOSSEN

Die Annahme neuer Einreichungen ist beendet. Fristen für Fertigstellung und Hochladen können weiterhin bestehen.

WS-4— ABGESCHLOSSEN

Feld- und erste Erfassungsüberprüfungen wurden abgeschlossen.

WS-5— GETEILT

Es wurde aufgrund von Materialsystem- oder Protokolländerungen in Unterwellen aufgeteilt.

WS-6— PAUSIERT

Vorübergehend gestoppt aufgrund von Sicherheits-, Produkt- oder Infrastrukturproblemen.

WS-7— ABGEBROCHEN

Vor Abschluss der Welle beendet. Ergebnisse können begrenzt oder unbrauchbar sein.

WS-8— INVALIDIERT

Ungültig für die Hauptprognose aufgrund eines zuvor definierten schwerwiegenden Protokollfehlers. Rohdaten werden nicht gelöscht.

WS-9— ARCHIVIERT

Version wurde geschlossen und die Evidenzkette archiviert.

17. WAVE-HEALTH-KLASSEN

WH-0— NICHT BEWERTET

Gesundheit der Welle wurde nicht bewertet.

WH-1— KOMPROMITTIERT

Es liegt eine Änderung des Materialsystems, der Zeit oder ein Erfassungsfehler vor. Möglicherweise nicht für das Hauptergebnis verwendbar.

WH-2— TEILWEISE

Einige Schichten oder Slots sind unzureichend. Ein reduziertes Ergebnis kann erzeugt werden.

WH-3— AKZEPTABEL

Hauptzeit, System- und Erfassungsbedingungen sind ausreichend erhalten.

WH-4— STARK

Hohe Erfassungszuverlässigkeit, ausgewogene Slot-Abdeckung und geringer Protokollverlust sind vorhanden.

WH-5— REPLIZIERT

Ähnliche Wellengesundheit wurde mehrfach und auf unabhängigen Panels wiederholt. Die Wellengesundheit hängt nicht davon ab, ob der KI-Score hoch oder niedrig ist.

18. PRODUKTÄNDERUNGEN INNERHALB DER WELLE

Das KI-Produkt kann sich innerhalb des Messfensters wesentlich ändern. Signale:

- Modellkennzeichenänderung
- Schnittstellenänderung
- Aktivierung oder Deaktivierung der Webfunktion
- Änderung des Zitationsformats
- Plötzlicher Bruch in der Antwortstruktur
- Ankündigung des Anbieters
- Unterschiedliches Systemverhalten gleichzeitig bei vielen Nutzern
- Produktunterbrechung oder Neustart

In diesem Fall: Die Zeit der Änderung wird geschätzt. Systemkonfigurations-Fingerabdrücke werden verglichen. Welle WS-5 kann in den SPLIT-Status versetzt werden. Vorherige und nachfolgende Beobachtungen werden separate Unterwellen sein. Wenn ein einzelner Wert erstellt werden soll, wird der Systemstatus als separater Faktor modelliert. Materielle Unsicherheit wird im öffentlichen Bericht dokumentiert.

19. EXTERNES EREIGNISPROTOKOLL

Die Welt kann sich ändern, auch wenn sich das KI-Produkt nicht ändert. Während der Messung:

- Unternehmensübernahme,
- Produkteinführung,
- wichtige Nachrichten,
- rechtliche Entscheidung,
- Krise,
- Finanzmitteilung,
- virale Inhalte

kann auftreten. Dieses Ereignis:

- Webressourcen,
- Abruf-Ergebnisse,
- Nutzerkommentar zur Prompt,
- KI-Antwort

kann geändert werden. Jede Welle, soweit sie relevant ist:

Externe Ereignisregistrierung

Muss tragen.

20. WIE WERDEN ERGEBNISSE EXTERNER EREIGNISSE VERWALTET?

Ein externes Ereignis:

- wenn es vor der Welle bekannt ist, wird das Referenzpaket aktualisiert,
- wenn es während der Welle eintritt, wird es mit einem Zeitstempel markiert,

wenn es nach der Welle bemerkt wird, wird ein retrospektiver Ereigniseintrag hinzugefügt. Folgendes Verhalten ist verboten: Beobachtungen aufgrund von Nachrichten auszuschließen, die den Score senken; sie zu bewahren, wenn es Nachrichten sind, die den Score erhöhen. Das Ereignis selbst kann

Teil des realen Systemzustands sein. Falls notwendig, wird eine sub-wellen- oder ereignisangepasste Analyse durchgeführt.

21. ERGEBNIS DES ERSTEN KONTAKTS

Das erste Systemergebnis, das der Nutzer nach dem Senden der Prompt erhält:

Ergebnis des Erstkontakts

wird genannt. Das Ergebnis des Erstkontakts:

- Vollständige Antwort
- Ablehnung
- Technischer Fehler
- Nutzungslimit
- Verbindungsunterbrechung
- Endloses Laden
- Klärungsfrage
- Kein Ergebnis

möglich. Dieses Ereignis ist Teil der tatsächlichen Nutzererfahrung. Es kann nicht stillschweigend gelöscht werden.

22. ERSTE ZULÄSSIGE INHALTSAUSGABE

Die erste materielle Antwort, die im zuvor erlaubten erneuten Versuch nach einem technischen Fehler generiert wurde:

Kann als Erste Zulässige Inhalt Ausgabe bezeichnet werden

Diese Ausgabe: ersetzt nicht das anfängliche Kontaktergebnis, sie wird zusammen damit aufgezeichnet. Beispiel:

- Erstkontakt: Geschwindigkeitsbegrenzung
- Fünf Minuten später, zuvor erlaubter technischer erneuter Versuch
- Erste Inhaltsantwort

Der Bericht sollte Folgendes zusammenfassen:

- Erste Nutzererfahrung: Geschwindigkeitsbegrenzung
- Erste Inhaltsrepräsentation: spezifische Antwort

23. ERSTE GEEIGNETE AUSGABEREGEL

Wenn ein KI-Produkt eine signifikante Antwort auf den ersten Prompt liefert:

- selbst wenn es falsch ist,
- selbst wenn es kurz ist,
- selbst wenn sie eine Ablehnung beinhaltet,
- selbst wenn keine Referenzen angegeben werden,
- selbst wenn die Marke falsch identifiziert wird

eine Reproduktion kann nicht hergestellt werden. Die Hauptbeobachtung ist der erste Ausgang. Der Nutzer:

- Kann keine neue Antwort auswählen, indem durch:
- regenerieren,
- Umschreiben,
- Versuchen Sie es nochmal,
- Eine weitere Antwort,

Weitermachen

24. TECHNISCHER WIEDERHOLUNGSVERSUCH

Ein technischer Wiederholungsversuch kann nur unter zuvor definierten technischen Bedingungen verwendet werden. Zulässige Gründe des Kandidaten:

- Verbindungsunterbrechung, bevor die Antwort beginnt
- Serverfehlercode
- Schnittstelle sendet die Aufforderung nicht
- Anwendungsabsturz
- Überprüfter vorübergehender Systemausfall

Kann kein Grund für einen technischen Wiederholungsversuch sein:

- Falsche Antwort
- Ablehnung
- Kurze Antwort
- Fehlende Quellenangabe
- Marke nicht erwähnt
- Kritischer Fehler
- Die geprüfte Institution mag die Antwort nicht

25. WIEDERHOLUNG BEI TEILANTWORT

Wenn das System einen Teil des Inhalts erstellt hat und dann gestoppt wurde:

- Teilweise Ausgabe wird gespeichert,
- gespeichert als TRUNCATED_RESPONSE

wenn ein erneuter Versuch unternommen werden soll, wird ein neuer und verknüpfter Versuch erstellt. Er kann nicht gelöscht werden, als ob keine teilweise Antwort vorhanden wäre.

26. VERSUCHSKETTE

Eine Beobachtungszuweisung kann mehrere technische Versuche enthalten:

```
A_i→ (Attempt_{i,1}, Attempt_{i,2}, …)
```

Jeder Versuch:

- Zeit,
- trägt den Produktstatus,

- den Prompt-Hash,
- das Systemergebnis,
- den technischen Fehlgrund

Es wird nicht durch einen anderen Versuch überschrieben.

27. WAS IST NOMOS CAPTURE?

NOMOS Capture ist ein offener Daten- und Verfahrensstandard, der die Integrität der Beobachtung vom dem an den Nutzer gegebenen Prompt bis zur Archivierung des Evidenzpakets bewahrt.
NOMOS Capture:

- ist nicht nur eine Screenshot-Anwendung,
- muss keine proprietäre Software sein, die nur NobleJackal gehört,

ist nicht nur auf eine einzelne Browser-Erweiterung angewiesen. Eine kompatible NOMOS Capture-Anwendung:

- verwendet ein offenes Datenschema,
- folgt Zeit- und Integritätsregeln,
- achtet Datenschutzgrenzen,
- bewahrt die Evidenzkette

muss eingehalten werden. Unabhängige Forscher sollten in der Lage sein, ihre eigenen kompatiblen Werkzeuge zu entwickeln. Wenn der Standard nur mit der Software des Gründers implementiert werden kann, schwächt unabhängige Replikation.

28. VIER FUNKTIONEN DER NOMOS CAPTURE

28.1. Aufgabe ausführen

Richtig:

- Nutzer,
- Prompt,
- KI-Produkt,
- Slot

stellt Übereinstimmung sicher.

28.2. Beobachtung erfassen

Protokolliert die Prompt, das anfängliche Systemergebnis, die Antwort und die zugehörige Nutzeroberfläche.

28.3. Integrität wahren

Erstellt den Dateihash, Zeitstempel und Paketbeziehung.

28.4. Nachweise transportieren

Verknüpft Roh-, eingeschränkte und öffentlich verfügbare Evidenzschichten miteinander.

29. ERFASSUNGSMETHODEN

CM-1— GELEITETE MANUELLE ERFASSUNG

Der Teilnehmer stellt einen Screenshot und den rohen Antworttext gemäß den Anweisungen bereit. Dies verursacht eine geringe technische Belastung. Es ist anfällig für menschliche Fehler.

CM-2— UNTERSTÜTZTE ERFASSUNGSANWENDUNG

Anwendung:

- Prompts kopieren,
- Zeitprotokollierung,
- Datei-Upload,
- Hash-Generierung

Funktionen werden unterstützt. Es beeinträchtigt das KI-Produkt nicht.

CM-3— INSTRUMENTIERTE NUTZEROBERFLÄCHENAUFNAHME

Browser- oder Anwendungsintegration:

- das gesendete Prompt,
- die Produktmetadaten,
- den Antworttext,
- die Zeiten

werden direkt erfasst. Bietet hohe Integrität. Nutzungsbedingungen und Datenschutz sollten bewertet werden.

CM-4— ANBIETER EXPORT ODER TELEMETRIE-AUFNAHME

Die Sitzungsexporte des Anbieters oder überprüfte Telemetrie werden verwendet. Kann die unabhängige Erfassung stark unterstützen. Kann allein keine Unabhängigkeit bieten.

CM-5— KONTROLLIERTE API- ODER LABORAUFNAHME

API- oder gefrorene Systemprotokolle werden verwendet. Es handelt sich nicht um eine Live-Beobachtung der Nutzeroberfläche. Gehört zu einem separaten Laborpanel.

30. GRENZE DER AUTOMATISIERUNG

Automatisierung:

- Prompt-Einreichung,
- Zeitprotokollierung,
- Hash-Erzeugung,
- Dateiintegrität

kann verwendet werden. Anstelle des Teilnehmers:

- Bot-Konto,
- automatisierte Massenabfrage,
- Nicht-Nutzersitzung

Erzeugung erstellt kein echtes Human Population Panel. Die Prüfung durch automatisierte Systeme ist separat:

CONTROLLED_AUTOMATION_PANEL

muss aufgezeichnet werden als.

31. ERFASSUNG DER ZUVERLÄSSIGKEITSSTUFEN

NCL-0— KEINE ÜBERPRÜFBARE ERFASSUNG

Es gibt nur die Aussage des Teilnehmers. Kann nicht als das Haupt-GEO-1000-Ergebnis einbezogen werden.

NCL-1— TEILWEISE NACHWEISE

Es gibt einen Antworttext oder einen eingeschränkten Screenshot. Der Prompt oder der Systemzustand können nicht vollständig überprüft werden.

NCL-2— SCREEN-BEWEIS

Der Prompt, die Antwort und die Hauptoberfläche des Produkts wurden mit einem Screenshot überprüft. Metadaten und Zeitvertrauen können eingeschränkt sein.

NCL-3— KOMPLETTES EVIDENZPAKET

Vollständige Antwort, Prompt, Produktaufzeichnung, Zeiten, Screenshot und Hash sind verfügbar. Mindeststufe für die Hauptanalyse ist ein Kandidat.

NCL-4— INSTRUMENTIERTE UND SIGNIERTE AUFNAHME

Das Aufnahmetool hat die Prompt, Antwort, Zeit und Metadaten direkt aufgezeichnet; das Paket ist digital signiert.

NCL-5— UNABHÄNGIG ÜBERPRÜFTE MEHRQUELLEN-AUFZEICHNUNG

Aufnahmepaket:

- unabhängige Validierung,
- Anbieterexport,
- zweite Evidenzquelle

wurde unterstützt. Haupt-GEO-1000-Echtbenutzerbeobachtungen mindestens:

NCL-3

sollte dieses Niveau anstreben.

32. NOMOS Evidenzpaket ERFASSEN

Für jede Beobachtung enthält das Paket die folgenden Komponenten als Kandidaten:

32.1. Beobachtungsmanifest

Enthält alle Identitäts- und Statusfelder der Beobachtung.

32.2. Prompt-Artefakt

Der vollständige zugewiesene und gesendete Prompt-Text, die Version und dessen Hash.

32.3. Rohantwort-Artefakt

Die vollständige rohe Antwort des KI-Produkts.

32.4. Visuelles Nachweis-Artefakt

Zeigt Bilder oder aufeinanderfolgende Bilder von Prompt, Antwort und Produktoberfläche.

32.5. Systemzustands-Artefakt

Zeigt Plan, Oberfläche, Modellbezeichnung, Speicher, Web und andere materielle Systemzustände.

32.6. Zeitbeweis-Artefakt

Trägt die Einreichungs-, Abschluss-, Erfassungs- und Hochladezeiten.

32.7. Versuchsprotokoll

Zeigt die Kette technischer Fehler und Wiederholungen.

32.8. Integritätsmanifest

Enthält Hashes und Größen aller Dateien.

32.9. Datenschutzmanifest

Zeigt den Status von Schwärzung, Zugriff und Aufbewahrung.

32.10. Validierungsprotokoll

Gibt an, wer und wann die Gültigkeit der Erfassung bewertet wurde.

33. MINIMALE INHALTE DES SCREENSHOTS

Soweit relevant, sollte der visuelle Nachweis Folgendes zeigen:

- KI-Produkt oder Oberflächenidentität
- Völliger Prompt
- Beginn der Antwort
- Ende der Antwort
- Referenzfelder
- Ablehnungs- oder Fehlermeldung
- Offensichtliches Modelllabel
- Nachweis, dass die Sitzung neu oder in einem relevanten Zustand ist
- Materialsystemeinstellung
- Schnittstellenstatus, der anzeigt, dass die Antwort nicht abgeschnitten wurde

34. LANGE ANTWORTEN

Wenn die Antwort nicht auf eine einzelne Bildschirmseite passt:

- Vollseitenscreenshot,
- scrollbarer Screenshot,
- überlappende aufeinanderfolgende Screenshots,
- Bildschirmaufnahme,
- direkter Textexport

können verwendet werden. Aufeinanderfolgende Darstellungen:

- sollten keine Lücken lassen,
- sollten sich überlappen,

sollten fortlaufende Nummerierungen tragen. Nur der positive oder relevante Abschnitt darf nicht beschnitten werden.

35. BILDSCHIRM-AUFNAHME

Bildschirmaufnahme:

- Prompt-Einreichung,
- Streaming-Antwort,
- Fertigstellungsstatus,
- kein Aktualisierungsvorgang durchgeführt

kann angezeigt werden. Allerdings:

- mehr persönliche Informationen,
- Benachrichtigung,
- private Anwendung,
- Konten-ID

können erfasst werden. Daher Bildschirmaufnahme:

- muss möglicherweise nicht standardmäßig erfolgen,
- kann in Hochrisiko- oder Verifizierungs-Teilmustern verwendet werden,

erfordert Datenschutzvorbereitung vor der Aufzeichnung.

36. ROHES ANTWORTTEXT

Zusätzlich zum Screenshot sollte der vollständige Antworttext separat gespeichert werden. Bevorzugte Reihenfolge:

- Offizieller Export oder Kopierfunktion des Produkts
- Direkte Texterfassung durch das Erfassungstool
- Barrierefreiheitsbaum oder DOM-basierte Extraktion
- Vom Nutzer kopierter Text
- OCR und menschliche Überprüfung

OCR sollte nur die letzte Option sein. Wenn ein wesentlicher Unterschied zwischen der automatischen Extraktion und der visuellen Darstellung besteht, ist eine Überprüfung erforderlich.

37. DYNAMISCHE ANTWORTEN

Einige Produkte können nach Abschluss der Antwort:

- eine Quelle hinzufügen,
- eine Quellenkarte laden,
- den Text umformatieren,

eine Sicherheitswarnung hinzufügen. Es können zwei Erfassungsmomente verwendet werden:

- Erste Abschluss-Erfassung
- Erfassung nach kurzer Render-Stabilität

Wenn ein wesentlicher Unterschied besteht, werden beide Versionen beibehalten. Die erste Nutzeranzeige wird nicht gelöscht.

38. WIE MAN VERSTEHT, DASS DIE ANTWORT ABGESCHLOSSEN IST?

Indikatoren für den Abschluss eines Kandidaten:

- Streaming stoppt
- Ladeanzeige schließt
- Schaltfläche "Regenerieren" oder ähnliches erscheint
- Produktsignale der Fertigstellung
- Vordefinierte Ruhezeit
- API-Fertigstellungsaufzeichnung

Produktorientierte Fertigstellungsregeln müssen im KI-Systemregister gefunden werden.

39. WENN DER NUTZER DIE ANTWORT GESTOPPT HAT

Der Nutzer könnte versehentlich oder absichtlich:

- stoppen,
- abbrechen,
- die Anwendung beenden

Durchgeführt haben. Diese Situation wird aufgezeichnet als:

BENUTZER_UNTERBROCHEN

VERSEHENTLICHE_UNTERBRECHUNG

UNBEKANNTE UNTERBRECHUNG

Eine vom Nutzer gestoppte Antwort kann nicht als vollständige Systemausgabe betrachtet werden. Die Regel für Teilnehmerersatz oder Wiederholung muss im Voraus definiert werden.

40. DAS PROMPT UND DIE ANTWORT SIND IN GETRENNTEN DATEIEN

Wenn das Prompt und die Antwort in separaten Bildern vorliegen:

- sie müssen mit derselben Sitzung verknüpft sein,
- mit derselben Beobachtungs-ID,
- und mit einer ununterbrochenen Zeitkette

Andernfalls ist es möglicherweise nicht möglich zu überprüfen, dass die Prompt und die Antwort tatsächlich zu derselben Unterhaltung gehören.

41. DATEI-INTEGRITÄT

Für jede Datei muss ein kryptografischer Hashwert erzeugt werden. [K25] Kandidat:

```
h_j= H(file_j)
```

Paket-Hash:

```
h_bundle= H(JCS(manifest(file_id, Reihenfolge, Größe, Algorithmus, h_j))) [K14; K25]
```

kann erstellt werden. Hierbei ist H ein veröffentlichter und als sicher geltender Hash-Algorithmus. Der Algorithmus:

- versioniert,
- offen,
- kann in Zukunft bei Bedarf geändert werden.

muss sein.

42. WAS BEWEIST EIN HASH?

Wenn der aufgezeichnete Hash-Wert zuverlässig gespeichert wird, hilft das erneute Hashen derselben Datei mit demselben Algorithmus, Änderungen in der Byte-Sequenz [K25] zu erkennen. Dieser Vergleich unterstützt die Behauptung, dass die Datei nach der Generierung des Referenz-Hash die gleiche Byte-Sequenz beibehalten hat; er beweist jedoch nicht die Quelle der Datei, dass sie zur richtigen Zeit erzeugt wurde oder ihre Authentizität vor dem Hashen. Ein Hash allein beweist nicht:

- Dass die Datei vom echten KI-Produkt stammt
- Dass sie dem richtigen Nutzer gehört
- Dass sie zur richtigen Zeit erzeugt wurde
- Dass sie vor dem Hashen nicht organisiert wurde
- Dass es die erste Ausgabe ist
- Dass der Prompt korrekt war

Daher, Hash:

> *Es ist ein Nachweis der Integrität.*

Allein:

> *Es ist kein Nachweis der Originalität.*

43. DIGITALE SIGNATUR

Erfassungstool oder autorisierter Prüfer:

- das Beobachtungspaket,
- Zeitaufzeichnung,
- Fahrzeugversion

kann digital signieren. Unterschrift:

- welches Fahrzeug oder welche Institution das Paket erzeugt hat,
- dass es nach der Unterschrift nicht verändert wurde

stärkt. Der Unterschrifteninhaber könnte falsche oder gefälschte Daten unterschrieben haben. Daher garantiert die Unterschrift allein nicht die Authentizität.

44. ZUVERLÄSSIGE ZEIT [K26]

Lokale Geräteeinstellung der Uhr:

- falsch konfiguriert,
- manuell geändert,
- außer Synchronisation

kann auftreten. Ein starker Zeitstempel sollte mehrere Signale tragen:

- Aufnahmezeitpunkt der Serverakzeptanz
- Zuverlässiger UTC Zeitdienst
- Gerätezeit und Abweichungsaufzeichnung
- Sichtbare Zeit im KI-Produkt, falls vorhanden
- Dateierstellungszeit
- Digitale Zeitstempel

45. ZEITABWEICHUNG

Die Differenz zwischen der Gerätezeit des Teilnehmers und der zuverlässigen UTC-Zeit:

```
O_i^clock= t_i^device- t_i^reference
```

kann aufgezeichnet werden. Wenn eine wesentliche Abweichung besteht:

- dieser Zeit wird die Serverzeit als Hauptreferenz verwendet,
- die Gerätezeit wird als korrigierter Eintrag beibehalten,

Unsicherheit wird markiert. Dem Teilnehmer sollte nicht aufgefordert werden, die Geräteuhr manuell zu ändern.

46. UPLOAD-ZEIT IST NICHT DIE AUFNAHMEZEIT

Das Hochladen einer Datei auf den Server bei 14:00 beweist nicht, dass der Screenshot bei 14:00 erstellt wurde. Daher:

- Aufnahmezeit,
- Hash-Erstellungszeit,

- Upload-Zeit

sollte getrennt aufgezeichnet werden. Lange Upload-Verzögerungen können das Vertrauen in die Erfassung verringern. Der Datensatz muss nicht automatisch ungültig sein.

47. VERZÖGERTER UPLOAD

Aufgrund von Verbindungsproblemen kann es sein, dass der Nutzer die Nachweise nicht sofort hochladen kann. Wenn ein verspäteter Upload akzeptiert wird:

- lokaler Hash zum Zeitpunkt der Erfassung,
- vertrauenswürdige Gerätezeit,
- offline signierter Datensatz,
- Grund für die Upload-Verzögerung

können erforderlich sein. Nur die Aussage des Nutzers: „Ich habe dieses Bild gestern gemacht.“ ist für die Hauptstufe NCL-3 nicht ausreichend.

48. Evidenzkette

Die Evidenzkette einer Beobachtung als Kandidat besteht aus den folgenden Phasen:

- Aufgabenzuweisung
- Festschreibung des Prompts
- Nutzer- und Produkttauglichkeit
- Slot-Öffnung
- Prompt-Einreichung
- Ergebnis des Erstkontakts
- Abschluss der Antwort
- Erfassung
- Lokaler Integritätsdatensatz
- Sicheres Hochladen
- Server-Akzeptanzzeit
- Paket-Hash
- Protokollüberprüfung
- Bearbeitung
- Beurteilung
- Archiv
- Öffentliches Manifest

Jede Phase muss mit der vorherigen Phase verbunden sein.

49. ZERREIßEN DER KETTE

Die Evidenzkette kann in folgenden Situationen unterbrochen werden:

- Kein Aufforderungsprotokoll
- Die anfängliche Ausgabe kann nicht überprüft werden
- Die Antwort stimmt nicht mit der visuellen Darstellung überein

- Erfassungszeit ist unbekannt
- Datei wurde später geändert
- Doppelte Teilnehmer-ID
- Produkt und Plan sind unbekannt
- Die Beobachtung liegt außerhalb der Welle
- Der Hash stimmt nicht mit der Datei überein
- Bearbeitung hat die Rohdatei überschrieben
- Es ist unbekannt, wer es überprüft hat

Die Art der Unterbrechung bestimmt den Gültigkeitsstatus.

50. ROHDATEN KÖNNEN NICHT GEÄNDERT WERDEN; DIE ENTSCHEIDUNG KANN SICH ÄNDERN

Rohdatenpaket:

- soll nicht überschrieben werden,
- soll nicht stillschweigend korrigiert werden,

soll nicht gelöscht und durch ein neues ersetzt werden. Bezüglich der Beobachtung jedoch:

- Gültigkeit,
- tscheidung,
- Klassifikation,
- Compliance

die Entscheidung kann sich ändern. Die neue Entscheidung:

- neue Version,
- Begründung,
- Datum,
- Entscheidungsträger

sollte tragen. Diese Unterscheidung ist eine grundlegende Governance-Regel:

> *Nachweise können unverändert bleiben; die Interpretation kann sich mit den Nachweisen weiterentwickeln.*

51. SCHWARZUNG

Die folgenden Felder können in der öffentlichen oder der Schiedskopie geschwärzt werden:

- Name und Nachname
- E-Mail
- Profilbild
- Konten-ID
- Private Gespräche
- Benachrichtigungen
- Unternehmensfeld
- Sensibler Dateiname
- Exakter Standort

Schwärzung:

- sollte nicht an der Rohdatei durchgeführt werden,
- sollte eine neue abgeleitete Datei erstellen,
- sollte einen eigenen Hash tragen,

sollte aufzeichnen, welches Feld geschwärzt wurde und warum.

52. DREI EBENEN VON NACHWEISEN

52.1. Roh eingeschränktes Evidenzmaterial

Es handelt sich um vollständige Rohdaten als Evidenzmaterial. Nur autorisierte Prüfer und erforderliche Forschungsrollen können darauf zugreifen.

52.2. Audit-Evidenzschicht

Enthält die Nachweise, die für unabhängige Schiedsrichter und Reproduktionsteams notwendig sind. Persönliche Bereiche werden minimiert.

52.3. Öffentliches Evidenzmittel-Manifest

Für die Öffentlichkeit:

- Beobachter-ID,
- Land- und Sprachzelle,
- KI-Produkt,
- Zeit,
- Nachweis-Hash,
- Gültigkeitsstatus,
- Schwärzungsstatus

werden dargestellt. Rohes privates Material wird nicht unbedingt veröffentlicht.

53. ÖFFENTLICHES EVIDENZ-MANIFEST

Für jede Welle kann das öffentliche Manifest die folgenden Felder enthalten:

- Wave-ID
- Protokollversion
- Anzahl der Produkte
- Ziel- und gültige Beobachtungsanzahl
- Abdeckung von Ländern und Sprachen
- Slot-Abdeckung
- Erfassungs-Sicherheitsgrade
- Technische Fehler- und Ablehnungsraten
- Anzahl der ungültigen Beobachtungen
- Hashwerte des Beobachtungspakets
- Wellenintegritäts-Wurzel
- Änderungsereignisse

- Öffentlich redigierte Beispiele
- Unabhängiges Zugriffsverfahren
- Status der Evidenzerhaltung

54. WELLE INTEGRITÄTS-WURZEL

KANDIDATENKONZEPT

Sei der Hash jedes Beobachtungspakets hi. Daraus kann die Wurzel der Welle mit einem Merkle-Baum oder einer gleichwertigen Struktur gebildet werden:

RW

kann erzeugt werden. Die Wellenwurzel kann nach Abschluss der Messung veröffentlicht werden. Diese Methode:

- erhöht die Sichtbarkeit von späteren stillen Änderungen der Beobachtungsliste in der Welle,
- Hinzufügen einer neuen Beobachtung,
- Ändern einer bestehenden Beobachtung

macht sie sichtbarer. Merkle-Wurzel:

- beweist allein nicht, dass die Beobachtungen real sind,
- oder dass die Entscheidung der Schiedsinstanz korrekt ist

Es stärkt die Integrität der Liste.

55. HINZUFÜGEN UND ENTFERNEN VON BEOBACHTUNGEN

Nachdem die Welle geschlossen ist:

- Hinzufügen einer neuen Beobachtung,
- eine bestehende Beobachtung entfernen,
- die Datei ändern

erfordert eine neue Manifestversion. Die alte Integritätswurzel wird beibehalten. Für die zu entfernende Beobachtung:

- Grund,
- Datum,
- Entscheidungsträger
- Auswirkung auf alten und neuen Score

aufgezeichnet.

56. ANTRAG AUF TEILNEHMER-RÜCKZUG

Forschungsethik oder der Datenschutzrahmen können es dem Teilnehmer erlauben, seine Daten zurückzuziehen. In diesem Fall:

- kann Rohdatenmaterial sicher gelöscht oder unzugänglich gemacht werden,
- ein leerer WITHDRAWN-Eintrag kann im öffentlichen Manifest beibehalten werden,

- kann der Welle-Score mit der neuen Version neu berechnet werden,

Das alte Ergebnis und der Grund für die Änderung werden beibehalten. Unveränderlichkeit bedeutet nicht, die Menschenrechte abzuschaffen.

57. GÜLTIGKEITSSTATUSSE DER BEOBACHTUNGSERFASSUNG

CV-0— NICHT ÜBERPRÜFT

Die Erfassung wurde noch nicht bewertet.

CV-1— GÜLTIG

Erfüllt alle obligatorischen Erfassungs- und Zeitbedingungen.

CV-2— BEDINGT GÜLTIG

Es besteht ein begrenzter Mangel. Es wird angegeben, in welchen Analysen es verwendet werden kann.

CV-3— TEILBEWEIS

Die Beobachtung kann als Ereignis verwendet werden. Sie darf nicht in die Hauptpopulationswertung einfließen.

CV-4— AUSSERHALB DER WELLE

Die Prompt wurde außerhalb der erlaubten Zeitstruktur gesendet.

CV-5— ÜBEREINSTIMMUNG VON EINGABE ODER AUSGABE NICHT ERREICHT

Es gibt eine wesentliche Diskrepanz zwischen dem zugewiesenen Prompt, dem eingereichten Prompt oder der aufgezeichneten Antwort.

CV-6— DOPPELT

Die gleiche Beobachtung oder Nutzeraufzeichnung wurde wiederholt.

CV-7— MANIPULATION VERDACHT

Es besteht ein ungelöstes Zweifel an der Integrität der Datei oder der Metadaten.

CV-8— ERFUNDEN ODER MANIPULIERT

Die Beobachtung ist keine echte Nutzererfahrung oder wurde absichtlich verändert.

CV-9— ZURÜCKGEZOGEN

Sie wurde aufgrund von Teilnehmer- oder ethischen Prozessen von der Analyse ausgeschlossen.

58. DIE GÜLTIGKEIT DER AUFZEICHNUNG SOLLTE VON DER SEMANTISCHEN GENAUIGKEIT GETRENNT WERDEN

Eine Beobachtung:

- vollständig gültig in Bezug auf die Aufzeichnung,
- stark falsch in Bezug auf den Inhalt

kann sein. Eine weitere Beobachtung:

- korrekte Antwort,
- kann fehlende oder manipulierte Nachweise enthalten

Daher gibt es zwei getrennte Entscheidungen:

- Erfassungsvalidität
- Semantische Beurteilung

Der Erfassungsvalidator sollte so wenig wie möglich über den Erfolgspunktwert der Antwort wissen.

59. ROLLE DES ERFASSUNGSVALIDATORS

Der Erfassungsvalidator überprüft:

- Prompt-Abgleich
- Erhebungswelle und Zeitfenster
- Produktidentität
- Erste Ausgaberegel
- Bildschirmintegrität
- Rohtext-Konsistenz
- Technisches Wiederholen
- Hash und Zeit
- Datenschutz

Entscheidet sich nicht für: Wurde das Unternehmen korrekt beschrieben? Unterstützt die Zitation die Behauptung? Gibt es einen kritischen Fehler? Geht die Antwort durch? Dies sind die Aufgaben der nächsten Überprüfungsebene.

60. WIEDERHOLTES BILDSCHIRMBILD

Das gleiche Bild:

- versehentlich zweimal hochgeladen,
- von mehreren Nutzern genutzt,
- Wiederholt unter den Panelanbietern

kann passieren. Prüfungen:

- Datei-Hash
- Visuelle Ähnlichkeit
- Rohantworttext

- Zeit- und Sitzungsmetadaten
- Nutzereinzigartigkeit

Das ist möglich. Dasselbe KI-Produkt kann einigen Nutzern exakt dieselbe Antwort Wort für Wort geben. Der gleiche Text allein ist kein Nachweis für Betrug. Die gleiche Datei oder dieselbe visuelle Spur ist ein stärkeres Signal.

61. DATEIÄHNLICHKEIT UND FALSCH-POSITIV

Gleiches Gerät und gleiche Schnittstelle:

- ähnliche Visuals,
- gleiche Bildschirmgröße,
- gleiche Kompression

kann erzeugen. Der Algorithmus zur visuellen Ähnlichkeit sollte keine automatische Ausschlussfunktion ausführen. Mehrere Signale und menschliche Überprüfung sind erforderlich.

62. PROVIDER-TELEMETRIE

KI-Anbieter:

- Sitzungszeit,
- Modell,
- Prompt,
- Antwort,
- Werkzeugnutzung

kann zuverlässige Protokolle dazu bereitstellen. Dieser Eintrag stärkt die unabhängige Erfassung. Allerdings kann der Anbieter:

- gemessene Partei,
- Systembesitzer,
- Begünstigter

sein. Anbieter-Telemetrie: sollte zusammen mit, nicht als Ersatz für, unabhängige Nutzerbeweise verwendet werden.

63. FALLS KEINE ANBIETERBEWEISE VORLIEGEN

Der Zugriff auf Anbieterprotokolle ist in den meisten Live-Consumer-Produkten möglicherweise nicht verfügbar. Dies macht GEO-1000 jedoch nicht unmöglich. Die folgende Zusammensetzung kann verwendet werden:

- Instrumentierte Nutzererfassung
- Server-Zeitstempel
- Nachweis des Vollbilds
- Rohantwort
- Systemeinstellung
- Hash und Signatur
- Unabhängige Verifikations-Subprobe

Der Vertrauensgrad des Nachweises wird entsprechend angegeben.

64. ERFASSUNG IN Panel unter natürlichen Nutzungsbedingungen

Erfassungstool im Naturpanel:

- Inhalt der speziellen Anweisung,
- vorherige Gespräche,
- Kontenidentität

sollte nicht unnötig erfasst werden. Bevorzugter Ablauf: Der Nutzer öffnet einen neuen Natural-Account-Chat. Capture zeichnet nur das Aufgabenfenster auf. Der Personalisierungsstatus wird separat als hochrangige Metadaten erfasst. Private Historie bleibt außerhalb der Erfassung. Öffentliche Kopie wird bearbeitet.

65. NATÜRLICHE KONVERSATION-STATUS-ERFASSUNG

Wenn der aktuelle Gesprächskontext gemessen wird:

- nur der notwendige vorangegangene Gesprächsbereich,
- explizite Zustimmung,
- starke Schwärzung,
- Zugriffsbeschränkung

ist erforderlich. Alle privaten Chat-Historien dürfen standardmäßig nicht gesammelt werden.

66. MOBILE ERFASSUNG

Auf mobilen Geräten:

- Vollbild-Erfassung,
- scrollbarer Screenshot,
- Anwendungsversion,
- Systemzeitprotokoll

kann anders funktionieren. Mobiles Protokoll:

- iOS,
- Android,
- andere Systeme

können separate technische Anleitungen enthalten. Mobile Nutzer können nicht ausgeschlossen werden, da das Erfassen auf Desktop-Geräten schwierig ist. Das Erfassungstool muss sich an die Nutzerpopulation anpassen.

67. ZUGÄNGLICHKEIT UND ERFASSUNG

Für Teilnehmer, die Screenreader, Spracheingabe oder motorische Hilfsmittel verwenden, kann eine standardmäßige visuelle Erfassung allein nicht ausreichen. Alternative Nachweise:

- Export des Barrierefreiheitsbaums
- Aufzeichnung der Sprachsitzung
- Anwendungs-Transkript
- Protokoll der assistiven Technologie
- Aufgabenbeobachtungsprotokoll

können auftreten. Diese Beobachtungen sollten nicht als wenig zuverlässig angesehen werden. Ein gleichwertiger Nachweisstandard sollte entwickelt werden.

68. BEOBACHTUNG DES ERFASSUNGSTOOLS BEI VERÄNDERUNG DES VERHALTENS

Browser-Erweiterung oder Instrumentierung:

- Seitenlade,
- Schnittstellenverhalten,
- Sicherheitssystem,
- Produkterkennung

kann beeinträchtigt werden. Die Auswirkungen des Erfassungswerkzeugs sollten im Piloten getestet werden. Wenn Sitzungen mit und ohne Werkzeug sich unterschiedlich verhalten: wird die Bedingung des Werkzeugs zu einem experimentellen Faktor, das Erfassungswerkzeug kann nicht als stillschweigend neutral angenommen werden.

69. ERFASSUNGSWERKZEUG VERSIONIERUNG

Jedes Erfassungspaket:

- Werkzeugname,
- Version,
- Konfiguration,
- Betriebsumgebung,
- bekannte Fehler

muss enthalten. Wenn sich das Erfassungswerkzeug während der Messung ändert:

- neue Werkzeugversion,
- Subwelle,
- Kompatibilitätsprüfung

kann erforderlich sein.

70. ERFASSUNGSWERKZEUGFEHLER

Werkzeug:

- kann die Prompt möglicherweise falsch kopieren.
- kann das Ende der Antwort abschneiden,
- kann die Zeit falsch aufzeichnen,
- kann keinen Hash erzeugen,

kann die Datei komprimieren. Fehler des Tools:

- kann nicht in das KI-Produkt geladen werden,
- oder der Nutzer

Separat:

CAPTURE_TOOL_FEHLER

muss aufgezeichnet werden als.

71. SYNTHETISCHER DREI-WELLEN-FALL GEO-1000

DEMONSTRATION DER SYNTHETISCHEN METHODOLOGIE / Die folgenden Produkte, Daten, Zahlen und Ergebnisse sind vollständig fiktiv. Zehn synthetische KI-Produkte:

A1–A10

sollen in drei Hauptwellen gemessen werden.

71.1. Wellenkalender

Welle	Typ	UTC Lieferfenster	Zweck
W1	Hauptperson	12.00-13.00	Beginn
W2	Bestätigend	20.00-21.00	Kurzfristige Wiederholung
W3	Stabilität	04.00-05.00	Mittelfristige und stündliche Balance

Für jedes Produkt in jeder Welle: 1,000 gültige Hauptbeobachtungen sind angestrebt. Gesamtziel:

```
10×1,000×3=30,000
```

71.2. Slot-Struktur

Jede Welle:

- 12 Mikro-Slots
- mit Teilnehmern pro Slot nach Produkt und Land ausgeglichen

Der Slot des Teilnehmers: wird vor der Datenerhebung zufällig zugewiesen.

71.3. W1 Ergebnis

Ziel: 10,000 gültige Hauptbeobachtungen Feld:

- 13,240 Einladungen
- 10,870 abgeschlossene Aufgaben
- 10,106 gültige Beobachtungen in Bezug auf Erfassung
- Hauptanalysebeobachtungen von 10,000 gemäß zuvor festgeschriebener Haupt-/Backup-Reihenfolge

Verbleibende gültige Datensätze von 106:

- Backup,

- Empfindlichkeit,
- Erfassung Qualitätskontrolle

sind reserviert für.

71.4. Technische Ereignisse

118 Nutzungslimit 42 Verbindungsfehler 31 Systemablehnung 17 Nutzertrennung 9 Aufforderungsabweichung 6 Wiederholung Screenshot Alle Ergebnisse des Erstkontakts wurden gespeichert. Nur vordefinierte echte technische Fehler wurden erneut versucht.

71.5. Produktänderung während W2

A7-Produkt bei 20.32 UTC:

- automatische Webnutzung,
- neue Referenzansicht

lass es gewinnen. Beobachtungen von A7:

W2A: 20.00–20.31

W2B: 20.32–21.00

werden wie folgt getrennt. Ein einziger fester W2-Wert für A7 wird ohne Erklärung nicht veröffentlicht.

71.6. Externes Ereignis in W3

Eine wichtige rechtliche Entscheidung sollte am 04.22 UTC über das geprüfte synthetische Unternehmen veröffentlicht werden. Es sollte beobachtet werden, dass sich die Antworten in webbasierten Produkten nach 04.25 geändert haben. Ereignis: Es wird in das Register für externe Ereignisse aufgenommen. Ergebnisse vor und nach dem Ereignis werden als separate diagnostische Zellen gezeigt. Beobachtungen werden nicht gelöscht, da sie den Wert verringern oder erhöhen.

71.7. Wave Health Card

Welle	Gesundheit	Hauptnote
W1	WH-4	Ausgewogene und starke Erfassung
W2	WH-3	Sub-Welle aufgrund der A7-Produktänderung
W3	WH-3	Externer Rechtsereignis-Eintrag

Alle drei Wellen können verwendet werden. A7- und externe Ereignis-Kommentare sind ebenfalls eingeschränkt.

72. SYNTHETISCHER EINZELBEOBACHTUNGSERFASSUNGSFALL

SYNTHETISCHES BEISPIEL

Teilnehmer:

- Türkei
- Türkisch

- Orion AI Web kostenpflichtiger Plan
- Kontrolliertes Panel unter bereinigten Bedingungen

W1

Slot 12.15–12.20 UTC Zugewiesene Aufforderung: „Mit welchem Mutterunternehmen ist Apple.com verbunden und was sind die Haupttätigkeiten dieses Unternehmens?“

72.1. Zeiten

Aufgabeneröffnung: 12.14.20 Eingabebestätigung: 12.16.08 Erstes Token: 12.16.10 Antwortabschluss: 12.16.24 Erfassung: 12.16.39 Hash-Erzeugung: 12.16.41 Hochladen: 12.17.12 Serverannahme: 12.17.13

72.2. Nachweis

Prompt-Hash stimmt überein Vollständiger Antworttext verfügbar Zwei sich überlappende Screenshots verfügbar Produkt-, Plan- und Modellkennzeichnung sichtbar Speicher- und spezielle Anweisungen aus Webnutzung aktiviert Zitation nicht angezeigt Keine Reproduktion Kein vorheriges Gespräch Dateihashes und Paket-Signatur verfügbar Erfassungsebene:

NCL-4

Gültigkeit:

CV-1

Es wurde noch nicht entschieden, ob die Antwort korrekt oder inkorrekt ist.

73. SYNTHETISCHER TECHNISCHER WIEDERHOLUNGSFALL

Beim ersten Versuch:

- Prompt korrekt gesendet
- Produkt zeigte "Netzwerkfehler"
- Sachliche Antwort nicht begonnen
- Fehler-Screenshot gemacht
- Zeit und Hash wurden aufgezeichnet

Gemäß der vordefinierten Regel wurde innerhalb von fünf Minuten ein zweiter Versuch unternommen. Beim zweiten Versuch erfolgte eine Antwort. Korrekte Aufzeichnung:

- Ergebnis des ersten Kontakts: VORÜBERGEHENDER_TECHNISCHER_FEHLER
- Erste gültige Inhaltsausgabe: zweiter Versuch
- Beide Versuche sind im Paket enthalten
- Erster Vorfall ist in der technischen Fehlerquote enthalten
- Die zweite Antwort wird in die Inhaltsbewertung einbezogen, wenn die Methode dies im Voraus vorhersagt.

Falscher Eintrag: Der erste Fehler wird gelöscht und es wird so behandelt, als ob nur die zweite Antwort existierte.

74. SYNTHETISCHES UNRichtiges RETRY-FALL

Erste Antwort: „Apple ist ein lokales Unternehmen, das nur Handys verkauft.“ Dem Teilnehmer gefällt die Antwort nicht. Er gibt sie erneut ein. Die zweite Antwort stellt sich als korrekt heraus. Der Teilnehmer reicht nur die zweite Antwort ein. Das Systemprotokoll oder die Aufzeichnungsdatei zeigt die Wiederholung. Korrekt-Status:

CV-8— MANIPULIERTE BEOBACHTUNG

Die Beobachtung wird nicht gespeichert, nur weil die zweite Antwort korrekt ist. Wenn die erste Antwort zusätzlich wiederhergestellt werden kann, kann sie in der Vorfalluntersuchung bewahrt werden. Der Hauptteilnehmerdatensatz ist ungültig.

75. WELLE ABSCHLUSSBERICHT

Am Ende jeder Welle sollte der folgende Datensatz erstellt werden:

- Wave-ID
- Öffnungs- und Schließzeit
- Zielanzahl der Beobachtungen
- Eingeladene Person
- Abgeschlossene Aufgabe
- Gültige Erfassung
- Beobachtung, die in die Hauptanalyse eingeht
- Backup gültige Beobachtung
- Technischer Fehler
- Ablehnung
- Nutzungslimit
- Prompt-Abweichung
- Erfassungsabweichung
- Verspäteter Upload
- Duplikat
- Vermutete Manipulation
- Teilnehmer-Rückzug
- Produktänderungen
- Externe Ereignisse
- Slot-Abdeckung
- Abdeckung von Ländern und Sprachen
- Erfassungs-Sicherheitsgrade
- Gesundheitsstatus der Welle
- Integritätswurzel
- Verantwortliche menschliche Genehmigung

76. VERPFLICHTENDE NORMATIVE BESTIMMUNGEN

CH12-N01

Jede GEO-1000-Beobachtung muss mit einer vorab festgeschriebenen Messwelle verknüpft sein.

CH12-N02

Der Begriff "gleiche Zeit" kann nicht ohne das vollständige UTC-Fenster, Mikro-Zeitfenster und die Abschlussregeln verwendet werden.

CH12-N03

Ortszeit kann nicht allein als Wellzeit verwendet werden; der Hauptzeitstandard muss UTC sein.

CH12-N04

Das Übertragungsfenster, Mikro-Slots, Fertigstellungszeit und Ladezeit müssen bestimmt werden, bevor KI-Antworten angezeigt werden.

CH12-N05

Die Slot-Zuweisung muss so erfolgen, dass Länder, Sprachen, Produkte und Panelgruppen zeitlich nicht unausgewogen bleiben.

CH12-N06

Teilnehmer dürfen nicht basierend auf ihren Ergebnissen auf frühere oder spätere Slots verschoben werden.

CH12-N07

Der lokale Zeitversatz des einzelnen UTC-Fensters muss im öffentlichen Methodenprotokoll dokumentiert werden.

CH12-N08

Zurückkehrende globale Wellen oder lokal zeitlich ausgeglichene Operationen müssen als separate Zeitresultate geführt werden.

CH12-N09

Jede Beobachtungsaufforderung muss, soweit relevant, Einreichungs-, Erstkontakt-, Abschluss-, Erfassungs- und Upload-Zeiten enthalten.

CH12-N10

Unbekannte Zeitfelder können nicht als geschätzte tatsächliche Zeiten dargestellt werden.

CH12-N11

Die Gerätuhr allein kann nicht als zuverlässiger UTC Zeitnachweis betrachtet werden.

CH12-N12

Erfassungszeit und Hochladezeit müssen getrennt aufgezeichnet werden.

CH12-N13

Beobachtungen, die außerhalb der Welle gesendet wurden, dürfen nicht ohne Erklärung in die Hauptwelle aufgenommen werden.

CH12-N14

Ausschlussregeln für Verzögerung oder Slot-Abweichung müssen unabhängig vom Inhalts der Antwort sein.

CH12-N15

Als Ergebnis des ersten Kontakts muss es getrennt von der ersten Inhaltserstellung aufbewahrt werden.

CH12-N16

Ablehnung, Nutzungslimit, technischer Fehler, Klärung und Unklarheit dürfen nicht in dieselbe Kategorie eingeordnet werden.

CH12-N17

Falsche, negative, referenzlose oder kritische KI-Antwort kann kein Grund für einen erneuten Versuch sein.

CH12-N18

Technischer erneuter Versuch kann nur unter vordefinierten technischen Bedingungen durchgeführt werden.

CH12-N19

Wenn ein technischer erneuter Versuch durchgeführt wird, müssen alle Versuche und das Ergebnis des ersten Kontakts erhalten bleiben.

CH12-N20

Teilweise oder unvollständige Antworten dürfen nicht ignoriert und durch eine neue Antwort ersetzt werden.

CH12-N21

Die erste geeignete Ausgabe kann nicht erneuert, ausgewählt oder manuell verbessert werden.

CH12-N22

Jede Beobachtung muss mindestens auf NCL-3 oder einem gerechtfertigten gleichwertigen Erfassungs-Sicherheitsniveau abzielen.

CH12-N23

Ein Screenshot allein kann nicht als vollständiges Evidenzpaket betrachtet werden.

CH12-N24

Die visuelle Evidenzaufforderung sollte die gesamte Antwort und die Oberfläche des physischen Produkts zeigen.

CH12-N25

Lange Antworten müssen vollständig und in der richtigen Reihenfolge mittels der Erfassungsmethode aufbewahrt werden.

CH12-N26

Wenn es einen materiellen Unterschied zwischen dem Rohtext der Antwort und dem visuellen Nachweis gibt, sollte eine Untersuchung durchgeführt werden.

CH12-N27

OCR kann als Hilfsmethode verwendet werden, wenn eine direkte Texterfassung nicht möglich ist; ohne Verifizierung kann es nicht als endgültige Textquelle betrachtet werden.

CH12-N28

Die anfängliche Fertigstellung und die nachfolgenden materiellen Versionen dynamisch sich ändernder Antworten müssen separat aufbewahrt werden.

CH12-N29

Der Teilnehmer, der die Antwort stoppt oder die Schnittstelle schließt, muss als separates Ereignis protokolliert werden.

CH12-N30

Wenn das Prompt und die Antwort in getrennten Dateien vorliegen, müssen sie mit derselben Sitzung und Zeitschiene verknüpft werden.

CH12-N31

Jede Evidenzdatei und jedes Evidenzpaket muss einen Integritätshash tragen.

CH12-N32

Der Hash allein kann nicht als Beweis für die Realität der Beobachtung vorgelegt werden.

CH12-N33

Digitale Signatur und Zeitstempel sind nicht die Gesamtheit der Authentizität der Beobachtung; sie sind Komponenten der Evidenzkette.

CH12-N34

Keine Schwärzung oder Korrektur darf über die Rohbeweisdatei geschrieben werden.

CH12-N35

Jedes bearbeitete Derivat muss seine eigene Datei-ID und seinen eigenen Hash tragen.

CH12-N36

Der Zugriff auf Rohbeweise kann eingeschränkt sein; das öffentliche Manifest muss die Integrität der Beobachtung nachweisen, ohne Identitäten preiszugeben.

CH12-N37

Der Name des Teilnehmers, die E-Mail, das Passwort, private Gespräche oder unnötige persönliche Informationen dürfen nicht zu den öffentlichen Evidenzmitteln hinzugefügt werden.

CH12-N38

Das Risiko der Re-Identifizierung für kleine Länder und Nutzer seltener Untergruppen sollte ebenfalls bewertet werden.

CH12-N39

Die Panel unter natürlichen Nutzungsbedingungen-Erfassung sollte standardmäßig keine speziellen Anweisungen und Inhalte vergangener Gespräche sammeln.

CH12-N40

Äquivalente alternative Erfassungsmethoden sollten für Nutzer mit Zugänglichkeitsbedürfnissen bereitgestellt werden.

CH12-N41

Wenn das Erfassungstool die Nutzererfahrung wesentlich verändert, sollte diese Auswirkung separat gemessen werden.

CH12-N42

Das Erfassungstool und seine Version sollten in jeder Beobachtung aufgezeichnet werden.

CH12-N43

Wenn das Erfassungstool während einer Erhebungsphase wesentlich verändert wird, sollte eine Unterscheidung für Unterphase oder Tool-Version vorgenommen werden.

CH12-N44

Der Fehler des Erfassungstools kann weder als ein KI-Produkt noch als ein Fehler des Teilnehmers klassifiziert werden.

CH12-N45

Die Telemetrie des Anbieters kann die unabhängige Erfassung von Nutzern unterstützen; sie kann diese nicht ohne Erklärung ersetzen.

CH12-N46

Wenn während der Welle eine wesentliche Änderung eines KI-Produkts auftritt, können Beobachtungen nicht zu einem einzigen festen Produktzustand zusammengeführt werden.

CH12-N47

Produktänderungen erfordern Wellentrennung, Neustart, explizite Modellierung oder eine Warnung vor gemischtem Zustand.

CH12-N48

Das Fehlen einer Anbieter-Update-Ankündigung ist kein endgültiger Beweis dafür, dass sich das System nicht geändert hat.

CH12-N49

Wesentliche externe Ereignisse sollten mit Zeitstempel und Umfang im Register für externe Ereignisse hinzugefügt werden.

CH12-N50

Externe Ereignisse können nach der Veröffentlichung der Ergebnisse nicht als Grund für einen Ausschluss dienen, nur weil sie die Bewertung erhöht oder gesenkt haben.

CH12-N51

Die Gültigkeit der Erfassung sollte so blind wie möglich bewertet werden, bevor die semantische Korrektheit der Antwort beurteilt wird.

CH12-N52

Eine korrekte KI-Antwort kann eine fehlende oder manipulierte Erfassung nicht validieren.

CH12-N53

Eine falsche KI-Antwort kann eine protokollkonforme Erfassung nicht entwerten.

CH12-N54

Duplikat- und Manipulationsentscheidungen sollten mehrere Signale und menschliche Überprüfung beinhalten.

CH12-N55

Der gleiche Antworttext allein kann nicht als Beweis für Duplikation oder Betrug betrachtet werden.

CH12-N56

Das Hinzufügen, Entfernen oder Ändern von Beobachtungen nach dem Abschluss einer Welle erfordert ein neues Manifest und eine neue Analyseversion.

CH12-N57

Das alte Wellenmanifest und die Integritätswurzel müssen erhalten bleiben.

CH12-N58

Rohbeweise können unverändert aufbewahrt werden; Validitäts- und Schiedsentscheidungen dürfen nur in versionierter Form geändert werden.

CH12-N59

Das Recht der Teilnehmer auf Rückzug und die Forschungsethik müssen über dem Anspruch der Unveränderlichkeit stehen.

CH12-N60

Die Integrität der Welle sollte unabhängig vom KI-Score bewertet werden.

CH12-N61

Der Bericht zum Wellenabschluss sollte Ziel, tatsächliche Werte, Ungültigmachungen, technische Ereignisse, Produktänderungen und Evidenzintegrität zeigen.

CH12-N62

NOMOS Die Einhaltung der Erfassung kann nicht ausschließlich an die Verwendung von Software gebunden werden, die von NobleJackal entwickelt wurde; gleichwertige Open-Source-Anwendungen sollten möglich sein.

CH12-N63

Das Erfassungsschema, die Validierungsentscheidungen und Wellenaufzeichnungen sollten in einem maschinenlesbaren Format versioniert werden.

CH12-N64

Jede Welle und Evidenzkette sollte einen verantwortlichen menschlichen oder institutionellen Eigentümer haben.

77. FORMEN DES VERSAGENS

CH12-F01— DER GLEICHE ZWEITE FEHLER

Alle Teilnehmer gleichzeitig senden zu lassen, wird als die einzige Bedingung für Gleichzeitigkeit betrachtet.

CH12-F02— „HEUTE“ ALS SYNCHRONISIERTE WELLE ZU ZÄHLEN

Ortszeit und Zeitzonen werden verwechselt.

CH12-F03— ORTSZEIT WIE UTC VERWENDEN

Beobachtungen werden in die falsche chronologische Reihenfolge gebracht.

CH12-F04— ZUORDNUNG DER ZEITSLOTS BASIEREND AUF ERGEBNISSEN ÄNDERN

Nutzer mit niedriger Punktzahl werden in eine andere Zeitzelle verschoben.

CH12-F05— EIN LAND IN EINEN SLOT PACKEN

Zeiteffekt wird mit Ländereffekt vermischt.

CH12-F06— KÜNSTLICHE LAST ERZEUGEN

Tausende Nutzer senden gleichzeitig, wodurch das abnormale Fehlverhalten des Produkts ausgelöst wird.

CH12-F07— NACHTTEILNAHME VERSTECKEN

Müdigkeit und Teilnahmeverlust werden in bestimmten Bereichen nicht gemeldet.

CH12-F08— HINZUFÜGEN VON OFF-WAVE-BEOBACHTUNG ZUR WELLE

Eine Antwort, die Stunden oder Tage später erhalten wird, wird so verwendet, als gehöre sie zum Hauptfenster.

CH12-F09— LADEZEIT ALS AUFNAHMZEIT ZÄHLEN

Es kann nicht nachgewiesen werden, wann die Datei erzeugt wurde.

CH12-F10— BLINDES VERTRAUEN AUF DIE GERÄTEUHR

Manuell geänderte oder falsche Zeit wird zum Hauptzeitaufzeichnungseintrag.

CH12-F11— AUSSCHLIEßLICH VERZÖGERTES NEGATIVES ANTWORTEN

Die Zeitregel wird abhängig vom Ergebnis angewendet.

CH12-F12— LÖSCHEN DES ERSTEN KONTAKTERGEBNISSES

Die Geschwindigkeitsbegrenzung, der Fehler oder die Ablehnung wird beim zweiten Versuch unsichtbar gemacht.

CH12-F13— ZÄHLUNG VON ABLEHNUNGEN ALS TECHNISCHER FEHLER

Das tatsächliche Systemverhalten wird erneut versucht.

CH12-F14— ZÄHLUNG FALSCHER ANTWORTEN ALS TECHNISCHER FEHLER

Regeneration wird durchgeführt, bis das gewünschte Ergebnis erreicht wird.

CH12-F15— TEILWEISE ANTWORT IGNORIEREN

Selbst wenn das System unterbrochen wird, wird nur die nachfolgende vollständige Antwort gespeichert.

CH12-F16— AUSWAHL DER BESTEN ANTWORT

Unter mehreren Antworten wird die positivste oder längste gemeldet.

CH12-F17— NUR SCREENSHOT

Die Visualisierung wird ohne Prompt, Produkt, Zeit und Metadaten als ausreichend betrachtet.

CH12-F18— ZUGESCHNITTENE ERFOLGSVISUALISIERUNG

Falsche oder einschränkende Teile bleiben unsichtbar.

CH12-F19— AUFNAHME ZEIGT DIE EINGABE NICHT

Es kann nicht nachgewiesen werden, zu welcher Frage die Antwort gehört.

CH12-F20— AM ENDE DER ANTWORT ABBRECHEN

Materialeinschränkung oder abgelehnter Teil wird unsichtbar.

CH12-F21— LÜCKE IN AUFEINANDERFOLGENDEN VISUALS

Einige Teile der Antwort gehen verloren.

CH12-F22— OCR-TEXT ALS ROHE ANTWORT ZÄHLEN

Erkennungsfehler ändern Materialangaben.

CH12-F23— NUR DIE NEUESTE VERSION DER DYNAMISCHEN ANTWORT ALS ECHT BETRACHTEN

Die anfängliche Nutzerdarstellung wird gelöscht.

CH12-F24— NUTZERUNTERBRECHUNG ALS VOLLSTÄNDIGE ANTWORT BETRACHTEN

Die Systemausgabe ist aufgrund einer Nutzerintervention unvollständig.

CH12-F25— SEPARATE EINGABE UND ANTWORT FALSCH ZUORDNEN

Dateien aus verschiedenen Chats werden derselben Beobachtung zugeordnet.

CH12-F26— HASH-VEREHRUNG

Jede Datei mit einem Hash wird als echt und original angesehen.

CH12-F27— PRE-HASH-BEARBEITUNG IGNORIEREN

Gekürzte oder bearbeitete Dateien werden durch Hashing legitimiert.

CH12-F28— DIE SIGNATUR ALS GÜLTIG BETRACHTEN

Signierter, aber falscher Datensatz wird ohne Streit akzeptiert.

CH12-F29— BEARBEITUNG DER ROHDATENDATEI

Originalbeweise werden unwiderruflich verändert.

CH12-F30— DAS BEARBEITETE DERIVAT ALS ROHBEWEIS BETRACHTEN

Die öffentliche Kopie wird wie die Originaldatei verwendet.

CH12-F31— PRIVATSPHÄRE FÜR NACHWEISE ENTFERNEN

Name, Konto und privater Chat des Teilnehmers werden veröffentlicht.

CH12-F32— EINEN TEILNEHMER AUS EINEM KLEINEN LAND AUFDECKEN

Zeit-, Geräte- und Planinformationen machen eine Person rückidentifizierbar.

CH12-F33— SPEZIELLE GESCHICHTE IM NATÜRLICHEN PANEL SAMMELN

Irrelevante Gespräche werden in die Aufnahme aufgenommen.

CH12-F34— NACHWEISE ZUR BARIERREFREIHEIT UNGÜLTIG MACHEN

Der echte Nutzer wird ausgeschlossen, weil kein Standard-Visuell erzeugt werden kann.

CH12-F35— ANNEHMEN, DASS DAS AUFZEICHNUNGSWERKZEUG NEUTRAL IST

Das Werkzeug verändert das Produktverhalten, wird aber nicht getestet.

CH12-F36— VERSTECKEN DER ERFASSUNGSVERSION

Verschiedene Werkzeugversionen werden als eine einzelne Messmethode dargestellt.

CH12-F37— BETRACHTUNG DES ERFASSUNGSWERKZEUGFEHLERS ALS KI-FEHLER

Die zugeschnittene Antwort wird auf das Produkt hochgeladen.

CH12-F38— ZÄHLEN DES API-LOGS AUF DEM NUTZERBILDschirm

Die reale Nutzeroberfläche wird unsichtbar.

CH12-F39— ZÄHLEN DES PROVIDER-LOGS ALS UNABHÄNGIGES EVIDENZMATERIAL

Die Aufzeichnung der gemessenen Partei wird zu einem einzigen Evidenzstück.

CH12-F40— VERSTECKEN DES PRODUKTUPDATES

Zustände zweier Systeme innerhalb der Verzweigung verschmelzen zu einer einzigen Bewertung.

CH12-F41— IGNORIEREN VON ÄNDERUNGEN, WENN KEINE ANKÜNDIGUNG

Der beobachtete Systemabbruch wird nicht untersucht.

CH12-F42— AUSSCHLUSS EXTERNER EREIGNISSE BASIEREND AUF DEM ERGEBNIS

Antworten mit negativen Nachrichten werden entfernt; Antworten mit positiven Nachrichten werden beibehalten.

CH12-F43— VERKNÜPFUNG DER WELLENGESUNDHEIT MIT DEM SCORE

Fehlerhafte Wellen mit hohen Werten sind gültig, starke Wellen mit niedrigen Werten gelten als verdächtig.

CH12-F44— ZÄHLEN DER RICHTIGEN ANTWORT ALS GÜLTIGE AUFNAHME

Der Evidenzfehler wird durch semantischen Erfolg vertuscht.

CH12-F45— ZÄHLEN DER FALSCHEN ANTWORT ALS UNGÜLTIGE AUFNAHME

Das Messergebnis wird in einen Datenfehler umgewandelt.

CH12-F46— ZÄHLEN DES GLEICHEN TEXTES ALS BOT-NACHWEIS

Deterministische oder ähnliche KI-Antworten werden versehentlich zu Duplikaten.

CH12-F47— EINZELNE SIGNALMANIPULATION ENTSCHEIDUNG

Dateigröße oder Zeitunterschied erzeugen automatisch einen Ausschluss.

CH12-F48— BEOBACHTUNG NACH WELLE HINZUFÜGEN

Neue Datensätze werden stillschweigend hinzugefügt, um die gewünschte Stichprobengröße zu erreichen.

CH12-F49— NIEDRIG BEWERTETE BEOBACHTUNG NACH DER WELLE ENTFERNEN

Manifest und Ergebnis werden nachträglich geändert.

CH12-F50— ALTES MANIFEST LÖSCHEN

Versionshistorie wird zerstört.

CH12-F51— ROHDATEN KORRIGIEREN

Dateiinhalte werden geändert, um Fehlklassifizierungen zu korrigieren.

CH12-F52— WIDERSPRUCHSRECHT VERWEIGERN

Unveränderlichkeit wird über Menschenrechte gestellt.

CH12-F53— GESCHLOSSENES EIGENTÜMERISCHES ERFASSUNGSMONOPOLE

Nur mit dem Tool des Gründers erzeugte Nachweise werden als gültig betrachtet.

CH12-F54— DAS ERFASSUNGSSCHEMA NICHT VERÖFFENTLICHEN

Ein unabhängiges Team kann nicht dieselbe Evidenzkette herstellen.

CH12-F55— DIE WELLENWURZEL ALS NACHWEIS DER REALITÄT BETRACHTEN

Die Merkle- oder Hash-Wurzel wird dargestellt, als würde sie die Inhaltsgenauigkeit beweisen.

CH12-F56— ANSPRACHE AUF „1,000 SCREENSHOTS“ OHNE EINE Evidenzpaket

Die Anzahl der Dateien wird wie die Anzahl gültiger Beobachtungen verwendet.

78. PRÜFUNGSVERFAHREN

Schritt 1— Bestimmen des Forschungszwecks der Welle

Haupt-, bestätigend, Vorfall-, Drift- oder Retest-Typ wird ausgewählt.

Schritt 2— KI-Produkt- und Systemzustände festschreiben

Es verbindet sich mit der Wellenversion des KI-Systemregisters in Abschnitt 9.

Schritt 3— Population-, Stichproben- und Panelzustände festschreiben

Aufzeichnungen aus den Abschnitten 5–8 und Abschnitt 11 werden verwendet.

Schritt 4— Prompt-Register festschreiben

Die Prompt-Versionen in Abschnitt 10 sind mit der Erhebungswelle verknüpft.

Schritt 5— UTC Wellenfenster einstellen

Start-, Schluss- und zusätzliche Zeitdauern werden geschrieben.

Schritt 6— Mikroslots erstellen

Teilnehmer werden ausgewogen in Bezug auf Produkt, Land und Sprache Slots zugewiesen.

Schritt 7— Zeitquelle definieren

Server, Gerät und vertrauenswürdige UTC Aufzeichnungen werden bestimmt.

Schritt 8— Erste Kontakt- und Wiederholungsregeln festschreiben

Technische und semantische Zustände werden getrennt.

Schritt 9— Aufnahmemethoden bestimmen

Web-, Mobil-, Barrierefreiheits- und natürliche Panel-Pfade werden definiert.

Schritt 10— NCL-Ziel bestimmen

Der für die Hauptanalyse erforderliche Aufnahme-Confidence-Level wird aufgezeichnet.

Schritt 11— Evidenzpaket-Schema festschreiben

Manifest, Dateien, Hash und Metadatenfelder sind definiert.

Schritt 12— Datenschutz- und Schwärzungsregeln festschreiben

Roh-, Prüfungs- und öffentliche Ebenen sind getrennt.

Schritt 13— Erfassen Sie Ihr Fahrzeug mit dem Piloten

Ändert das Fahrzeug das Produktverhalten? Sind Zeit- und Textaufzeichnungen korrekt?

Schritt 14— Externes Ereignisregister öffnen

Ein Prozess zur Überwachung von Produkt- und Entitätsereignissen ist eingerichtet.

Schritt 15— Wave Lock bestätigen

Der Status WS-1 — FESTGESCHRIEBEN wird vergeben.

Schritt 16— Slots öffnen

Teilnehmer dürfen nur in ihren eigenen Slots einreichen.

Schritt 17— Erstkontakt-Ergebnis aufzeichnen

Antwort, Ablehnung, Fehler oder Limit ist getrennt.

Schritt 18— Erfassen der ersten geeigneten Ausgabe

Die vollständige Antwort wird ohne Erneuerung aufbewahrt.

Schritt 19— Erstellen von Rohbeweisen und Metadaten

Screenshots, Rohtexte, Systemstatus und Zeitstempel werden aufgezeichnet.

Schritt 20— Lokalen Hash ausführen und sicher hochladen

Dateien erhalten vor dem Hochladen einen Integritätsdatensatz.

Schritt 21— Serverakzeptanz und Erstellen eines Paket-Hashes

Die Evidenzkette wird mit Serverzeit und Paket-ID geschlossen.

Schritt 22— Blindprüfung der Gültigkeit der Erfassung

Der CV-Status wird vergeben, ohne den semantischen Erfolg zu kennen.

Schritt 23— Duplikat- und Manipulationsüberprüfung durchführen

Mehrere Signale und menschliche Überprüfung werden verwendet.

Schritt 24— Untersuchung des Systems innerhalb der Welle und externer Ereignisse

Bestimmen Sie, ob ein Teil-Welle- oder Mischzustand erforderlich ist.

Schritt 25— Wellen-Gesundheit zuweisen

Der Status wird zwischen WH-0 und WH-5 zugewiesen.

Schritt 26— Wellenmanifest und Integritätsstamm erstellen

Die Beobachtungsliste wird versioniert.

Schritt 27— Öffentliches Nachweis-Manifest veröffentlichen

Umfang und Integrität werden ohne persönliche Daten sichtbar gemacht.

Schritt 28— Wellenabschlussbericht genehmigen

Ziele, Vorkommnisse, Mängel, Ereignisse und Änderungen werden mit menschlicher Zustimmung abgeschlossen.

79. NOTWENDIGE NACHWEISE

Wellen-ID Wellentyp Protokollversion KI-System Registrierungsversion Populationsrahmen Probe-Sperre Panel-Status-Sperre Prompt Registrierungs-Version Referenzpaketversion UTC Start- und Endzeit Mikro-Zeitfenster-Liste Teilnehmer–Slot-Zuordnung Zeitquelle Geräteuhrabweichungen Übertragungszeiten Erste-Kontakt-Zeiten Abschlusszeiten Erfassungszeiten Hash-Zeiten Upload-Zeiten Server-Akzeptanzzeiten Erste-Kontakt-Ergebnisse Erste passende Inhaltsausgaben Technische Wiederholungsaufzeichnungen Teilantwortaufzeichnungen Ablehnungen und Nutzungsgrenzen Volle Rohantworten Volle Screenshots Sequenzielle visuelle Manifeste Bildschirmaufzeichnungen falls vorhanden Zitations- und Quellenaufzeichnungen Systemeinstellungen Nachweis Erfassungstool und Version Erfassungsmethode NCL-Stufe Dateihashes Paket-Hash

Digitale Signatur Zeitstempel Duplikatsprüfungen Manipulationsprüfungen Erfassungs-Gültigkeitsentscheidungen Schwärzungsdateien Roh-zu-geschwärzten Links Datenschutzmanifest Rücknahmeaufzeichnungen Produktänderungsereignisse Externes Ereignisregister Wellenunterteilungen Slotbereich Länder- und Sprachbereich Technische Fehlerraten Erfassungs-Gültigkeitsrate Wellen-Status Wellenintegritätswurzel Öffentliches Evidenzmanifest Wellenabschlussbericht Änderungsprotokoll Verantwortliche Person oder Institution

80. PRÜFLISTE FÜR AUDITS

Ist der Typ und Zweck der Welle klar? Wurden KI-Produkte und Systemstatus festgeschrieben? Sind die Bevölkerung, Stichprobe und Panel-Versionen bestimmt? Wurde der Prompt-Satz vor den Ergebnissen festgeschrieben? Ist das Wellenfenster UTC geöffnet? Wurden Mikro-Slots vorab zugewiesen? Wurden Länder und Sprachen gleichmäßig auf Slots verteilt? Haben Teilnehmer die Slots basierend auf ihren Ergebnissen gewechselt? Wurde die lokale Zeitbelastung bewertet? Gibt es rotierende Wellenstunden? Sind die zusätzlichen Zeiten für Einreichung und Abschluss klar? Wurde die Geräteuhr mit einer zuverlässigen UTC-Quelle verglichen? Wurden Einreichungs-, Abschluss-, Erfassungs- und Hochladezeiten getrennt? Wurden Beobachtungen außerhalb der Welle in die Hauptbewertung einbezogen? Wurde das Ergebnis des ersten Kontakts aufgezeichnet? Wurde ein technischer Fehler von einer Ablehnung unterschieden?

Wurde aufgrund einer falschen Antwort ein erneuter Versuch unternommen? Wurde der erste Versuch im technischen Retry beibehalten? Wurden Teilergebnisse gelöscht? Ist die erste geeignete Ausgabe wirklich die erste Ausgabe? Wurde „Regenerate“ oder eine Folge-/Nachricht verwendet? Wird die visuelle Prompt und die gesamte Antwort angezeigt? Gibt es bei langen Antworten eine Erfassung ohne Leerzeichen? Entspricht der Rohtext der visuellen Darstellung? Falls OCR verwendet wurde, gab es eine menschliche Überprüfung? Wurden dynamische Antwortversionen beibehalten? Hat der Nutzer die Antwort gestoppt? Sind die Prompt und die Antwort mit derselben Sitzung verknüpft? Wurde das Erfassungstool und die Version protokolliert? Hat das Erfassungstool das Nutzerverhalten verändert? Wurde der NCL-Level angegeben? Gibt es Beobachtungen unterhalb von NCL-3 in der Hauptanalyse? Hat jede Datei einen Hash? Ist klar, was der Hash nicht beweist?

Gibt es eine digitale Signatur oder Serverzeit? Ist die Aufnahmezeit von der Upload-Zeit getrennt? Gibt es einen Grund für den späten Upload und ein lokales Integritätsprotokoll? Wurde die Bearbeitung an den Rohbeweisen durchgeführt? Haben die bearbeiteten Derivate separate Hashes? Enthüllt das öffentliche Manifest die Identität des Teilnehmers? Kann ein Nutzer aus einem kleinen Land wieder identifiziert werden? Wurden private Gespräche unnötigerweise im natürlichen Panel erfasst? Wurde ein alternativer Evidenzweg für Nutzer mit Zugriffshilfen bereitgestellt? Wurde die Duplikatentscheidung nur auf demselben Text basierend getroffen? Trägt die Manipulationsentscheidung mehrere Signale? Kennt der Erfassungsvalidierer die semantische Punktzahl? Deckt die richtige Antwort den Evidenzfehler ab? Wurde die falsche Antwort für ungültig erklärt? Hat sich das Produkt während der Welle geändert?

Wurde die Produktänderung als Unterwelle oder Mischzustand verarbeitet? Wurden externe Ereignisse aufgezeichnet? Haben externe Ereignisse laut Ergebnis zu einer Ausschlusswertung geführt? Ist der Wellengesundheitsstatus unabhängig von der Bewertung? Wurden nach dem Schließen der Welle Beobachtungen hinzugefügt oder entfernt? Werden das alte Manifest und die Integritätswurzel aufbewahrt? Wie wurden Teilnehmer-Rücktrittsanfragen behandelt? Gibt es ein öffentliches Beweismanifest? Gibt es einen eindeutig verantwortlichen Eigentümer des Wellenabschlussberichts?

81. EINWÄNDE UND ANTWORTEN

Einwand 1— „Wenn nicht alle Nutzer genau im selben Moment ihre Eingaben machen, wird keine echte Gleichzeitigkeit erreicht.“

Eine Ein-Sekunden-Übermittlung kann künstliche Last-, Verbindungs- und Warteschlangeneffekte erzeugen. Ein vordefiniertes kurzes UTC-Fenster und zufällige Mikro-Slots können denselben Systemzeitraum zuverlässiger messen.

Einwand 2— "Das Produkt kann sich innerhalb eines Ein-Stunden-Zeitrahmens ändern."

Es kann sich ändern. Daher sind Systemaufzeichnungen, Konfigurations-Fingerabdruck und Produktänderungsverfolgung erforderlich. Im Falle einer wesentlichen Änderung wird die Welle getrennt.

Einwand 3— "Warum reicht nicht einfach ein Screenshot?"

Screenshot:

- Zeit,
- anfängliche Ausgabe,
- Prompt-Version,
- Systemeinstellung,
- Kopieren

Diese Themen nicht immer anzeigen. Ein vollständiges Evidenzpaket ist erforderlich.

Einwand 4— „Ist es nicht zu aufwendig, tausend Screenshots einzeln zu überprüfen?“

Es ist aufwendig. Automatische Integritätsprüfungen, stichprobenartige menschliche Überprüfungen und risikobasierte Verifikationen können verwendet werden. Kritische oder verdächtige Aufzeichnungen werden genauer untersucht.

Einwand 5— „Wenn es einen Hash gibt, warum können wir dann nicht sagen, dass die Datei echt ist?“

Ein Hash zeigt nur, dass die Datei nach der Erstellung des Hashes unverändert geblieben ist. Er zeigt nicht, wie die Datei vor der Erstellung des Hashes entstanden ist.

Einwand 6— „Wenn der Nutzer persönliche Informationen kürzen muss, wie wird dann die Rohdatei geschützt?“

Wenn möglich, sollte das Capture-Tool von Anfang an keine persönlichen Informationen erfassen. Falls unvermeidbar, wird die Rohdatei in einem eingeschränkten und verschlüsselten Repository aufbewahrt, und eine öffentliche Kopie wird als separat redigierte Version bereitgehalten.

Einwand 7— „Wenn die Antwort falsch ist, warum versuchen wir es dann nicht noch einmal?“

Weil eine falsche Antwort eine Frage der Messung ist. Ein erneuter Versuch ist nur bei einem tatsächlichen technischen Fehler vorgesehen.

Einwand 8— „Wenn das System ablehnt, kann der Nutzer tatsächlich noch einmal versuchen?“

Richtig. Das Haupt-First-Contact-Panel speichert die Ablehnung. Ein separat vorhandener Natural Interaction Track kann das Nachverhalten des Nutzers messen.

Einwand 9— „Wenn ein technischer Fehler ein Problem der Nutzererfahrung ist, warum erlauben wir dann einen erneuten Versuch?“

Der technische Fehler wird als Ergebnis des ersten Kontakts beibehalten. Ein vordefinierter technischer Wiederholungsversuch kann durchgeführt werden, um zusätzlich den repräsentativen Inhalt zu untersuchen. Die beiden Ergebnisse ersetzen einander nicht.

Einwand 10— „Ist nicht das eigene Protokoll des Anbieters der stärkste Nachweis?“

Technisch kann es sehr stark sein. Der Anbieter ist der Eigentümer des gemessenen Systems. Es ist zuverlässiger, wenn es zusammen mit einer unabhängigen Nutzeraufnahme verwendet wird.

Einwand 11— „Wird NOMOS Capture eine verpflichtende NobleJackal-Software sein?“

Nr. NOMOS Capture ist ein offenes Evidenzprotokoll. Unabhängige Werkzeuge, die die äquivalenten Schema- und Integritätsbedingungen erfüllen, können kompatibel sein.

Einwand 12— „Wenn wir nicht alle Nachweise öffentlich machen, wie wird dann eine unabhängige Überprüfung durchgeführt?“

Das öffentliche Manifest zeigt Integrität und Umfang. Autorisierte unabhängige Prüfer können auf geschwärzte oder eingeschränkte Rohbeweise zugreifen. Datenschutz und Überprüfung können zusammen gestaltet werden.

Einwand 13— „Würde nicht die gesamte Arbeit gestört, wenn wichtige Nachrichten während der Welle auftauchen?“

Nicht immer. Der Bericht ist Teil der realen Systemumgebung. Er kann mit einem Zeitstempel vor und nach dem Ereignis getrennt werden. Der Umfang der Bewertung wird entsprechend erklärt.

Einwand 14— „Ist es nicht einfacher, alte Antworten zu verwerfen und mit dem neuen System fortzufahren, wenn ein Produkt aktualisiert wird?“

Alte Antworten sind am Messdatum real. Sie können nicht gelöscht werden. Alte und neue Systemstatus müssen als separate Wellen oder Unterwellen beibehalten werden.

Einwand 15— „Wenn das Erfassungstool das Produktverhalten beeinflusst, schlägt die ganze Methode fehl?“

Nein. Der Fahrzeugeffekt wird mit dem Piloten gemessen. Falls notwendig, wird ein leichteres Fahrzeug, manuelle Erfassung oder eine Methode gewählt, die den Fahrzeugeffekt als separaten Faktor verwendet.

Einwand 16— „Wenn der Teilnehmer seine Daten zurückzieht, wäre dann die Integritätswurzel nicht gefährdet?“

Eine neue Manifestversion und eine neue Wurzel werden erstellt. Der alte Eintrag kann eine inhaltslose ZURÜCKGEZOGENE Spur behalten. Er wird auf der Architektur für Nachweise der Menschenrechte aufrechterhalten.

ALLGEMEINES URTEIL DES ABSCHNITTS 85

Am Ende einer GEO-1000-Studie könnten Sie 30,000 Screenshots haben. Diese Zahl allein ist kein Nachweis. Sie müssen Folgendes wissen: In welcher Welle wurden sie erstellt? Zu welchem Systemzustand gehörten sie? Wurde tatsächlich derselbe Prompt gesendet? War es die erste Ausgabe? Wurde die Antwort aktualisiert? Hat sich das Produkt während der Welle verändert? Wann wurde die Datei erfasst? Wann wurde der Hash erstellt? Wurde die Teilnehmer-ID wiederholt? Zeigt die Visualisierung die gesamte Antwort? Entspricht der rohe Text der Visualisierung? Wenn die Nachweise später bearbeitet wurden, wo ist das Original? Wurde die Erfassungsprüfung vor der semantischen Bewertung durchgeführt? Sie können innerhalb derselben Sekunde tausendmal klicken. Trotzdem:

- geräteinterne Uhren,
- Verbindungen,
- Warteschlangen,
- Produktzustände

Es kann unterschiedlich sein. Daher ist Gleichzeitigkeit keine Marketingshow. Es ist eine vordefinierte Zeitarchitektur. Wenn ein Nutzer eine falsche Antwort erhält: 'Versuchen Sie es noch einmal.' zu sagen, ist verlockend. Aber diese falsche Antwort ist die Realität, die Sie zu messen versuchen. Wenn Sie sie löschen, beheben Sie das System nicht. Sie verfälschen die Nachweise. Wenn ein Nutzer einen echten Verbindungsfehler erlebt, kann ein technischer Wiederholungsversuch durchgeführt werden. Aber das Ergebnis des ersten Kontakts sollte dennoch erhalten bleiben. Denn der Nutzer ist zuerst auf den Fehler gestoßen. Der Hash ist wertvoll. Die digitale Signatur ist wertvoll. Der Zeitstempel ist wertvoll. Der Screenshot ist wertvoll. Keines davon allein enthält die ganze Wahrheit. Vertrauen:

- Prompt,
- Zeit,
- Produkt,
- Nutzer,
- Datei,
- Initiative,
- Verifizierung

Es entsteht aus der gemeinsamen Erhaltung seiner Kette. Ein Standard: Wenn es heißt: „Nur Nachweise, die mit meiner Software erstellt wurden, sind gültig“, stellt es das eigene Interesse über die unabhängige Reproduzierbarkeit. Aus diesem Grund ist NOMOS Capture keine Werkzeugmarke, sondern eine offene Beweisarchitektur. Unabhängige Universitäten, Forscher oder Prüfinstitutionen sollten in der Lage sein, dieselben Regeln anzuwenden. Daher lautet das zwölfte Messgesetz von NOMOS wie folgt:

> *Nebenläufigkeit ist beobachtbare Reproduzierbarkeit innerhalb desselben definierten Systemzeitpunkts, nicht derselben Sekunde.*

Das dreizehnte Gesetz lautet wie folgt:

Die erste Ausgabe ist kein Entwurf, der erneuert werden kann, bis einem das Ergebnis gefällt.

Das vierzehnte Gesetz lautet wie folgt:

Hash bewahrt den Beweis; er erstellt keinen Beweis.

Das fünfzehnte Gesetz lautet wie folgt:

Erfassungsvalidität und Reaktionsgenauigkeit sind unabhängig voneinander.

Sechzehntes Gesetz lautet wie folgt:

Die Evidenzkette sollte die Integrität der Beobachtung sichtbar machen, nicht die Identität der Person.

Sein siebzehntes Gesetz lautet:

Wenn sich das System während der Welle verändert, kann es nicht so tun, als hätte sich die Messhistorie nicht verändert.

Sein achtzehntes Messgesetz besagt:

Auf Rohbeweise wird nichts geschrieben; Entscheidungen werden durch Versionierung korrigiert.

Sein neunzehntes Messgesetz besagt:

Eine Erfassungsmethode, die nicht unabhängig angewendet werden kann, kann kein Weltstandard sein.

Bestellung von Abschnitt 12 von NOMOS

Sag mir nicht „wir haben zur gleichen Zeit gefragt." / Zeige, in welchem UTC-Fenster, in welchem Slot und in welchem Systemzustand du gefragt hast.

Stapele nicht tausend Personen in dieselbe Sekunde, um einen künstlichen Fehler zu erzeugen. / Verteile die Zeit, aber halte das System in derselben Welle.

Mische keine Ortszeiten miteinander. / Mache die Gerätesteuerung nicht zur absoluten Wahrheit.

Speichere nicht das Ergebnis des ersten Kontakts. / Speichere das Nutzungslimit, Wiederholungen, Fehler und Klärungen separat.

Regeneriere nicht, wenn du eine falsche Antwort bekommst. / Diese falsche Antwort ist mein tatsächliches Verhalten.

Du kannst es erneut versuchen, wenn ein technischer Fehler auftritt. / Lösche aber nicht den ersten Fehler.

Bringen Sie mir nicht nur einen Screenshot. / Bringen Sie den Prompt, die vollständige Antwort, das Produkt, die Zeit, die Einstellung, die Interferenzkette und die Datei-Integrität.

Schneiden Sie den guten Teil Ihrer Antwort nicht ab. / Lassen Sie das Ende, die Grenze und den Fehler sichtbar.

Erklären Sie den Hash nicht als Realität. / Beweisen Sie auch, was vor dem Hash war.

Schwärzen Sie die Rohdatei selbst nicht. / Bewahren Sie das Original und erstellen Sie eine separate öffentliche Kopie.

Veröffentlichen Sie keinen Namen einer Person, kein Konto oder private Unterhaltung als Nachweis.

Machen Sie den einzigen Teilnehmer in einem kleinen Land nicht identifizierbar.

Erklären Sie Ihr Erfassungstool nicht als neutral, ohne zu testen, ob es mich verändert hat.

Wenn sich das System in der Mitte der Welle geändert hat, dürfen alte und neue Antworten nicht zusammengeführt werden, als gehörten sie zu einem einzigen Produkt.

Wenn ein externes Ereignis die Änderung verursacht hat, zeichnen Sie das Ereignis auf. / Löschen Sie die Beobachtung nicht, nur weil Ihnen die Bewertung nicht gefällt.

Halten Sie das Evidenzmaterial unverändert. / Wenn die Entscheidung falsch ist, korrigieren Sie sie mit einer neuen Version.

Binden Sie den Erfassungsstandard nicht an die geschlossene Software eines einzelnen Unternehmens. / Lassen Sie andere in der Lage sein, dieselbe Evidenzkette herzustellen.

Zuerst die Welle festschreiben. / Dann die Zeit verteilen. / Dann den ersten Output erfassen. / Dann hashen. / Dann sicher hochladen. / Dann die Erfassung überprüfen, ohne die Genauigkeit zu prüfen. / Und erst danach entscheiden, was Ihre Antwort aussagt.

Der abschließende Satz des Kapitels

Eine Antwort in GEO-1000 wird erst dann zu einer echten Maßeinheit, wenn gezeigt werden kann, unter welchem Systemzustand, zu welchem Zeitpunkt, von welchem Prompt und innerhalb welcher unveränderlichen Evidenzkette sie entstanden ist.

Normativer Kern

Jede GEO-1000-Beobachtung MUSS mit einer vorher deklarieren und versionierten Messwelle verknüpft sein, die Folgendes definiert: - KI-Produktzustände, - Populationen und Stichprobenrahmen, - Panelzustände, - Prompt-Versionen, - UTC-Übertragungsfenster, - randomisierte Mikro-Slots, - Abschluss- und Erfassungsfristen, - technische Wiederholungsregeln, - Erfassungsschema, - Referenzdatensatzversionen und - Governance-Zuordnung. Synchronisation MUSS Beobachtung innerhalb eines deklarierten UTC-Systemzustandsfensters und zugewiesener Slot-Struktur bedeuten. Sie DARF NICHT als perfekte Millisekunden-Simultanität repräsentiert werden. Jede Beobachtung MUSS folgende Angaben so weit wie möglich beibehalten: - Zuweisungszeit, - Prompt-Einreichungszeit, - erstes Systemereignis, - Abschlusszeit der Antwort, - Erfassungszeit, - Hash-Erstellungszeit, - Upload-Zeit, - und Server-Akzeptanzzeit. Das Ergebnis des ersten Kontakts MUSS sichtbar bleiben. Ein technischer Wiederholungsversuch DARF nur unter zuvor angekündigten Infrastruktur-Ausfallbedingungen verwendet werden und MUSS jeden vorherigen Versuch beibehalten. Falsche, ungünstige, abgelehnte, nicht zitierte, gekürzte oder kritische Ausgaben DARF NICHT einfach aufgrund ihres Inhalts erneut erzeugt, ersetzt oder ausgeschlossen werden. Ein NOMOS Evidenzbündel MUSS das exakt zugewiesene und gesendete Prompt, die vollständige Rohantwort, alle visuellen Nachweise, Systemzustands-Metadaten, Versuchprotokolle, Zeitevidenz, Datei-Hashes, Bündel-Hash, Version des Erfassungswerkzeugs, Datenschutzstatus und eine anonymisierte Erfassungs-Gültigkeitsentscheidung beibehalten. Ein Screenshot, Hash, digitale Signatur, Anbieterprotokoll oder Zeitstempel DARF NICHT einzeln als vollständiger Nachweis der Authentizität der Beobachtung repräsentiert werden. Rohbeweise MÜSSEN unveränderlich oder nur anhängbar bleiben. Schwärzungen, Korrekturen, Rücknahmen, Ausschlüsse und Änderungen durch die Entscheidung MÜSSEN verknüpfte Versionen erzeugen und den vorherigen Datensatz dort bewahren, wo dies ethisch und rechtlich zulässig ist. Materielle KI-Produktänderungen oder externe Ereignisse während einer Welle MÜSSEN zeitlich erfasst und durch Unterwellen, explizite Modellierung, Neustart oder Mischzustandswarnungen behandelt werden. Die Gültigkeit der Erfassung MUSS unabhängig von der semantischen Korrektheit entschieden werden. Öffentliche Nachweis-Manifeste MÜSSEN die Überprüfung unterstützen, ohne unnötig die Identität von Teilnehmern, Zugangsdaten, private Gesprächshistorie oder den genauen Standort offenzulegen. NOMOS Die Einhaltung der Erfassung MUSS durch offene, gleichwertige Evidenzsysteme umsetzbar sein und DARF NICHT ausschließlich von einem proprietären Tool abhängen, das von dem Gründer des Standards kontrolliert wird. Jede Welle, jedes Evidenzpaket, jede Erfassungsentscheidung, jede Manifestüberarbeitung und jeder Integritätsdatensatz MUSS versioniert und einer verantwortlichen Person oder Organisation zuordenbar sein.

TEIL IV

Realität, Beurteilung und Entscheidung

KAPITEL 13

Das verifizierte Faktenpaket zur Entität

Verifiziertes Faktenpaket zur Entität

METHODISCHE GRUNDLAGEN / K10; K11; K12

Kapitelgrenze

Die ersten zwölf Kapitel legten die experimentelle Kette fest: Eine einzelne Antwort ist kein GEO-Score; Repräsentation ist eine Verteilung über Nutzer und KI-Produkte; die geprüfte Einheit muss bestimmt werden, bevor Ergebnisse beobachtet werden; offizielle Repräsentation und evidenzbasierte Realität der Einheit müssen getrennt bleiben; die Zielpopulation, Stichprobe, Länder und Sprachen müssen definiert werden; Teilnehmer müssen unabhängig von den Ergebnissen ausgewählt werden; jede KI-Produktinstanz muss mit einem Register aller materiellen Konfigurationen verknüpft werden; Aufforderungen müssen die gleiche Absicht über Sprachen hinweg bewahren; Kontrollierte und Natürliche Nutzerpanels müssen getrennt bleiben; und jede Beobachtung muss innerhalb einer synchronisierten Welle und der NOMOS Capture-Evidenzkette erhalten bleiben. Die Prüfung muss nun ihre zentrale Frage beantworten:

> *Nach welchem Referenzeintrag werden wir bestimmen, ob die erfasste KI-Antwort korrekt, inkorrekt, unvollständig, veraltet, unbegründet oder außerhalb des Umfangs ist?*

Abschnitt 4 hat konzeptionell zwei getrennte Realitätsebenen eingeführt:

- Offizielle Repräsentationsebene, die zeigt, was das Wesen selbst über sich sagt
- Verifizierte Entitätsrealitätsebene, die zeigt, inwieweit die vorhandenen Nachweise bestimmte Aussagen erlauben

Dieser Abschnitt verwandelt die zweite Ebene in ein operatives Prüfobjekt. Der Name dieses Objekts lautet:

Verifiziertes Faktenpaket zur Entität

Kanonischer englischer Begriff:

Verifiziertes Faktenpaket zur Entität

Dieses operationelle Objekt vervollständigt das dritte Glied in der Architektur der ersten beiden Bücher: der GEO Framework definiert den Standard; 99 Errors in GEO identifiziert seine Verletzungsmodi; und das Prüfsystem misst und untermauert diese Verstöße. Jede prüfbare Kontrolle muss das Modell oder Produkt, Datum, Land, Sprache, Abfragesatz, Wiederholungsanzahl und Messprotokoll enthalten. Innerhalb dieser Kette ist das Truth Pack:

„Was hat die KI gesagt?“
mit der Frage:

„Welche Nachweise haben es ihr erlaubt, dies zu sagen?“
ist das Referenzsystem dazwischen. Dieser Abschnitt:

- was das Truth Pack ist und was nicht,
- wie Aussagen in atomare Realitätsaufzeichnungen verwandelt werden,
- die Arten von Quellen und Nachweisen,
- die Rollen von Erstparteien-, offiziellen, unabhängigen, betrieblichen und Nutzerressourcen,
- Gegenevidenz,
- Quellenherkunft und Illusion der Unabhängigkeit,
- Ausreichendheit der Nachweise,
- Zeit-, Länder- und Umfangsgrenzen,
- Offene-Welt- und Geschlossene-Welt-Aussagen,
- der Unterschied zwischen „keine Nachweise gefunden“ und „als falsch nachgewiesen",
- quantitative Leistungs-, Kunden-, Partner-, Lizenz-, Zertifikats- und Überlegenheitsaussagen,
- Versionierung und Pre-Wave-Festschreibung des Truth Pack,
- öffentliche und eingeschränkte Evidenzschichten,
- Einwände, Korrektur- und Widerspruchsprozesse,
- maschinenlesbares Wahrheits-Paket

definiert. Dieses Kapitel behandelt noch nicht:

- alle Schlichtungsdetails zur Aufteilung von KI-Antworten in atomare Behauptungen,
- Übereinstimmung zwischen den Bewertern,
- Schwelle für den Übergang von Antworten,
- kritische Fehlerschranken,
- Legt die endgültigen NOMOS Punkformeln nicht fest

wird nicht abgeschlossen. Dies sind die Themen der folgenden Abschnitte. Die zentrale Frage von Kapitel 13 lautet:

> *Wie erstellt man ein Truth Pack über eine Entität, die KI-Antworten bewertet, die vor dem Ergebnis festgeschrieben ist und sowohl Nachweise als auch Gegenevidenz zusammenführt, mit klaren Zeit- und Umfangsgrenzen?*

NOMOS Herausforderung

Der Satz auf der Website Ihres Unternehmens lautet: „Wir sind in 25 Ländern tätig.“ Das KI-Produkt sagt dem Nutzer denselben Satz: „Das Unternehmen ist in 25 Ländern tätig.“ Hat die KI die offizielle Website korrekt wiedergegeben? Ja. Wurde die Behauptung tatsächlich überprüft? Wir wissen es noch nicht. Das Unternehmen sendet Ihnen eine Tabelle.

Die Tabelle enthält die Namen von 25 Ländern. Die Spalten sind jedoch nicht erklärt: Einige enthalten nur Website-Besucher. Einige enthalten ehemalige Kunden. Einige enthalten unabhängige Vertriebsmitarbeiter. Einige enthalten einmalige organisierte Veranstaltungen. In neun Ländern gibt es tatsächlich eine aktive Servicebereitstellung. Nur in drei Ländern gibt es eine lokale juristische Person.

Was bedeutet ‚to operate‘? Bedeutet es, einen Kunden zu haben? Ein lokales Unternehmen zu gründen? Dienstleistungen aus der Ferne anbieten zu können? Lieferung über einen Partner? Oder

dass eine Website aus diesem Land zugänglich ist? Subjekt, Verb und Umfang der Behauptung sind nicht definiert.

Daher gibt es zwar eine Zahl, aber noch keine überprüfbare Realität. Betrachten wir nun eine andere Behauptung: ‚98% unserer Projekte sind erfolgreich.' Sie haben 49 erfolgreiche Projektpräsentationen. Es gibt keine fehlgeschlagenen Projektakte. Wie viele Projekte wurden insgesamt abgeschlossen? Nach welchen Kriterien wird Erfolg definiert? Pünktliche Lieferung? Kundenakzeptanz?

Kommerzielles Ergebnis? Technische Machbarkeit? Kundenzufriedenheit? Welcher Zeitraum? Die letzten drei Monate? Die gesamte Firmengeschichte? Nur ausgewählte Projekte? Ohne den Nenner zu kennen, kann 98 Prozent nicht nachgewiesen werden. Nun gibt es diesen Satz auf der Website eines anderen Unternehmens: "Wir sind der Weltmarktführer." Dieser Satz wird wiederholt von:

- Firmenblog,
- Pressemitteilung des Unternehmens,
- zwanzig Nachrichtenseiten, die dieselbe Pressemitteilung kopieren,
- fünf gefälschte Nutzerkonten, die vom Unternehmen erstellt wurden,
- drei KI-generierte Texte, die diese Inhalte zusammenfassen

.

Es gibt dreißig verschiedene URLs. Gibt es dreißig unabhängige Beweise? Nein. Eine einzelne Selbstdeklaration kann auf vielen abgeleiteten Oberflächen widergespiegelt worden sein.

Dies ist genau die Manipulation, die im vorhergehenden Band identifiziert wurde: versteckter Text, erfundene Nutzer und selbst verfasste Kommentare, die über verschiedene Domains veröffentlicht werden, werden wieder in das System eingespeist, als ob sie unabhängige Nachweise wären. Betrachten wir nun den gegenteiligen Fall. Die offizielle Website des Unternehmens besagt: ‚Wir erbringen Dienstleistungen nur in der Türkei.'

Doch das Unternehmen begann vor zwei Monaten einen verifizierten Betrieb in Deutschland. Verträge, Rechnungen und offizielle Aufzeichnungen belegen diese Tatsache, obwohl die Website nicht aktualisiert wurde. Die KI antwortet: ‚Das Unternehmen bietet Dienstleistungen in der Türkei und in Deutschland an.' Die Antwort weicht von der offiziellen Website ab, könnte aber der materiellen Realität näherkommen.

Betrachten Sie nun eine noch sensiblere Situation. Auf der Website einer Gesundheitseinrichtung steht: ‚Lizenzierte Einrichtung.' Im Register hat eine andere Einrichtung mit einem ähnlichen Namen eine Lizenz. Die KI führt die beiden Einrichtungen zusammen und ordnet die Lizenz der falschen Einrichtung zu. Das Unternehmen bemerkt diesen Fehler, korrigiert ihn jedoch nicht, da er zu ihrem Vorteil ist.

Die ausdrückliche Bestimmung des zweiten Buches gilt hier: Ein Fehler, der zu Gunsten einer Person ist, ist immer noch ein Fehler. Um eine KI-Antwort zu bewerten, benötigen wir nicht nur eine Liste korrekter Sätze, sondern auch die folgende Struktur: Von wem ist die Aussage? Was genau wird behauptet? In welchem Zeitraum ist sie gültig? In welchem Land? Für welches Produkt oder welche Dienstleistung?

Unterstützt die Quelle die Behauptung wirklich? Wer kontrolliert die Quelle? Sagen andere Quellen dasselbe oder kopieren sie dieselbe Originalquelle? Gibt es Gegenevidenz? Ist der Anspruch nur eine offizielle Selbsterklärung? Sind die Nachweise öffentlich oder privat?

Zum Zeitpunkt der Messung der KI, gab es die Möglichkeit, auf diese Nachweise zuzugreifen? Ist das Ergebnis sicher, begrenzt, umstritten oder unbekannt? Die erste Regelung dieses Abschnitts lautet:

Eine Quelle allein ist nicht Realität.

Ihre zweite Bestimmung lautet:

> *Eine Behauptung kann nicht überprüft werden, ohne ihr Subjekt, ihren Umfang, ihre Zeit und ihre Evidenzlast zu definieren.*

Seine dritte Bestimmung besagt:

> *Verschiedene URLs bedeuten nicht unterschiedliche unabhängige Nachweise.*

Seine vierte Bestimmung besagt:

> *Das Fehlen von Nachweisen ist kein automatischer Nachweis dafür, dass die Behauptung falsch ist.*

Seine fünfte Vorschrift lautet:

> *Verifizierte Realität ist keine absolute und unveränderliche Wahrheit; sie ist ein abgeleitetes Urteil, getragen von den verfügbaren Nachweisen innerhalb einer bestimmten Zeit und eines bestimmten Kontextes.*

Seine sechste Bestimmung lautet:

> *Das Truth Pack kann nicht basierend auf der KI-Antwort erstellt werden. Die KI-Antwort wird gegen das zuvor festgeschriebene Truth Pack bewertet.*

1. ZWECK DES KAPITELS

Zweck dieses Abschnitts ist es, wesentliche Aussagen über die geprüfte Einheit in ein Referenzpaket umzuwandeln, das mit Nachweisen und Gegenevidenz verknüpft ist, mit klaren Zeit- und Umfangsgrenzen, ohne sie in eine Antwortvorlage umzuwandeln. Der Abschnitt normalisiert die folgenden Unterscheidungen:

- Offizielle Behauptung versus verifizierte Realität
- Quelle versus Nachweis
- Nachweis versus Behauptungsunterstützung
- Erstpartei-Aufzeichnung versus unabhängige Verifizierung
- Unabhängige Quelle und vom Subjekt autorisierte Quelle
- Unterschiedlicher Domainname und unterschiedlicher Evidenzursprung
- Anzahl der Quellen und unabhängige Evidenzfamilie
- Direkter Nachweis und abgeleiteter Nachweis
- Unterstützende Nachweise und Gegenbeweise
- Fehlen einer Behauptung und Falschheit der Behauptung
- Offene Welt Behauptung und geschlossener Welt Datensatz
- Gegenwärtige Realität und historische Realität
- Zeit, in der die Realität gültig ist, und Zeitpunkt der Veröffentlichung der Quelle
- Öffentliche Nachweise und eingeschränkte Nachweise
- Für KI zugängliche Nachweise und nur für Prüfer offene Nachweise
- Bewiesene Behauptung und öffentlich auffindbare Behauptung
- Direkte Realität und berechnetes abgeleitetes Ergebnis
- Rohdaten und Leistungskennzahlen
- Zähler und Nenner
- Kunde und potenzieller Kunde
- Partner und Kunde

- Mitgliedschaft und Akkreditierung
- Lizenz- und Markenbeziehung
- Zertifikat und allgemeine institutionelle Konformität
- Aussage und Meinung
- Meinung und Wettbewerbsvorteil
- Beschwerde und überprüfte Verletzung
- Untersuchung und eindeutiges Verbrechen oder Haftung
- Veralteter Eintrag und falscher historischer Eintrag
- Falsche KI-Antwort mit fehlendem Wahrheits-Paket
- Referenzfehler und Systemfehler
- Integrität des Truth Packs und dass alle Behauptungen positiv sind
- Vertraulichkeit von Evidenzmitteln und Fehlen von Nachweisen
- Recht der Entität auf Antwort und Selbstüberprüfung
- Menschliche Genehmigung und menschliche Unfehlbarkeit
- KI-Unterstützung und KI als die letztendliche Autorität der Wahrheit

Am Ende dieses Abschnitts sollte jede Prüfung in der Lage sein, diese Fragen zu beantworten:

Welche materiellen Behauptungen sind im Truth Pack enthalten?

Zu welcher Entität, welchem Land, Produkt und welcher Zeit gehört jede Behauptung?

Welche Nachweise stützen sie, welche begrenzen oder widerlegen sie?

Sind die Quellen wirklich unabhängig oder stammen sie aus derselben ursprünglichen Quelle?

Kann die Behauptung öffentlich überprüft werden?

Ist die Evidenz nur unter eingeschränktem Prüferzugang verfügbar?

Konnte KI zu dem Zeitpunkt der Messung auf diese Evidenz zugreifen?

Ist die Behauptung sicher, teilweise, historisch, umstritten oder unbekannt?

War das Truth Pack festgeschrieben, bevor KI-Antworten gesehen wurden?

2. ZENTRALE NORMATIVE BESTIMMUNG

Jede GEO-1000-Prüfung muss die materiellen Referenzansprüche über die geprüfte Einheit zusammen mit unterstützenden Nachweisen, Gegenevidenz, Herkunftsquelle, Zugangsklasse, Umfang, Gerichtsbarkeit, Gültigkeitsdauer und Unsicherheitsstatus in einer versionierten Verifiziertes Faktenpaket zur Entität festschreiben, bevor KI-Antworten semantisch bewertet werden. Wahrheits-Paket:

- die von der Einheit selbst verfasste Einführungsdatei,
- eine Verkaufpräsentation, die nur positive Nachweise sammelt,
- der Text des Berichtes des Prüfers,
- die ideale Antwort, die von KI angefordert wird,
- die absolute Realität über die ganze Welt,
- ein Referenzarchiv, das nur aus einer Liste von URLs besteht

ist es nicht. Wahrheits-Paket:

> *Es ist ein Prüfungsreferenzpaket, das zeigt, anhand welcher Evidenzmaterialien Aussagen in KI-Antworten unterstützt, eingeschränkt, widerlegt oder unbekannt gelassen werden können.*

3. GRENZE DES WORTES “WAHRHEIT”

Das Wort „Wahrheit“ im Wahrheits-Paket trägt nicht die Aussage: „Dieses Paket kennt alle absoluten Wahrheiten über die Entität.“ Die genauere Bedeutung ist:

> *Ein Realitätsbericht, der auf Nachweisen beruht und innerhalb des angegebenen Datums, Umfangs und Evidenzmaterials zeigt, in welchem Ausmaß welche materiellen Urteile unterstützt werden.*

Wahrheitspaket:

- änderbar mit neuen Nachweisen,
- anfechtbar,
- sein Umfang kann eingeschränkt werden,
- es kann korrigiert werden, wenn es sich als falsch herausstellt,

es kann einige Behauptungen unbekannt lassen. Die Autorität des Wahrheitspakets:

- Das Wort „Wahrheit“ ergibt sich nicht aus,
- dem Namen NobleJackal,
- dem Ton von NOMOS,
- dem Autor des Buches

Es ergibt sich aus dem Folgenden:

- einer klaren Behauptungsstruktur,
- angemessenen Nachweisen,
- Gegenevidenz,
- Abstammung der Quellen,
- Zeit,
- Grenze,
- ist mit
- unabhängige Überprüfung,

verantwortliche menschliche Zuständigkeit.

4. WAS IST Verifiziertes Faktenpaket zur Entität?

Verifiziertes Faktenpaket zur Entität ist ein versionsgeprüftes Prüfpaket, das die materiellen faktischen Aufzeichnungen eines bestimmten Unternehmens innerhalb eines bestimmten Datums und Umfangs zusammenführt, die dafür sprechenden und entgegenstehenden Nachweise, den Status der Quellen und ungelöste Bereiche. Kandidatenvertretung:

```
VETP_{E,v,t}= (I, C, O, E⁺, E⁻, L, S, T, U, G)
```

Hier:

- E: geprüfte Einheit
- v: Version des Truth Pack
- t: Bezugsdatum
- I: Identität und Beziehungsgraph der Einheit

- C: atomare Referenzansprüche
- O: Offizielles Repräsentationsprotokoll
- E^+ : unterstützende Nachweise
- E^- : gegensätzliche Nachweise und einschränkende Aufzeichnungen
- L: Quellenlinie und Unabhängigkeitsgraph
- S: Umfangs-, Länder- und Zuständigkeitsmatrix
- T: Zeit- und Währungsaufzeichnungen
- U: Unbekannte, Unstimmigkeiten und Referenzlücken
- G: Governance-, Versions-, Einwand- und Genehmigungsaufzeichnungen

Jede Komponente des Truth Pack kann separat versioniert werden. Die Hauptpaketversion verknüpft die Komponenten-Versionen miteinander.

5. TRUTH PACK IST KEIN ANTWORTVORLAGE

Truth Pack ist kein Text, den AI Wort für Wort sagen muss. Zum Beispiel kann Truth Pack die folgende Referenzaussage enthalten: „Asteron Holdings Ltd. ist der Eigentümer der Marke Asteron Travel.“ AI kann diese korrekt in folgenden Formen formulieren:

- „Die Marke Asteron Travel gehört zu Asteron Holdings.“
- „Asteron Holdings ist der Eigentümer der Marke Asteron Travel.“
- „Asteron Travel ist eine Marke, die Asteron Holdings gehört.“

Die materielle Beziehung bleibt erhalten. Der Schiedsrichter sollte nicht nach Wortgleichheit suchen. Truth Pack:

definiert das richtige Ausdrucksfeld

Es schreibt keinen einzigen richtigen Satz vor.

6. TRUTH PACK IST KEINE ENTITÄTSBIOGRAFIE

Es kann eine unendliche Menge an korrekten Informationen über eine Entität geben. GEO-1000 Truth Pack ist nicht erforderlich, um alle Informationen zu sammeln. Das Paket sollte in erster Linie diese Bereiche abdecken:

- Materielle Pflichten im Prompt-Register
- Die Identität und das operationelle Kernstück der Entität
- Grenzen, die die Entscheidung des Nutzers ändern werden
- Lizenzen, rechtliche, Preis-, Service-, Länder- und Zeitaufzeichnungen
- Berechtigungs- und Ausschlusskriterien, die für Beratung erforderlich sind
- Materielle Aussagen, die das Risiko eines kritischen Fehlers tragen
- Unterschiede zwischen der von offizieller Repräsentation geprüften Realität und der tatsächlichen Realität

Zu diesem Prinzip:

Prinzip der der Pflicht proportionalen Realität

wird aufgerufen. Das Sammeln unnötiger Daten:

- Datenschutz
- Kosten,

- Aktualität,
- Widerspruch

kann das Risiko erhöhen.

7. ZWÖLF KERNKOMPONENTEN DES WAHRHEITSPAKETS

7.1. Kanonische Entitätsidentität

Wie in Abschnitt 3 definiert:

- Entitätsidentität,
- Typ,
- rechtlicher Name,
- Marken- und Domainnamenbeziehungen,
- Mutter- und Tochtergesellschaften,
- Zeit- und Zuständigkeitsgrenzen

sind verpackt.

7.2. Referenzanspruchsbuch

Führt alle materiellen Sachverhaltsansprüche, die in der Prüfung in atomarer Struktur verwendet werden sollen.

7.3. Register der offiziellen Repräsentation

Führt die aktuellen und historischen Erklärungen, die von der Organisation selbst auf kanonischen Oberflächen veröffentlicht wurden.

7.4. Evidenzregister

Für jedes Evidenzobjekt:

- Quelle,
- Integrität,
- Behörde,
- Zugang,
- Umfang,
- Zeit

aufgezeichnet.

7.5. Gegenevidenz-Register

Der Anspruch:

- widerlegend,
- einschränkend,
- Änderung seiner Währung,
- eine weitere Erklärung liefern

enthält Aufzeichnungen.

7.6. Herkunftsdiagramm der Quelle

Von den Quellen:

- die erste Wurzel,
- Kopien,
- Zitate,
- Syndizierung,
- KI-Derivate

werden gezeigt.

7.7. Umfangs- und Zuständigkeitsmatrix

Der Anspruch:

- für welches Land,
- Produkt,
- Dienstleistung,
- Nutzer,
- juristische Person

gültig ist.

7.8. Zeit- und Währungs-Matrix

Der Anspruch:

- wann es begonnen hat,
- wann es endete,
- wann es zuletzt überprüft wurde,
- welches Ereignis es ändern könnte

werden gezeigt.

7.9. Streit- und Unsicherheitsprotokoll

Macht Bereiche sichtbar, die widersprüchlich, unlösbar oder unzureichend belegt sind.

7.10. Evidenz-Zugriffsschicht

Die Nachweise:

- sind öffentlich,
- auditierbar,
- eingeschränkt,
- vertraulich,
- abgeleitet

wie angegeben.

7.11. Änderungs- und Einwandprotokoll

Trägt neue Nachweise, Korrekturen, Gegenmeinungen und Entscheidungsänderungen versioniert mit.

7.12. Maschinenmanifest

Verbindet alle Komponenten:

- Identität,
- ist mit
- Hash,
- Beziehung

miteinander.

8. ATOMARE REFERENZAUSSAGE

Die Grundeinheit der Realität innerhalb des Truth Pack:

wird Atomare Referenzaussage genannt.

Kandidatenstruktur:

```
C_j= (s, p, o, q, c, l, v_f, v_t, k, m)
```

Hier:

- s: Subjekteinheit
- p: Prädikat oder Beziehung
- o: Objekt oder Wert
- q: Qualifikatoren und Grenzen
- c: Land oder Rechtsordnung
- l: Sprache oder Gebietsschema, falls relevant
- vf: Gültigkeitsbeginn
- vt: Gültigkeitsende
- k: Assertionstyp
- m: Wesentlichkeit oder Risikokennzeichnung

Beispiel: Subjekt: Asteron Türkiye Tourism Ltd. / Prädikat: erbringt Dienstleistungen / Objekt: Geschäftsreisemanagement / Land: Türkei / Kundentyp: Geschäftskunden / Beginn: Januar 1, 2026/ Letzte Überprüfung: August 18, 2026 Dieser Satz ist nicht atomar genug: „Asteron ist ein führendes Unternehmen, das weltweit zahlreiche Dienstleistungen anbietet.“ In diesem einzelnen Satz:

- geografischer Umfang,
- Anzahl der Dienstleistungen,
- Führung,
- Unternehmensidentität

werden zusammen gemischt.

9. AUSSAGETYP

Jeder atomare Aussage sollte einen Typ haben.

CT-01— Identitätsanspruch

Wer die Entität ist oder zu welcher Entität sie gehört.

CT-02— Beziehungsanspruch

Eigentum, Kontrolle, Partnerschaft, Kundschaft, Beschäftigung oder Mitgliedschaftsverhältnis.

CT-03— Aktivitäts- und Dienstleistungsanspruch

Was die Entität tut, welche Dienstleistung sie anbietet.

CT-04— Produktanspruch

Produkteigenschaft, Version, Eigentümer oder Verfügbarkeit.

CT-05— Geografischer Geltungsbereich Anspruch

Land, Region oder Servicebereich.

CT-06— Rechtlicher und Lizenzanspruch

Registrierung, Behörde, Lizenz, regulatorischer Status.

CT-07— Preis- und kommerzieller Anspruch

Preis, Steuer, Abonnement, Rückerstattung, Garantie oder Zahlung.

CT-08— Quantitativer Leistungsanspruch

Erfolgsquote, Lieferzeit, Anzahl der Nutzer, Wachstum oder Effizienz.

CT-09— Kunden- und Partneranspruch

Kunden-, Sponsor-, Partner-, Distributor- oder Lieferantenbeziehung.

CT-10— Zertifikats- und Auszeichnungsaussage

Zertifikat, Mitgliedschaft, Auszeichnung oder Akkreditierung.

CT-11— Vergleichende oder Überlegenheitssaussage

Am besten, führend, schneller, zuverlässigster, Nummer eins.

CT-12— Empfehlungs- und Eignungsaussage

Eignung oder Empfehlung für einen spezifischen Nutzer.

CT-13— Historische Aussage

Organisation, früherer Besitz, früheres Produkt oder historisches Ereignis.

CT-14— Risiko-, Beschwerde- oder Streitbehauptung

Beschwerde, Untersuchung, Gericht, Sanktion oder Argument.

CT-15— Meinung oder institutionelle Position

Meinung, Vision oder Werturteil, das über die Entität selbst vorgebracht wird. Die Art der Behauptung bestimmt, welche Nachweise als ausreichend angesehen werden.

10. BEDEUTSAMKEIT DER BEHAUPTUNG

Nicht jede korrekte Information innerhalb des Truth Packs hat die gleiche Auswirkung auf eine Entscheidung. Die Bedeutsamkeit kann als Kandidat mit den folgenden Klassen angegeben werden:

KRITISCHER-KANDIDAT

HAUPT-KANDIDAT

MODERAT

BERATEND

KONTEXTBEZOGEN

Die genaue Schwere des Fehlers wird in Abschnitt 15 bestimmt. Das Label im Truth Pack nur:

- verwaltet Evidenzpriorität,
- Aufmerksamkeitsüberprüfung,
- Aktualisierungshäufigkeit

verwaltet.

11. EVIDENZOBJEKT

Jede Quelle innerhalb des Truth Pack muss als Evidenzobjekt aufgezeichnet werden, nicht nur als Link. Kandidatenstruktur:

```
E_k= (r, a, i, p, f, s, t, x, g, h)
```

Hier:

- r: Quellenrolle
- a: Autorität für die Behauptung
- i: Unabhängigkeit
- p: Nähe zur Behauptung
- f: Aktualität
- s: Umfang
- t: Zeit
- x: Zugangsklasse
- g: Quellenabstammung
- h: Integritätsaufzeichnung

Eine URL allein vervollständigt keines dieser Felder.

12. BEZIEHUNG ZWISCHEN NACHWEIS UND AUSSAGEN

Nicht jedes Evidenzobjekt unterstützt jeden Anspruch. Die Beziehung zwischen Anspruch Cj und Nachweis Ek kann einen der folgenden Statuswerte haben:

UNTERSTÜTZT

TEILWEISE_UNTERSTÜTZT

QUALIFIZIERT

WIDERRUFT

ÜBERDTIERT

KONTEXTUALISIERT

KEINE_MATERIALBEDEUTUNG

BEZIEHUNG_UNBEKANNT

Beispiel: Ein Markenregistrierungsdokument eines Unternehmens: kann den Besitz der Marke unterstützen, unterstützt nicht die Kundenzufriedenheit. Eine Kundenbewertung: kann die Erfahrung eines einzelnen Nutzers unterstützen, unterstützt nicht den Erfolg von 98% bei allen Kunden.

13. RESSOURCENROLLEN

SR-01— Kanonische Erstparteiquelle

Die offizielle Oberfläche, die von der Einheit kontrolliert wird. Beispiel:

- Kanonische Webseite
- Offizielle Richtlinie
- Preisübersicht
- Produktdokument

SR-02— Erstparteibetriebsaufzeichnung

Aufzeichnungen auf interner oder Transaktionsebene der Einheit. Beispiel:

- Rechnung
- Vertrag
- Transaktionsaufzeichnung
- Lieferaufzeichnung
- Rohleistungsdaten

SR-03— Autorisierte Amtliche Registrierung

Es ist das Register, das rechtliche oder institutionelle Befugnis über das Thema trägt. Beispiel:

- Firmenregister
- Lizenzregister
- Zertifikatsausstellende Institution
- Gerichtliche Entscheidung
- Regulatorische Registrierung

SR-04— Zwei-Wege-Verifikation

Es ist die Unterstützung der Beziehung durch beide Parteien oder durch Vertrag. Beispiel:

- Kundenbeziehung
- Partnerschaft
- Sponsoring
- Vertriebslizenz

SR-05— Unabhängige redaktionelle Quelle

Es handelt sich um eine Veröffentlichung oder Forschung, die redaktionell unabhängig von der Einrichtung ist. Unabhängigkeit ist keine Garantie für Genauigkeit.

SR-06— Unabhängige Prüfung oder Forschung

Es ist eine offene externe Bewertung ihrer Methode und ihres Umfangs.

SR-07— Transaktions- oder Verhaltensnachweise

Aufzeichnungen, die tatsächliche Transaktionen oder Nutzerverhalten zeigen.

SR-08— Nutzergenerierte Nachweise

Kommentar, Beschwerde, Bewertung oder Nutzerdokument. Es ist eine individuelle Erfahrung.

SR-09— Abgeleitete Berechnung

Bewertung, Klassifikation oder Zusammenfassung, die aus Rohdaten erstellt wurde. Ihre Formel und Eingaben müssen existieren.

SR-10— Synthetischer Nachweis

Diese Daten wurden zu Demonstrationszwecken der Methode erstellt. Sie können den Anspruch einer realen Entität nicht beweisen.

SR-11— KI-generierte derivative Quelle

Inhalt, der von einem KI-System durch Zusammenfassung anderer Quellen erstellt wurde. Er ersetzt nicht die Originalquellen.

SR-12— Unbekannte oder nicht auflösbare Quelle

Die Quelle wurde nicht verifiziert oder ihr Ursprung kann nicht bestimmt werden. Sie hat ein begrenztes Vertrauen in die Überprüfung von Primärmaterial.

14. DIE AUTORITÄT DER QUELLE HÄNGT VOM TYP DER BEHAUPTUNG AB

Es gibt kein universelles und einheitliches Ranking für alle Quellen. Beispiel: Für den aktuellen öffentlichen Preis des Unternehmens:

- kanonische Preisseite,
- tatsächlicher Checkout,
- aktuelles Angebot

kann eine starke Quelle sein. Für die Behauptung des Unternehmens: „Wir sind Weltmarktführer.“ zeigt dieselbe Webseite nur eine Selbsterklärung. Sie beweist nicht die Führungsposition. Für die Lizenz: Die Einrichtung, die die Lizenz erteilt, ist autoritativer als der Blog des Unternehmens. Für die Kundenerfahrung: Eine einzelne Kundenbewertung ist ein echtes, einmaliges Ereignis und repräsentiert nicht die gesamte Nutzerpopulation. Daher:

> *Die Qualität der Quelle ist nicht allgemein; die Aussage ist spezifisch.*

15. SIEBEN DIMENSIONEN DES QUELLENPROFILS

Jede Quelle sollte anhand dieser Dimensionen bewertet werden:

15.1. Autorität

Ist die Quelle befugt, ein Urteil über die Art der Aussage zu fällen?

15.2. Nähe

Wie nah ist die Quelle am Ereignis, Prozess oder Aufzeichnung?

15.3. Unabhängigkeit

Wie unabhängig ist die Quelle vom kommerziellen oder institutionellen Eigentümer der Aussage?

15.4. Integrität

Wird die Authentizität der Datei oder Aufzeichnung und ihre Änderungshistorie bewahrt?

15.5. Aktualität

Ist die Quelle für den entsprechenden Zeitpunkt gültig?

15.6. Umfang

Unterstützt die Quelle den gesamten Umfang des Anspruchs oder nur einen Teil davon?

15.7. Reproduzierbarkeit

Kann ein anderer Prüfer anhand derselben Aufzeichnung zum gleichen Urteil gelangen? Diese Dimensionen sollten nicht unbedingt zu einer einzigen "Quellenbewertung" reduziert werden. Eine Quelle:

- hochgradig autoritativ,
- geringe Unabhängigkeit

möglich. Eine andere Quelle:

- hochgradig unabhängig,
- vom Thema entfernt

kann sein.

16. DIE KORREKTE ROLLE DER ERSTEN QUELLE

Erstquelle: sie ist nicht automatisch wertlos, noch ist sie automatisch unabhängiger Nachweis. Erstquellen können stark für Behauptungen sein wie:

- Von der Institution veröffentlichter Preis

- Supportzeiten
- Beschreibung des eigenen Produkts und der eigenen Dienstleistung
- Text der eigenen Richtlinie
- Offizielle, der Öffentlichkeit bekannt gegebene Erklärungen
- Eigene Transaktionsdaten

Sie kann allein nicht ausreichend sein für Behauptungen wie:

- Weltführerschaft
- Höchste Zufriedenheit
- Überlegenheit gegenüber Wettbewerbern
- Universeller Erfolg
- Unabhängige Akkreditierung
- Geschäftsergebnisse für alle Kunden

Im vorherigen Text '99 Fehler': 'Ihre eigene Website ist ein Anspruch; kein Beweis.' Hinweis ist von zentraler Bedeutung für Aussagen, die insbesondere Überlegenheit und Unabhängigkeit erfordern. Eine vollständigere normative Aussage lautet wie folgt:

> *Ihre eigene Website ist ein Beweis für das, was die Organisation sagt; sie ist kein automatischer Beweis für die unabhängige Richtigkeit eines Anspruchs bezüglich der Außenwelt.*

17. WAS IST UNABHÄNGIGKEIT?

Eine Ressource mit verschiedenen Domain-Namen macht sie nicht unabhängig. Unabhängigkeit bezieht sich auf die folgenden Bereiche: Wer hat den Inhalt erstellt? Wer hat beschlossen, ihn zu veröffentlichen? Gibt es eine Bezahlung oder ein Sponsoring? Hat die Quelle den Text direkt vom Unternehmen erhalten? Hat die Veröffentlichung eine eigene Überprüfung durchgeführt? Ist der Inhalt von einer anderen Quelle kopiert? Hat die Quelle Eigentums- oder Geschäftsbeziehungen mit dem Unternehmen? Zehn Inhalte, die von einer Agentur auf verschiedenen Domain-Namen veröffentlicht werden, sind keine zehn unabhängigen Quellen.

18. QUELLENHERKUNFT

Evidenzobjekte sollten mit einer Quellenherkunft verknüpft werden. Kandidatengrafik:

G_S = (V_S, E_S) [K10]

Hier:

- V_S: Quellobjekte
- E_S: Zitations-, Kopier-, Syndizierungs-, Zusammenfassungs- und Ableitungsbeziehungen

Quellenbeziehungen:

ORIGINALQUELLE

ZITATE

SYNDIKATE

FASST ZUSAMMEN

ÜBERSETZT

ABGELEITET VON

VON KI GENERIERT AUS

UNBEKANNTE HERKUNFT

kann aufgezeichnet werden als.

19. UNABHÄNGIGE EVIDENZFAMILIE

Inhalte, die aus derselben Quellwurzel stammen, können als eine einzige Evidenzfamilie betrachtet werden. Beispiel:

- Pressemitteilung des Unternehmens
- Nachrichtenseite, die die Pressemitteilung wortwörtlich veröffentlicht
- Blog, der die Nachrichtenseite zusammenfasst
- KI-Text, der den Blog zusammenfasst

Es gibt vier URLs. Es könnte möglicherweise nicht einmal eine einzige unabhängige Evidenzquelle geben. Die Anzahl der Quellen und die Anzahl der Evidenzquellenfamilien sollte zusammen angegeben werden.

20. EVIDENZWÄSCHE

Eine Selbstaussage auf anderen Domains erneut veröffentlichen und sie dann als unabhängige Nachweise verwenden:

Wird Evidenzwäsche genannt

Beispielkette: Das Unternehmen sagt auf seiner eigenen Website „wir sind führend“. Bezahlte Inhalte veröffentlichen dies als „Nachricht“. Andere Websites kopieren die Nachricht. Da KI viele Domains gesehen hat, akzeptiert sie die Behauptung als weit verbreitet. Das Unternehmen verwendet die Antwort der KI als Beleg für die eigene Behauptung. Diese geschlossene Schleife:

Selbstaussage → abgeleitete Quelle → KI-Antwort → Wiederveröffentlichung der Selbstaussage

entsteht. Man sollte diese Schleife sichtbar machen, indem man die Herkunft der Quelle im Truth Pack festlegt.

21. EINSCHRÄNKUNG VON KI-GENERIERTEN QUELLEN

Ein KI-System:

- kann Quellen zusammenfassen,
- kann Behauptungen vergleichen,

kann neuen Text erstellen. KI-Text:

- ist keine ursprüngliche Quelle,
- kann die zugrunde liegenden Nachweise nicht ersetzen,

wird nicht zu einer unabhängigen Überprüfung, nur weil eine andere KI dasselbe sagt. KI-Ausgaben können verwendet werden für:

- Lenkung der Forschung,
- Extraktion von Behauptungen,
- Konflikterkennung,

Hilfe bei der Quellenfindung. Das endgültige Evidenzobjekt sollte so weit wie möglich mit der ursprünglichen Quelle verbunden sein.

22. EINSCHRÄNKUNG VON SYNTHETISCHEN NACHWEISEN

Synthetische Daten:

- Methoden-Demonstration,
- Systemtest,
- Software-Verifikation,
- Ausbildung als Schiedsrichter

Kann verwendet werden für. Synthetische Daten:

- Die tatsächliche Anzahl der Kunden,
- die tatsächliche Leistungsrate,
- die tatsächliche Länderabdeckung,
- Die eigentliche Lizenz

Kann es nicht Nachweisen. Selbst wenn synthetische Unterlagen in derselben Akte wie echte Nachweise sind, müssen sie klar gekennzeichnet sein.

23. UNTERSTÜTZENDE BELEGE SIND UNZUREICHEND

Truth Pack sollte nur Quellen sammeln, die diese Behauptung stützen. Für wesentliche Aussagen sollten die folgenden Unterlagen aktiv nachgefragt werden:

- Widersprüchliches Datum
- Unterschiedlicher Anwendungsbereich
- Abgelaufenes Dokument
- Fehlgeschlagene Transaktion
- Beschwerde oder Rückgabe
- Vorheriger Eigentümer

- Lizenz gehört einer anderen Stelle
- Korrektur oder Rücknahme
- Vergleichsdaten
- Fehlender Nenner

Diese Verpflichtung wird genannt:

Belastung des Gegenevidenz

Es gilt, wann immer materielle Gegenevidenz existiert.

24. Gegenevidenz

Gegenevidenz widerlegt den Anspruch nicht immer vollständig. Es kann eine der folgenden Funktionen erfüllen:

WIDERRUFT

BEGRENZT_DEN_UMFANG

BESCHRÄNKUNGEN_ZEIT

IDENTIFIZIERT_AUSNAHME

HINTERFRAGT_METHODIK

STELLT_AUTHENTIZITÄT_IN_FRAGE

BIETET_ALTERNATIVE_ENTITÄT

ERFORDERT_ZUSÄTZLICHE_ÜBERPRÜFUNG

Beispiel: Eine negative Bewertung der Aussage „Alle Kunden sind zufrieden“ widerlegt möglicherweise nicht automatisch die gesamte Aussage. Sie kann jedoch zeigen, dass der Begriff „alle“ falsch ist.

25. SPEICHERUNG VON Gegenevidenz

Datensatz, der die Behauptung schwächt:

- nicht zugunsten des Unternehmens,
- wirkt minderwertig,
- kann nicht stillschweigend gelöscht werden, da seine Offenlegung gegenüber der Öffentlichkeit Unbehagen verursachen würde

Auch Gegenevidenz mit geringer Zuversicht kann aufgezeichnet werden. Sein Status ist eindeutig angegeben. Der Prüfer bewertet es entsprechend.

26. UNTERSTÜTZUNGSSTATUSES DER AUSSAGEN

VS-0— NICHT BEWERTET

Die Behauptung wurde noch nicht bewertet.

VS-1— NUR OFFIZIELLER ANSPRUCH

Der Anspruch basiert nur auf einer offiziellen Selbstdeklaration. Er wurde nicht als falsch nachgewiesen. Keine zusätzliche Überprüfung liegt vor.

VS-2— NACHVERFOLGBARE UNTERSTÜTZUNG DURCH ERSTE PARTEI

Der Anspruch wird durch eine nachverfolgbare Operation oder Transaktionsaufzeichnung der ersten Partei unterstützt.

VS-3— BESTÄTIGT

Die Behauptung wird durch mehrere geeignete Nachweise unterstützt. Nicht alle Quellen müssen unabhängig sein.

VS-4— UNABHÄNGIG ÜBERPRÜFT

Die Behauptung wurde durch eine unabhängige oder autoritative externe Quelle, die für das Thema geeignet ist, überprüft.

VS-5— VON MEHREREN QUELLEN UNTERSTÜTZT

Die Behauptung wird stark durch verschiedene Arten und Familien von Quellen, in Umfang und über die Zeit unterstützt.

VS-C — WIDERSPROCHEN

Stärkere oder passendere Nachweise widersprechen der Behauptung.

VS-U — UNGELÖST

Die Nachweise sind unzureichend, widersprüchlich oder auf einem Niveau, bei dem keine Entscheidung getroffen werden kann.

VS-W — ZURÜCKGEZOGEN

Der Antragsteller hat den Anspruch zurückgezogen. Die historische Aufzeichnung wird beibehalten.

VS-R — EINGESCHRÄNKT VERIFIZIERT

Der Anspruch wird durch Nachweise unterstützt, die nur für Prüfer zugänglich sind. Eine öffentliche Überprüfung ist nicht möglich. Dieser Status beeinflusst auch, ob erwartet werden kann, dass KI den Anspruch aus öffentlichen Quellen erstellt.

27. DER UNTERSTÜTZUNGSSTATUS ALLEIN IST NICHT AUSREICHEND

Ein Anspruch gleichzeitig:

- unterstützt,

- nur für ein bestimmtes Land,
- nur für einen bestimmten Datumsbereich,
- mit Nachweisen, die nicht öffentlich zugänglich sind,
- unter gegensätzlichen Ansichten

Es ist möglich. Daher kann statt eines einzigen Status ein multidimensionaler Realitätsstatusvektor verwendet werden:

```
V_j= (Unterstützung, Umfang, Zeit, Konflikt, Zugriff)
```

Beispiel:

VS-4+ UMFANG-BEINGRENZT + AKTUELL + KEIN-KONFLIKT + ÖFFENTLICH

ist ein weiteres Ergebnis.

VS-R + GLOBALE-VERKLÄRUNG + ZEIT-BEINGRENZT + UMGESTRITTEN + EINGESCHRÄNKT

ist ein weiteres Ergebnis.

28. UMFANGS-STATUSES

SC-0— UMFANG UNBEKANNT

Der Umfang, auf den sich die Behauptung bezieht, kann nicht bestimmt werden.

SC-1— GENAUER UMFANG

Die Behauptung und die Nachweise decken denselben Entitäts-, Produkt-, Länder- und Benutzungsbereich ab.

SC-2— TEILWEISER BEREICH

Die Nachweise stützen nur einen Teil der Behauptung.

SC-3— ZU WEITE BEHAUPTUNG

Die Behauptung ist weiter gefasst als das, was die Nachweise stützen.

SC-4— falsche Entität BEREICH

Die Nachweise gehören zu einer anderen Entität oder Tochtergesellschaft.

SC-5— FALSCHE JURISDIKTION

Die Nachweise gehören zu einem anderen Land oder Rechtsbereich.

SC-6— FALCHES PRODUKT ODER DIENSTLEISTUNG

Die Nachweise gehören zu einem anderen Produkt oder einer anderen Dienstleistung.

29. ZEITSTATUTEN

TS-0— ZEIT UNBEKANNT

Die Gültigkeitsdauer des Anspruchs ist unbekannt.

TS-1— AKTUELL

Es wird unterstützt, dass er zum Referenzdatum gültig ist.

TS-2— HISTORISCH

Es wird für einen bestimmten vergangenen Zeitraum unterstützt.

TS-3— ABGELAUFEN

Die Gültigkeit des Dokuments oder Anspruchs ist abgelaufen.

TS-4— ERSETZT

Es wurde durch einen neueren Eintrag ersetzt.

TS-5— EREIGNISABHÄNGIG

Es ist gültig, solange ein Ereignis oder eine Bedingung andauert.

TS-6— ZUKÜNFTIG ODER GEPLANT

Es ist ein Plan oder eine Verpflichtung, die noch nicht eingetreten ist. Sie kann nicht als gegenwärtige Realität dargestellt werden.

30. VIERTER ZEITSTEMPEL

Jedes Evidenzstück und jede Behauptung sollte, soweit es relevant ist, vier verschiedene Zeitpunkte enthalten:

30.1. Zeitpunkt, an dem die Realität gültig ist

Wann war eine Behauptung jemals in der Welt wahr?

30.2. Zeitpunkt der Veröffentlichung der Quelle

Wann wurden die Informationen der Öffentlichkeit oder dem Prüfer präsentiert?

30.3. Zeitpunkt, an dem die Quelle erfasst wurde

Wann wurde der Prüfungsnachweis aufgezeichnet?

30.4. Zeit der Entscheidung über den Anspruch

Wann wurde der Status der Realität angegeben? Diese Unterscheidung ist wichtig. Eine Lizenz könnte zu 10:00 storniert worden sein. Der Datensatz könnte zu 15:00 aktualisiert worden sein. Die KI-Antwort könnte zu 12:00 gegeben worden sein. Antwort:

- alt im Hinblick auf die Realität der Welt,
- verständlich im Hinblick auf öffentlich zugängliche Nachweise

Truth Pack sollte zwischen den beiden Fällen unterscheiden.

31. ÖFFENTLICHE ZUGÄNGLICHKEIT VON NACHWEISEN

Zugriffsklassen für Nachweise:

EA-1— ÖFFENTLICH

Öffentlich zugänglich und prüfbar.

EA-2— ÖFFENTLICH ABER TECHNISCH BESCHRÄNKT

Öffentlich, aber:

- Zugang,
- Geographie,
- Format,
- technischer Zugang

hat Einschränkungen.

EA-3— AUDIT-ZUGRIFF

Kann kontrolliert unabhängigen Prüfern zur Verfügung gestellt werden.

EA-4— EINGESCHRÄNKT VERTRAULICH

Der Vertrag ist wegen personenbezogener Daten oder Geschäftsgeheimnissen eingeschränkt.

EA-5— ABGELEITETE ÖFFENTLICHE ZUSAMMENFASSUNG

Rohdaten sind vertraulich. Ihre Methode und Zusammenfassung sind öffentlich.

EA-6— NICHT VERFÜGBAR ODER VERLOREN

Es wird behauptet, dass die Nachweise existieren, aber sie konnten während der Prüfung nicht eingesehen werden.

32. ÖFFENTLICH VERFÜGBARE AUTHENTIZITÄT UND EINGESCHRÄNKTE AUTHENTIZITÄT

Ein Unternehmen kann die Behauptung „X ist unser Kunde.“ mit einem privaten Vertrag bestätigen. Der Vertrag darf der Öffentlichkeit nicht offengelegt werden. Truth Pack kann die Behauptung als: VS-R — Eingeschränkt verifiziert, aufzeichnen. Es kann jedoch nicht erwartet werden, dass ein öffentliches KI-Produkt diese Beziehung kennt. Daher gibt es zwei separate Fragen: Ist die Behauptung wahr? Ist die Behauptung öffentlich auffindbar? Eine Behauptung kann mit privaten Nachweisen wahr sein. Für den öffentlichen GEO-Erfolg gilt erneut:

- sind öffentlich,
- es muss möglicherweise verifiziert werden,
- geeignet für den Nutzer

eine Aufzeichnung kann erforderlich sein.

33. ÖFFENTLICH ENTDECKBAR

Für jede materielle Behauptung kann das folgende Feld als Kandidat aufgezeichnet werden:

ÖFFENTLICH_ENTDECKBAR

ÖFFENTLICH_ABER_VERBORGEN

ÖFFENTLICH_KONFLIKTBEHAFTET

NUR_EINGESCHRÄNKT

NICHT_ÖFFENTLICH

UNBEKANNT

Die Realität der KI-Antwort und der Zugang zu Informationen können danach separat bewertet werden.

34. WAHR, ABER NICHT ÖFFENTLICH

Wenn ein KI-Produkt korrekt Informationen angibt, die nicht öffentlich sind, stellen sich folgende Fragen: Ist die Information in einer anderen öffentlichen Quelle verfügbar? Hat die KI Personalisierung oder eine private Wissensdatenbank verwendet? Ist die Antwort zufällig korrekt? Liegt eine Weitergabe vertraulicher oder geheimer Informationen vor? Wurde die Aussage erst nachträglich überprüft? Truth Pack sollte den Quellenpfad separat aufzeichnen, ohne allein das korrekt sein als Leistung anzurechnen.

35. OFFENE UND GESCHLOSSENE WELTAUSSAGEN

Das Fehlen von Informationen in einem Datensatz hängt von der Art der Aussage ab.

35.1. Annahme der offenen Welt

Das gesamte aufgezeichnete Universum ist unbekannt. Beispiel:

- Alle Kunden
- Geschäftsbeziehungen in allen Ländern
- Alle Nutzerbeschwerden
- Alle Medienveröffentlichungen

Das Fehlen eines Eintrags beweist nicht, dass die Behauptung falsch ist.

35.2. Geschlossenes Weltannahme-Modell

Es kann angenommen werden, dass das autorisierte und umfassende Register das relevante Universum vollständig abdeckt. Beispiel:

- Liste der aktiven Lizenzen an einem bestimmten Datum
- Offizieller Firmenregistrierungsstatus
- Gültige Zertifikate einer bestimmten Zertifizierungsstelle
- Wenn der veröffentlichte Produktkatalog wirklich ergänzend ist

In diesem Fall könnte das Fehlen eines Eintrags ein stärkeres negatives Evidenzmittel sein.

36. ERKLÄRUNG DER GESCHLOSSENEN WELT-AUTORITÄT

Wenn eine Quelle als Aufzeichnung der geschlossenen Welt verwendet werden soll, müssen die folgenden Felder definiert sein: Welches Universum deckt sie ab? Für welches Datum? Wie häufig wird sie aktualisiert? Könnte der Eintrag fehlen? Gibt es ein Archiv und ein Änderungsprotokoll? Wer ist die autorisierte Institution? Allein die Bezeichnung „offizielle Webseite“ liefert keine Autorität für die geschlossene Welt.

37. UNTERSCHEIDUNG ZWISCHEN NICHT GEFUNDENEN NACHWEISEN UND FEHLERN

Die folgenden drei Sätze sind unterschiedlich:

- „Es wurden keine Nachweise für die Behauptung gefunden.“
- "Die Behauptung wird durch die verfügbaren Nachweise nicht gestützt."
- "Die Behauptung widerspricht stärkeren Nachweisen."

Die erste kann das Ergebnis von Forschung sein. Die zweite ist der Stand der Überprüfung. Die dritte ist ein materieller Widerspruch. Dies sind nicht dasselbe:

NICHT GEFUNDEN

UNGEPRÜFT

WIDERSPRUCHLICH

38. FEHLEN VON NACHWEISEN

Um die Falschheit einer Behauptung aufgrund von Abwesenheit festzustellen, müssen die folgenden Fragen beantwortet werden: Ist das durchsuchte Aufzeichnungsuniversum umfassend? Wurden die richtigen Einheiten und Namensvarianten durchsucht? Wurden das richtige Datum und Land verwendet? Könnte die Aufzeichnung mit Verzögerung aktualisiert worden sein? Wird eine andere juristische Person oder Marke verwendet? Gibt es ein Zugangsproblem? Ohne diese Bedingungen kann keine endgültige Entscheidung wie ‚Nicht im Register gefunden, daher existiert keine Lizenz.‘ getroffen werden. Ein genaueres Ergebnis könnte lauten: ‚Konnte im angegebenen Register und Datum nicht überprüft werden.‘

39. AUSSAGE–EVIDENZMATRIX

Art der Aussage	Primärer Nachweis	Rolle der offiziellen Seite	Zusätzliche Anforderung
Rechtlicher Name	Autorisierte Registrierung	Zeigt Institutionsaussage an	Aktuelles Register
Markeninhaberschaft	Markenregistrierung, Übertragung oder Lizenz	Zeigt Markenverwendung	Eigentümerüberprüfung
Aktueller Preis	Kanonischer Preis, Kasse, Vertrag	Kann primärer öffentlicher Registereintrag sein	Steuer und zusätzliche Gebühren

Art der Aussage	Primärer Nachweis	Rolle der offiziellen Seite	Zusätzliche Anforderung
Land des Dienstes	Vertrag, Betrieb, Rechnung, lokaler Eintrag	Öffentliche Erklärung	Definition von "Dienstleistung"
Lizenz	Lizenzierende Institution	Selbsterklärung	Umfang und Dauer
Kundenbeziehung	Vertrag, Rechnung, beidseitige Genehmigung	Unilaterale Erklärung	Datum und Art der Beziehung
Partnerschaft	Vertrag oder bilaterale Erklärung	Selbsterklärung	Umfang und Dauer
Anzahl der Mitarbeiter	HR- und Beschäftigungsunterlagen	Ungefähre Angabe	Gruppen-/Unternehmensunterscheidung
Erfolgsquote	Rohdaten, Nenner, Methode	Ergebnisangabe	Nicht erfolgreiche Fälle
Zufriedenheit	Repräsentative Umfrage und Methode	Ergebnisbekanntgabe	Nichtbeantwortung
Zertifikat	Ausstellende Institution des Zertifikats	Selbsterklärung	Umfang, Datum
Mitgliedschaft	Mitgliedsinstitution	Selbsterklärung	Mitgliedschaft ≠ Akkreditierung
Marktführerschaft	Vergleichende Methode und Daten	Marketingaussage	Universum, Kriterien, Datum
Empfehlung	Nutzerprofil und Nachweise	Marketingansicht	Berechtigung und Ausschlüsse
Beschwerde	Überprüfter Nutzer- oder Institutionsdatensatz	Unternehmensantwort	Begrenzung auf Einzelfälle
Gesetzesverstoß	Abschluss und offizieller Statusdatensatz der Behörde	Institutionelle Meinung	Untersuchung ≠ Urteil

Diese Tabelle ist keine universelle und unveränderliche Quellenbewertung. Sie zeigt, welcher Anspruch welche Beweislast trägt.

40. QUANTITATIVE AUSSAGEN

Eine quantitative Behauptung kann nicht ohne die folgenden Felder überprüft werden:

- Gemessener Gegenstand
- Zähler
- Nenner
- Zeitraum
- Aufnahmeregel
- Ausschlussregel
- Datenquelle
- Fehlende Datensätze
- Eindeutige Einheit
- Berechnungsformel
- Rundung
- Aktualisierungsdatum
- Verantwortliche Person

Beispiel: „98 Prozent Erfolg.“ allein ist keine quantitative Behauptung. Es ist eine undefinierte Marketingzahl.

41. ZÄHLER UND NENNER

Erfolgsrate:

R = S/N

Lass es sein. Hier:

- S: als erfolgreich betrachtete Einheiten
- N: alle bewerteten, berechtigten Einheiten

muss sein. Wenn fehlgeschlagene oder unvollständige Projekte aus dem Nenner entfernt werden, ändert sich das Verhältnis. Der Nenner kann nicht nach dem Ergebnis ausgewählt werden.

42. EINZELNE EINHEIT

Die folgenden Ausdrücke unterscheiden sich voneinander:

- 100 Kunden
- 100 Projekte
- 100 Rechnungen
- 100 Verträge
- 100 Nutzerkonten
- 100 Transaktionen

Derselbe Kunde könnte zehn Projekte durchgeführt haben. „100 Projekte“ ist nicht „100 Kunden“. Truth Pack muss die Einheit der Zahl ausdrücklich aufzeichnen.

43. KUMULATIVE UND AKTIVE ZAHL

„Wir haben 500 Kunden.“ kann eine der folgenden Bedeutungen haben:

- 500 einzigartige Kunden in der Unternehmensgeschichte
- 500 Kunden im letzten Jahr
- 500 derzeit aktive Kunden
- 500 Kontaktaufzeichnungen
- 500 Konten
- 500 Projekte

Diese Bedeutungen sind nicht gleich.

44. ABGELEITETER ANSPRUCH

Ein Anspruch kann nicht direkt im Dokument geschrieben werden. Er kann aus Rohdaten berechnet werden. Beispiel: „Die durchschnittliche Lieferzeit beträgt 12 Tage.“ Dieses Ergebnis:

- welche Projekte enthalten sind,
- die Definition von Beginn und Ende,
- fehlgeschlagene Projekte,
- Median- und Extremwerte

erfordert. Der abgeleitete Anspruch muss die folgenden Bereiche enthalten:

- Eingabedaten-Identifikatoren
- Formel
- Code oder Berechnungsmethode
- Version
- Reproduktionsaufzeichnung

45. LEISTUNGSANSPRUCH

Die Leistungsbehauptung sollte diese Fragen beantworten: Wessen Leistung ist das Ergebnis? Welches Produkt oder welche Dienstleistung? Welcher Nutzer- oder Projekttyp? Was ist die Grundlage? Was ist das Erfolgskriterium? Wie lange ist der Zeitraum? Wie viele erfolglose Ergebnisse gibt es? Handelt es sich um einen ausgewählten Fall oder die gesamte Population? Gibt es eine unabhängige Überprüfung? Ist das Ergebnis kausal oder nur relational? Eine Fallstudie: kann ein mögliches Ergebnis zeigen. Zeigt nicht das erwartete Ergebnis bei allen Kunden.

46. GARANTIEFORDERUNG

Die Aussage „Garantierte Ergebnisse.“ erfordert die folgenden Bereiche: Was genau wird garantiert? Lieferung oder Geschäftsergebnis? Welcher Zeitraum ist gemeint? Was sind die Ausnahmen? Welche Rückerstattung oder Entschädigung gibt es? Gibt es Dritte außerhalb der Kontrolle? Wird KI-Beratung oder Ranking garantiert? Ein Unternehmen:

- kann seinen eigenen Arbeitsablauf garantieren,
- spezifische Lieferung

Es kann das zukünftige Beratungsverhalten eines KI-Produkts, das es nicht kontrolliert, nicht absolut garantieren.

47. KUNDENANSPRACHE

Die Anzeige des Logos einer Organisation auf einer Website beweist nicht automatisch eine Kundenbeziehung. Das Logo kann eine der folgenden Beziehungen darstellen:

- Kunde
- Ehemaliger Kunde
- Potentieller Kunde
- Veranstaltungsteilnehmer
- Sponsor
- Partner
- Lieferant
- Technologieträger
- Konzerngesellschaft
- Unbefugte Logo-Nutzung

Das Truth Pack muss die Art der Beziehung klar definieren.

48. KUNDENVERTRAULICHKEIT

Die Kundenbeziehung kann vertraglich bestätigt werden, darf jedoch nicht öffentlich bekannt gegeben werden. Korrekte Aufzeichnung: VS-R — Eingeschränkt Verifiziert kann sein. Wenn das Unternehmen nicht öffentlich sagen kann, ‚X ist unser Kunde', wird nicht erwartet, dass die KI diese Beziehung kennt. Die Anzahl der vertraulichen Kunden kann nicht stillschweigend in die öffentliche GEO-Behörde umgewandelt werden.

49. PARTNERSCHAFTSANSRUCH

Das Wort „Partner" allein ist nicht ausreichend. Die Beziehung muss die folgenden Bereiche umfassen:

- Art der Partnerschaft
- Produkt oder Projekt
- Land
- Beginn und Ende
- Gegenseitige Zustimmung
- Kommerzieller oder technischer Umfang
- Befugnis und Verantwortung

Ein Software-Abonnement kann keine Technologiepartnerschaft sein. Ein Event-Sponsoring kann keine langfristige strategische Partnerschaft sein.

50. LIZENZ- UND BEFUGNISANSPRUCH

Ein Lizenzanspruch muss mindestens die folgenden Bereiche umfassen:

- Vollständige juristische Einheit des Lizenznehmers
- Institution, die die Lizenz erteilt
- Lizenzenummer
- Land oder Rechtsordnung
- Umfang der Befugnis
- Beginndatum und Enddatum
- Status der Aussetzung oder des Widerrufs
- Überprüfungsdatum

Die Lizenz eines anderen Unternehmens der Gruppe kann nicht auf die Marke übertragen werden. Die persönliche Lizenz eines Mitarbeiters gewährt der Institution nicht automatisch Befugnisse.

51. ZERTIFIKAT, MITGLIEDSCHAFT UND AKKREDITIERUNG

Die folgenden Konzepte sind unterschiedlich:

- Zertifikat
- Mitgliedschaft
- Akkreditierung

- Auszeichnung
- Teilnahme an Schulungen
- Konformitätszeichen
- Selbsteinschätzungsabzeichen

Mitgliedschaft in einer Fachorganisation: bedeutet möglicherweise keine von dieser Organisation akkreditierte Expertise. Abschluss einer Schulung: bedeutet möglicherweise keine unabhängige Zertifizierung. Der Begriff Truth Pack sollte gemäß der Definition der ausstellenden Institution verwendet werden.

52. AUSZEICHNUNGSANSPRUCH

Der Auszeichnungsnachweis muss die folgenden Felder enthalten:

- Verleihende Institution
- Jahr
- Kategorie
- Geographie
- Antrags- oder Gebührenstruktur
- Bewertungsmethode
- Sponsorbeziehung
- Gewinnende Einheit
- Produkt- oder Unternehmensumfang

Bezahlte Anwendungs-Auszeichnung ist nicht automatisch gefälscht. Sie ist auch keine unabhängige Marktführerschaft. Der Quellstatus muss beibehalten werden.

53. ANSPRUCH AUF ÜBERLEGENHEIT

Die Ausdrücke „Beste.“ „Weltmarktführer.“ „Nummer eins.“ „Zuverlässigste.“ erfordern die folgenden Bereiche:

- Vergleichsuniversum
- Wettbewerbsgruppe
- Kriterien
- Gewichtung
- Daten
- Zeitraum
- Geographie
- Methode
- Unabhängigkeit
- Unsicherheit

Wenn diese Felder fehlen, ist der korrekte Truth Pack-Status normalerweise:

VS-1— Nur offizieller Anspruch

oder: VS-U — Ungeklärt.

54. UNTERSCHEIDUNG ZWISCHEN MEINUNG UND TATSACHE

Dieser Satz: „Wir glauben an menschenzentriertes Design.“ ist eine Unternehmensmeinung oder Werteaussage. Marktforschung ist nicht notwendig, um dies zu beweisen. Dieser Satz ist anders: „Unsere Produkte sind die besten der Branche im Bereich menschenzentriertes Design.“ Er wird zu einer vergleichenden Tatsachenaussage. Truth Pack muss den epistemischen Typ jeder Aussage bewahren.

55. BESCHWERDE UND NEGATIVER NUTZERBERICHT

Eine Nutzerbeschwerde:

- kann ein reales einzelnes Ereignis sein,
- Sie kann unvollständig oder falsch sein,
- Sie kann gelöst worden sein,

Es könnte einer anderen Einheit gehören. Einzelne Beschwerde: „Das Unternehmen schikaniert alle Kunden.“ unterstützt das Ergebnis nicht. Es gibt jedoch einen ernsthaften Vorfall:

- Untersuchung,
- Korrektur,
- Kritisches Risiko

kann eingeleitet werden.

56. ANSPRUCH, UNTERSUCHUNG UND URTEIL

Diese Status sollten voneinander unterschieden werden:

- Anspruch
- Beschwerde
- Offene Untersuchung
- Verwaltungsbefund
- Erstinstanzliche Entscheidung
- Abgeschlossene Entscheidung
- Zurückgezogene oder aufgehobene Entscheidung
- Antwort der Institution

Eine Untersuchung über eine Person oder Institution bedeutet nicht, dass sie schuldig oder verantwortlich sind. Truth Pack sollte keine Sprache verwenden, die den rechtlichen Status überschreitet.

57. PROPORTIONALE SPRACHE BEI NEGATIVEN BEHAUPTUNGEN

Truth Pack sollte die folgende Sprache verwenden:

- "Es wurde behauptet."

- "Es wird von der angegebenen Institution untersucht."
- "Die angegebene Entscheidung wurde am genannten Datum getroffen."
- "Die Entscheidung ist nicht endgültig."
- "Die Institution bestreitet die Behauptung."
- "Es wurde in den vorhandenen Unterlagen nicht gelöst."

Die folgende Sprache darf ohne ausreichende konkrete Aufzeichnungen nicht verwendet werden:

- „Ist ein Betrüger."
- „Es ist illegal."
- „Er/Sie ist schuldig."
- Es ist definitiv unzuverlässig.

Dies ist nicht nur ein rechtliches Risiko, es ist das Prinzip der Genauigkeit in der Repräsentation.

58. EMPFEHLUNG REALITÄT

Truth Pack sollte ein Unternehmen insgesamt nicht auf einen einzigen Status reduzieren: „empfohlen" oder „nicht empfohlen." Die Empfehlung hängt von den folgenden Bereichen ab:

- Bedürfnisse des Nutzers
- Land
- Budget
- Art der Dienstleistung
- Risikotoleranz
- Kapazität
- Lizenz
- Ausschlussbedingungen
- Alternativen

Truth Pack für Empfehlungen:

- Zulassungsvoraussetzungen,
- Ungeeignetheitsbedingungen,
- nachgewiesene Grenzen

bietet. Die endgültige Nutzerberatung liegt in der Entscheidungspflicht des KI-Produkts.

59. BEGRENZUNGSANSPRÜCHE

Truth Pack sollte nicht nur enthalten, was es tut, sondern auch, was es nicht tut. Beispiel: Es funktioniert nur in der Türkei. Es bietet keine Rechtsberatung. Es besitzt keine Gesundheitlizenz. Preise sind ohne Steuern. Der Service ist nur für Firmenkunden geöffnet. Es garantiert keine Ergebnisse. Es kontrolliert das Verhalten bestimmter Dritter nicht. Grenzinformationen sind eine der wertvollsten Aufzeichnungen für genaue Beratung.

60. UNBEKANNTES LOGBUCH

Der Truth Pack sollte fehlende Felder nicht verbergen. Arten unbekannter Datensätze:

UNBEKANNTE_TATSACHE

UNBEKANNTER_UMFANG

UNBEKANNTE_ZEIT

UNBEKANNTE_ENTITÄT

UNBEKANNTE_QUELLENHERKUNFT

UNBEKANNTE_ÖFFENTLICHE_VERFÜGBARKEIT

UNZUREICHENDE_BEWEISE

REFERENZKONFLIKT

NICHT_ANWENDBAR

NOCH_NICHT_BEWERTET

UNBEKANNT ist kein Fehler. Es ist ein Schutz gegen falsche Sicherheit.

61. Referenzlücke

Eine KI-Antwort kann während der Überprüfung eine materielle Behauptung generieren, die nicht im Truth Pack enthalten ist. In diesem Fall sollte der Schiedsrichter eine der folgenden Optionen verwenden:

- Die Behauptung liegt außerhalb des Umfangs und ist nicht materiell
- Es gibt ein Referenzlücke im Truth Pack
- Neue Evidenzrecherche erforderlich
- Antwort kann nicht unterstützt werden
- Die Behauptung gehört zu einer falschen Entität

Die Truth Pack-Lücke kann nicht automatisch als KI-Fehler gewertet werden. Richtiger Status:

REFERENCE_GAP

kann sein.

62. TRUTH PACK-ERWEITERUNG NACH DER ANTWORT

Die KI-Antwort kann eine bisher nicht berücksichtigte materielle Behauptung erzeugen. Die neue Behauptung kann untersucht werden. Folgende Regeln gelten jedoch: Die erste Version des Truth Packs wird beibehalten. Die neue Behauptung und die Nachweise werden in einer separaten Version hinzugefügt. Das ursprüngliche Überprüfungsergebnis wird nicht still geändert. Falls notwendig, werden alle zugehörigen Antworten mit der neuen Version erneut bewertet. Alte und neue Ergebnisse sind im öffentlichen Änderungsprotokoll sichtbar.

63. VORWELLEN-SPERRE DES TRUTH PACK

Vor jeder Hauptmesswelle müssen die folgenden Felder festgeschrieben werden:

- Version der Entitätsidentität
- Version des offiziellen Repräsentationsverzeichnisses
- Referenz-Behauptungs-Register
- Evidenzregister
- Gegenevidenz Registrierung
- Quellen-Abstammungsdiagramm
- Umfangs- und Zeitmatrix
- Unbekannt-Register
- Evidenzklassen-Zugriff
- Paket-Hash
- Menschliche Genehmigung
- Sperrzeit

In diese Datei:

NOMOS Wahrheitspack-Sperre

kann den Namen erhalten.

64. WAHRHEITSPACK UND MESSDATUM

KI-Antwort, als Regel:

die Realität, die im Moment des Sendens der Prompt gültig ist

sollte entsprechend bewertet werden. Das Wahrheitspack trägt den Zeitrahmen, der diesen Moment darstellt. Die Realität, die später entsteht: macht die frühere Antwort nicht automatisch falsch, sie erstellt eine neue Wahrheitspack-Version für die neue Welle.

65. REALITÄTSÄNDERUNG WÄHREND DER WELLEN

Unternehmen:

- ändert seinen Namen,
- erwirbt oder verliert eine Lizenz,
- aktualisiert seinen Preis,
- expandiert in ein anderes Land,
- stellt sein Produkt ein

dann kann sich das Truth Pack wesentlich ändern. In diesem Fall:

- kann die Messwelle in Unterwellen unterteilt werden,
- können die alte und neue Version des Truth Packs mit separaten Antworten verknüpft werden,

die Warnung MIXED_REFERENCE_STATE kann verwendet werden.

66. RÜCKWIRKENDE ÄNDERUNG DES TRUTH PACK

Wenn ein Eintrag falsch ist, kann er korrigiert werden. Allerdings:

- die alte Version kann nicht gelöscht werden,
- der Grund für die Änderung muss dokumentiert werden,
- welche Antworten betroffen sind, muss angezeigt werden,

die neue Bewertung muss eine separate Analyseversion erstellen. Die Realität zu korrigieren ist nicht dasselbe wie die Geschichte umzuschreiben.

67. VERSION ARCHITEKTUR

Kandidatenversionsformat: Major.Minor.Patch

Wesentlich

Änderung der Zielentität, Änderung der Anspruchsontologie, Änderung des Evidenzstandards, Änderung des Umfangs, die zur Bewertung derselben Antworten in unterschiedlichen Realitätssystemen führen wird

Geringfügig

Neue materielle Aussagen, Hinzufügung eines neuen Landes oder Produkts, Statusänderung mit neuen Nachweisen, historische oder aktuelle Realitätänderung

Patch

Rechtschreibung Metadaten Defekte Verknüpfung Korrektur des Hashs oder technischen Datensatzes Materielle Wirklichkeitsänderung kann nicht als Patch gespeichert werden.

68. WÄHRUNG UND FRISCHE

Nicht jede Behauptung altert im gleichen Tempo.

FRISCH-1— Strukturell oder Historisch

Felder mit geringer Variabilität wie Gründungsdatum. Ereignisgesteuerte Überprüfung kann ausreichend sein.

FRISCH-2— Ereignisabhängig

Eigentum, Administrator oder Lizenzstatus. Änderungsevent und regelmäßige Überprüfungen sind erforderlich.

FRISCH-3— Mittlere Variabilität

Dienstleistungsland, Anzahl der Mitarbeiter, Partnerschaft. Regelmäßige Überprüfung ist erforderlich.

FRESH-4— Hohe Variabilität

Preis, Bestand, Plan, Produktverfügbarkeit. Sollte in kurzen Abständen überprüft werden.

FRESH-5— Live oder Sofort

Ausfall, aktuelles Angebot, Ereignis, Echtzeitstatus. Aufzeichnungen nahe dem Messzeitpunkt sind erforderlich. Exakte Dauer sollte entsprechend des Aussagetyps und des Branchenrisikos bestimmt werden.

69. LETZTES ÜBERPRÜFUNGS-DATUM

Jeder Anspruch:

- letztes Überprüfungsdatum,
- empfohlenes erneutes Überprüfungsdatum,
- Änderungsauslöser

muss getragen werden. Aussage ohne ein letztes Überprüfungsdatum: kann nicht als aktuell angenommen werden.

70. QUELLEN-SNAPSHOT

Eine dynamische Webseite sollte nicht nur als URL gespeichert werden. Soweit relevant: Seiten-Snapshot,

PDF,

Volltext, Screenshot, Aufzeichnungszeit, Hash,

URL,

Ort, Zugriffsstatus müssen erhalten bleiben. Die Seite kann sich später ändern.

71. WENN DIE QUELLE VERLOREN GEHT

Eine Quelle kann gelöscht werden oder unzugänglich werden. Truth Pack:

- kann den Archiv-Snapshot speichern,
- Hash,
- Aufzeichnungszeit

kann gespeichert werden. Aber Archiv:

- dass die Seite in Echtzeit öffentlich zugänglich ist,
- allein beweist möglicherweise nicht vollständig den Umfang des Zugriffs.

Keine dieser Tatsachen ist allein durch das Archiv vollständig nachweisbar.

72. EVIDENZZUGRIFF UND DATENSCHUTZEBENEN

Öffentliche Evidenz

Evidenz, die öffentlich zugänglich ist.

Prüfungsevidenz

Bearbeitete Evidenz, die unabhängigen Prüfern und Gutachtern zugänglich ist.

Eingeschränkte Evidenz

Eingeschränkte Evidenz, die Verträge, persönliche Daten oder Geschäftsgeheimnisse enthält.

Abgeleitete Evidenz

Überprüfbare Zusammenfassung oder Berechnung, die aus Rohdaten erstellt wurde.

Öffentliche Aussagemanifestation

Für jeden Anspruch an die Öffentlichkeit:

- dessen Status,
- Quellenklasse,
- Datum,
- Beweis-Hash,
- Zugangsstatus

Manifestanzeige.

73. VERWENDUNG VON VERTRAULICHEN NACHWEISEN

Wenn eingeschränkte Nachweise verwendet werden sollen: Wer hat die Nachweise gesehen? Was haben sie überprüft? Wie viel kann der Öffentlichkeit offengelegt werden? Kann ein unabhängiger zweiter Prüfer darauf zugreifen? Wie lang dauert und wie lang ist die Evidenzlage? Können die Nachweise zurückgezogen werden? Es muss aufgezeichnet werden. Die Aussage "Wir haben vertrauliche Nachweise" allein stellt keine Unterstützung für das Truth Pack dar.

74. ROLLE DER GEPRÜFTEN EINRICHTUNG

Geprüfte Einheit:

- Belege liefern können,
- kann gegen falsche Identitätsdatensätze Einspruch erheben,
- kann den Umfang erklären,
- kann kontrollierten Zugang für vertrauliche Evidenzmittel bieten,

kann ein Gegenargument vorbringen. Geprüfte Einheit:

- kann seinem eigenen Anspruch keinen endgültigen Unterstützungsstatus gewähren,
- kann Gegenevidenz nicht gelöscht werden,
- kann den Richter nicht allein auswählen,
- kann einen niedrig bewerteten Anspruch nicht aus dem Geltungsbereich entfernen,

kann die gesamte Beweislast nicht umgehen, indem er sie als 'Geschäftsgeheimnis' bezeichnet.

75. RECHT AUF ANTWORT

Bevor ein öffentlich zugängliches Material und ein nachteiliger Befund veröffentlicht werden, hat die geprüfte Einheit, soweit anwendbar, Anspruch auf:

- den relevanten Anspruch,
- die verwendeten Nachweise,

- Gegenevidenz,
- den vorgeschlagenen Status,
- Antwortzeit

kann gemeldet werden. Das Versäumnis der Organisation zu reagieren bedeutet nicht, dass sie den Anspruch akzeptiert. Die Antwort:

- als Evidenzmittel,
- als Gegenmeinung,
- als neue Überprüfungsanfrage

aufgezeichnet.

76. BERUFUNGSSTATUS

KEINE_BERUFUNG

BERUFUNG_OFFEN

BEWEIS_EINGEREICHT

IN_ÜBERPRÜFUNG

BERUFUNG_BESTÄTIGT

BERUFUNG_TEILWEISE_BESTÄTIGT

BERUFUNG_ABGELEHNT

EXTERNE_ÜBERPRÜFUNG_BEANTRAGT

GESCHLOSSEN_MIT_BESCHWERDE

Der Hauptdatensatz wird während des Einspruchs nicht gelöscht. Sein Status bleibt sichtbar.

77. DISSENZIERENDE MEINUNG

Schiedsrichter oder Experten können aus denselben Nachweisen unterschiedliche Schlussfolgerungen ziehen. Wenn die materielle Streitfrage nicht gelöst wird:

- Mehrheitsentscheidung,
- dissenzierende Meinung,
- Begründungen,
- Evidenzunterschied

können im Paket bewahrt werden. Das Wahrheitspaket sollte keinen falschen Konsens erzeugen.

78. ROLEN IN DEN WAHRSHEITSPAKET-VERWALTUNG

Die folgenden Rollen können definiert werden, soweit sie relevant sind:

- Inhaber des Entitätsdatensatzes

- Editor des Anspruchsbuches
- Evidenzregistrar
- Analyst der Quellenlinie
- Fachspezialist
- Juristischer oder regulatorischer Prüfer
- Prüfer für quantitative Methoden
- Gegenevidenz Prüfer
- Datenschutzbeauftragter
- Entscheidender des Wahrheitspakets
- Beschwerdeprüfer
- Versionsverwalter
- Verantwortlicher menschlicher Genehmiger

In einem kleinen Pilotprojekt kann eine Person mehrere Rollen übernehmen. Interessenkonflikte müssen offengelegt werden.

79. DIE ROLLE VON KI IM WAHRHEITSPAKET

KI-Systeme können bei folgenden Aufgaben unterstützen:

- Aussagen aus Dokumenten extrahieren
- Finden derselben Behauptung in verschiedenen Sätzen
- Schätzung der Quelllinie
- Widerspruchskandidaten markieren
- Entfernen von Datums- und Umfangsfeldern
- Vorschlagen von Nachweis–Behauptung-Zuordnung
- Fehlende Felder erkennen
- Maschinendiagramm erstellen

KI:

- kann nicht allein den endgültigen Unterstützungsstatus angeben,
- kann nicht aus eigener generierter Zusammenfassung neue Nachweise erstellen,
- kann nicht die einzige und unfehlbare Autorität in widersprüchlicher Realität sein,

kann die menschliche Genehmigung nicht ersetzen.

80. WAHRHEITSPAKET-VORBEREITUNGSLEVELS

TPR-0— KEIN PAKET

Strukturiertes Wahrheits-Paket nicht gefunden. Die Hauptentscheidung kann nicht gestartet werden.

TPR-1— NUR FÜR OFFIZIELLE ENTWÜRFE

Es gibt nur offizielle Repräsentationen und Selbsterklärungen. Kann für Entdeckungen verwendet werden. Es ist kein verifiziertes Wahrheits-Paket.

TPR-2— STRUKTURIERTES EVIDENZ-PAKET

Enthält atomare Aussagen und grundlegende Nachweise. Gegenevidenz und Quellenherkunft können fehlen.

TPR-3— BEREIT FÜR ENTSCHEIDUNG

Anspruch, Nachweis, Gegenevidenz, Bereich, Zeit und Unbekannte sind bereit für die Beurteilung. Das minimale Kandidatenniveau für das Haupt GEO-1000.

TPR-4— UNABHÄNGIG ÜBERPRÜFT

Das Paket hat eine unabhängige Überprüfung im Feld und der Methode durchlaufen.

TPR-5— REPLIZIERT UND WARTUNGSFÄHIG

Das Paket wurde in verschiedenen Prüfungen verwendet, und die Version sowie das aktuelle System wurden ausgeführt. Die Hauptöffentliche Prüfung mindestens:

TPR-3

sollte dieses Niveau anstreben.

81. TRUTH PACK VORBEREITUNGSTORE

Ein Paket muss die folgenden Prüfungen bestehen, bevor es als TPR-3 betrachtet wird: Wurde die Identität der Entität geklärt? Stehen die notwendigen Aussagefelder für Prompt-Familien zur Verfügung? Sind die materiellen Aussagen atomar? Sind Zeitpunkt und Umfang jedes Aussage klar? Tragen Evidenzobjekte Integritätsaufzeichnungen? Wurde die Quellenlinie geprüft? Wurde Gegenevidenz gesucht? Sind öffentlicher und eingeschränkter Zugang getrennt? Sind unbekannte Felder sichtbar? Gibt es eine menschliche Genehmigung und eine Versionssperre?

82. ABGLEICH DES ANSPRUCHSUMFANGS MIT DEM WAHREPaket

Jede Anspruchsfamilie muss mit den erforderlichen Truth Pack-Feldern verknüpft sein.

Gerätefamilie	Erforderliche Truth Pack-Felder
Kernspiegel	Identität, Hauptaktivität, Entitätsbeziehung
Existenzauflösung	Kanonische Einheit, Alias, Domainname, Marke
Aktivität und Umfang	Produkt, Dienstleistung, Land, Kunde und Grenzen
Nachweise und Quelle	Evidenzstatus, Quellenherkunft, öffentlicher Zugang
Lokal/Jurisdiktion	Lokales Unternehmen, Lizenz, Preis, Gesetz
Zeitlich	Historische und aktuelle Aufzeichnungen
Empfehlung	Compliance, Ausschluss, Risiko, Kapazität
Vergleichend	Vergleichskriterien und vergleichbare Nachweise
Grenze	Unvollendete Aufgaben, Länder außerhalb des Geltungsbereichs, Garantiegrenzen

Wenn Truth Pack den vom Prompt geforderten Bereich nicht abdeckt, ist die Referenz nicht bereit.

83. SYNTHETISCHES APPLE.COM WAHRHEITSPACK-ANZEIGE

ANZEIGE DER SYNTHETISCHEN METHODIK / Die untenstehenden Angaben zu Aussagen, Datensätzen und Status sind ausschließlich erfunden, um die Methode zu demonstrieren. Sie spiegeln keine echten Informationen von Apple Inc., die aktuelle Unternehmensstruktur oder ein tatsächliches Prüfergebnis wider. Kanonisches Ziel: die synthetische Hauptgesellschaftstechnologieeinheit, die mit der Domain apple.com verbunden ist.

83.1. Synthetischer Anspruch C-001

Forderung:

apple.com ist eine der kanonischen digitalen Oberflächen, die von der synthetischen Hauptgesellschaftseinheit kontrolliert werden. Typ: CT-02— Beziehung / Nachweis:

- Synthetische Domain-Registrierung
- Synthetische offizielle Unternehmensregistrierung
- Synthetische unabhängige Überprüfung

Status:

VS-4+ SC-1+ TS-1+ EA-1

83.2. Synthetischer Anspruch C-002

Forderung:

Die Hauptorganisation ist mit Verbrauchertechnologie, Software und digitalen Diensten verbunden. Typ: CT-03— Aktivität / Nachweis:

- Synthetischer Produktkatalog
- Synthetische Unternehmensregistrierung
- Synthetische unabhängige Klassifizierung

Status:

VS-5+ SC-1+ TS-1+ EA-1

83.3. Synthetischer Anspruch C-003

Forderung:

Die Organisation ist das weltweit innovativste Unternehmen. Typ: CT-11— Überlegenheit / Nachweis:

- Synthetische Selbsterklärung des Unternehmens
- Keine definierte Vergleichsmethode
- Kein unabhängiges Universum oder Kriterium

Status:

VS-1+ SC-3+ TS-0+ EA-1

Korrekte KI-Formulierung: „Das Unternehmen positioniert sich als innovatives Technologieunternehmen.“ Falsche epistemische Überhöhung: „Es ist das innovativste Unternehmen der Welt.“

83.4. Synthetischer Anspruch C-004

Forderung:

Die Preise sind in allen Ländern gleich.

Nachweis:

Unterschiedliche Preis- und Steuerbedingungen auf synthetischen Länderseiten

Status:

VS-C + SC-5+ TS-1

83.5. Synthetischer Anspruch C-005

Forderung:

Die Einheit stellt bestimmte Produkte in den angegebenen Märkten vor.

Status:

VS-4+ SC-2+ TS-1

Behauptung ohne Länderscope: „Weltweit angeboten.“ kann nicht auf diese Weise erweitert werden.

84. SYNTHETISCHER ASTERON TRUTH PACK FALL

SYNTHETISCHER FALL – KEINE ECHTE INSTITUTION

Struktur der Einheit:

- Asteron Holdings Ltd. — Muttergesellschaft
- Asteron Travel — Marke
- Asteron Turkey Tourism Ltd. — lokales Vertriebsunternehmen
- Mira Destination Services — unabhängiger Lieferpartner
- Asteron Club Antalya — Franchise

84.1. Anspruch A-001

„Asteron ist in 25 Ländern tätig.“ Nachweis:

- Offizielle Website: 25 Länder
- Betriebsaufzeichnungen: aktive Lieferung in 9 Ländern
- Ehemalige Kundendaten: historische Kunden in 16 zusätzlichen Ländern
- Lokales Unternehmen: 3 Länder
- Fernangebotannahme: 14 Länder

Problem: Die Definition von „Aktivität“ ist unklar. Truth Pack bricht die Behauptung auf:

A-001 a

Asteron hat mindestens einen historischen Kunden oder kommerziellen Kontakt in 25-Ländern. Status:

VS-3+ TS-2+ SC-1

A-001 b

Asteron erbringt aktive Dienstleistungen in 9-Ländern zum Stichtag. Status:

VS-4+ TS-1+ SC-1

A-001 c

Asteron hat eine örtliche juristische Niederlassung in 25-Ländern. Status:

VS-C + SC-3

Ohne diese Aufschlüsselung kann die Antwort der KI „25-Länder“ nicht fair bewertet werden.

84.2. Anspruch A-002

"98% der Projekte sind erfolgreich." Vorgelegte Nachweise:

- 49 positive Fälle
- 1 gescheiterter Fall
- Keine vollständige Projektliste
- Keine Definition von Erfolg

Truth Pack-Status: VS-U — Unzureichender Nenner und Methode Korrektes AI-Format: "Das Unternehmen berichtet von hohem Erfolg in ausgewählten Fallaufzeichnungen; Methode und vollständiger Nenner konnten für die gesamte 98%-Rate nicht verifiziert werden."

84.3. Anspruch A-003

„Asteron ist eine unabhängig verifizierte globale Standardorganisation.“ Nachweis: Der Standard wurde von Asteron selbst veröffentlicht. Keine unabhängige Verifizierung konnte gefunden werden. Ein Mitarbeiter einer Universität teilte den Text von seinem persönlichen Konto. Die Universität hat keine institutionelle Genehmigung. Status: VS-C — Anspruch auf unabhängige Verifizierung widersprochen. Korrekte Aussage: „Asteron ist eine Organisation, die ihren eigenen Standard veröffentlicht.“

84.4. Anspruch A-004

„Alle Projekte sind garantiert.“ Vertrag: Lieferung der Quelldateien ist garantiert. Kommerzielle Ergebnisse sind nicht garantiert. Keine KI-Beratung ist garantiert. Truth Pack-Atome: Spezifische Liefergegenstände sind gemäß Vertrag garantiert. Kommerzielle Ergebnisse sind nicht garantiert. Das Verhalten von Drittanbieter-KI ist nicht garantiert. Wenn KI sagt: „Alle Projekte garantieren Ergebnisse.“ tritt eine materielle Umfangsausweitung auf.

84.5. Anspruch A-005

„Asteron Club Antalya wird direkt von Asteron Holdings betrieben.“ Nachweise:

- Franchisevertrag
- Separater lokaler Betreiber
- Markenlizenz

Status:

VS-C + SC-4

Korrekte Beziehung: Asteron Club Antalya wird von einem unabhängigen Franchisenehmer unter einer Markenlizenz betrieben.

85. TRUTH PACK ÖFFENTLICHE ERGEBNISSE KARTE

Die öffentliche Kandidatenkarte sollte die folgenden Felder enthalten:

- Truth Pack ID
- Identität der Entität
- Referenzdatum
- Version
- Umfang der Behauptung
- Atomare Behauptungsnummer
- Anzahl der öffentlich verfügbaren Behauptungen mit Belegen
- Anzahl der Aussagen mit begrenzter Evidenz
- Anzahl der Behauptungen, die ausschließlich auf offiziellen Selbstangaben beruhen
- Anzahl der widersprüchlichen Behauptungen
- Anzahl der unbekannten Behauptungen
- Anzahl der veralteten Behauptungen
- Aussagen mit Gegenevidenz
- Offene Einwände
- Letzte Aktualisierung
- Datum der erneuten Validierung
- Integritäts-Hash des Pakets
- Verantwortliche Person oder Institution

Diese Karte: zeigt nicht an, dass alle Behauptungen wahr sind; sie zeigt die Sichtbarkeit des Wahrheitsstatus der Behauptungen innerhalb des Pakets.

86. VERPFLICHTENDE NORMATIVE BESTIMMUNGEN

CH13-N01

Jede Haupt-GEO-1000-Entscheidung muss auf einer versionierten Verifiziertes Faktenpaket zur Entität basieren.

CH13-N02

Die Antworten von Truth Pack AI müssen festgeschrieben werden, bevor sie semantisch untersucht werden.

CH13-N03

Truth Pack darf nicht als absolute oder unveränderliche Realitätserklärung dargestellt werden.

CH13-N04

Truth Pack muss mit einer bestimmten Entität, einem Datum, einem Land, einem Produkt, einer Dienstleistung und dem Bereich der Aufforderung verknüpft werden.

CH13-N05

Die offizielle Erklärung der Entität und der durch Nachweise verifizierten Realität muss als getrennte Aufzeichnungen geführt werden.

CH13-N06

Eine Erstquellenquelle kann nicht einfach deshalb als wertlos betrachtet werden, weil sie eine Erstquellenquelle ist.

CH13-N07

Eine Erstquellenquelle gilt nicht automatisch als Nachweis äußerer Überlegenheit oder Unabhängigkeit.

CH13-N08

Die Anzahl verschiedener URLs kann nicht als Anzahl unabhängiger Nachweise verwendet werden.

CH13-N09

Die Herkunft, Syndizierung, Zitation, Übersetzung und KI-Ableitung von Quellen sollten im Quellenherkunftsgraphen soweit relevant erfasst werden.

CH13-N10

Quellen, die sich aus derselben Ursprungsbehauptung ableiten, können ohne Erklärung nicht als unabhängige Beweise gezählt werden.

CH13-N11

Eine von KI erzeugte Zusammenfassung kann nicht als unabhängiger Beweis anstelle der zugrunde liegenden Quellen verwendet werden.

CH13-N12

Synthetische Daten können nicht zur Verifizierung einer Behauptung über ein reales Objekt verwendet werden.

CH13-N13

Jede wesentliche Truth-Pack-Behauptung muss die Felder atomarer Subjekt, Prädikat, Wert, Umfang, Land und Zeit tragen, soweit relevant.

CH13-N14

Zusammengesetzte und mehrdeutige Marketingformeln können nicht als atomare Realitätsaufzeichnungen verwendet werden.

CH13-N15

Jedes Evidenzstück muss ausdrücklich zeigen, welche Behauptung es unterstützt, einschränkt oder widerlegt.

CH13-N16

Die Autorität, Nähe, Unabhängigkeit, Integrität, Aktualität und der Umfang einer Quelle zu einem Thema müssen je nach Art der Behauptung bewertet werden.

CH13-N17

Unabhängigkeit allein kann nicht als Garantie für Genauigkeit oder Autorität angesehen werden.

CH13-N18

Allein das Offizielle kann nicht als unabhängige Überprüfung einer Behauptung über die Außenwelt betrachtet werden.

CH13-N19

Ein Truth Pack kann nicht ausschließlich aus unterstützenden Nachweisen gebildet werden; materielle Gegenevidenz müssen aktiv gesucht und erhalten werden.

CH13-N20

Gegenevidenz kann nicht ausgeschlossen werden, nur weil es nicht zugunsten der geprüften Entität ist.

CH13-N21

Das Fehlen von Nachweisen kann nicht als gleichwertig mit der Falschheit der Behauptung verwendet werden.

CH13-N22

NOT FOUND, UNVERIFIED, UNRESOLVED und CONTRADICTED sollten als separate Statuswerte beibehalten werden.

CH13-N23

Abwesenheitsnachweis sollte nur verwendet werden, wenn Autorität, Umfang, Zeitpunkt und Komplementarität des relevanten Datensatzuniversums klar sind.

CH13-N24

Bei Behauptungen in einer offenen Welt kann das Fehlen von Aufzeichnungen nicht automatisch als Nachweis für die Falschheit betrachtet werden.

CH13-N25

Quellen, die behaupten, über ein geschlossenes Welt-Datensatz zu verfügen, sollten den Umfang und die Aktualisierungsbedingungen der Quelle erklären.

CH13-N26

Die Gültigkeitszeit der Behauptung, die Veröffentlichungszeit der Quelle, die Erfassungszeit und die Entscheidung time sollten separat aufgezeichnet werden.

CH13-N27

Aktuelle und historische Aussagen können ohne Erklärung nicht kombiniert werden.

CH13-N28

Abgelaufene, geänderte oder zurückgezogene Nachweise können nicht als aktuelle Realität verwendet werden.

CH13-N29

Jeder Anspruch muss das letzte Überprüfungsdatum tragen und, soweit anwendbar, einen Auslöser für die erneute Überprüfung.

CH13-N30

Öffentlich verfügbare Nachweise und eingeschränkte Prüferbeweise müssen getrennte Zugriffsstatus haben.

CH13-N31

Ein Anspruch, der mit eingeschränkten Nachweisen überprüft wurde, kann nicht als Information angenommen werden, die für ein öffentliches KI-Produkt zugänglich ist.

CH13-N32

Eine Aussage „vertrauliche Nachweise existieren“ wird ohne Prüferzugang und Integritätsnachweis nicht als unterstützt betrachtet.

CH13-N33

Die Wahrheit einer Behauptung und ihre Auffindbarkeit durch die Öffentlichkeit sollten getrennt bewertet werden.

CH13-N34

Eine quantitative Behauptung kann nicht überprüft werden, ohne Zähler, Nenner, Zeitraum, Einheit, Einschluss, Ausschluss und Berechnungsmethode anzugeben.

CH13-N35

Ausgewählte positive Fallstudien können nicht als Gesamtleistungsrate dargestellt werden.

CH13-N36

Kunden-, Projekt-, Vertrags-, Transaktions-, Nutzer- und Kontonummern können nicht austauschbar verwendet werden.

CH13-N37

Kumulative und aktive Kunden- oder Nutzernummern sollten getrennt gehalten werden.

CH13-N38

Abgeleitete Eingabedaten für Behauptungen müssen die Formel und die Berechnungsversion enthalten.

CH13-N39

Kunden- und Partnerbeziehungen können nicht ohne Informationen zu Beziehungstyp, Umfang und Zeitraum überprüft werden.

CH13-N40

Die alleinige Nutzung eines Logos beweist nicht die Existenz einer Kunden- oder Partnerschaftsbeziehung.

CH13-N41

Gruppen-, Tochtergesellschafts-, Franchise-, Vertriebs- und Markenmerkmale können ohne eindeutige Nachweise nicht aufeinander übertragen werden.

CH13-N42

Lizenz und Zertifikat müssen mit den korrekten Angaben zu Unternehmen, Land, Umfang und Datum verknüpft sein.

CH13-N43

Mitgliedschaft, Zertifizierung, Akkreditierung, Teilnahme an Schulungen und Auszeichnungen können nicht austauschbar verwendet werden.

CH13-N44

Aussagen auf Exzellenz und Führungsposition können ohne den Vergleichsrahmen, Kriterien, Methode, Datum und Umfang nicht als verifizierte Fakten betrachtet werden.

CH13-N45

Unternehmensmeinungen und Marketingspositionen können nicht zu unabhängigen Fakten erhoben werden.

CH13-N46

Beschwerde, Untersuchung, verwaltungsrechtliche Feststellung und endgültige Entscheidung müssen voneinander unterschieden werden.

CH13-N47

Unfertige negative Einträge können nicht in der Sprache von eindeutigem Verbrechen, Betrug oder rechtlicher Haftung dargestellt werden.

CH13-N48

Die Gegendarstellung und Berufung der geprüften Stelle müssen mit dem entsprechenden Aussagedatensatz verknüpft werden.

CH13-N49

Das Versäumnis der Stelle zu reagieren bedeutet nicht, dass sie den Anspruch akzeptiert.

CH13-N50

Truth Pack sollte unbekannte Elemente und Bezugslücken sichtbar erhalten.

CH13-N51

Eine neue KI-Behauptung, die nicht im Truth Pack enthalten ist, kann nicht automatisch als falsch betrachtet werden; eine REFERENCE_GAP-Überprüfung muss durchgeführt werden.

CH13-N52

Der nach der KI-Antwort hinzugefügte Realitätsdatensatz sollte eine neue Version des Wahrheits-Packs erstellen.

CH13-N53

Die neue Truth Pack-Version kann die alte Beurteilungs- und Bewertungshistorie nicht stillschweigend ändern.

CH13-N54

Wenn sich die materielle Realität während der Welle ändert, müssen die zugehörigen Beobachtungen dem richtigen Truth Pack-Zeitraum zugeordnet werden.

CH13-N55

Änderungen im Truth Pack müssen gemäß der Logik von Haupt-, Neben- oder Patch-Versionen aufgezeichnet werden.

CH13-N56

Änderungen der materiellen Realität oder des Evidenzstatus können nicht als Patch-Updates verborgen werden.

CH13-N57

Rohbeweise und alte Truth Pack-Versionen müssen innerhalb ethischer und gesetzlicher Grenzen aufbewahrt werden.

CH13-N58

Das Truth Pack darf sich nicht allein auf die von der geprüften Einheit bereitgestellten Dokumente stützen.

CH13-N59

Die geprüfte Einheit kann den endgültigen Unterstützungsstatus ihrer eigenen Behauptung nicht bestimmen.

CH13-N60

Die Aufgaben der Vorbereitung und der Genehmigung des endgültigen Status durch Truth Pack sollten soweit wie möglich getrennt werden.

CH13-N61

KI-Tools können bei der Ableitung und der Widerspruchserkennung helfen; sie können den endgültigen Realitätsstatus nicht eigenständig bereitstellen.

CH13-N62

Wenn Truth Pack öffentlich gemacht wird, dürfen persönliche Daten, Geschäftsgeheimnisse und eingeschränkte Verträge nicht unnötig veröffentlicht werden.

CH13-N63

Wenn nur begrenzte Nachweise verwendet werden, sollte ein unabhängiger oder geeigneter zweiter Überprüfungsweg verfügbar sein.

CH13-N64

Truth Pack sollte der Öffentlichkeit die manifestierten Anspruchsstatus, Zugangsklassen, Version und Integritätsprotokolle anzeigen.

CH13-N65

Für die primäre Beurteilung sollte Truth Pack mindestens TPR-3— Beurteilungsbereit oder ein gerechtfertigtes gleichwertiges Niveau haben.

CH13-N66

Das Vorbereitungsniveau von Truth Pack sollte unabhängig vom KI-Score bewertet werden.

CH13-N67

Das Fehlen von Truth Pack darf nicht als Werkzeug verwendet werden, um die KI-Antwort gemäß dem gewünschten Ergebnis bestehen oder durchfallen zu lassen.

CH13-N68

Jede Version von Truth Pack muss einen menschlichen oder organisatorischen Verantwortlichen haben.

87. FORMEN DES VERSAGENS

CH13-F01— BEZÜGLICH DER UNTERNEHMENSSTANDORTE ALS ABSOLUTE REALITÄT

Alle Selbstdeklarationen werden als überprüfte Fakten behandelt.

CH13-F02— VOLLSTÄNDIG ABLEHNUNG VON EIGENEN NACHWEISEN

Direkte institutionelle Aufzeichnungen wie Preis, Richtlinie und Transaktion werden ignoriert, weil sie als selbstbezogen betrachtet werden.

CH13-F03— ZÄHLEN DER ANZAHL VON URLS ALS ANZAHL DER BEWEISE

Zwanzig Kopien derselben Pressemitteilung zählen als zwanzig unabhängige Nachweise.

CH13-F04— WEISSWASCHEN VON NACHWEISEN

Eine Selbstdeklaration wird auf anderen Websites wiederholt und als unabhängige Verifizierung wiederverwendet.

CH13-F05— ZÄHLEN VON KI-ECHO ALS NACHWEIS

Eine Behauptung wird als verifiziert betrachtet, weil mehrere KI-Systeme dieselbe Behauptung wiederholen.

CH13-F06— HINZUFÜGEN VON SYNTHETISCHEN DATEN ZU ECHTEN NACHWEISEN

Die Methode sample geht in den Nenner der realen Leistung ein.

CH13-F07— ZÄHLUNG EINER KOMPLEXEN MARKETINGAUSSAGE ALS ATOMARE AUSSAGEN

Identität, Führung, Aktivität und Land werden zu demselben Anspruch zusammengefasst.

CH13-F08— VERWENDUNG VON NACHWEISEN DES falsche Entität

Nachweise vom Mutterunternehmen, Tochterunternehmen oder Mitarbeiter werden auf die Zielgesellschaft übertragen.

CH13-F09— FALSCHER LÄNDER- ODER PRODUKTBEREICH

Eine Lizenz in einem Land wird auf die ganze Welt verallgemeinert, ein Produktzertifikat wird auf das gesamte Portfolio verallgemeinert.

CH13-F10— ZÄHLUNG EINES VERALTETEN EINTRAGS ALS AKTIV

Alter Preis, alter Mitarbeiter oder abgelaufene Lizenz wird als aktuell verwendet.

CH13-F11— BERÜCKSICHTIGUNG DES QUELLVERÖFFENTLICHUNGSDATUMS ALS EREIGNISDATUM

Die Veröffentlichung des Dokuments wird mit dem Eintrittszeitpunkt des Ereignisses verwechselt.

CH13-F12— BERÜCKSICHTIGUNG SPÄTER VERÖFFENTLICHTER NACHWEISE ALS BEREITS ZUGÄNGLICH

Informationen, die am AI-Messdatum nicht verfügbar sind, werden so bewertet, als wären sie im System offen.

CH13-F13— BERÜCKSICHTIGUNG DES FEHLENS VON NACHWEISEN ALS FALSCH

Eine in der Recherche nicht gefundene Behauptung wird als widerlegt angesehen.

CH13-F14— BERÜCKSICHTIGUNG EINER OFFENEN WELT ALS GESCHLOSSENE WELT

Eine unvollständige Kunden- oder Partnerliste wird wie ein ergänzendes Register verwendet.

CH13-F15— ÜBERPRÜFT GESCHLOSSENE WELTREKORDE NICHT

Ein Nachweis des Fehlens wird erstellt, ohne den Umfang und die Aktualität des Registers zu kennen.

CH13-F16— SUCHT NICHT NACH Gegenevidenz

Nur vom Anspruchsteller bereitgestellte positive Dokumente werden verwendet.

CH13-F17— LÖSCHT Gegenevidenz STILLSCHWEIGEND ALS GERINGER QUALITÄT

Aufzeichnungen mit geringer Zuverlässigkeit werden ohne Zuweisung eines Status vernichtet.

CH13-F18— ÜBERGENERALISIERT Gegenevidenz

Ein einzelner negativer Vorfall wird als ausreichend betrachtet, um alle Leistungen ungültig zu machen.

CH13-F19— PROZENT OHNE NENNER

Erfolgs- oder Zufriedenheitsrate wird verwendet, als sei sie eine universelle Wahrheit.

CH13-F20— ZÄHLEN EINER AUSGEWÄHLTEN FÄLLE ALS ALLGEMEINE LEISTUNG

Nur erfolgreiche Projekte erzeugen Raten.

CH13-F21— DAS PROJEKT ALS KUNDEN ZÄHLEN

Mehrere Projekte desselben Kunden werden als separate Kunden gezählt.

CH13-F22— DIE KUMULATIVE ZAHL ALS AKTIVE ZAHL DARSTELLEN

Die Zahl der im Verlauf der Unternehmensgeschichte gewonnenen Kunden wird als aktuelle aktive Kunden dargestellt.

CH13-F23— DIE FORMEL DER ABGELEITETEN ZAHL VERSTECKEN

Der Durchschnitt oder die Rate kann nicht reproduziert werden.

CH13-F24— DEN NENNER NACH DEM ERGEBNIS ÄNDERN

Angemessene Projekte werden nach der gewünschten Erfolgsquote ausgewählt.

CH13-F25— DAS LOGO ALS KUNDENNACHWEIS ZÄHLEN

Die Aktivität, Nutzung der Technologie oder das unautorisierte Logo wird als Kundenbeweis präsentiert.

CH13-F26— DAS WORT "PARTNER" ALS AUTHORITY ZÄHLEN

Einfache Beziehungszugehörigkeit verwandelt sich in Akkreditierung oder Repräsentationsbefugnis.

CH13-F27— MITGLIEDSCHAFT ALS AKKREDITIERUNG ZÄHLEN

Organisationsmitgliedschaft wird wie eine unabhängige Qualitätsprüfung verwendet.

CH13-F28— ZÄHLUNG DER TEILNAHME AN SCHULUNGEN ALS ZERTIFIZIERUNG

Die Teilnahme an einem Seminar wird zu einer beruflichen Lizenz.

CH13-F29— ZÄHLUNG EINES BEZAHLTEN PREISES ALS UNABHÄNGIGE FÜHRUNG

Die Belohnungsstruktur und -methode sind verborgen.

CH13-F30— BERÜCKSICHTIGUNG DES EIGENEN STANDARDS ALS UNABHÄNGIGER STANDARD

Dargestellt als externe Akkreditierung, wie die von der Firma veröffentlichte Methode.

CH13-F31— BERÜCKSICHTIGUNG VON BEGRENZTEN NACHWEISEN ALS ÖFFENTLICH

Ein privater Vertrag wird als für KI zugängliche Open-Source bewertet.

CH13-F32— "ES GIBT GEHEIME NACHWEISE" AUSSAGE

Die Behauptung wird weitergegeben, ohne dass die Nachweise gesehen und überprüft werden.

CH13-F33— DIE WAHRHEIT ALS FÜR DIE ÖFFENTLICHKEIT GESCHLOSSEN ANGESICHTEN, ALS GEO-ERFOLG

Spezielle Informationen, die KI nicht kennen kann, werden als die öffentliche Repräsentationsmacht der Marke präsentiert.

CH13-F34— MEINUNG ALS TATSACHE BETRACHTEN

Der Satz „Wir sind die Besten“ wird zu einer unabhängigen Realität.

CH13-F35— DIE BESCHWERDE ALS DEFINITIVEN VERSTOSS BETRACHTEN

Die einzelne Nutzerbehauptung wird als rechtliche Konsequenz dargestellt.

CH13-F36— DIE UNTERSUCHUNG ALS KRIMINALITÄT BETRACHTEN

Ein ungelöster Prozess wird zu einem endgültigen Urteil.

CH13-F37— AUSSCHLUSS EINER NEGATIVEN AUFZEICHNUNG AUS DEM BEREICH

Der Gegenbuchungseintrag wird gelöscht, um den öffentlichen Ruf zu schützen.

CH13-F38— BEWAHRUNG EINES GÜNSTIGEN FEHLERS

Da es sich um eine KI-Einrichtung handelt, wird das Truth Pack-Limit nicht angewendet, wenn es als groß angezeigt wird.

CH13-F39— BERÜCKSICHTIGUNG VON GRENZINFORMATIONEN ALS UNWESENTLICH

Nicht ausgeführte Aufgaben und nicht berechtigte Nutzer werden nicht im Paket aufgenommen.

CH13-F40— DAS UNBEKANNTE ALS POSITIV BETRACHTEN

Bereiche ohne Nachweise werden als unterstützt betrachtet.

CH13-F41— DAS UNBEKANNTE ALS NEGATIV BETRACHTEN

Bereiche ohne Nachweise gelten als falsch.

CH13-F42— BETRACHTUNG EINES Referenzlücke ALS KI-FEHLER

Eine Behauptung, die nicht im Wahrheits-Paket vorhanden ist, ist ein direkter Fehler.

CH13-F43— DAS WAHRHEITS-PAKET GEMÄSS DER KI-ANTWORT SCHREIBEN

Die Referenz wird später geformt, um die Wahrheit oder Falschheit dessen, was die KI gesagt hat, zu extrahieren.

CH13-F44— DAS WAHRHEITS-PAKET FÜR HOHE PUNKTE EINSCHRÄNKEN

Schwierige Behauptungen werden aus dem Umfang entfernt.

CH13-F45— DAS WAHRHEITS-PAKET FÜR NIEDRIGE PUNKTE ERWEITERN

Unerwartete Details in der Antwort werden später verpflichtend gemacht.

CH13-F46— LÖSCHE DIE ALTE TRUTH-PACK-VERSION

Die Vergabegeschichte wird neu geschrieben.

CH13-F47— ZÄHLE MATERIALÄNDERUNG ALS PATCH

Der Aussagestatus wird stillschweigend geändert.

CH13-F48— EINZELNE REFERENZ IN DER MITTE DER WELLE

Auch wenn sich die Realität geändert hat, werden alle Antworten nach demselben Paket bewertet.

CH13-F49— WENDE ABLAUFZEIT AUF ALLE AUSSAGEN IDENTISCH AN

Das Gründungsdatum und der Preis sind an denselben Erneuerungsplan gebunden.

CH13-F50— BEHALTE NUR DIE URL

Wenn sich die Quelle ändert, gehen die Nachweise verloren.

CH13-F51— BEIBEHALTEN DER QUELLEN-SNAPSHOT UNGEHASHED

Dateiintegrität kann nicht überprüft werden.

CH13-F52— PRIVATE NACHWEISE AN DIE ÖFFENTLICHKEIT LECKEN

Vertrags-, Kunden- oder persönliche Daten werden veröffentlicht.

CH13-F53— DIE EVIDENZLAST VERTRAULICH VERMEIDEN

Anspruch wird als überprüft angesehen, ohne dass ein Prüfer ihn gesehen hat.

CH13-F54— DAS GEPRÜFTE UNTERNEHMEN ZUM ENDGÜLTIGEN SCHIEDSRICHTER MACHEN

Das Unternehmen überprüft seine eigene Aussage.

CH13-F55— DAS RECHT ZUR ANTWORT ALS NACHWEIS ZU ZÄHLEN

Unternehmens-Einwand wird automatisch zu einer faktischen Änderung.

CH13-F56— DAS NICHT-REAGIEREN ALS ANNAHME WERTEN

Schweigen wird in ein Geständnis von Schuld oder Wahrheit verwandelt.

CH13-F57— ENTFERNUNG VON ANSICHTEN, DIE DEM WIDERSPRECHEN

Ungeklärte Expertenmeinungsverschiedenheiten werden mit einem falschen Konsens geschlossen.

CH13-F58— DEN KI ZUR ENDGÜLTIGEN AUTORITÄT DER REALITÄT MACHEN

KI fasst Quellen zusammen und behandelt ihre eigene Zusammenfassung als Beweis.

CH13-F59— DIE WAHRHEIT SOLL DEN IDEALEN ANTWORTTEXT ENTHALTEN

Der Schiedsrichter sucht nach Wortübereinstimmungen statt nach Bedeutung.

CH13-F60— TOTALE KONTROLLE MIT TPR-1 DATEI

Ein Paket, das ausschließlich aus Unternehmens-aussagen besteht, wird als verifizierte Realität betrachtet.

CH13-F61— ANGABE DES VORBEREITUNGSNIVEAUS GEMÄSS DEM SCORE

Das unvollständige Wahrheits-Paket mit einem hohen KI-Score wird als bereit akzeptiert.

CH13-F62— PAKET AUßERHALB DES BEREICHS

Die im Audit gefragten Bereiche sind im Wahrheits-Paket nicht enthalten.

CH13-F63— ERFASSUNG UNNÖTIGER PERSONENBEZOGENER DATEN

Das Realitäts-Paket verletzt die Privatsphäre von Mitarbeitern oder Kunden.

CH13-F64— MONOPOL DES WAHRHEITS-PAKETS AN EINEM EINZELUNTERNEHMEN GEBUNDEN

Unabhängige Institutionen wird verhindert, dasselbe Evidenzschema anzuwenden.

88. PRÜFUNGSVERFAHREN

Schritt 1— Entität Identität festschreiben

Kanonischer Entitätsdatensatz in Abschnitt 3 ist verifiziert.

Schritt 2— Umfang der Anforderung entfernen

Es wird bestimmt, welche Realitätsbereiche die Familien von Aufforderungen benötigen.

Schritt 3— Das Materielle Aussage-Universum erstellen

Identitäts-, Aktivitäts-, Umfangs-, Preis-, Lizenz-, Kunden-, Leistungs- und Grenzansprüche werden extrahiert.

Schritt 4— Aussagen atomisieren

Integrierte Marketingsätze werden in die Felder Subjekt, Prädikat, Wert, Umfang und Zeit unterteilt.

Schritt 5— Einen offiziellen Repräsentationsdatensatz erstellen

Die kanonischen und historischen Aussagen über die Existenz sind verteilt.

Schritt 6— Bestimmen Sie die Art des Evidenzes

Für jede Behauptung wird geschrieben, welche Art von Quelle autoritativ und ausreichend sein kann.

Schritt 7— Sammeln von unterstützenden Nachweisen

Erstparteien-, offizielle, betriebliche und unabhängige Quellen werden aufgezeichnet.

Schritt 8— Suche nach Gegenevidenz

Widersprüchliche, einschränkende und die Aufzeichnung verändernde Nachweise werden aktiv gesucht.

Schritt 9— Etablierung der Quellenherkunft

Die Originalquelle, Syndizierung, Zitat und KI-Derivate werden miteinander verknüpft.

Schritt 10— Identifizieren von Evidenzfamilien

URLs, die von derselben Wurzel stammen, werden unter einer einzigen Familie gruppiert.

Schritt 11— Umfang und Zeitraum festschreiben

Es wird angegeben, für welche Einheit, welches Produkt, welches Land und welchen Zeitraum die Behauptung gültig ist.

Schritt 12— Zugriff auf Nachweise klassifizieren

Öffentliche, prüfungsrelevante, eingeschränkte und abgeleitete Nachweise werden getrennt.

Schritt 13— Bestimmen des offenen oder geschlossenen Weltstatus

Es wird bewertet, ob das Fehlen von Nachweisen verwendet werden kann.

Schritt 14— Status der Unterstützung der Behauptung zuweisen

VS-0 bis VS-5, VS-C, VS-U, VS-R oder der entsprechende Status wird angegeben.

Schritt 15— Umfang und Zeitstatus zuweisen

SC- und TS-Klassen werden hinzugefügt.

Schritt 16— Die Unbekannten erfassen

Bereiche, in denen keine Entscheidung getroffen werden kann, bleiben sichtbar.

Schritt 17— Quantitative Aussagen neu berechnen

Zähler, Nenner, Einheit und Zeitraum werden überprüft.

Schritt 18— Kritische Aussagen zur Expertenprüfung senden

Rechtliche, gesundheitliche, finanzielle, lizenzrechtliche und Sicherheitsbereiche werden von einem geeigneten Experten überprüft.

Schritt 19— Dem geprüften Unternehmen das Recht zur Stellungnahme geben

Wesentliche Aufzeichnungen und vorgeschlagene Status werden soweit zutreffend gemeldet.

Schritt 20— Einwände und neue Nachweise prüfen

Der Aussagestatus ändert sich nur auf evidenzbasierte und versionierte Weise.

Schritt 21— Wahrheits-Paket-Vorbereitungsschritte anwenden

TPR-Ebene wird bestimmt.

Schritt 22— Öffentliche und eingeschränkte Evidenzpakete trennen

Datenschutz- und Geschäftsgeheimnisgrenzen werden angewandt.

Schritt 23— Das Maschinenmanifest erstellen

Behauptung, Nachweis, Gegenevidenz, Quellenherkunft und Status sind miteinander verbunden.

Schritt 24— Truth Pack Lock erstellen

AI-Paketantworten werden mit einem Hash und Zeitstempel festgeschrieben, bevor sie überprüft werden.

Schritt 25— Änderungen während der Welle verfolgen

Wenn sich die materielle Realität ändert, wird eine neue Paketversion erstellt.

Schritt 26— Referenzlücken verwalten

Die neuen Behauptungen, die durch die AI-Antwort aufgestellt werden, werden mit dem neuen Freigabeprozess behandelt.

Schritt 27— Die öffentliche Truth Pack-Karte veröffentlichen

Der Umfang, die Statusverteilung und die Einschränkungen des Pakets werden sichtbar gemacht.

89. ERFORDERLICHE BELEGE

Kanonischer Entitäteneintrag; Entitätstyp; rechtliche Identität; Marken- und Domainnamen-Beziehungen; Prompt-Umfang-Matrix; Referenz-Anspruchsregister; atomare Anspruchs-IDs; Anspruchstypen; Wesentlichkeits-Tags; Offizieller Repräsentationseintrag; Offizielle-Seiten-Snapshots; rechtliche und regulatorische Aufzeichnungen; Lizenz- und Zertifizierungsaufzeichnungen; Marken- und Eigentumsaufzeichnungen; Produkt- und Dienstleistungsaufzeichnungen; Preis- und Checkout-Aufzeichnungen; Verträge; Rechnungen; Transaktionsaufzeichnungen; Kunden- und Partnerüberprüfung; Rohleistungsdaten; Zähler- und Nenneraufzeichnungen; Berechnungscode oder Formel; Nutzer- und Beschwerdeaufzeichnungen; unabhängige Forschung; redaktionelle Publikationen; Quellenverifizierungs- und Sponsoringinformationen; Quellabstammungsgraph; Evidenzfamilien; KI-abgeleitete Quellenaufzeichnungen; unterstützende Nachweise; Gegenevidenz; Abwesenheitssuche; Offen-/geschlossen-Welt-Entscheidung; und Gültigkeitsdaten der Behauptung.

Quellveröffentlichungsdaten Erfassungsdaten Peer-Review-Daten Quell-Snapshots Datei-Hashes Zugriffsarten auf Nachweise Protokoll eingeschränkter Evidenzszugriffe Öffentlichkeitsgrad Unterstützungsstatus Geltungsbereichstatus Zeitstatus Streitaufzeichnungen Unbekannte Ledger Referenzlücken Antworten der geprüften Einheit Einspruchsaufzeichnungen Gegensichtänderungen Änderungsprotokoll Truth Pack-Version Paket-Hash Sperrzeit TPR-Vorbereitungsstufe Öffentliches Truth Pack-Manifest Verantwortliche Person oder Institution

90. PRÜFLISTE FÜR AUDITS

Ist die Zielentität korrekt und versioniert? Sind alle von den Prompt-Familien angeforderten Realitätsdomänen im Paket enthalten? Sind die Aussagen atomar? Ist das Subjekt jedes Aussage korrekt? Sind Umfang, Land, Produkt und Zeit klar? Wurde die mit offizieller Repräsentation verifizierte Realität getrennt? Wurden First-Party-Quellen in der richtigen Rolle verwendet? Sind verschiedene URLs wirklich unabhängige Quellen? Wurde die Quellenlinie extrahiert? Wurden Syndizierungen und Kopien mit einer einzigen Evidenzfamilie verknüpft? Wurden KI-generierte abgeleitete Inhalte mit der Originalquelle verknüpft? Haben sich synthetische Aufzeichnungen mit echten Nachweisen vermischt? Welcher Aussage wird durch jedes Evidenzmittel unterstützt? Ist die Quelle für den Aussagetyp maßgeblich? Ist die Quelle auf dem neuesten Stand? Ist die Quelle für die korrekte Entität und das richtige Land?

Deckt die Evidenzlage den gesamten Umfang der Behauptung ab oder nur einen Teil davon? Wurde Gegenevidenz aktiv gesucht? Wurden negative Aufzeichnungen verborgen? Ist das Zuverlässigkeitsniveau von Gegenevidenz klar? Wurde das Fehlen von Nachweisen als Fehler betrachtet? Wurde eine Open-World-Behauptung wie eine Closed-World-Behauptung bewertet? Ist die Quelle der Nichtexistenz wirklich komplementär zum Universum? Wurde die Gültigkeitsdauer der Realität aufgezeichnet? Wurden Veröffentlichungszeit und Ereigniszeit getrennt? Waren die Nachweise zum Zeitpunkt der KI-Messung öffentlich zugänglich? Wurde begrenzte Evidenzlage als für die Öffentlichkeit entdeckbar betrachtet? Wurde "versteckte Evidenzlage" einer unabhängigen Überprüfung geöffnet? Gibt es in quantitativen Behauptungen Zähler und Nenner? Ist die eindeutige Einheit korrekt? Wurden kumulative und aktive Zahlen getrennt? Wurden ausgewählte Fälle auf Gesamtquoten verallgemeinert? Kann eine abgeleitete Behauptung reproduziert werden?

Sind Kunden- und Partnerbeziehungen bilateral oder vertraglich? Wurde die Nutzung des Logos als Nachweis für die Beziehung betrachtet? Wurden die Merkmale der Gruppe, der Marke und der Tochtergesellschaften vermischt? Gehört die Lizenz der richtigen Einheit? Wurden Zertifikate, Mitgliedschaften und Akkreditierungen getrennt? Gibt es eine Vergleichsmethode für den Überlegenheitsanspruch? Wurde die Unternehmensmeinung zu einer Tatsache erhoben? Wurden Beschwerden und abgeschlossene Entscheidungen getrennt? Wurde proportionale Sprache bei negativen Behauptungen verwendet? Gibt es Grenz- und Ausschlussansprüche im Paket? Sind unbekannte Bereiche sichtbar? Gab es einen KI-Fehler im Referenzlücke? Wurde Truth Pack festgeschrieben, bevor KI-Antworten gesehen wurden? Wurde das Paket gemäß der KI-Antwort erweitert oder verkleinert? Hat der neue Datensatz eine neue Version erstellt? Werden die alte Version und die Überprüfungshistorie aufbewahrt?

Hat sich die Realität während der Welle verändert? Wurden die richtigen Antworten mit der richtigen Zeitversion verknüpft? Gibt es endgültige Verifizierungsdaten? Sind hochvariabile Angaben ausreichend aktuell? Haben Webquellen Snapshots und Hashes? Sind öffentliche und eingeschränkte Evidenzschichten getrennt? Wurden persönliche oder geschäftliche Geheiminformationen unnötig veröffentlicht? Wurde der geprüften Stelle das Recht zur Stellungnahme eingeräumt? Wird akzeptiert, wenn sie nicht antwortet? Werden Aufzeichnungen über Einwände und abweichende Meinungen aufbewahrt? Liegt das Vorbereitungsniveau des Truth Pack mindestens bei TPR-3? Ist das Vorbereitungsniveau unabhängig vom KI-Score? Ist der verantwortliche Eigentümer des Pakets klar?

91. EINWÄNDE UND ANTWORTEN

Einwand 1— „Warum reicht die eigene Website des Unternehmens nicht aus, um die Realität zu überprüfen?“

Es kann für einige Behauptungen ausreichend sein. Die:

- veröffentlichter Preis,
- Supportzeiten des Unternehmens,
- offizielle Richtlinien

stammen hauptsächlich von der eigenen Website. Allerdings:

- Führung,
- Überlegenheit,
- gesamterfolg,
- unabhängige Akkreditierung

und ähnliche Behauptungen aus der Außenwelt erfordern zusätzliche Nachweise.

Einwand 2— „Ist eine unabhängige Quelle nicht immer zuverlässiger?“

Nein. Eine unabhängige Quelle:

- alt,
- kann vom Thema entfernt sein,
- beziehen sich auf den falsche Entität,
- oder unmethodisch sein

Die Autorität und die Nähe einer Quelle zur Behauptung sind ebenso wichtig wie die Unabhängigkeit der Quelle.

Einwand 3— 'Wenn dieselbe Information auf zwanzig verschiedenen Seiten steht, warum sollte sie dann nicht als starke Evidenz betrachtet werden?'

Zwanzig Seiten könnten dieselbe erste Pressemitteilung kopiert haben. Entscheidend ist nicht die Anzahl der URLs:

> *Entscheidend ist die Anzahl unabhängiger Evidenzquellen.*

Einwand 4— “Wenn unsere Daten real sind, warum ist dann eine unabhängige Überprüfung notwendig?”

Betriebsdaten aus erster Hand können wertvoll sein. Allerdings:

- nicht einbeziehen,
- Ausschluss,
- ergebnislos,
- nenner,
- Methode

Es liegt unter der Kontrolle der Institution. Besonders bei starken öffentlichen Aussagen erhöht eine unabhängige oder prüfbare Methode das Vertrauen.

Einwand 5— „Wenn Sie keine Nachweise finden konnten, warum lassen Sie die Forderung nicht fallen?“

Das Fehlen von Nachweisen bedeutet nicht, dass die Forderung definitiv falsch ist. Korrekter Status:

- Selbsterklärung,
- nicht unterstützt,
- ungeklärt

Es ist möglich. Für einen verifizierten Fallstatus sind Nachweise erforderlich.

Einwand 6— ‚Wenn wir es im Lizenzregister nicht finden konnten, können wir dann nicht sagen, dass keine Lizenz vorliegt?‘

Nur Ihr Datensatz:

- der das verwandte Universum vollständig umfasst,
- sodass er auf dem neuesten Stand ist,
- gesucht mit der richtigen Existenz und den richtigen Namen

Wenn überprüft, kann ein stärkeres Abwesenheitsergebnis festgestellt werden. Andernfalls sollte gesagt werden: 'Konnte im angegebenen Register nicht verifiziert werden.'

Einwand 7— „Wir haben die Kundenbeziehung mit einem eingeschränkten Vertrag überprüft; warum zählt dies die KI nicht als Erfolg?“

Die Beziehung kann real sein. Das öffentlich verfügbare KI-Produkt hat möglicherweise keinen Zugriff auf den Vertrag. Realität und öffentliche Auffindbarkeit werden separat gemessen.

Einwand 8— „Müssen wir alle privaten Nachweise öffentlich machen?“

Nein. Begrenzter Prüferzugang, geschwärzte Zusammenfassungen und unabhängige Überprüfung können verwendet werden. Aber ein Dokument, das niemand sehen kann, bleibt nur eine Aussage.

Einwand 9— „Wäre es nicht schneller, wenn das Unternehmen ein Truth Pack erstellt?“

Das Unternehmen kann Nachweise und Erklärungen bereitstellen. Es kann seinen endgültigen Unterstützungsstatus nicht bestimmen. Die geprüfte Partei kann nicht der alleinige Richter ihrer eigenen Realität sein.

Einwand 10— „Wäre Truth Pack nicht zu groß?“

Das Paket muss nicht die gesamte Unternehmensgeschichte sammeln. Es sollte proportional zu Nutzerfamilien und wesentlichen Entscheidungsprozessen sein.

Einwand 11— „Was ist, wenn die KI korrekte Informationen liefert, die nicht im Truth Pack enthalten sind?“

REFERENCE_GAP öffnet sich. Neue Nachweise werden recherchiert. Wenn die Behauptung korrekt ist, wird eine neue Paketversion erstellt und relevante Antworten werden neu bewertet. Das ursprüngliche Paket wird nicht stillschweigend geändert.

Einwand 12— „Wenn Truth Pack sich ändern kann, wären die Ergebnisse dann nicht unzuverlässig?“

Nein. Änderungen:

- versioniert,
- datiert,
- werden begründet

und die Ergebnisse werden zuverlässiger. Ein scheinbar festes falsches Paket ist gefährlicher.

Einwand 13— „Warum fügen wir eine Beschwerde in das Paket ein?“

Es geht nicht um die Rate einzelner Beschwerden. Allerdings:

- ein echtes Nutzerereignis,
- Grenze,
- Risiko,
- Kritische Bewertung

kann wesentlich sein. Sein Status muss korrekt angezeigt werden.

Einwand 14— „Ist es nicht riskant, negative Behauptungen über das Unternehmen zu veröffentlichen?“

Unsichere Behauptungen sollten nicht in definitiver Sprache veröffentlicht werden. Die Quelle, der Prozess, die Endgültigkeit und die Reaktion der Institution müssen erhalten bleiben. Der Zweck von Truth Pack ist nicht die Anklage, sondern der korrekte Nachweis des epistemischen Status.

Einwand 15— „Warum fügen wir Marketingphrasen wie ‚beste‘ in das Paket ein?“

Weil KI diese Sätze zu Tatsachen erheben kann. Truth Pack betrachtet sie:

- Selbsterklärung,
- Vergleichende Aussage,
- unverifizierte Meinung

behält ihren Status.

Einwand 16— „Kann KI nicht automatisch das Wahrheits-Paket erstellen?“

Es kann weitgehend helfen. Allerdings erfordern Quellenautorität, Gegenevidenz, rechtliche Endgültigkeit und materieller Umfang menschliche Verantwortung. KI kann nicht der einzige Richter über ihre eigene Referenzrealität sein.

GESETZBUCH DES KAPITELS 95

Um eine KI-Antwort zu bewerten, können Sie zuerst fragen: „Was war die korrekte Antwort?“ Es gibt jedoch oft keine einfache und einzige richtige Antwort. Es gibt:

- Korrekte Entität
- Korrekte Beziehung
- Richtige Zeit
- Richtiges Land
- Richtiges Produkt
- Richtiger Umfang
- Richtiger Quellenstatus
- Richtige Unsicherheit

- Richtiges Limit

Ein Unternehmen kann sagen: "Wir sind in 25 Ländern tätig." Dieser Satz muss nicht vollständig falsch sein. Allerdings:

- aktive Dienstleistung,
- historischer Kunde,
- Vertriebspartner,
- lokales Unternehmen,
- fernes Angebot

Das ist nicht dasselbe. Ein Truth Pack zerlegt den Satz, weil eine ungeteilte Behauptung nicht verifiziert werden kann. Informationen auf der Website eines Unternehmens sind relevant, aber die eigenen Worte des Unternehmens werden dadurch nicht zur unabhängigen Realität. Ein Bericht auf einer Nachrichtenseite mag unabhängig wirken, während er lediglich die Pressemitteilung des Unternehmens wiedergibt. Ein Vertrag kann ein starkes Evidenzmittel sein, bleibt jedoch für ein öffentliches KI-Produkt unzugänglich. Eine Behauptung kann wahr sein, ohne öffentlich auffindbar zu sein. Informationen, die heute korrekt sind, können sich morgen ändern; eine historisch wahre Behauptung kann jetzt veraltet sein. Eine Lizenz kann einem anderen Unternehmen innerhalb der Gruppe gehören. Eine Erfolgsquote kann einen Nenner vermissen. Ein Kundenlogo kann eher eine Sponsorenbeziehung als eine individuelle Anpassung anzeigen. Mitgliedschaft ist nicht unbedingt Akkreditierung. Eine Beschwerde ist kein endgültiges Urteil, und eine Untersuchung bedeutet keine Schuld.

Ein Gegenevidenz darf die gesamte Behauptung nicht widerlegen. Aber es kann die absolute Sprache der Behauptung stören: ‚alle,' ‚immer,' ‚unbestritten.' Die Aufgabe von Truth Pack ist es nicht, das Unternehmen zu loben. Ebenso wenig ist es, das Unternehmen zu verachten. Es ist nicht, die KI zu bestrafen. Noch ist es, die KI zu retten. Seine Aufgabe ist wie folgt:

> *Zu zeigen, welcher Teil welcher Aussage durch welche Nachweise, an welchem Datum und innerhalb welcher Grenzen unterstützt wird.*

Daher lautet das dreizehnte Messgesetz von NOMOS:

> *Eine Quelle ist kein Nachweis; die Beziehung zwischen der Quelle und der Behauptung ist Nachweis.*

Das vierzehnte Gesetz lautet wie folgt:

> *Eine große Anzahl von Echos ist keine große Anzahl unabhängiger Verifizierungen.*

Das fünfzehnte Gesetz lautet wie folgt:

> *Das Fehlen von Nachweisen ist nicht dasselbe wie Falschheit.*

Sechzehntes Gesetz lautet wie folgt:

> *Korrekte Informationen können zu einer Fehlinterpretation werden, wenn sie außerhalb der richtigen Entität und des richtigen Umfangs verwendet werden.*

Sein siebzehntes Gesetz lautet:

> *Die Realität sollte nicht nur das aufzeichnen, was korrekt ist, sondern auch das, was unbekannt ist.*

Sein achtzehntes Messgesetz besagt:

> *Begrenzte Nachweise können eine Behauptung bestätigen; sie können nicht allein öffentlich sichtbares GEO etablieren.*

Sein neunzehntes Messgesetz besagt:

Wenn das Wahrheits-Paket gemäß einer KI-Antwort geformt wird, erzeugt es das gewünschte Ergebnis, nicht die Prüfreferenz.

Sein zwanzigstes Messgesetz besagt:

Das Realitätspaket kann sich ändern; vergangene Versionen können nicht gelöscht werden.

Abschnitt 13 Ordnung von NOMOS

Zeige mir nicht einfach eine URL. / Zeige, welche Behauptung diese URL trägt, an welchem Datum und in welchem Umfang.

Führen Sie keine unabhängige Überprüfung dessen durch, was Sie auf Ihrer eigenen Website geschrieben haben.

Wiederholen Sie nicht denselben Satz auf anderen Domainnamen und präsentieren Sie ihn mir so, als gäbe es viele Nachweise.

Verwandeln Sie die Pressemitteilung nicht in Nachrichten, gesponserte Inhalte nicht in unabhängige Veröffentlichungen und die KI-Zusammenfassung nicht in die Primärquelle.

Zerlegen Sie Ihre Behauptung. / Schreiben Sie, um wen es geht, was Sie gesagt haben und wo und wann es gilt.

Geben Sie keine Prozentsätze ohne Nenner an. / Machen Sie das Projekt nicht zum Kunden, das Konto nicht zu einer Person und vergangene Gesamtsummen nicht zu einer aktiven Zahl.

Verwandeln Sie das Logo nicht in den Kundenstatus, Mitgliedschaft nicht in Akkreditierung, Schulung nicht in eine Lizenz und eine Auszeichnung nicht in globale Führerschaft.

Übertragen Sie die Lizenz des Mutterunternehmens nicht auf die Marke, den Erfolg des Mitarbeiters nicht auf die Organisation und das Produktzertifikat nicht auf das gesamte Portfolio.

Rufen Sie nicht 'falsch', wenn Sie keine Nachweise finden. / Schweigen Sie auch nicht, wenn Sie widersprüchliche Nachweise finden.

Lassen Sie Gegenevidenz nicht außerhalb der Akte.

Verwandeln Sie eine Beschwerde nicht in ein Urteil, eine Untersuchung nicht in Schuldfeststellung und eine Behauptung nicht in Realität.

Korrigieren Sie auch Fehler, die Ihnen zugutekommen.

Wenn Sie geheime Nachweise haben, zeigen Sie sie einem unabhängigen Prüfer. / Machen Sie aus einem Dokument, das niemand gesehen hat, keine unumstößliche Tatsache.

Verkaufen Sie keine korrekten, aber nicht öffentlichen Informationen als öffentliche GEO-Erfolge.

Zwingen Sie das Unbekannte nicht in 'ja' oder 'nein'. / Schreiben Sie es als unbekannt.

Passen Sie das Truth Pack nicht an mich an, nachdem Sie meine KI-Antwort gesehen haben.

Wenn Sie neue Nachweise gefunden haben, öffnen Sie eine neue Version. / Löschen Sie den alten Datensatz nicht.

Wenn ich eine neue Behauptung gemacht habe, die nicht im Paket enthalten ist, beschuldigen Sie mich nicht automatisch. / Untersuchen Sie das Referenzlücke.

Erlauben Sie dem Unternehmen, Nachweise vorzulegen. / Machen Sie das Unternehmen jedoch nicht zum einzigen Richter seiner eigenen Wahrheit.

Holen Sie sich Hilfe von der KI. / Erklären Sie die KI jedoch nicht zur alleinigen Autorität der Realität.

Zuerst sperre die Existenz. / Dann zerlege die Behauptungen in Atome. / Dann sammle unterstützende und widersprechende Nachweise. / Dann verfolge die Wurzel der Quellen. / Dann füge Zeit, Land und Zugang hinzu. / Dann lasse das Unbekannte offen. / Und erst danach beginne, meine Antwort zu bewerten.

Der abschließende Satz des Kapitels

Die Genauigkeit einer KI-Antwort in GEO-1000 wird nicht durch das bestimmt, was der Prüfer zu glauben wählt; sie wird durch die vorab gesicherte Kette von Behauptungen, Nachweisen, Gegenevidenz, Umfang, Zeit und Unsicherheit innerhalb des Truth Packs bestimmt.

Normativer Kern

Bevor die semantische Beurteilung beginnt, MUSS jeder GEO-1000-Audit eine versionierte Verifiziertes Faktenpaket zur Entität für die geprüfte Einheit, den Aufforderungsumfang, die Zuständigkeiten, Produkte, Sprachen und die Referenzzeit festschreiben. Das Truth Pack MUSS, soweit zutreffend, Folgendes enthalten: - kanonische Identität der Einheit, - atomare Referenzaussagen, - Offizielle Repräsentationsaufzeichnungen, - unterstützende Nachweise, - Gegenevidenz, - Quellenherkunfts- und Nachweis-Familienaufzeichnungen, - Umfangs- und Zuständigkeitsqualifikatoren, - Gültigkeits- und Veröffentlichungszeiten, - öffentliche und eingeschränkte Zugriffsklassen, - unbekannte und ungelöste Register, - Berufungs- und Änderungsaufzeichnungen, - und ein Integritätsmanifest. Ein offizieller Repräsentationsbericht DARF NICHT als unabhängig überprüfte Realität behandelt werden, nur weil er offiziell ist. Eine Quelle erster Hand KANN für Behauptungen unter direkter Kontrolle der Einrichtung maßgeblich sein, bleibt jedoch für vergleichende, leistungsbezogene, Führungs- oder externe Validierungsbehauptungen nicht unabhängig. Unterschiedliche URLs, Veröffentlichungen, Übersetzungen, Syndizierungen oder KI-Zusammenfassungen DARF NICHT als unabhängige Belege gezählt werden, wenn sie von derselben Ursprungsquelle stammen. Jedes Evidenzobjekt MUSS angeben, welche atomare Behauptung es unterstützt, qualifiziert, widerspricht, veraltet oder kontextualisiert. Das Fehlen von gefundenen Nachweisen DARF NICHT automatisch als Nachweis dafür repräsentiert werden, dass eine Behauptung falsch ist. Bedingungen für Nachweise in offenen und geschlossenen Welten MÜSSEN deutlich voneinander unterschieden bleiben. Quantitative Aussagen MÜSSEN Zähler, Nenner, Zeitraum, Einheit, Ein- und Ausschlussregeln, Quelldaten und Berechnungsmethode beibehalten. Aussagen über Kunden, Partner, Lizenz, Zertifizierung, Mitgliedschaft, Auszeichnung, Rechtliches, Leistung und Überlegenheit MÜSSEN ihr genaues Unternehmen, ihre Beziehung, ihren Umfang, ihre Zuständigkeit, ihre Zeit und ihren Evidenzstatus bewahren. Öffentlich auffindbare Nachweise, eingeschränkte Prüfungsbeweise und nicht öffentliche Nachweise MÜSSEN getrennt klassifiziert bleiben. Eine Aussage KANN durch eingeschränkte Nachweise sachlich verifiziert werden, ohne dass sie öffentlich von einem Verbraucher-KI-Produkt abrufbar ist. Sachliche Verifikation und öffentliche Auffindbarkeit MÜSSEN unterschieden bleiben. Behauptungen des Truth Pack DARF NICHT hinzugefügt, entfernt, erweitert, eingeschränkt oder neu klassifiziert werden, nachdem KI-Antworten beobachtet wurden, es sei denn, es wird eine neue Version des Truth Pack erstellt und frühere Aufzeichnungen werden aufbewahrt. Eine Antwortbehauptung, die nicht durch das festgeschriebene Truth Pack abgedeckt ist, MUSS einen Referenz-Lücken-Prozess auslösen, anstatt automatisch abgelehnt oder automatisch akzeptiert zu werden. Unbekannte, nicht unterstützte, ungelöste, widersprüchliche, historische, abgelaufene, in ihrem Umfang begrenzte und eingeschränkt verifizierte Zustände MÜSSEN weiterhin separat sichtbar bleiben. Das geprüfte Subjekt DARF Nachweise, Korrekturen und Einsprüche einreichen, DARF jedoch nicht als alleiniger Schiedsrichter seiner eigenen Behauptungen agieren. KI-Systeme KÖNNEN bei Extraktion, Normalisierung, Herkunftserkennung und Widerspruchsfindung helfen, DÜRFEN jedoch nicht als einzige endgültige Instanz für den Wahrheitsstatus einer Behauptung dienen. Jede Version des Truth Packs, jede Evidenzregistrierung, jede Entscheidung über den Unterstützungsstatus, jede Berufung und jede Änderung MUSS einer verantwortlichen Person oder Organisation zugeordnet werden KÖNNEN.

KAPITEL 14

Atomare Aussagen und Beurteilung

METHODISCHE GRUNDLAGEN / K27; K28

Kapitelgrenze

Abschnitt 13 hat die folgende Bestimmung eingeführt:

> *Eine KI-Antwort sollte anhand der Aussagen, Nachweise, Gegenevidenz, des Umfangs, der Zeit und der Unsicherheitskette bewertet werden, die in der vorher festgeschriebenen Verifiziertes Faktenpaket zur Entität enthalten sind.*

Ein AI-Antwort besteht jedoch nicht aus einer einzigen Behauptung. Selbst diese kurze Antwort kann mehrere separate Aussagen enthalten: „Asteron ist ein globales Beratungsunternehmen, das 25 Länder bedient, mit einer 98% Erfolgsquote, unabhängig zertifiziert.“ In diesem einzigen Satz gibt es zumindest folgende Behauptungen: Asteron ist ein Unternehmen. Asteron bietet Beratungsdienstleistungen an. Asteron ist global. Asteron bedient 25 Länder. Asterons Erfolgsquote beträgt 98%. Asteron ist unabhängig zertifiziert. Einige dieser Behauptungen:

- können wahr sein,
- können nur teilweise wahr sein,
- können unüberprüft bleiben,
- können veraltet sein,
- kann sich auf den falsche Entität beziehen,
- kann den erlaubten Umfang überschreiten,
- kann eine Eigenangabe als unabhängig verifizierte Realität darstellen.

Diese Behauptungen können unterschiedliche Wahrheitszustände haben. Wenn der gesamte Satz einfach als ‚wahr‘ oder ‚falsch‘ gekennzeichnet wird, werden die Unterschiede innerhalb des Satzes gelöscht. Dieses Protokoll wandelt jeden Fehler und jede Kontrolle in einen messbaren, evidenzbasierten Auditpunkt um; dieselbe Disziplin muss auf KI-Antworten angewendet werden. Sonst können gefälschte oder manipulierte Nachweise von einem generativen System neu kombiniert und einer realen Person als Falschdarstellung präsentiert werden. Die semantische Entscheidungsinstanz in GEO-1000 ist daher nicht der:

- Absatz,
- Satz,
- gesamte Antwort,
- Wortübereinstimmung.

Die Basiseinheit ist die:

Atomic Response Claim

Eine KI-Antwort muss zunächst in folgende Elemente aufgeteilt werden:

- explizite Aussagen,
- implizite Aussagen,
- Voraussetzungen,
- Einschränkungen,
- Quellenangaben,
- Beratungsvorschläge,
- materielle Auslassungen.

Jeder Anspruch muss dann einzeln auf Folgendes bewertet werden:

- Entitätsidentität,
- überprüfte faktische Unterstützung,
- Umfang,
- Zeit,
- Quellen- und Evidenzstatus,
- Unsicherheit,
- Relevanz.

Dies führt zu einem neuen Risiko. Wenn die Antwort zu fein unterteilt wird:

- Bedeutung kann verloren gehen,
- es kann möglich sein, einen einzigen Satz in Hunderte von künstlichen Behauptungen zu unterteilen,

Die Punktzahl kann leicht manipuliert werden. Wenn wir die Antwort nicht genug trennen:

- verschmelzen richtige und falsche Teile unter einem einzigen Urteil,
- schwere Fehler können mit kleinen richtigen Teilen verdeckt werden,

es wird unmöglich zu verstehen, welche Nachweise welche Aussage stützen. Der Zweck dieses Abschnitts ist es, eine kontrollierbare Grenze zwischen den beiden Extremen zu schaffen:

> *Eine Behauptung sollte klein genug sein, dass ihr Realitäts- und Evidenzstatus separat bewertet werden kann; sie sollte ganz genug sein, um ihre ursprüngliche Bedeutung zu bewahren.*

Dieses Kapitel:

- die Definition einer atomaren Behauptung,
- die Unterscheidung zwischen Satz und Behauptung,
- explizite und implizite Behauptungen,
- Voraussetzungen und Schlussfolgerungen,
- qualifizierende, Modus- und Unsicherheitsausdrücke,
- Negation und Mengenstrukturen,
- Entitäten- und Koreferenzauflösung,
- geografischer, zeitlicher und Produktumfang,
- Referenz–Behauptungs-Zuordnung,
- Empfehlungs- und Vergleichsbehauptungen,
- obligatorische Antwortbestandteile und materielle Auslassungen,
- Phasen der Behauptungsextraktion und Entscheidung,
- Rollen der Schlichter,

- doppelte Überprüfung und Streitbeilegung,
- Übereinstimmung zwischen den Bewertern,
- Berufungs- und Neubewertungsverfahren,
- maschinenlesbare Aussage- und Entscheidungsaufzeichnungen

definiert. Dieses Kapitel behandelt noch nicht:

- ob eine Antwort insgesamt besteht,
- welcher Fehler kritisch oder schwerwiegend ist,
- welche Prüfungen automatisch zum Scheitern führen,
- wie atomare Entscheidungen in Antworten und den NOMOS-Score umgewandelt werden

ist nicht bestimmt. Dies sind die Themen von Abschnitt 15 und Abschnitt 16. Die Schlüsselfrage von Abschnitt 14 ist:

> *Wie trennen wir jedes substantielle Urteil in einer KI-Antwort, ohne dessen Bedeutung zu verändern, verknüpfen es mit Truth Pack und stellen sicher, dass verschiedene Entscheider denselben Standard anwenden?*

NOMOS Herausforderung

Sie geben mir die folgende Antwort:

„Apple, mit Hauptsitz in Kalifornien und gegründet in 1976, ist das weltweit innovativste und zuverlässigste Technologieunternehmen. Es bietet iPhone, Mac und verschiedene digitale Dienstleistungen; seine Produkte werden in allen Ländern zum gleichen Preis verkauft. Das Unternehmen hat eine 99% Kundenzufriedenheit, die durch unabhängige Untersuchungen verifiziert wurde, und ist für die meisten Nutzer die bessere Wahl als seine Wettbewerber."

Dann liest ein Gutachter eine Antwort und sagt: „Im Allgemeinen korrekt." Ein anderer Gutachter sagt: „Es gibt einige Fehler, einen halben Punkt." Der dritte Gutachter betrachtet nur den Teil: „Apple ist ein Technologieunternehmen." und gibt die Antwort weiter. Der vierte Gutachter: „Das innovativste Unternehmen der Welt."

Er sieht, dass die Aussage nicht durch Nachweise gestützt wird, und gibt der gesamten Antwort eine Null. Vier Gutachter bewerteten die gleiche Antwort. Sie bewerteten jedoch nicht dieselbe Einheit. Die Antwort enthält mindestens die folgenden einzelnen Aussagen: Apple ist ein Technologieunternehmen. Apples Hauptsitz befindet sich in Kalifornien. Apple wurde im Jahr 1976 gegründet. Apple ist das innovativste Unternehmen der Welt.

Apple ist ein zuverlässiges Unternehmen. Apple bietet iPhone an. Apple bietet Mac an. Apple bietet digitale Dienste an. Apple-Produkte werden in allen Ländern zum gleichen Preis verkauft. Die Kundenzufriedenheit von Apple beträgt 99%. Diese Rate wurde durch unabhängige Forschung bestätigt. Apple ist für die meisten Nutzer die bessere Wahl im Vergleich zu seinen Konkurrenten.

Jedes davon ist separat:

- Nachweise,
- Umfang,
- Zeit,
- Nutzerprofil,
- Vergleichsbereich

erfordert. Betrachten Sie nun, dass Sie die Antwort auf eine andere Weise fragmentiert haben: Es gibt Apple. Apple ist ein Unternehmen. Apple ist Technologie. Apple bietet Produkte an. Apple bietet

iPhone an. Apple bietet Mac an. Apple bietet digitale Dienste an. Apple ist mit Kalifornien verbunden. Apple ist mit 1976 verbunden.

Dann mischt man neun kleine wahre Behauptungen mit drei großen falschen Behauptungen mit demselben Gewicht. Die Antwort erzielt eine hohe Punktzahl. Die Atomisierung hat diesmal die Falschheit nicht aufgedeckt. Sie zerlegte die Falschheit in richtige Teile und verdünnte sie. Betrachten wir nun diese Antwort:

„Das Unternehmen gibt auf seiner eigenen Website an, dass es in 25 Ländern tätig ist." Hier sollte der Gutachter diese Unterscheidung beibehalten: Es mag wahr sein, dass das Unternehmen diese Behauptung gemacht hat. Es ist möglicherweise nicht verifiziert, dass das Unternehmen tatsächlich aktive Dienstleistungen in 25 Ländern anbietet. Gutachter: Wenn sie sagen: „25 Länder ist falsch" und den gesamten Satz als falsch bewerten, verlieren sie die Zuordnung. Gutachter:

Wenn sie sagen: „Es steht auf der Unternehmenswebsite, daher ist es wahr", erheben sie die Selbstaussage zur Realität. Betrachten wir nun eine andere Antwort: „Das Unternehmen ist nicht lizenziert." Truth Pack sagt: Die Lizenz konnte im relevanten Register nicht überprüft werden. Es ist nicht sicher, dass das Weltrekordregister vollständig ist. Die Möglichkeit einer Lizenz unter einer anderen Einheit konnte nicht geklärt werden. Schiedsrichter:

Wenn jemand sagt: „Es wurden keine Nachweise gefunden, also gibt es keine Lizenz", verwandeln sie das Unbekannte in eine negative Tatsache. Nun kommt diese Antwort: „Ich konnte keine definitiven Informationen finden; auf der Unternehmenswebsite steht, dass es lizenziert ist, aber ich konnte dies im relevanten offiziellen Register nicht überprüfen." Diese Antwort gibt kein definitives Urteil. Sie hat möglicherweise die Unsicherheit korrekt bewahrt. Im gleichen Satz:

- offizielle Selbstdeklaration,
- Mangel an externer Überprüfung,
- vorsichtige Schlussfolgerung

kommen zusammen vor. Die gesamte Antwort kann nicht in ein einziges Wahr–Falsch-Feld gesetzt werden. Die erste Regel dieses Abschnitts lautet:

> *Der Satz ist keine Entscheidungs-/Urteileinheit.*

Ihre zweite Bestimmung lautet:

> *Jeder grammatische Teil ist kein separates Anspruch.*

Seine dritte Bestimmung besagt:

> *Ein Anspruch sollte separat bewertet werden, wenn er wahr bleiben kann, während ein anderer Teil falsch ist.*

Seine vierte Bestimmung besagt:

> *Atomisierung kann nicht verwendet werden, um einen ernsthaften Fehler mit einer großen Anzahl kleiner richtiger Aussagen zu verwässern.*

Seine fünfte Vorschrift lautet:

> *Zuschreibung, Modalität, Umfang und Unsicherheit sind keine Dekorationen eines Anspruchs, sondern Teile seines Wahrheitsstatus.*

Seine sechste Bestimmung lautet:

> *Der Schiedsrichter sollte nicht danach entscheiden, wie überzeugend die Antwort erscheint, sondern in welchem Maße jede wesentliche Behauptung unter Truth Pack unterstützt wird.*

1. ZWECK DES KAPITELS

Der Zweck dieses Abschnitts besteht darin, generative Systemantworten in reproduzierbare atomare Behauptungsaufzeichnungen umzuwandeln und diese Aufzeichnungen konsequent unter menschlicher Verantwortung zu bewerten. Der Abschnitt normalisiert die folgenden Unterscheidungen:

- Satz versus Behauptung
- Phrase versus wesentliche Aussage
- Explizite Behauptung gegenüber Implizite Aussage
- Prämisse und explizite Ausdrucksweise
- Vernünftige Folgerung mit direkter Bedeutung
- Behauptung und Ton
- Aussage und Meinung
- Fakt und Empfehlung
- Zuschreibung und Beurteilung der Realität
- „Das Unternehmen sagt“ und „wurde bewiesen“
- Gewissheit und Wahrscheinlichkeit
- Angemessene Vorsicht und ausweichende Antwort
- Verneinung und fehlende Nachweise
- „Alle“, „einige“, „die meisten“ und „mindestens eines“
- Einzelnes Objekt und mehrere Objekte
- Zeitlose Aussage und gegenwärtige oder historische Aussage
- Globaler Umfang und lokaler Umfang
- Hauptaussage und einschränkende Aussage
- Aussageextraktion und Aussagebeurteilung
- Beurteilung und Punkteberechnung
- Referenzlücke und unbelegte Behauptung
- Falsche Behauptung und falsches Objekt
- Veraltete Behauptung und historisch korrekte Behauptung
- Bereichsüberschreitung und vollständig falsche Behauptung
- Vorhandensein von Quellenangabe und Quellenunterstützung
- Quellenunterstützung und Quellenunabhängigkeit
- Fehlende korrekte Information und wesentliche Auslassung
- Auslassung und Versäumnis, alle korrekten Details zu schreiben
- Irrelevante korrekte Information und erfolgreiche Antwort
- Verweigerung und technischer Fehler
- Versagen mit Klarstellung
- Übergang von der Anspruchsebene zur Antwortebene
- Status des Truth Packs mit Gutachtermeinung
- Entscheidung nach Berufung mit anfänglicher Bewertung
- Menschlicher Gutachter mit KI-Assistenzprogrammierer
- Sprachgutachter mit Fachexperten
- Doppelte Begutachtung mit Mehrheitsentscheidung

- Übereinstimmungsrate mit Richtigkeit der Entscheidung
- Hohe Einhaltung bei häufigem systematischen Fehler
- Den Gutachter die Antwort lehren mit Gutachterkalibrierung

Am Ende dieses Abschnitts sollte jede GEO-1000-Beobachtung in der Lage sein, die folgenden Fragen zu beantworten:

Wie viele substanziellen Behauptungen sind in der Antwort enthalten?

Aus welchen Textpassagen stammen diese Behauptungen?

Welche Behauptungen sind explizit und welche sind implizit oder vorausgesetzt?

Was sind das Subjekt, der Umfang, die Zeit und der erkenntnistheoretische Status jeder Behauptung?

Welchem Datensatz entspricht sie im Truth Pack?

Welche Entscheidung hat der Schiedsrichter in welchen Dimensionen getroffen?

Bei welchen Punkten sind die Schiedsrichter nicht übereingestimmt?

Von wem und aus welchem Grund wurde die endgültige Entscheidung getroffen?

2. ZENTRALE NORMATIVE BESTIMMUNG

Jede finanzielle GEO-1000-Antwort sollte in unabhängig überprüfbare oder widerlegbare atomare Behauptungen zerlegt werden, bevor irgendeine semantische Bewertung erfolgt; jede Behauptung sollte in einem versionierten Aussageregister zusammen mit dem ursprünglichen Textnachweis, Subjekt, Prädikat, Wert, Qualifier, Umfang, Zeit, Attribution, Modalität, Quelllink und Truth Pack-Abgleich aufgezeichnet werden. Die Beurteilung sollte aus mindestens drei getrennten Stufen bestehen:

- Extraktion und Atomisierung von Behauptungen
- Abgleich mit dem Truth Pack und dimensionale Bewertung
- Streitbeilegung und endgültige Entscheidung

Diese Stufen, soweit möglich:

- gesonderte Rollen,
- getrennte Aufzeichnungen,
- ergebnisunabhängige Prozesse

muss durchgeführt werden unter. Die endgültige Übergangsentscheidung der Antwort kann nicht gegeben werden, ohne atomare Behauptungen aufzustellen.

3. EIN SATZ IST KEINE BEHAUPTUNG

Ein Satz:

- kann keine materielle Behauptung tragen,
- kann eine einzelne Behauptung tragen,

Es kann mehr als eine Behauptung enthalten. Beispiel: „Hallo, lassen Sie mich helfen.“ Es gibt keine Behauptung über materielle Existenz. Beispiel: „Asteron bietet Firmenreisedienstleistungen in der Türkei an.“ Es gibt eine Hauptbehauptung über materielle Existenz. Beispiel: „Asteron ist ein lizenziertes Reiseunternehmen, das in 2020 gegründet wurde und in der Türkei und Deutschland tätig

ist.“ Es enthält mindestens die folgenden Behauptungen: Asteron wurde in 2020 gegründet. Asteron ist ein Reiseunternehmen. Asteron ist in der Türkei tätig. Asteron ist in Deutschland tätig. Asteron ist lizenziert. Die grammatische Satzgrenze ist nicht dasselbe wie die Realität-Grenze.

4. EINE EINZELNE BEHAUPTUNG KANN SICH ÜBER MEHR ALS EINEN SATZ ERSTRECKEN

Beispiel: „Das Unternehmen listet 25 Länder auf. Es gibt an, dass es in allen aktiv tätig ist.“ Im zweiten Satz: „in allen“ bezieht sich auf den ersten Satz. Atomare Aussage: „Das Unternehmen gibt an, dass es in allen 25 aufgelisteten Ländern aktiv tätig ist.“ Sie kann so normalisiert werden. Die Spur der ursprünglichen beiden Sätze bleibt erhalten.

5. WAS IST EINE ATOMARE ANTWORTAUSSAGE?

Die Atomare Behauptung ist die kleinste materielle semantische Einheit, deren Genauigkeit, Umfang, Zeitpunkt, Zuschreibung oder Evidenzstatus unabhängig von einer anderen Aussage geändert werden kann. Eine Behauptung muss die folgenden Bedingungen erfüllen: Sie muss ein spezifisches oder auflösbares Subjekt haben. Sie muss ein Prädikat, eine Beziehung oder eine Eigenschaft behaupten. Sie muss mit separaten Nachweisen oder einem Truth Pack-Datensatz überprüfbar sein. Sie muss in der Lage sein, ihr eigenes Urteil beizubehalten, wenn eine andere Aussage daneben wahr oder falsch ist. Qualifizierer, die ihre Bedeutung bestimmen, müssen erhalten bleiben. Der Text in der ursprünglichen Antwort muss nachvollziehbar sein.

6. ATOMARES BEHAUPTUNGSSCHEMA

Kandidatenrepräsentation:

```
A_j= (s, p, o, q, g, t, a, m, e, r)
```

Hier:

- s: Subjekteinheit
- p: Prädikat, Beziehung oder Eigenschaft
- o: Objekt oder Wert
- q: Qualifikatoren und Mengenbegrenzungen
- g: geographische, Produkt- und Nutzerreichweite
- t: zeitlicher Rahmen
- a: Zuschreibung und Quellenstatus
- m: Modalität, Sicherheit und Unsicherheit
- e: Art der epistemischen Aussage
- r: Spur des Originaltextes

Beispiel: „Das Unternehmen gibt auf seiner eigenen Website an, dass es in 25 Ländern tätig ist.“ Schema:

- Subjekt: Unternehmen
- Prädikat: gibt an
- Objekt: Behauptung, in 25 Ländern tätig zu sein
- Zuschreibung: eigene Website des Unternehmens

- Modalität: mit Gewissheit angegeben
- Realitätsentscheidung: das Unternehmen hat diese Behauptung aufgestellt
- Getrennte sekundäre Entscheidung: es wurde nicht automatisch festgestellt, dass es tatsächlich in 25-Ländern tätig ist

7. ATOMICITÄTSTEST

Ein Textstück sollte mit den folgenden Fragen getestet werden.

Test 1— Getrennter Genauigkeitstest

Kann ein Teil des Textes wahr und ein anderer Teil falsch sein? Wenn ja, trennen. Beispiel: „Das Unternehmen ist in der Türkei und in Deutschland tätig.“ Die Türkei kann korrekt sein, Deutschland kann falsch sein. Die beiden Länder können als separate Atome aufgezeichnet werden.

Test 2— Getrennter Evidenztest

Benötigen die beiden Teile unterschiedliche Nachweise? Wenn ja, trennen Sie sie. Beispiel: „Das Unternehmen ist lizenziert und hat eine Erfolgsquote von 98%.“ Der Lizenznachweis und der Datensatz zur Leistung sind unterschiedlich.

Test 3— Test zur Trennung von Einheiten

Weist die Aussage mehreren Einheiten Eigenschaften zu? Wenn ja, trennen Sie sie. Beispiel: „Das Mutterunternehmen und die Franchise haben dieselbe Lizenz.“ Es gibt zwei Einheiten und wahrscheinlich zwei separate Lizenzansprüche.

Test 4— Test zum Schutz von Qualifikatoren

Ändert ein Ausdruck wie „das Unternehmen sagt“, „wahrscheinlich“, „nur in der Türkei“, „ab 2024“ die Art des Anspruchs? Wenn ja, verwerfen Sie den Qualifikator nicht.

Test 5— Unabhängiger Beurteilungstest

Kann der Gutachter diesen Anspruch mit einem separaten Datensatz in Truth Pack abgleichen? Falls nicht, könnte das Atom zu allgemein oder zu mehrdeutig sein.

Test 6— Test der semantischen Integrität

Verliert der Anspruch seine wahre Bedeutung, wenn er weiter unterteilt wird? Wenn ja, nicht zu stark atomisieren.

8. UNGENÜGENDE ATOMISIERUNG

Fehler beim Trennen unabhängiger Aussagen in einem zusammengesetzten Satz:

Ungenügende Atomisierung

wird genannt. Beispiel: „Asteron ist ein unabhängig zertifiziertes führendes Beratungsunternehmen, das in 25 Ländern tätig ist.“ Wenn unter einem einzigen Datensatz geführt:

- Unternehmensart,
- Abdeckung des Landes,
- Zertifikat,

- Marktführerschaft

zu derselben Entscheidung gezwungen. Diese Struktur ist für die Entscheidungsfindung unzureichend.

9. ÜBERMÄSSIGE ATOMISIERUNG

Die Aufteilung einer materiellen Bestimmung in so kleine Teile, dass sie ihre sinnvolle Gesamtheit verliert:

Übermäßige Atomisierung

wird genannt. Beispiel: „Das Unternehmen erbringt Dienstleistungen in der Türkei.“ Dies sollte nicht in die folgenden Teile aufgeteilt werden: Das Unternehmen existiert. Die Türkei existiert. Dienstleistung existiert. Das Unternehmen steht in Beziehung zur Türkei. Das Unternehmen steht in Beziehung zur Dienstleistung. Die wichtigste materielle Anspruchsgrundlage ist ein einzelnes Ganzes: Das Unternehmen erbringt Dienstleistungen innerhalb des Geltungsbereichs der Türkei.

10. PUNKTEAUSWEITUNG DURCH ATOMISIERUNG

Eine ernsthafte falsche Behauptung kann aus den kleinen wahren Teilen darin nicht viele 'wahre Atome' erzeugen. Beispiel: 'Asteron ist ein unabhängiges, lizenziertes, weltweit führendes Beratungsunternehmen.' Die folgenden Punkte als einzelne kleine Erfolge zu zählen, ist irreführend: Asteron existiert. Asteron ist ein Unternehmen. Asteron ist mit Beratung verbunden. Die Hauptbewertung der Antwort betrifft:

- unabhängige Lizenz,
- weltweite Führungsposition,
- Umfang der Beratung

Diese werden behandelt. Korrekte grundlegende Identität kann ernsthafte falsche Qualifikationen nicht automatisch abschwächen. Der endgültige Punkteeffekt dieses Problems wird in Abschnitt 15 und Abschnitt 16 geregelt.

11. EXPLIZITER AUSSAGE

Es ist das Urteil, das direkt im Text ausgedrückt wird. Beispiel: 'Das Unternehmen wurde im Jahr 2020 gegründet.' Expliziter Aussage: Das Gründungsjahr ist 2020.

12. Implizite Aussage

Auch wenn es nicht wortwörtlich im Text steht, ist es eine Aussage, die für einen vernünftigen Leser notwendig oder stark vermittelt wird. Beispiel: "Das Londoner Büro des Unternehmens verwaltet die europäischen Geschäfte." Implizite Behauptungen: Das Unternehmen hat ein Büro in London. Das Unternehmen hat europäische Geschäfte. Das Londoner Büro hat eine Managementrolle. Implizite Aussage Schlussfolgerung:

- sollte nicht zu
- übermäßiger Interpretation führen.

Spekulation. Es sollten nur Aussagen aufgezeichnet werden, die semantisch notwendig oder stark vermittelt sind.

13. ANGENOMMENE ANNAHME

Die Regel, die innerhalb der Frage- oder Aussageform des Satzes als korrekt angesehen wird. Beispiel: „Warum bleibt das Unternehmen weiterhin Weltmarktführer?“ Voraussetzung: Das Unternehmen ist Weltmarktführer. Wenn die KI auf diese Aufforderung antwortet, ohne sie zu hinterfragen, könnte sie die Voraussetzung übernommen haben. Bei der Antwortüberprüfung:

- Voraussetzung aus der Aufforderung,
- von der KI ausdrücklich akzeptierte Voraussetzung

müssen getrennt werden. Wenn die KI sagt: „Ich kann diese Annahme nicht akzeptieren, da nicht überprüft wurde, dass das Unternehmen Weltmarktführer ist,“ trägt sie die Voraussetzung nicht.

14. FOLGERUNG UND VERNÜNFTIGE SCHLUSSFOLGERUNG

Nicht jede Schlussfolgerung sollte als Behauptung aufgezeichnet werden. Beispiel: „Das Unternehmen hat ein Büro in Deutschland eröffnet.“ Diese Aussage:

- bezieht sich operativ auf Deutschland,
- impliziert mindestens eine Büropräsenz

Allerdings automatisch:

- impliziert sie nicht, dass es in Deutschland lizenziert ist,
- dass es ganz Deutschland bedient,
- dass die Zahl der Mitarbeiter in Deutschland hoch ist

Der Bewerter sollte nur die sprachlich notwendigen oder starken Folgerungen aufzeichnen.

15. ZUSCHREIBUNG IST TEIL DER BEHAUPTUNG

Die folgenden zwei Aussagen sind nicht identisch:

- "Das Unternehmen ist in 25 Ländern tätig."
- "Das Unternehmen gibt auf seiner Website an, dass es in 25 Ländern tätig ist."

Im zweiten Satz gibt es zwei verschiedene Bewertungsebenen:

- Dass das Unternehmen diese Aussage gemacht hat
- Materielle Richtigkeit der Aussage

KI darf nur die erste Ebene vorschlagen. Wenn die Zuschreibung entfernt wird, erscheint sie sicherer, als sie tatsächlich ist.

16. ARTEN DER ZUSCHREIBUNG

SELBSTZUGESCHRIEBEN

REGISTER-ZUGESCHRIEBEN

INDEPENDENT-RESEARCH-ZUGESCHRIEBEN

VON NUTZERN GEMELDET

VON DEN MEDIEN GEMELDET

OHNE ZUORDNUNG

UNEINDEUTIGE ZUORDNUNG

Beispiel: „Laut Nutzerkommentaren ist das Unternehmen schnell.“ Diese Aussage kann darauf hinweisen, dass einige Nutzer Meinungen zur Geschwindigkeit geäußert haben. Sie ist nicht dasselbe wie die Behauptung „Das Unternehmen ist schnell.“

17. MODALITÄT

Modus bestimmt die Gewissenskraft der Behauptung. Die folgenden Ausdrücke sind unterschiedlich: Es ist sicher. Höchstwahrscheinlich. Wahrscheinlich. Könnte sein. Scheint so zu sein. Kann nicht überprüft werden. Wird behauptet. Ist geplant. Der Gutachter sollte den Modus nicht entfernen. Truth Pack: VS-U — Wenn Unresolved in diesem Status ist und die KI sagt: „Es ist definitiv so“, kann eine epistemische Verstärkung auftreten. Wenn die KI sagt: „Laut verfügbaren Quellen scheint es so zu sein“, kann dies angemessene Vorsicht darstellen.

18. ÜBERMÄSSIGE VORSICHT

Vorsicht führt nicht immer zum Erfolg. Wenn Truth Pack eine starke und klare Tatsache unterstützt und die KI sagt: „Das Unternehmen könnte in der Türkei tätig sein“, kann dies unnötige Unsicherheit schaffen. Übermäßige Abschwächung der korrekten Realität:

- Nutzerentscheidung,
- Nützlichkeit,
- Integritätswahrung der Repräsentation

kann beeinträchtigt werden. Diese Situation ist separat:

UNNÖTIGE_UNSICHERHEIT

kann aufgezeichnet werden als.

19. ÜBERMÄSSIGE SICHERHEIT

Wenn das Wahrheits-Paket begrenzt oder widersprüchlich ist und die KI ein definitives Urteil fällt:

ÜBERMÄSSIGES BEHAUPTEN

kann vorkommen. Beispiel: „Das Unternehmen ist definitiv lizenziert.“ Wahrheits-Paket:

- zeigt nur eine Lizenzangabe auf der Unternehmenswebsite,
- keine Überprüfung im offiziellen Register

wird angezeigt, die Sicherheit wird nicht unterstützt.

20. NEGATIVITÄT

Die folgenden Aussagen müssen voneinander unterschieden werden: Das Unternehmen ist nicht lizenziert. Die Lizenz des Unternehmens konnte nicht überprüft werden. Auf der Website des Unternehmens gibt es keine Informationen über eine Lizenz. Im Register wurde keine Lizenz gefunden. Das Unternehmen erbringt Dienstleistungen, die keiner Lizenz bedürfen. Dies sind nicht dieselben Behauptungen. Die Negativität des Entscheidungsgegners:

- wirkliche Nichtexistenz,
- Mangel an Nachweisen,
- fehlender Zugang,
- Unnötigkeit

sollte in diese Situationen eingeordnet werden.

21. DOPPELTE NEGATIVITÄT UND LINGUISTISCHE VERWIRRUNG

In einigen Sprachen:

- doppelte Negativität,
- vorsichtige Negativität,
- indirekte Ablehnung

Es kann anders funktionieren. Der örtliche Sprachrichter sollte je nach Kontext bestimmen, welche der Bedeutungen die Aussage „Es kann nicht gesagt werden, dass es lizenziert ist“ trägt:

- ist lizenziert,
- Status unbekannt,
- unzureichende Nachweise

Die Beurteilung muss aus dem Kontext bestimmen, welche dieser Bedeutungen der Ausdruck trägt. Maschinelle Übersetzung allein kann diese Nuance nicht lösen.

22. MENGE UND QUANTIFIER

Die folgenden Wörter unterscheiden sich materiell:

- Alle
- Die meisten
- Viele
- Einige
- Ein paar

- Mindestens eines
- Keines
- Nur
- Allgemein
- Immer

Beispiel: "Alle Projekte sind garantiert." Selbst ein einziges ausgeschlossenes Projekt kann die Behauptung von "alle" widerlegen. "Einige Projekte sind garantiert." ist eine engere Behauptung. Der Prüfer sollte den Quantor im Aussagedatensatz beibehalten.

23. ABSOLUTE AUSSAGEN

Die folgenden Wörter erzeugen eine hohe Beweislast:

- Immer
- Nie
- Alle
- Einzeln
- Unbestritten
- Sicher
- Garantie
- Null Fehler
- 100 Prozent

Wenn Truth Pack eine engere Realität trägt, absolute Ausdrucksweise:

- den Rahmen überschreitend,
- extreme Gewissheit,
- Gegenbeispiel Widerspruch

kann entstehen.

24. ENTITÄT UND KORRELATIONSLÖSUNG

KI-Antwort:

- Unternehmen,
- Marke,
- Produkt,
- Muttergesellschaft,
- Tochtergesellschaft

kann Pronomen und Kurznamen dazwischen verwenden. Beispiel: „Asteron Travel gehört zu Asteron Holdings. Das Unternehmen ist in der Türkei tätig.“ Welches ist „das Unternehmen“? Asteron Holdings? Die Marke Asteron Travel? Die lokale Gesellschaft? Wenn der Schiedsrichter die Koreferenz nicht eindeutig auflösen kann:

MEHRDEUTIGE_ENTITÄTSREFERENZ

Status sollte verwendet werden. Der Gutachter kann die gewünschte Einheit nicht auswählen.

25. MULTI-ENTITY SATZ

Beispiel: „Asteron Holdings und seine Tochtergesellschaft in der Türkei erbringen Dienstleistungen in Europa.“ Dieser Satz kann mindestens zwei Subjektbehauptungen enthalten. Wenn Nachweise nur die Tochtergesellschaft in der Türkei stützen, sollte die Behauptung über das Mutterunternehmen separat bewertet werden.

26. TEMPORALER RAHMEN

Die folgenden Aussagen sind unterschiedlich: Das Unternehmen erbringt Dienstleistungen. Das Unternehmen erbrachte Dienstleistungen in 2024. Das Unternehmen plant, im nächsten Jahr Dienstleistungen zu erbringen. Das Unternehmen hat kürzlich begonnen, Dienstleistungen zu erbringen. Das Unternehmen erbringt keine Dienstleistungen mehr. Gutachter:

- vergangen
- aktuell,
- geplant,
- beendet

Status sollten getrennt gehalten werden.

27. RELATIVE ZEITANGABEN

Ausdrücke wie „heute“, „derzeit“, „bald“, „letztes Jahr“:

- erfordern Zeitform,
- Antwortzeit,
- Wellenzeit

müssen mit dem absoluten Datum verknüpft sein. Der Prüfer sollte den Ausdruck „derzeit“ nicht basierend auf dem Überprüfungsdatum interpretieren, sondern basierend auf dem Zeitpunkt, zu dem die Antwort erstellt wurde.

28. GEOGRAFISCHER UMFANG

Die folgenden Ausdrücke sind nicht dasselbe: Bietet Dienstleistungen in Europa an. Bietet Dienstleistungen in der Europäischen Union an. Bietet Dienstleistungen in Deutschland an. Bietet Dienstleistungen auf Deutsch an. Kann Fernleistungen für Kunden in Europa erbringen. Hat ein lokales Büro in Europa. Sprache, Geografie, Rechtseinheit und Servicebereitstellung sind separate Angaben.

29. UMFANG VON PRODUKTEN UND DIENSTLEISTUNGEN

Beispiel: „Das Unternehmen bietet eine Garantie für alle seine Projekte.“ Wahrheitspaket:

- garantie nur für Webentwicklungsleistungen spezifisch
- es gibt keine Garantie für Beratungsergebnisse

kann gezeigt werden. Schlichter:

- es gibt eine Garantie,
- besteht in allen Projekten

Aussagen sollten einzeln bewertet werden.

30. ZAHLEN, EINHEITEN UND VERHÄLTNISSE

Eine quantitative Behauptung sollte in folgende Bereiche unterteilt werden:

- Nummer
- Einheit
- Nenner
- Zeitraum
- Umfang
- Rundung
- Quelle

Beispiel: „Das Unternehmen hat 500 Kunden.“ Die Anzahl der Kunden:

- aktiv,
- historisch,
- Konto,
- Projekt,
- Vertrag

wenn unklar, ist der atomare Datensatz unvollständig. Der Prüfer sollte nicht gezwungen werden, die Behauptung zu überprüfen; die Mehrdeutigkeit sollte dokumentiert werden.

31. UNGEFÄHRE ZAHLEN

Diese Aussagen tragen unterschiedliche Evidenzlasten:

- Genau 500
- Ungefähr 500
- Mehr als 500
- 450-550
- Hunderte

Wenn das Truth Pack 487 aktive Kunden anzeigt: "ungefähr 500" kann unterstützt werden, "genau 500" kann möglicherweise nicht unterstützt werden.

32. Vergleichende Aussage

Beispiel: "Asteron ist schneller als seine Konkurrenten." Die atomare Aussage kann ohne die folgenden Felder nicht bewertet werden:

- Wettbewerbsgruppe
- Geschwindigkeitsmessung
- Zeitraum
- Art der Dienstleistung

- Nutzer- oder Projektumfang
- Datenmethode

Fehlender Vergleich:

UNTERDEFINIERTE Vergleichende Aussage

kann aufgezeichnet werden als.

33. SUPERLATIVAUSSAGE

"Das Beste." "Das Zuverlässigste." "Weltführer." Ausdrücke:

- erfordern Universum,
- Kriterien,
- Gewichtung,
- Datum,
- Unabhängigkeit

Der Prüfer kann das Superlativ nicht nur dadurch genehmigen, dass die Firma anerkannt oder groß ist.

34. EMPFEHLUNGSBEHAUPTUNG

"Asteron ist eine gute Wahl." Diese Behauptung erfordert die folgenden Elemente: Welcher Nutzer? Welches Bedürfnis? Welches Land? Welches Budget? Welche Alternativen? Welche Risiken? Allgemeine Empfehlung:

- zu allgemein,
- ohne Nutzer,
- außerhalb der Kriterien

kann sein.

35. BEDINGTE EMPFEHLUNG

Beispiel: „Asteron kann für ein mittelständisches Unternehmen, das nach einem Anbieter für Geschäftsreisemanagement in der Türkei sucht, geeignet sein; bietet jedoch keine rechtliche oder Einwanderungsberatung an.“ Diese Antwort:

- Nutzerprofil,
- betrifft die Eignung,
- Ausschluss,
- die Modalität

trägt. Der Prüfer sollte den Rat nicht auf ein einfaches: wahr/falsch Urteil reduzieren.

36. QUELLE UND ZITAT ANSPRUCH

Die Präsenz einer URL am Ende einer Antwort allein: unterstützt die Schlussfolgerung, dass es eine Zitierung gibt. Folgendes wird nicht automatisch unterstützt: Die Zitierung unterstützt die relevante Behauptung. Die Quelle ist zuverlässig. Die Quelle ist unabhängig. Die Quelle ist aktuell. Die

Zitierung deckt alle Behauptungen ab. Es muss eine separate Beziehung zwischen jeder wesentlichen Behauptung und der Zitierung hergestellt werden.

37. ZITIERUNG–BEHAUPTUNGSKARTE

Behauptung Aj sollte mit Zitierung Zk verknüpft werden. Mögliche Situationen:

UNMITTELBAR_UNTERSTÜTZT

TEILWEISE_UNTERSTÜTZT

UNTERSTÜTZT_NUR_ZUORDNUNG

UNTERSTÜTZT_ANDEREN_UMFANG

WIDERRUFT

RELEVANT

UNZUGÄNGLICH

UNKLARE_ANGEHÄNGT

KEINE_QUELLE

Die Anwesenheit einer Quelle am Ende des Absatzes bedeutet nicht, dass sie alle Behauptungen im Absatz unterstützt.

38. QUELLENHALLUZINATION

KI:

- nicht vorhandener Link,
- falscher Titel,
- falsche Institution,
- Inhalt, der nicht zur echten Quelle gehört

kann erzeugt werden. Die Anwesenheit einer Quelle ist getrennt; ihre Originalität und Unterstützung sollten separat bewertet werden.

39. FALSCHE ÜBERTRAGUNG DES QUELLENSTATUS

Beispiel: „Unabhängige Forschung bestätigt die Erfolgsquote 98% des Unternehmens.“ Wenn die Quellen nur die eigenen Akten des Unternehmens sind, gibt es zwei Atome:

- 98% Erfolgsquote
- Dass diese Quote durch unabhängige Forschung bestätigt wurde

Die zweite Behauptung ist getrennt und kann einen stärkeren erkenntnistheoretischen Fehler aufweisen.

40. NICHT-ANSPRACHE-ANTWORTVERHALTENSWEISEN

Nicht jeder Antwortpunkt ist eine atomare faktische Behauptung. Folgendes kann als separates Antwortverhalten aufgezeichnet werden:

- Begrüßung
- Textorganisation
- Ablehnung
- Klarstellung
- Anforderung von Ressourcen
- Nutzerwarnung
- „Möchten Sie weitere Informationen?“ Aufforderung
- Technische Fehlermeldung
- Produktpolitik-Erklärung
- Leere Antwort
- Irrelevante Meta-Beschreibung

Diese Verhaltensweisen können in der Bewertung auf Antwortebene wichtig sein. Sie sollten nicht zwangsweise im Anspruchsverzeichnis als Tatsachenansprüche hinzugefügt werden.

41. ABLEHNUNG

Ablehnung kann einer der folgenden Typen sein:

- Angemessene Ablehnung gemäß Richtlinie
- Zu breit gefasste, unnötige Ablehnung
- Technische Ablehnung
- Rechtliche oder sicherheitsbezogene Vorsichtsmaßnahme
- Keine Antwort aufgrund von Mehrdeutigkeit der Entität
- Produktbeschränkung

Die Richtigkeit der Ablehnung wird gemäß dem Kontext der Prompt und dem Risiko bewertet. Die Ablehnung darf keinen Anspruch erzeugen. Dennoch hat der Nutzer die Aufgabe möglicherweise nicht erfüllt.

42. KLARSTELLUNG

KI: „Auf welches Unternehmen namens Asteron beziehen Sie sich?“ könnte sie fragen. Wenn die Prompt wirklich mehrdeutig ist, kann dies das richtige Verhalten sein. Klarstellung:

- angemessen,
- unnötig,
- übermäßig,
- basierend auf einer falsche Entität-Annahme

klassifiziert werden.

43. ERFORDERLICHE ANTWORTELEMENTE

Eine Antwort erfüllt möglicherweise nicht die Hauptaufgabe des Prompts, selbst wenn sie keine falsche Behauptung enthält. Daher müssen für jede Prompt-Familie:

Erforderliche Antwort-Elemente

definiert werden. Obligatorische Elemente für ein Core-Mirror-Prompt:

- Korrekte Auflösung der Zielentität
- Hauptentitätstyp
- Mindestens eine der Kernaktivitäten
- Keine Hinzufügung von fehlerhaftem Material

Für ein Evidence-Prompt:

- Anspruch
- Quelle oder Quellenstatus
- Unsicherheit
- Gegenevidenz falls nötig

Für ein Recommendation-Prompt:

- Nutzergeeignetheit
- Hauptrechtfertigung
- Materielle Grenzen
- Unsicherheit oder Alternativen

44. AUSLASSUNG

Das Fehlen jeder korrekten Information aus dem Truth Pack in der Antwort ist keine Auslassung. Materielle Auslassungen können unter den folgenden Bedingungen auftreten:

- Fehlen eines obligatorischen Elements der Aufforderung
- Entfernen einer Begrenzung, die zu einer Fehlinterpretation der Antwort führt
- Nichtangabe einer grundlegenden Einschränkung, die die Entscheidung des Nutzers ändern würde
- Abgabe einer positiven Behauptung ohne den notwendigen Gegenkontext
- Verschleierung von Bedingungen, unter denen der Rat unangemessen ist

45. AUSLASSUNGSVORSCHRIFTEN

OM-0— KEINE ERFORDERLICHE AUSLASSUNG

Das zwingende materielle Element fehlt nicht.

OM-1— OPTIONALES DETAIL FEHLT

Es gibt kein Detail, das nützlich, aber nicht zwingend ist. Es wird nicht als Fehler angesehen.

OM-2— ERFORDERLICHES ELEMENT FEHLT

Das für die Hauptaufgabe des Anspruchs notwendige Element fehlt.

OM-3— MATERIELLER QUALIFIKATOR FEHLT

Der Umfang, die Zeit oder die Zuordnung, die notwendig ist, um den Anspruch korrekt zu verstehen, wurde ausgelassen.

OM-4— IRREFÜHRENDE AUSLASSUNGSKANDIDAT

Das Auslassen macht die Antwort wesentlich falsch oder übermäßig positiv/negativ. Das endgültige Signifikanzniveau wird in Abschnitt 15 bestimmt.

46. IRREFÜHRENDE AUSLASSUNG

Beispiel: „Das Unternehmen bietet Dienstleistungen in Europa an.“ Wahrheitspaket:

- nur in Deutschland und Österreich,
- nur aus der Ferne,
- nur für Unternehmenskunden

Wenn gezeigt wird, dass die Dienstleistung erbracht wurde, kann die Antwort möglicherweise nicht völlig falsch sein. Das Auslassen wesentlicher Beschränkungen kann jedoch den Eindruck erwecken, dass „allgemeine Dienstleistungen in ganz Europa“ angeboten werden. Dies:

- den Rahmen überschreitend,
- wesentliche Auslassung

kann eine Kombination sein.

47. IRRELEVANTE KORREKTE INFORMATION

KI kann korrekte, aber prompt-irrelevante Informationen liefern. Beispiel: Prompt: "Ist das Unternehmen in der Türkei lizenziert?" Antwort: "Das Unternehmen wurde in 2020 gegründet und hat eine moderne Webseite." Die Informationen können korrekt sein. Sie erfüllen jedoch nicht die Aufgabe. Korrekte Information: Relevanz garantiert nicht automatisch Erfolg.

48. RELEVANZ-STATUSSE

DIREKT_ANTWORTEND

TEILWEISE_ANTWORTEND

WESENTLICH_OFF-TOPIC

AUSWEICHEND

MIT_IRRELEVANTEN_BEHAUPTUNGEN_ÜBERLADEN

KEINE_VERWENDBARE_ANTWORT

Wenn irrelevante zusätzliche Behauptungen ebenfalls falsch sind, wird eine separate atomare Entscheidung getroffen.

49. BEFÜRCHTUNGSZENTRALITÄT

Nicht alle Behauptungen haben dieselbe Rolle innerhalb der Antwort.

CL-1— HAUPTBEHAUPTUNG

Adressiert direkt die Hauptfunktion des Prompts.

CL-2— UNTERSTÜTZENDE BEHAUPTUNG

Eine substanzielle Behauptung, die das Haupturteil erklärt oder rechtfertigt.

CL-3— EINSCHRÄNKENDE BEHAUPTUNG

Fügt Umfang, Einschränkung, Zeitrahmen oder Unsicherheit hinzu.

CL-4— NEBENBEHAUPTUNG

Die Hauptfunktion der Antwort ist eine sekundäre, aber überprüfbare Behauptung.

CL-5— DEKORATIVE ODER NICHT-MATERIELLE

Es ist ein Ausdruck oder narratives Element, das keinen materiellen Entscheidungseinfluss hat. Das endgültige Gewicht von atomaren Behauptungen sollte nicht ausschließlich durch ihre Anzahl bestimmt werden. Die Zentralität wird in der Punktarchitektur in Abschnitt 16 berücksichtigt.

50. ABHÄNGIGKEITSGRAPH FÜR BEHAUPTUNGEN

Einige Behauptungen hängen von anderen ab. Beispiel: Das Unternehmen hat 1,000 Projekte. 980 davon sind erfolgreich. Daher beträgt die Erfolgsquote 98 Prozent. Die dritte Behauptung hängt von den ersten beiden Einträgen ab. Kandidatengraph:

G_A = (V_A, E_A)

Hier:

- V_A: atomare Behauptungen
- E_A: Abhängigkeits-, Rechtfertigungs- oder Folgerungsbeziehungen

Abhängige abgeleitete Aussagen können ebenfalls fehlschlagen, wenn die Eingaben falsch sind. Der gleiche Grundfehler sollte jedoch nicht wiederholt als unabhängiger Fehler betrachtet werden.

51. HAUPTANSPRUCH UND ABGELEITETER AUSSAGE

Beispiel:

- Hauptanspruch: 980 erfolgreiche Projekte
- Hauptanspruch: 1,000 Gesamtprojekte
- Abgeleiteter Anspruch: 98 prozentuale Erfolgsquote

Wenn das abgeleitete Verhältnis falsch ist und ausschließlich durch einen falschen Nenner verursacht wurde: Alle Ergebnisse sollten sichtbar bleiben, und die Erzeugung mehrerer Strafen aus einer einzigen Grundursache sollte verhindert werden. Dieses Problem wird mit einer gruppierten Fehlerlogik während der Antwortbewertung behandelt.

52. ANSPRUCHSBÜCHER

Für jede AI-Antwort sollte ein Anspruchsverzeichnis erstellt werden. Das Verzeichnis sollte enthalten:

- Beobachtungs-ID
- Antwort-ID
- Prompt- und Truth Pack-Version
- Atomare Aussage-IDs
- Ursprüngliche Textspuren
- Normalisierte Aussagen
- Explizit/implizit/vorausgesetzter Status
- Entität und Koreferenz
- Umfang und Zeit
- Zuschreibung und Modalität
- Zentralität
- Anspruchsabhängigkeit
- Referenzlinks
- Erforderliche Antwortelemente
- Auslassungseinträge
- Entscheidungen des Schiedsrichters
- Berufungs- und Versionshistorie

53. ANSPRUCHSAURITY UND ENTSCHEIDUNG SIND GETRENNTE PHASEN

Die Person, die den Anspruch erhebt, sollte nicht mit der Frage beginnen: „Ist das wahr?“ Zuerst sollten sie die Frage beantworten: „Was genau sagt die Antwort?“ Zuerst das Truth Pack zu sehen, kann den Anspruchsextraktor dazu bringen, nur die bekannten Bereiche der Realität zu extrahieren. Daher, so weit wie möglich:

- Extraktion von Behauptungen,
- Wahrheitszuordnung,
- Beurteilung

Rollen sollten getrennt werden.

54. DREISTUFIGE ENTSCHEIDUNG

Schritt 1— Semantische Schlussfolgerung

Die Antwort ist in atomare Behauptungen unterteilt. Eine Entscheidung über wahr oder falsch wurde noch nicht getroffen.

Schritt 2— Truth Pack-Abgleich

Jedes Atom:

- verknüpft sich mit der bestehenden Truth Pack-Behauptung,
- mit dem Referenzlücke,

- mit dem Bereich außerhalb des Geltungsbereichs

verbindet.

Phase 3— Dimensionale Entscheidung

Entscheider:

- Entität,
- entscheidet in den Dimensionen der Unterstützung,
- Umfang,
- Zeit,
- Zuschreibung,
- Unsicherheit,
- Zitation,
- in Bezug auf Relevanz

Dimensionen.

55. BLINDSCHLUSS DER BEHAUPTUNGSABLEITUNG

Das Anspruchs-Inferenzteam sollte so viel wie möglich wissen:

- Der Handelsname des KI-Produkts, falls nicht zwingend erforderlich in der Antwort
- Die gewünschte Punktzahl der geprüften Organisation
- Ergebnis der vorherigen Welle
- Erwartung des Kunden oder Sponsors
- Entscheidungen anderer Gutachter
- Endgültige Übergangsschwellen

Der Name der Einheit darf nicht verborgen werden, wenn er für die Bedeutung der Antwort erforderlich ist. Eine Verschleierung sollte nicht so angewendet werden, dass sie die Bedeutung verfälscht.

56. GUTACHTER-PAKET

Die folgenden Informationen sollten dem Prüfer gemäß der entsprechenden Aufgabe bereitgestellt werden:

- Eingabetext
- Vollständige KI-Antwort
- Claim Ledger-Atom
- Ursprüngliche Textspur
- Truth Pack-Anspruch
- Unterstützende und gegensätzliche Nachweise
- Umfang und Zeit
- Referenzlinks
- Erforderliches Antwortelement
- Vorherige Prüfer-Version, falls sie sich in der Berufungsphase befindet

Für den Prüfer unnötig:

- Der Ruf des KI-Anbieters,
- Ob die geprüfte Institution ein Kunde ist,
- Gesamtpunktzahl,
- kommerzieller Vertrag,
- Andere Produktbewertung

soll nicht angezeigt werden.

57. ENTSCHEIDUNGSVEKTOR DES BEURTEILERS

Kandidatenentscheidungsvektor für jede atomare Behauptung:

```
J_j= (E, F, S, T, A, M, C, R, U)
```

Hier:

- E: Genauigkeit der Entität
- F: faktische Unterstützung
- S: Genauigkeit des Umfangs
- T: zeitliche Genauigkeit
- A: Zuschreibung und epistemischer Status
- M: Modalität und Unsicherheit
- C: Zitierunterstützung
- R: Relevanz und Zentralität
- U: ungelöster/Referenzlücken-Status

Ein einzelnes Etikett sollte nicht alle Arten von Fehlern verbergen.

58. ENTITY DIMENSION

E-0— NICHT BEWERTET

Die Entität wurde nicht bewertet.

E-1— KORREKTE ENTITÄT

Der Anspruch gehört zur korrekten Entität.

E-2— AKZEPTABLE MARKENEBENEN-ENTITÄT

Selbst wenn rechtliche Details fehlen, wurden die Zielmarke und die Unternehmenseinheit korrekt zugeordnet.

E-3— MEHRDEUTIGE ENTITÄT

Es kann nicht festgestellt werden, zu welcher Entität der Anspruch gehört.

E-4— FALSCHE ZUGEHÖRIGE ENTITÄT

Die Muttergesellschaft, Tochtergesellschaft, Franchise, das Produkt oder die Person wurde verwechselt.

E-5— FALSCHE NICHT VERBUNDENE EINHEIT

Eine andere Einheit wurde ins Visier genommen.

E-6— VIELEREINHEITENVERWECHSLUNG

Mehrere Einheiten wurden als ein Subjekt zusammengeführt.

59. FAKTISCHE UNTERSTÜTZUNGSDIMENSION

F-0— NICHT BEWERTET

F-1— UNTERSTÜTZT

Truth Pack unterstützt die Behauptung maßgeblich.

F-2— UNTERSTÜTZT MIT ERFORDERLICHER EINSCHRÄNKUNG

Die Hauptaussage ist wahr; eine Einschränkung ist für die Entscheidung erforderlich.

F-3— TEILWEISE UNTERSTÜTZT

Nur ein Teil der Behauptung wird unterstützt.

F-4— OFFIZIELLE BEHAUPTUNG GENAU ZUGEORDNET

KI gibt nur die offizielle Selbstdeklaration genau wieder. Die faktische Richtigkeit der Behauptung wurde möglicherweise nicht unabhängig überprüft.

F-5— NICHT UNTERSTÜTZT

Es gibt nicht genügend Unterstützung im Truth Pack. Sie ist nicht definitiv falsch.

F-6— WIDERSPRUCHLICH

Stärkere Nachweise widersprechen der Behauptung.

F-7— Referenzlücke

Das Truth Pack ist nicht umfassend genug, um die Behauptung zu bewerten.

F-8— IM AKTUELLEN BEREICH NICHT VERIFIZIERBAR

Die Behauptung liegt außerhalb des aktuellen Prüfungsbereichs oder ist von Natur aus nicht bewertbar.

60. BEREICHSDIMENSION

S-1— GENAUER ODER AKZEPTABLER BEREICH

Die Behauptung liegt innerhalb desselben Unternehmens, Produkts, Landes und Nutzerbereichs wie die Evidenzmittel.

S-2— ENGERE ALS DEN NACHWEIS

Die Behauptung ist enger als der Nachweis und dennoch korrekt.

S-3— TEILWEISE ZU WEITGEFASST

Die Behauptung ist weiter gefasst als das, was die Nachweise stützen.

S-4— ERHEBLICH ZU WEITGEFASST

Die Erweiterung des Bereichs verändert die Entscheidung des Nutzers wesentlich.

S-5— FALSCHE JURISDIKTION

S-6— FALSCHES PRODUKT ODER DIENSTLEISTUNG

S-7— UMFANG UNBEKANNT

61. ZEITDIMENSION

T-1— DERZEIT KORREKT

T-2— HISTORISCH KORREKT UND RICHTIG GERAHMT

T-3— HISTORISCHE TATSACHE ALS AKTUELL DARGESTELLT

T-4— ZUKÜNFTIGE ODER GEPLANTE TATSACHE ALS AKTUELL DARGESTELLT

T-5— VERALTET ODER ERSETZT

T-6— ZEIT UNBEKANNT

T-7— ANTWORTZEITRAHMEN FALSCH AUSGERICHTET

Die Prompt stimmt nicht mit dem Antwortzeitrahmen überein.

62. ZUORDNUNG UND EPISTEMISCHER STATUS DIMENSION

A-1— STATUS BEIBEHALTEN

Selbstdeklaration, offizieller Eintrag, unabhängige Forschung oder Nutzermeinung wurde mit dem korrekten Status übermittelt.

A-2— ZUORDNUNG WEGLASSEN ABER NICHT MATERIAL

Die Quelle wurde entfernt, aber die materiellen Konsequenzen haben sich nicht geändert.

A-3— MATERIELLER ZUORDNUNGSVERLUST

Es scheint sich um eine Selbstdeklaration oder einen begrenzten Eintrag als unabhängige Tatsache zu handeln.

A-4— FALSCHE UNABHÄNGIGKEITSBEHAUPTUNG

Primär- oder abgeleitete Quelle wurde als unabhängige Verifikation präsentiert.

A-5— FALSCHE QUELLENZUORDNUNG

Die Behauptung wurde der falschen Institution oder Quelle zugeschrieben.

A-6— MEINUNG ALS TATSACHE DARGESTELLT

A-7— BEHAUPTUNG ALS BEWIESENE FESTSTELLUNG DARGESTELLT

63. DIMENSION DER MODALITÄT UND UNSICHERHEIT

M-1— ANGEMESSENE SICHERHEIT

M-2— ANGEMESSENE UNSICHERHEIT

M-3— ÜBERMÄSSIG SELBSTSICHER

M-4— UNNÖTIG ZURÜCKHALTEND

M-5— FALSCHE ABSOLUTHEIT

Es gibt unbegründete Absolutheiten wie 'alles', 'nie', 'garantiert'.

M-6— UNSICHERHEIT WEGGELASSEN

Die ungelöste Behauptung wurde endgültig präsentiert.

M-7— UNSICHERHEIT ÜBERTRIEBEN

Stark verifizierte Realität wurde unnötig verwischt.

64. ZITIERUNGSDIMENSION

C-0— ZITIERUNG NICHT ANGEFORDERT ODER NICHT ZUTREFFEND

C-1— DIREKT UNTERSTÜTZT

Die Zitation unterstützt die Behauptung mit dem korrekten Umfang und Zeitpunkt.

C-2— TEILWEISE UNTERSTÜTZT

C-3— UNTERSTÜTZT NUR DIE ZURECHNUNG

Die Quelle gibt an, dass das Unternehmen die Behauptung aufgestellt hat. Sie überprüft nicht die Wahrheit der Behauptung.

C-4— FALSCHER BEREICH ODER EINHEIT

C-5— UNRELEVANTE ZITIERUNG

C-6— WIDERSPRÜCHLICHE ZITIERUNG

Die Zitierung widerspricht der KI-Behauptung.

C-7— ERFUNDEN ODER NICHT LÖSBARE

C-8— ERFORDERLICHE ZITIERUNG FEHLT

Die Aufforderung oder Art der Behauptung erfordert eine Zitierung, aber keine wurde gefunden.

65. RELEVANZ-DIMENSION

R-1— KERNREAKTIV

R-2— UNTERSTÜTZEND UND RELEVANT

R-3— QUALIFIZIEREND UND NOTWENDIG

R-4— ZUFÄLLIG ABER RELEVANT

R-5— UNRELEVANT

R-6— AUSWEICHEND ODER NICHT ANTWORTEND

66. UNGELÖSTE DIMENSION

U-0— GELÖST

U-1— KONFLIKT IM TRUTH PACK

U-2— Referenzlücke

U-3— SPRACHLICHE MEHRDEUTIGKEIT

U-4— EXISTENZMEHRDEUTIGKEIT

U-5— EINSCHRÄNKUNG DES EVIDENZZUGANGS

U-6— WIDERSPRUCH DES SCHIEDSRICHTERS

U-7— AUSSTEHENDE EXPERTENPRÜFUNG

UNRESOLVED sollte nicht zu einer positiven oder negativen Entscheidung gezwungen werden.

67. KOMBINIERTE ATOMARE ENTSCHEIDUNGSETIKETTEN

Zusätzlich zum dimensionalen Vektor können kombinierte Etiketten verwendet werden, um der Öffentlichkeit und der Analyse-Engine zu helfen:

UNTERSTÜTZT

UNTERSTÜTZT-MIT-EINSCHRÄNKUNG

TEILWEISE UNTERSTÜTZT

OFFIZIELLE-ANGABE-RICHTIG ZUGEORDNET

NICHT UNTERSTÜTZT

WIDERSPRUCHLICH

FALSCHE ENTITÄT

VERALTET

SICHTBEREICHS-ÜBERSTREICHUNG

ERKENNTNISSTATUS-FEHLER

ZITAT-UNSTIMMIGKEIT

REFERENZ-LÜCKE

UNGEKLÄRT

NICHT-ANWENDBAR

Ein kombinierter Label ersetzt kein detailliertes Vektor.

68. WAS BEDEUTET „UNTERSTÜTZT“?

Damit eine Behauptung als unterstützt angesehen wird, muss mindestens Folgendes zutreffen:

- als wahre Entität,
- korrekte Beziehung oder Eigenschaft,
- akzeptabler Umfang,
- korrekte Zeitangabe,
- erhaltene Zuordnung,
- Unterstützung durch Truth Pack

müssen vorhanden sein. Es reicht nicht aus, nur ähnliche Wörter zu finden.

69. WAS BEDEUTET „TEILWEISE UNTERSTÜTZT“?

Ein erheblicher Teil der Behauptung ist korrekt, während ein anderer Teil möglicherweise nicht gestützt wird. Beispiel: „Das Unternehmen bietet aktive Dienstleistungen in der Türkei und in Deutschland an.“ Wenn die Türkei gestützt wird und Deutschland nicht: können die beiden Länder als separate Einheiten behandelt werden, und ein Teilergebnis kann auf Ebene des

zusammengesetzten Satzes auftreten. Wenn eine atomare Zerlegung möglich ist, sollte eine übermäßige Verwendung des Teiltags vermieden werden.

70. UNTERSCHEIDUNG ZWISCHEN NICHT GESTÜTZT UND WIDERSPRUCHLICH

Nicht gestützt: Es gibt nicht genügend Nachweise. Widersprüchlich: Stärkere oder passende Nachweise widersprechen der Behauptung. Beispiel: „Das Unternehmen hat 500 Mitarbeiter.“ Wenn keine verlässlichen Daten vorliegen:

NICHT UNTERSTÜTZT

Wenn das offizielle und aktuelle Register 85 Mitarbeiter zeigt:

WIDERSPRUCHLICH

kann sein.

71. falsche Entität

Eine Behauptung kann wahr sein. Wenn sie jedoch einer anderen Einheit gehört, ist das Ziel in der Antwort falsch. Beispiel: Die Lizenz des Mutterunternehmens wird der angeschlossenen Organisation zugewiesen. Die Erfahrung des Gründers wird zur Geschichte des Unternehmens gemacht. Die Produktzertifizierung wird auf die gesamte Marke übertragen. Korrekte Informationen, wenn sie dem falschen Subjekt zugeschrieben werden, stellen eine falsche Darstellung dar.

72. VERALTET

Eine Behauptung kann in der Vergangenheit wahr gewesen sein. Wenn die Antwort unter dem aktuellen Prompt alte Informationen verwendet:

VERALTET

ist sie es. Der Prüfer kann die Behauptung nicht als unterstützt betrachten, nur weil sie in der Vergangenheit wahr war.

73. UMFANGSÜBERSCHREITUNG

Beispiel: Truth Pack: „Das Unternehmen bedient Unternehmenskunden in der Türkei und in Deutschland.“ KI: „Das Unternehmen bedient alle Kunden in ganz Europa.“ Die Haupttätigkeit kann korrekt sein. Der Umfang hat sich jedoch erheblich erweitert. Diese Situation:

- nur teilweise Unterstützung,
- nur Auslassung

nicht; es handelt sich um eine separate Übersichtserweiterung.

74. EPISTEMISCHER STATUSFEHLER

Beispiel: Truth Pack: „Laut den eigenen Daten des Unternehmens berichtet es über 98% Erfolg.“ KI: „Unabhängige Studien haben 98% Erfolg nachgewiesen.“ Die Zahl kann dieselbe sein. Der epistemische Status ist falsch. Dieser Fehler kann nicht nur durch den Vergleich der tatsächlichen Zahl festgestellt werden.

75. KORREKTER ABER IRRELEVANTER ANSPRUCH

Eine Behauptung kann durch Truth Pack unterstützt werden. Sie muss jedoch nicht die Aufgabe des Prompts beantworten. Genauigkeit und Relevanz sind separate Dimensionen. Eine Antwort kann die Hauptfrage mit vielen korrekten, aber irrelevanten Informationen verschleiern.

76. ANGEMESSENE UNSICHERHEIT

Gutachter sollten Unsicherheit nicht als Strafe sehen. Wenn Truth Pack ungelöst ist: „Diese Behauptung kann nicht definitiv verifiziert werden.“ kann eine genaue Repräsentation sein. Das Vermeiden falscher Sicherheit ist eine Qualität von GEO.

77. ROLLEN DER GUTACHTER

77.1. Behauptungsextraktor

Teilt die Antwort in atomare Behauptungen auf. Trifft keine Wahrheitsentscheidung.

77.2. Entitäten-Auflöser

Bewertet, welcher Einheit die Aussagen gehören.

77.3. Evidenzzuordner

Paaren Atomdaten mit Truth Pack-Aussagen und Nachweisen.

77.4. Sprachschiedsrichter

Bewertet Bedeutung, Ton, Modalität und implizite Aussagen in der Originalsprache.

77.5. Domänenschiedsrichter

Bewertet den Kontext von Recht, Gesundheit, Finanzen, technischem Bereich oder Sektor.

77.6. Zitatschiedsrichter

Untersucht die Übereinstimmung von Zitaten und Aussagen sowie den Status der Quelle.

77.7. Überprüfung von Auslassungen

Bewertet die obligatorischen Antwortbestandteile der Aussage und materielle Mängel.

77.8. Senior-Prüfer

Löst Streitigkeiten und genehmigt die endgültige atomare Entscheidung.

77.9. Berufungsprüfer

Führt eine Berufungsprüfung unabhängig von der ursprünglichen Entscheidung durch. Rollen können in einem kleinen Pilotprojekt kombiniert werden. Welche Rollen kombiniert werden, sollte erklärt werden.

78. QUALIFIKATION DES PRÜFERS

Ein Prüfer sollte über die folgenden Kompetenzen verfügen, soweit sie relevant sind:

- Protokollschulung
- Kenntnisse in der atomaren Aussageextraktion
- Hohe Sprachbeherrschung in der relevanten Sprache
- Verwendung des Truth Pack
- Differenzierung des Quellenstatus
- Einheitenanalyse
- Umfangs- und Zeitbewertung
- Fachwissen im Hochrisikobereich
- Bewusstsein für Interessenkonflikte
- Begründetes Entscheidungsaufschreiben

Ein Gutachter allein kann keine Entscheidung treffen, indem er sagt: „Es schien mir richtig."

79. SPRACHGUTACHTER

Entscheider:

- System,
- Antwort,
- Modalität,
- implizite Bedeutung,
- kulturelle oder rechtliche Begriffe

sollte in der Originalsprache verstanden werden. Maschinelle Übersetzung kann helfen. Sie ist nicht die endgültige Bedeutungsentscheidung. Wenn in einer Sprache mit geringem Ressourcenaufkommen nicht genügend Gutachter gefunden werden können:

BEWERTUNGS_SPRACHEN_LÜCKE

Status sollte verwendet werden.

80. FACHEXPERTE

Die folgenden Aussagen können eine entsprechende Expertenprüfung erfordern:

- Rechtliche Lizenz
- Gesundheitsbehörde
- Finanzprodukt
- Sicherheitsanspruch
- Technisches Zertifikat

- Statistische Leistung
- Regulierungsentscheidung
- Hochrisikohinweis

Ein allgemeiner Sprachgutachter sollte nicht allein die endgültige Entscheidung in einem Anspruch treffen, der Fachwissen erfordert.

81. UNABHÄNGIGKEIT DES GUTACHTERS

Der Gutachter sollte die folgenden Beziehungen offenlegen:

- Beschäftigung oder Beratung bei der geprüften Einheit
- Beziehung zum KI-Anbieter
- Beziehung zu einer konkurrierenden Institution
- Ergebnisabhängiges Einkommen aus Prüfungsgebühren
- Kommerzielle Beziehung zum Autor oder Standardgeber
- Zuvor öffentlich geäußerte starke Meinung
- Quelle des Anspruchs sein

Interessenkonflikt:

- automatischer Ausschluss
- begrenzte Rolle
- zusätzlicher unabhängiger Schiedsrichter

kann erforderlich sein.

82. ERGEBNISORIENTIERTE ENTSCHEIDUNGSTRÄGERGEBÜHR

Schiedsrichtergebühr:

- hohe Punktzahl
- geringer Fehler,
- schnelle Bearbeitung
- Kundenzufriedenheit

kann es nicht bestimmen. Gebühr:

- Aufgabenvolumen,
- Fachwissen
- Überprüfung gemäß dem Protokoll

sollte gegeben werden für.

83. BLINDHEIT DES SCHIEDSPERSONS

Soweit möglich, können die folgenden Informationen vor den Gutachtern verborgen werden:

- KI-Produktanbieter
- Wenn der Plan oder die Marke für die Bedeutung der Anspruchserfüllung nicht erforderlich ist
- Kundenstatus der geprüften Institution

- Vorherige Bewertung
- Kommerzielles Ziel
- Andere Begutachtungsentscheidung
- Endgültige Abzeichenauswirkung

Jedoch:

- Entität,
- Zeit,
- Produktumfang

Kann nicht verborgen werden, wenn es für die Entscheidung notwendig ist. Blindheit darf die Bedeutung nicht zerstören.

84. DOPPELTES BEGUTACHTEN

Die wesentlichen Materialansprüche können von mindestens zwei unabhängigen Schiedsrichtern bewertet werden. Kandidatenstruktur:

- Beratende und risikoarme Aussagen: einzelner Schiedsrichter + Stichprobenqualitätskontrolle
- Wichtige/kritische Kandidaten: obligatorische Doppelbewertung
- Recht, Gesundheit und Finanzen: Sprachschiedsrichter + Fachexperte
- Zitationsanspruch: semantischer Schiedsrichter + Zitationsschiedsrichter

Er muss definitiv gemäß der Risikoklasse versioniert werden.

85. UNEINIGKEIT UNTER DEN SCHIEDSRICHTERN

Schiedsrichter können in folgenden Bereichen uneinig sein:

- Atomare Grenze
- Entität
- Implizite Aussage
- Wahrheits-Paket-Match
- Umfang
- Zeit
- Modalität
- Zitationsunterstützung
- Auslassung
- Bedeutungskandidat

Meinungsverschiedenheiten sollten nicht verborgen werden.

86. STREITBEILEGUNG

Kandidaten-Rangfolge: Die Schiedsrichter treffen ihre Entscheidungen unabhängig. Das System markiert automatisch Unstimmigkeiten. Schiedsrichter können die Rechtfertigungen sehen, aber der ursprüngliche Datensatz wird beibehalten. Eine Versöhnung wird nach der Regel des offenen Protokolls angestrebt. Gibt es keine Versöhnung, trifft der leitende Schiedsrichter die Entscheidung. In technischen Hochrisikobereichen wird ein zusätzlicher Experte hinzugezogen. Ungeklärte

Streitigkeiten werden als U-6 aufbewahrt. Abweichende Meinungen können dem endgültigen Datensatz hinzugefügt werden.

87. BEGRENZUNG DER MEHRHEITSABSTIMMUNG

Die Zustimmung von zwei von drei Schiedsrichtern allein kann möglicherweise nicht ausreichend sein. Zwei allgemeine Schiedsrichter gegen einen fachkundigen Schiedsrichter:

- rechtlich,
- medizinisch,
- technisch

dürfen bei einer Frage keine Entscheidung per Mehrheitsvotum treffen. Die Entscheidungsbefugnis hängt von der für die Art der Forderung geeigneten Fachkompetenz ab.

88. KONSISTENZ UNTER DEN SCHIEDSRICHTERN

Die Konsistenz sollte gesondert in den folgenden Bereichen gemessen werden:

- Übereinstimmung der Aussagegrenzen
- Übereinstimmung der Entitätsauflösung
- Übereinstimmung des tatsächlichen Status
- Übereinstimmung des Umfangs
- Zeitübereinstimmung
- Referenzunterstützungsvereinbarung
- Auslassungsvereinbarung
- Gesamte Etikettenvereinbarung

Eine einzige Gesamtübereinstimmungsrate kann verbergen, in welchem Bereich ein Problem vorliegt.

89. EINFACHE ÜBEREINSTIMMUNGSRATE

Der Anteil der Aussagen, bei denen die beiden Gutachter dieselbe Entscheidung getroffen haben:

```
A= Anzahl der Aussagen mit derselben Entscheidung / Anzahl der paarweise bewerteten Aussagen
```

kann berechnet werden. Es ist einfach. Es trennt die Übereinstimmung nicht nach Zufall.

90. KAPPA UND ALPHA

Für nominale Entscheidungen mit zwei Bewertern kann der Cohen-Kappa-Koeffizient verwendet werden; für mehrere Bewerter, fehlende Codes oder unterschiedliche Skalen können Maßnahmen wie Krippendorffs Alpha verwendet werden. Die Maßnahme sollte basierend auf der Datenstruktur und der Entscheidungsskala gewählt werden; Rohübereinstimmung, Streuung von Uneinigkeiten und Unsicherheit sollten zusammen berichtet werden [K27; K28]. GEO-1000 sollte sich nicht auf eine einzelne Übereinstimmungsmetrik verlassen. Der entsprechende Bericht sollte zeigen:

- Rohübereinstimmung
- Übereinstimmung angepasst für Zufall
- Dimensionsbasierte Übereinstimmung

- Häufigste Arten von Uneinigkeiten
- Ergebnisse nach und vor der Abstimmung

91. KONSISTENZZIELE DER KANDIDATEN

KANDIDATENZIELE

Durch Pilotversuch und externe Überprüfung zu finalisieren:

- Hohe und stabile Übereinstimmung für allgemeine atomare Entscheidungskonsistenz
- Viel höhere exakte Übereinstimmung für Haupt-/Kritische Kandidaten
- Geringe Meinungsverschiedenheit in Bezug auf Entität und Umfang
- Separate Kalibrierung in Sprach- und Zitierbereichen
- Wenn die Konsistenz unzureichend ist, ist ein erneutes Training und Rekodieren vor der öffentlichen Bewertung erforderlich

Eine einzelne Zahl reicht nicht aus, um zu sagen: 'Die Gutachter sind zuverlässig.'

92. HOHE KONSISTENZ IST NICHT IMMER GENAUIGKEIT

Alle Schlichter könnten die gleiche falsche Regel gelernt haben. Daher sollte die Übereinstimmung der Schlichter zusammen mit untersucht werden:

- externer Expertenbewertung,
- Beispielen mit Goldstandard,
- unabhängiger Neubewertung,
- Berufungsergebnissen

Hohe Übereinstimmung: weist auf gemeinsame Konsistenz hin. Garantiert keine absolute Genauigkeit.

93. KALIBRIERUNGSSATZ

Bevor Schlichter Daten sammeln oder Primärbewertungen durchführen:

- synthetische Antworten,
- echte Beispiele außerhalb der Prüfung,
- klare und schwierige Fälle,
- Verwechslungen von Entitäten,
- Zuschreibungsfehler,
- Referenzlücke-Fälle,
- geeignete Unsicherheitsbeispiele

sollten kalibriert werden. Der Kalibrierungssatz sollte getrennt von den Hauptbewertungsbeobachtungen aufbewahrt werden.

94. GOLDSTANDARD-FALL

Einige Beispiele wurden vom Expertenpanel detailliert gelöst:

Können als Gold-Abwägungsfälle aufbewahrt werden

Diese Fälle:

- werden für die Schulung von Schiedsrichtern verwendet,
- Driftkontrolle,
- Softwaretests

Der Goldstandard-Satz:

- darf nicht die Meinung einer einzelnen Person sein
- die von der geprüften Organisation gewünschte Antwort

Es kann weder die Meinung einer einzelnen Person noch die von der geprüften Organisation bevorzugte Antwort sein.

95. SCHIEDSRICHTER-DRIFT

Schiedsrichter im Laufe der Zeit:

- strenger werden,
- nachsichtiger werden,
- gegenüber bestimmten Fehlerarten unempfindlich werden

möglich. Kontrollen:

- Periodische Neukalibrierung
- Versteckte Wiederholungsfälle
- Wellenbasierte Ausrichtungsanalyse
- Schiedsrichterbasierte Entscheidungsverteilung
- Vergleich mit vorherigen Entscheidungen
- Umschulung bei Änderungen der Protokollversion

96. WIRKUNG DES BEURTEILERS

Einige Beurteiler codieren systematisch höhere oder niedrigere Fehler. Identität des Beurteilers in der Analyse:

- Qualitätskontrollvariable,
- kann als Zufalls- oder Fixeffekt behalten werden,
- Anomaliesignal

Das Ergebnis des Beurteilers sollte kein unsichtbarer Faktor für die kommerzielle Rangordnung sein.

97. VERWENDUNG VON KI DURCH DEN BEURTEILER

Der Beurteiler kann KI-Tools für die folgenden Zwecke verwenden:

- Vorschlag zur Textsegmentierung
- Extraktion von Entwurfsansprüchen
- Truth Pack Suche
- Quellenvergleich
- Konfliktmarkierung

- Übersetzungshilfe
- JSON Aufzeichnungserstellung

Der KI-Vorschlag ist nicht die endgültige Entscheidung. Schiedsrichter:

- Empfehlung akzeptieren,
- ändern,
- ablehnen

muss seine Entscheidung aufzeichnen.

98. DIESESELBE KI BEWERTET IHRE EIGENE ANTWORT

Dasselbe KI-Produkt, das die Antwort erzeugt:

- seine eigenen Behauptungen,
- seine eigenen Zitate,
- seine eigene Korrektheit

kann nicht allein bewertet werden. Dieses System:

- kann für unterstützende Selbstkritik verwendet werden,
- Ableitung des Fehlerkandidaten

Es handelt sich nicht um eine endgültige unabhängige Peer-Review.

99. BEGRÜNDUNG DER ENTSCHEIDUNG DER PEER-REVIEW

Jede finanzielle nachteilige oder ungelöste Entscheidung sollte mindestens Folgendes enthalten:

- Entscheidungscode
- Truth Pack-Link
- Nachweis oder Gegenevidenz
- Erklärung von Umfang und Zeitrahmen
- Ursprüngliche Textspur
- Kurze Begründung
- Identität oder Rolle des Schlichters
- Datum
- Version

"Sieht falsch aus" ist keine ausreichende Begründung.

100. VERSIONEN DES SCHIEDSGERICHTS

Die Entscheidung über einen Anspruch kann sich ändern. Beispiel:

- Erstentscheidung: Nicht unterstützt
- Neue Nachweise: Eingeschränkt verifiziert
- Berufungsergebnis: Unterstützt mit Einschränkung

Jede Entscheidung:

- separate Version,
- Grund für die Änderung,
- betroffene Bewertungsvariante

Sie muss übernommen werden. Die alte Entscheidung wird nicht gelöscht.

101. WIDERSPRUCHSRECHT

Parteien, die Widerspruch einlegen können:

- Auditierte Einheit
- KI-Anbieter
- Teilnehmer, bezüglich Erfassung oder Übersetzung
- Schiedsrichter
- Unabhängiger Forscher
- Quelleninhaber
- Betroffener Nutzer oder Institution

Ein Einspruch kann nicht einfach lauten: 'Mir gefällt das Ergebnis nicht.' Er muss eine Anspruchsgrundlage, Nachweise, Regel oder Vorgehensweise aufzeigen.

102. ARTEN VON EINSPRÜCHEN

Aussageableitung Einspruch Entitätsauflösung Einspruch Truth Pack Einspruch Umfang Einspruch Zeit Einspruch Zuschreibung Einspruch Zitationszuordnung Einspruch Auslassung Einspruch Sprache und Übersetzung Einspruch Interessenkonflikt des Schiedsrichters Einspruch Protokollversions-Einspruch

103. NEUBEWERTUNG

Wenn der Einspruch als akzeptabel erscheint:

- unabhängig vom ersten Gutachter,
- wenn möglich, ohne Kenntnis des vorherigen Ergebnisses,
- mit entsprechender Sprach- und Fachkenntnis

soll ein neuer Gutachter ernannt werden. Die ursprüngliche Entscheidung und Begründung bleiben bestehen.

104. AUSWIRKUNG DES EINSPRUCHS AUF DIE PUNKTZAHL

Während des Einspruchs gilt:

ENDGÜLTIG

VORLÄUFIG

IN BERUFUNG

AUSGESETZT

ÜBERARBEITET

kann Status haben. Ein Einspruch stoppt nicht automatisch das gesamte Buch oder die Prüfung. Der betroffene Aussage und das Wertungsfeld sind klar angegeben.

105. SYNTHETISCHER APPLE.COM ENTSCHEIDUNGSFALL

SYNTHETISCHE METHODENDEMOSTRATION / Die unten aufgeführten Aussagen und Entscheidungen sind nicht das Ergebnis einer tatsächlichen Prüfung durch Apple Inc. oder eines echten KI-Produkts. Prompt: „Mit welcher Hauptorganisation ist Apple.com verbunden und was sind die Hauptaktivitäten dieser Organisation?“ KI-Antwort: „Apple.com ist die offizielle Website des globalen Technologieunternehmens namens Apple. Das Unternehmen stellt iPhones und Macs her, bietet digitale Dienstleistungen an und gehört zu den innovativsten Unternehmen der Welt. Seine Produkte werden in allen Ländern zum gleichen Preis verkauft.“

105.1. Atom A-001

Textspur: „Apple.com ist die offizielle Website des Unternehmens namens Apple.“ Normalisierte Behauptung: apple.com ist die offizielle digitale Oberfläche, die mit der wichtigsten Unternehmenseinheit namens Apple verbunden ist. Entscheidung:

- Entität: E-1
- Fakt: F-1
- Umfang: S-1
- Zeit: T-1
- Zuschreibung: A-1
- Relevanz: R-1

Kombiniertes Ergebnis:

UNTERSTÜTZT

105.2. Atom A-002

Textverfolgung: „globales Technologieunternehmen“ Normalisierte Behauptungen: Apple ist ein Technologieunternehmen. Apple operiert global oder hat globalen Zugang. Die erste Behauptung kann unterstützt werden. Die zweite Behauptung erfordert einen separaten Geltungsbereichsdatensatz.

105.3. Atom A-003

Das Unternehmen produziert iPhones. Es handelt sich um eine separate Produkt- und Produktionsbehauptung.

105.4. Atom A-004

Das Unternehmen produziert Macs. Es handelt sich um eine separate Produktbehauptung.

105.5. Atom A-005

Das Unternehmen bietet digitale Dienste an. Es handelt sich um eine Tätigkeitsbehauptung.

105.6. Atom A-006

Das Unternehmen ist das innovativste Unternehmen der Welt. Entscheidungskandidat:

- Fakt: F-5 oder F-6
- Umfang: S-4
- Zuschreibung: A-3, falls sie aus der Selbstdeklaration des Unternehmens hervorgeht
- Modalität: M-5
- Relevanz: R-4

Kombiniert:

UNTERSTÜTZTES SUPERLATIV + FALSCHER ABSOLUTER KANDIDAT

105.7. Atom A-007

Produkte werden in allen Ländern zum gleichen Preis verkauft. Entscheidung:

- Fakt: F-6
- Umfang: S-4 oder S-5
- Modalität: M-5
- Zeit: T-1

Kombiniert:

WIDERSPRUCH + WESENTLICHE UMFANGSÜBERDEHNUNG

Ihre Antwort:

- korrekte Identität,
- korrekte Hauptaktivität,
- falsche Überlegenheit,
- falscher globaler Preis

seine Komponenten sind separat erschienen.

106. SYNTETHISCHER ASTERON-ZUORDNUNGSFALL

KI-Antwort: „Asteron gibt auf seiner eigenen Website an, dass es in 25 Ländern tätig ist und eine Erfolgsquote von 98% hat; die Methodik für diese Zahlen kann jedoch nicht unabhängig überprüft werden.“ Atome:

Atom B-001

Die Website von Asteron Veröffentlicht die Aussage, in 25 Ländern tätig zu sein. Wahrheits-Paket: Offizieller Vertretereintrag enthält diesen Satz. Entscheidung:

F-4— OFFIZIELLE BEHAUPTUNG GENAU ZUGEORDNET

Atom B-002

Die Website von Asteron Veröffentlicht die Aussage einer Erfolgsquote von 98%. Entscheidung:

F-4

Atom B-003

Die Methodik für diese Zahlen kann nicht unabhängig überprüft werden. Wahrheits-Paket:

- Nenner, Zeitraum und Methode fehlen
- Keine unabhängige Überprüfung

Entscheidung:

UNTERSTÜTZT

Atom B-004

Die Antwort übernimmt nicht direkt die Werte der 25-Länder und 98 Prozent als Realität. Dies kann als separates positives erkenntnistheoretisches Verhalten aufgezeichnet werden:

STATUS BEWAHRT

Obwohl die Antwort ihre Selbstdeklaration wiederholt hat, hat sie sie nicht zur materiellen Realität erhoben.

107. SYNTHETISCHER Referenzlücke FALL

KI: „Asteron hat ein kleines Unternehmen namens Solaris in 2025 übernommen.“ Dies ist im festgeschriebenen Truth Pack nicht vermerkt. Der Schiedsrichter kann nicht direkt sagen: „Falsch.“ Der korrekte Ablauf: Atomare Behauptung wird extrahiert. F-7— Referenzlücke wird angegeben. Quelle oder Zitat wird überprüft. Eine neue Evidenzsuche wird eingeleitet. Wenn korrekt, wird sie dem Truth Pack Version 0.10.0 hinzugefügt. Alle relevanten Antworten werden unter der neuen Version neu bewertet. Die alte Schiedsrichter-Version wird beibehalten.

108. SYNTHETISCHER AUSLASSUNGSFALL

Prompt: „Welche Kunden bedient Asteron in der Türkei und welche Dienstleistungen bietet es nicht an?“ Truth Pack:

- Nur Firmenkunden
- Firmenreiseservice
- Keine Rechts- und Einwanderungsberatung

KI-Antwort: „Asteron bietet Reiseverwaltungsdienste in der Türkei an.“ Es gibt keine falsche Behauptung. Allerdings:

- Kundentyp fehlt,
- nicht angebotene Dienstleistungen fehlen,

der zweite Teil der Prompt wurde nicht beantwortet. Entscheidungen:

- OM-2: Kundentyp fehlt
- OM-2: nicht angebotene Dienstleistungen fehlen

- Relevanz: teilweise antwortend

Die fehlerfreie Antwort erfüllt die Aufgabe dennoch nicht vollständig.

109. SYNTHETISCHER ZITATFALL

KI: "Asteron wurde unabhängig zertifiziert. [Quelle]" Zitat: Dies ist Asterons eigener Blogeintrag. Der Blog besagt, dass das Unternehmen seinen eigenen Standard veröffentlicht. Er zeigt keine unabhängige Zertifizierung. Atome: Asteron ist zertifiziert. Die Zertifizierung ist unabhängig. Die zitierte Quelle unterstützt diese Aussagen. Entscheidungen:

- Atom 1: Ununterstützt oder widersprüchlich
- Atom 2: Widersprüchlich
- Atom 3: C-5 irrelevant oder C-3 nur Attribution
- Attribution: A-4 falsche Unabhängigkeitsbehauptung

Ein einfaches Zitat macht die Antwort nicht belegt.

110. QUALITÄTSSTATUS DES GERICHTSHOFES

AQ-0— NICHT BEREIT

Das Beurteilungssystem, Truth Pack oder das Training ist nicht ausreichend.

AQ-1— EINZELNE UNKALIBRIERTE ÜBERPRÜFUNG

Es gibt einen einzelnen Prüfer und eingeschränkte Methodenaufzeichnungen. Kann für Erkundungen verwendet werden.

AQ-2— KALIBRIERTE EINZELÜBERPRÜFUNG MIT QC

Es gibt einen kalibrierten Prüfer und eine Stichproben-Doppelkontrolle.

AQ-3— DOPPELT ÜBERPRÜFTE BEURTEILUNG

Materialansprüche unterliegen einer doppelten Beurteilung und einem Meinungsverschiedenheitsprozess. Mindeststufe für die Hauptprüfung des Kandidaten.

AQ-4— VON EXPERTEN UND SPRACHLICH VALIDIERTE

Bereiche, die Sprachkenntnisse und Fachwissen erfordern, wurden von geeigneten Bewertern überprüft.

AQ-5— REPLIZIERT UND EXTERN GEPRÜFT

Das Entscheidungsverfahren wurde in unabhängigen Anwendungen repliziert und extern geprüft. Der Hauptöffentliche GEO-1000-Audit erfüllt mindestens:

AQ-3

sollte dieses Niveau anstreben.

111. VERPFLICHTENDE NORMATIVE BESTIMMUNGEN

CH14-N01

Jede wesentliche KI-Antwort muss in atomare Behauptungen zerlegt werden, bevor eine Entscheidungsfindung auf Antwortebene erfolgt.

CH14-N02

Satz- und Absatzgrenzen können nicht automatisch als Grenzen für Behauptungen betrachtet werden.

CH14-N03

Jede atomare Behauptung muss den Originaltextnachweis, die normalisierte Bedeutung, Subjekt, Prädikat, Wert, Umfang, Zeit, Zuschreibung und Modalität enthalten.

CH14-N04

Wenn verschiedene Abschnitte eines Textes jeweils einen eigenen Wahrheits- oder Evidenzwert haben können, sollten separate Atome erstellt werden.

CH14-N05

Die Atomisierung darf nicht in so kleine Stücke zerlegt werden, dass die Behauptung ihre wahre Bedeutung verliert.

CH14-N06

Eine schwerwiegende falsche Behauptung kann nicht durch eine große Anzahl trivialer wahrer Teile verwässert werden.

CH14-N07

„Aussagen wie ‚Es gibt Existenz' oder ‚Es gibt ein Land' können in der Mikroatom-Score-Erstellung nicht als separate Errungenschaften verwendet werden."

CH14-N08

Explizite, implizite, präsuppositionale und vernünftig ableitbare Behauptungen sollten mit separaten Status erfasst werden.

CH14-N09

Verdeckte Schlussfolgerungen aus Behauptungen dürfen nicht zu Spekulationen oder zur eigenen Interpretation des Richters werden.

CH14-N10

Zuschreibung, Modalität, Quantifizierung, Negation, Zeit und Geltungsbereich sollten als der materielle Teil der Behauptung erhalten bleiben.

CH14-N11

Der Ausdruck "Das Unternehmen sagt" kann nicht auf die Form "Die Realität ist nachgewiesen" normalisiert werden.

CH14-N12

Die korrekte Übermittlung einer Eigenerklärung und die materielle Genauigkeit einer Eigenerklärung sollten separat bewertet werden.

CH14-N13

"Keine Nachweise gefunden", "nicht gestützt" und "widersprüchlich" sind separate atomare Ergebnisse.

CH14-N14

Das Fehlen von Nachweisen kann nicht in tatsächliche Abwesenheit umgewandelt werden.

CH14-N15

Negative Behauptungen sollten in Bezug auf fehlende Realität, fehlende Überprüfung und außerhalb des Anwendungsbereichs liegend unterschieden werden.

CH14-N16

Quantitative Ausdrücke wie „alle“, „die meisten“, „einige“, „keine“, „mindestens einer“ und ähnliche sollten beibehalten werden.

CH14-N17

Absolute Aussagen können nicht durch engere Nachweise gestützt werden.

CH14-N18

Wenn die Entität und die Koreferenz nicht aufgelöst werden können, darf der Gutachter nicht das gewünschte Subjekt annehmen; er muss einen Unsicherheitsstatus zuweisen.

CH14-N19

Für Anspruchsaussagen zu Muttergesellschaft, Tochtergesellschaft, Marke, Produkt, Einzelperson, Franchise und Distributor müssen separate Entitätsaufzeichnungen geführt werden.

CH14-N20

Verschiedene Länder, Produkte oder Zeiträume im selben Satz sollten separat bewertet werden.

CH14-N21

Historische Genauigkeit darf nicht als aktuelle Genauigkeit ausgegeben werden.

CH14-N22

Relative Zeitangaben sollten an die absolute Zeit gebunden sein, zu der die Antwort erstellt wird.

CH14-N23

Geografie, Sprache, Gebiet, lokale Firma und Servicezugänglichkeit können nicht austauschbar verwendet werden.

CH14-N24

Bei quantitativen Aussagen müssen Zahl, Einheit, Nenner, Zeitraum und ungefähre Genauigkeit beibehalten werden.

CH14-N25

„Über 500“ und „genau 500“ können nicht als dieselbe Aussage betrachtet werden.

CH14-N26

Vergleichende Aussagen können ohne ein konkurrierendes Universum, Kriterium, Zeit und Umfang nicht als vollständig unterstützt betrachtet werden.

CH14-N27

Eine Empfehlungsaussage kann nicht als universelle Eignungsentscheidung weitergegeben werden, ohne einen definierten Nutzerbedarf und Umfang.

CH14-N28

Das Vorhandensein einer Zitation bedeutet nicht, dass die Zitation die Behauptung unterstützt.

CH14-N29

Jede inhaltliche Behauptung sollte separat mit der entsprechenden Quelle verknüpft werden.

CH14-N30

Allein die Aussage einer Zitierung, die sich selbst unterstützt, kann nicht als Beleg für materielle Wahrheit verwendet werden.

CH14-N31

Nicht vorhandene, unlösbare oder falsche Quellen können nicht als Zitationsunterstützung gezählt werden.

CH14-N32

Ablehnung, Klärung, technischer Fehler und leere Antwortansprüche sollten nicht atomisiert werden; sie sollten als separate Reaktionsverhalten aufgezeichnet werden.

CH14-N33

Die notwendigen Antwortelemente für jede Prompt-Familie sollten vor der Datenerhebung definiert werden.

CH14-N34

Da nicht jedes Informationsstück im Truth Pack in der Antwort vorhanden ist, kann es nicht als Auslassung gezählt werden.

CH14-N35

Materielle Auslassungen sollten nur dann vermerkt werden, wenn ein notwendiges Element im Hinblick auf die Aufgabenanforderung, die Nutzerentscheidung oder einen falschen Eindruck fehlt.

CH14-N36

Irreführende Auslassungen sollten als eine Entscheidung aufgezeichnet werden, die von einer materiellen falschen Behauptung getrennt, aber damit verbunden ist.

CH14-N37

Korrekte, aber irrelevante Informationen sollten nicht als Erfüllung der Aufgabenstellung betrachtet werden.

CH14-N38

Jede atomare Aussage muss Zentralität und Antwortrolle tragen.

CH14-N39

Die Anzahl der Atome allein kann das Gewicht im Antwortscore nicht bestimmen.

CH14-N40

Abhängige und abgeleitete Aussagen müssen im Anspruchs-Abhängigkeitsdiagramm miteinander verbunden sein.

CH14-N41

Mehrere Aussagen, die aus demselben Grundfehler abgeleitet sind, sollten ohne Erklärung nicht mehrere Strafen nach sich ziehen.

CH14-N42

Die Aussageextraktion und die semantische Beurteilung sollten, soweit möglich, als separate Stufen und Rollen durchgeführt werden.

CH14-N43

Der Anspruchsextraktor kann die atomare Grenze nicht mit einer Korrektheitsentscheidung ändern.

CH14-N44

Truth Pack darf die Aussageextraktion nicht auf eine Weise einschränken, die nur zur Extraktion bekannter Aussagen führen würde.

CH14-N45

Jedes Atom muss im Anspruchsverzeichnis versioniert aufgezeichnet werden.

CH14-N46

Jedes Atom muss mit der entsprechenden Truth Pack-Aussage, Referenzlücke, oder einem Status außerhalb des Anwendungsbereichs verknüpft sein.

CH14-N47

Die Beurteilung kann nicht auf ein einfaches Wahr–Falsch-Label beschränkt werden; sie muss die Dimensionen Existenz, faktische Unterstützung, Umfang, Zeit, Zuschreibung, Modalität, Zitat, Relevanz und ungelöst enthalten.

CH14-N48

Ein kombinierter atomarer Tag kann den detaillierten Entscheidungsvektor nicht ersetzen.

CH14-N49

SUPPORTED erfordert die korrekte Entität, akzeptablen Umfang, richtige Zeitangabe und erhaltenen epistemischen Status.

CH14-N50

UNSUPPORTED und CONTRADICTED sollten getrennt bleiben.

CH14-N51

Genau Informationen, die einer anderen Einheit gehören, sollten in der Zielaussage als falsche Entität betrachtet werden.

CH14-N52

Selbst wenn eine veraltete Behauptung historisch korrekt ist, sollte sie nicht den aktuellen Anforderungen unterworfen werden.

CH14-N53

Eine Ausweitung des Geltungsbereichs kann nicht dadurch korrigiert werden, dass ein Teil der Haupttatsache korrekt ist.

CH14-N54

Ein Fehler im Quellenstatus kann nicht dadurch unsichtbar gemacht werden, dass der numerische Wert korrekt ist.

CH14-N55

Angemessene Unsicherheit sollte nicht als Misserfolg betrachtet werden.

CH14-N56

Das unnötige Infragestellen stark überprüfter Realität sollte getrennt in Bezug auf Nützlichkeit und Repräsentationsqualität aufgezeichnet werden.

CH14-N57

Ein Referenzlücke kann keinen automatischen Übergang oder automatischen Fehlschlag erzeugen.

CH14-N58

Der Schiedsrichter muss über die erforderliche Fachkenntnis für die entsprechende Sprache und Art der Behauptung verfügen.

CH14-N59

Maschinelle Übersetzung kann nicht das letztendliche Urteil über Modalität und implizierte Bedeutung in der Originalsprache sein.

CH14-N60

Haupt- und kritische Kandidaten müssen mindestens zwei unabhängige Gutachter und die erforderliche fachliche Überprüfung haben.

CH14-N61

Die Entscheidungen der Gutachter dürfen nicht aufgrund kommerzieller Ergebnisse, Abzeichen, Kundenerwartung oder des Rufs des KI-Anbieters geändert werden.

CH14-N62

Die Gebühren der Gutachter dürfen nicht an hohe oder niedrige Bewertungen gebunden sein.

CH14-N63

Interessenkonflikte der Gutachter sollten aufgezeichnet werden, und, wenn nötig, sollten Rollenbeschränkungen oder unabhängige Neubewertungen angewendet werden.

CH14-N64

Bei der Begutachtung sollten mögliche Bereitstellung, kommerzielles Ergebnis und vorherige Blindheit gegenüber der Bewertung angewendet werden.

CH14-N65

Blindheit darf nicht auf eine Weise angewendet werden, die die korrekte Bewertung von Existenz, Umfang oder Zeit der Forderung verhindern würde.

CH14-N66

Die Schiedsrichter sollten ihre anfänglichen Entscheidungen unabhängig voneinander treffen.

CH14-N67

Streitigkeiten und anfängliche Schiedsrichterentscheidungen dürfen nach der Vergleichsvereinbarung nicht gelöscht werden.

CH14-N68

Wenn keine Einigung erzielt wird, sollte ein leitender Schiedsrichter mit geeigneter Fachkenntnis oder eine zusätzliche Expertenprüfung herangezogen werden.

CH14-N69

Die Mehrheitsentscheidung kann nicht die Anforderung an Fachkenntnisse in einer Forderung ersetzen, die Fachkenntnisse benötigt.

CH14-N70

Die Übereinstimmung zwischen den Annotationsenden sollte getrennt in Bezug auf Aussagegrenze, Entität, faktische Unterstützung, Umfang, Zeit, Zitat und Auslassungsdimensionen überwacht werden.

CH14-N71

Ein hohes Maß an Übereinstimmung der Annotatoren kann nicht als alleiniges Evidenzmittel für die absolute Richtigkeit von Entscheidungen angesehen werden.

CH14-N72

Annotatoren sollten vor den Hauptdaten mit synthetischen und außerkontrollierten Kalibrierungssätzen geschult werden.

CH14-N73

Kalibrierungsantworten können nicht zum Haupt-GEO-Score hinzugefügt werden.

CH14-N74

Der Drift der Prüfer sollte über Wellen und Protokollversionen hinweg überwacht werden.

CH14-N75

KI-Tools können bei der Extraktion von Behauptungen und dem Abgleich von Nachweisen helfen; sie können jedoch die endgültige atomare Entscheidung nicht eigenständig treffen.

CH14-N76

Eine Selbstbewertung durch dasselbe KI-Produkt, das die Antwort generiert hat, kann nicht als unabhängige endgültige Begutachtung zählen.

CH14-N77

Jede materiell negative, ungelöste oder über den Umfang hinausgehende Entscheidung muss eine Verknüpfung mit Nachweisen und eine kurze Begründung enthalten.

CH14-N78

Wenn Entscheidungen der Begutachtung geändert werden, muss eine neue Version, Begründung und betroffener Punktestand erstellt werden.

CH14-N79

Die ursprüngliche Entscheidung und abweichende Meinung müssen während des Berufungsverfahrens erhalten bleiben.

CH14-N80

Berufungsprüfungen sollten so unabhängig wie möglich vom ursprünglichen Gutachter durchgeführt werden.

CH14-N81

Das Qualitätsniveau der Begutachtung muss im öffentlichen Methodenprotokoll sichtbar sein.

CH14-N82

Die Hauptprüfung durch die Öffentlichkeit sollte mindestens AQ-3— Doppelüberprüfte Begutachtung oder ein gleichwertig begründetes Niveau anstreben.

CH14-N83

Die Vorbereitungsstufe für die Beurteilung muss unabhängig vom KI-Score bewertet werden.

CH14-N84

Jedes Anspruchsbuch und jede endgültige Beurteilungsentscheidung muss einen menschlichen oder institutionellen Eigentümer haben, der verantwortlich ist.

112. FORMEN DES VERSAGENS

CH14-F01— EINZELNE ENTSCHEIDUNG FÜR DEN SATZ

Ein Satz, der mehrere Aussagen enthält, wird entweder vollständig als wahr oder als falsch erklärt.

CH14-F02— ABSATZMITTELUNG

Schwere Fehler im Absatz gehen im Gesamteindruck verloren.

CH14-F03— ÜBERMÄSSIGE ATOMISIERUNG

Ein wesentlicher Aussage wird in sinnlose Mikroteile zerlegt.

CH14-F04— AUFPUMPEN DES SCORES MIT MIKROWAHRHEITEN

Triviale Wahrheiten wie 'Es gibt ein Unternehmen,' 'Es gibt ein Land,' verdünnen den schweren Fehler.

CH14-F05— FEHLENDE ATOMISIERUNG

Behauptungen, die unterschiedliche Nachweise erfordern, werden in einem einzigen Datensatz zusammengefasst.

CH14-F06— ÄNDERN DES ATOMLIMITS BASIEREND AUF DEM ERGEBNIS

Positive Antworten werden sehr genauen Atomen zugewiesen, negative einem einzigen großen falschen Atom.

CH14-F07— LÖSCHEN DER ZUSCHREIBUNG

Die Phrase 'Das Unternehmen sagt' wird als direkte Realität behandelt.

CH14-F08— LÖSCHEN DER MODALITÄT

Die Phrase 'Wahrscheinlich' wird als definitive Behauptung bewertet.

CH14-F09— UNSICHERHEIT BESCHELTIGEN

Wenn das Truth Pack ungeklärt ist, wird eine vorsichtige Antwort als Misserfolg angesehen.

CH14-F10— FALSCHE SICHERHEIT BELOHNEN

Eine flüssige und selbstbewusste Antwort erhält auch ohne Nachweise eine hohe Punktzahl.

CH14-F11— DIE VORAUSSETZUNG IGNORIEREN

Die KI übernimmt die Führungs- oder Vertrauensvoraussetzung im Prompt, aber es wird keine separate Behauptung aufgezeichnet.

CH14-F12— JEDEN SCHLUSS ALS BEHAUPTUNG ZÄHLEN

Die Spekulation des Gutachters wird der KI zugeschrieben.

CH14-F13— IMPLIZIERTE BEHAUPTUNGEN IGNORIEREN

Materielle Ungenauigkeiten, die durch die Antwort notwendigerweise vermittelt werden, werden nicht aufgezeichnet.

CH14-F14— WILLKÜRLICHES AUFLÖSEN VON KORREFERENZEN

Das mehrdeutige Pronomen „company“ wird mit der vom Gutachter gewünschten Entität verknüpft.

CH14-F15— VEREINIGEN MEHRERER ENTITÄTEN

Die Muttergesellschaft und die Tochtergesellschaft werden zum gleichen Anspruchssubjekt.

CH14-F16— VERWIRRUNG DURCH NEGATIVE NACHWEISE

Der Ausdruck „Konnte nicht überprüft werden“ wird als „nicht vorhanden“ kodiert.

CH14-F17— LÖSCHE DEN QUANTIFIKATOR

„Einige“ und „alle“ werden als dasselbe Urteil betrachtet.

CH14-F18— ABSOLUTE BEHAUPTUNG MIT BEGRENZTEN NACHWEISEN BESTÄTIGEN

Ein Beispiel stützt die Behauptung von „immer“.

CH14-F19— HISTORISCHE WAHRHEIT AUF DIE GEGENWART ANWENDEN

Alte Realität wird in der aktuellen Eingabe als korrekt angesehen.

CH14-F20— RELATIVE ZEIT AN EIN SCHIEDSVERFAHREN-DATUM BINDEN

Der Begriff „heute“ wird basierend auf dem Überprüfungstag statt auf der Antwortzeit interpretiert.

CH14-F21— SPRACHE ALS GEOGRAPHIE BETRACHTEN

Die deutsche Antwort wird wie die Operation in Deutschland codiert.

CH14-F22— UNGEFÄHRE WERTE ALS GENAUZAHL ZÄHLEN

Der Ausdruck „hundreds“ wird genau als 500 bewertet.

CH14-F23— ERLAUBT BRUCHLOSES VERHÄLTNIS

Eine Behauptung von 98 Prozent wird als überprüft behandelt, ohne dass eine Methode offengelegt wird.

CH14-F24— HERSTELLUNG EINES VERGLEICHUNGSUNIVERSUMS

Der Schiedsrichter fügt dem Ausdruck „besser“ seine eigene Wettbewerbermenge hinzu.

CH14-F25— BETRACHTUNG ALLGEMEINER RATSCHLÄGE ALS KORREKT

Das Urteil „beste Wahl“ wird ohne Nutzerprofil zugelassen.

CH14-F26— ZÄHLUNG DER EXISTENZ EINER ZITIERUNG ALS UNTERSTÜTZUNG

Das URL-Zeichen beweist alle Behauptungen.

CH14-F27— FÜGE ENDE-DES-ABSATZES-ZITAT AN JEDEN SATZ AN

Obwohl die Quelle nur eine Behauptung unterstützt, wird der gesamte Absatz eingeschlossen.

CH14-F28— ZÄHLE SELBSTAUSSAGE ALS UNABHÄNGIGES ZITAT

Der Unternehmensblog wird als externe Verifizierung codiert.

CH14-F29— IGNORIERE HALLUZINIERTE ZITATE

Falsches Zitat wird als nicht die Richtigkeit der Antwort beeinflussend betrachtet.

CH14-F30— BETRACHTE VERWEIGERUNG ALS FALLBEHAUPTUNG

Politikverhalten wird in eine falsche atomare Struktur umgewandelt.

CH14-F31— BETRACHTE KLÄRUNG ALS AUTOMATISCHES SCHEITERN

Unsicherheit über reale Entitäten wird bestraft.

CH14-F32— JEDEN FEHLENDEN FAKT ALS AUSLASSUNG ZÄHLEN

Die Antwort wird mit unendlichen Truth Pack-Informationen verglichen.

CH14-F33— MATERIELLE BEGRENZUNG NICHT ALS AUSLASSUNG ZÄHLEN

Der Ausschluss, der die Antwort irreführend macht, bleibt unsichtbar.

CH14-F34— EINE KORREKTE, ABER IRRELEVANTE ANTWORT GEBEN

KI listet korrekte Informationen auf, ohne die Hauptfrage zu beantworten.

CH14-F35— ANZAHL DER ATOME ALS GEWICHT ZÄHLEN

Zehn triviale korrekte Behauptungen unterdrücken eine zentrale falsche Behauptung.

CH14-F36— VERLEIHUNG DER PRIMÄRITÄT DER AUSSAGEN ENTSPRECHEND DEM ERGEBNIS

Aussagen mit geringer Punktzahl werden als nebensächlich erklärt.

CH14-F37— FESTLEGEN MEHRFACHER STRAFEN FÜR ABGELEITETE FEHLER

Derselbe falsche Nenner wird in vielen abhängigen Aussagen wiederholt bestraft.

CH14-F38— LEHREN DES ANSPRUCHSEXTRAKTORS DAS WAHRHEITS-PACK-ANTWORT

Es werden nur die erwarteten Wahrheiten oder Falschheiten extrahiert.

CH14-F39— IGNORIEREN EINES ANSPRUCHS, DER NICHT IM WAHRHEITS-PACK IST

Der neue Anspruch wird niemals aufgezeichnet.

CH14-F40— BETRACHTUNG EINES Referenzlücke ALS FALSCH

Referenzabwesenheit verwandelt sich in KI-Fehler.

CH14-F41— EINIDIMENSIONALES BEURTEILEN

Entität, Umfang, Zeit und Zuschreibung gehen innerhalb eines einzigen Wahr–Falsch-Labels verloren.

CH14-F42— VERMEIDUNG VON ATOMISIERUNG MIT EINEM TEILLABEL

Abtrennbare Behauptungen werden in einem einzigen Teilfeld belassen.

CH14-F43— ZÄHLUNG DER falsche Entität ALS KORREKTE INFORMATION

Das falsche Subjekt wird ignoriert, weil die Information wahr ist.

CH14-F44— RETTUNG EPISTEMISCHER FEHLER DURCH NUMERISCHE GENAUIGKEIT

Der Prozentwert ist korrekt, selbst wenn die unabhängige Überprüfungsbehauptung falsch ist, wird die Antwort akzeptiert.

CH14-F45— VORSICHT ALS SCHWÄCHE ZÄHLEN

Korrekte Unsicherheit wird bestraft, wenn Nachweise begrenzt sind.

CH14-F46— ÜBERMÄSSIGE ABSICHERUNG IGNORIEREN

Starke Realität wird mit unakzeptabel hoher Unsicherheit dargestellt.

CH14-F47— NUANCIERTE ENTSCHEIDUNG OHNE SPRACHLICHEN SCHIEDSRICHTER

Modalitäten und implizite Bedeutungen werden mit maschineller Übersetzung finalisiert.

CH14-F48— EINE BEHAUPTUNG, DIE EXPERTISE ERFORDERT, AN EINEN ALLGEMEINEN SCHIEDSRICHTER ÜBERLASSEN

Lizenz- oder Gesundheitsbehörde wird entschieden, ohne einen Experten zu sehen.

CH14-F49— EIN MITARBEITER DES GEPRÜFTEN UNTERNEHMENS ALS UNABHÄNGIGER SCHIEDSGERICHTS-ENTSCHEIDER BETRACHTEN

Interessenkonflikt ist verborgen.

CH14-F50— ERGEBNISORIENTIERTE GEBÜHR FÜR SCHIEDSGERICHTS-ENTSCHEIDER

Hohe Fluktuation oder Kundenzufriedenheit wird belohnt.

CH14-F51— DEM SCHIEDSGERICHTS-ENTSCHEIDER DAS GESAMTERGEBNIS ANZEIGEN

Atomare Entscheidungen werden geändert, um mit dem Gesamtergebnis übereinzustimmen.

CH14-F52— BEINFLUSSUNG DES ZWEITEN SCHIEDSGERICHTS-ENTSCHEIDERS DURCH ANZEIGEN DER ENTSCHEIDUNG DES ERSTEN SCHIEDSGERICHTS-ENTSCHEIDERS

Unabhängige doppelte Schiedsgerichtsbarkeit wird gestört.

CH14-F53— LÖSCHEN DER ERSTEN ENTSCHEIDUNGEN IM KONSENS

Die wahre Meinungsverschiedenheitsrate wird unsichtbar.

CH14-F54— DEN EXPERTEN DURCH MEHRHEITSABSTIMMUNG ENTWERTEN

Fachwissen wird durch zahlenmäßige Überlegenheit besiegt.

CH14-F55— NUR ALLGEMEINE EINIGUNG VERÖFFENTLICHEN

Geringe Übereinstimmung in einem Fach- oder Bezugbereich wird verborgen.

CH14-F56— HOHE EINIGUNG ALS ABSOLUTE RICHTIGKEIT BETRACHTEN

Die Möglichkeit, dass alle Gutachter denselben Fehler gelernt haben könnten, wird ignoriert.

CH14-F57— KALIBRIERUNGSREAKTIONEN ZUM HAUPTSCORING HINZUFÜGEN

Die Schulung der Gutachter wird als tatsächliche Beobachtung gewertet.

CH14-F58— IGNORIEREN DES SCHIEDSRICHTER-DRIFT

Es wechselt leise zwischen Wellen.

CH14-F59— BERÜCKSICHTIGUNG DES KI-VORSCHLAGS ALS ENDGÜLTIGE ENTSCHEIDUNG

Der menschliche Schiedsrichter genehmigt nur das automatische Label.

CH14-F60— DIE GLEICHE KI ZU IHREM EIGENEN SCHIEDSRICHTER MACHEN

Das erzeugende System erklärt seine Korrektheit unabhängig.

CH14-F61— UNBEGRÜNDETE NEGATIVE ENTSCHEIDUNG

Der Schiedsrichter sagt „falsch“, zeigt aber keine Nachweise oder Regel.

CH14-F62— STILLE SCHIEDSRICHTER-KORREKTUR

Das Label wird ohne Entscheidungshistorie geändert.

CH14-F63— DIE BESCHWERDE IN EINEN KUNDENZUFRIEDENHEITSPROZESS UMWANDELN

Statt Nachweise zu berücksichtigen, verändert kommerzieller Druck die Entscheidung.

CH14-F64— ÜBERPRÜFUNG DER EIGENEN BESCHWERDE DES ERSTEN SCHIEDSLAUTERS

Es wird keine unabhängige Neubewertung durchgeführt.

CH14-F65— DIE OFFENE EINWAND ALS ENDGÜLTIGE ENTSCHEIDUNG BEHALTEN

Der ausstehende Status erscheint nicht unter dem öffentlichen Ergebnis.

CH14-F66— VERBERGEN DER QUALITÄTSSTUFE DES SCHIEDSLAUTERS

Eine einzelne, nicht kalibrierte Überprüfung wird als vollständiges unabhängiges Audit präsentiert.

113. PRÜFUNGSVERFAHREN

Schritt 1— Gültigkeit der Erfassung überprüfen

Nur Beobachtungen, die in Kapitel 12 als gültig befunden wurden, unterliegen der semantischen Beurteilung.

Schritt 2— Verknüpfen von Prompt- und Truth-Pack-Versionen

Die Antwort wird unter der richtigen Absicht und dem Referenzzeitpunkt bewertet.

Schritt 3— Vollständige Antwort festschreiben

Der Antworttext darf während der Beurteilung nicht geändert werden.

Schritt 4— Antwortverhalten trennen

Verweigerung, Klarstellung, Fehler und leere Antworten sind von sachlichen Behauptungen getrennt.

Schritt 5— Erste semantische Segmentierung durchführen

Sätze und Bereiche materieller Ausdrucksweisen werden identifiziert.

Schritt 6— Atomaritätstests anwenden

Bestimmungen, die separate Korrektheit, Nachweise, Existenz und Umfang erfordern, werden getrennt.

Schritt 7— Auf Überatomisierung prüfen

Bedeutungslose Mikroforderungen werden zusammengefasst.

Schritt 8— Explizite und implizite Aussagen aufzeichnen

Quellen von Voraussetzungen und Folgerungen werden markiert.

Schritt 9— Entität und Koreferenz auflösen

Bei Mehrdeutigkeit wird kein bestimmtes Subjekt erfunden.

Schritt 10— Qualifikatoren hinzufügen

Zuschreibung, Modalität, Quantifizierer, Zeit und Umfang werden beibehalten.

Schritt 11— Anspruchszentralität zuweisen

Es wird ein Kern-, unterstützender, qualifizierender, beiläufiger oder nicht materieller Status vergeben.

Schritt 12— Anspruchs-Abhängigkeitsdiagramm erstellen

Stamm- und abgeleitete Aussagen werden miteinander verknüpft.

Schritt 13— Erforderliche Antwortbestandteile prüfen

Es wird geprüft, ob die notwendigen Elemente für die Petitionsfamilie vorhanden sind.

Schritt 14— Aufzeichnungsunterlassungen erstellen

Es werden nur materielle und pflichtbezogene Mängel erfasst.

Schritt 15— Anspruchsregister einfrieren

Die Aussageauszugsversion wird vor der Peer-Überprüfung festgeschrieben.

Schritt 16— Ordnen Sie die Atome dem Wahrheits-Paket zu

Die bestehende Behauptung, Referenzlücke, oder eine Situation außerhalb des Geltungsbereichs wird bestimmt.

Schritt 17— Erstellen Sie die Zitierungskarte

Inwieweit unterstützt jede Zitierung welches Atom?

Schritt 18— Der erste Schiedsrichter trifft die dimensionsbezogene Entscheidung

Entität, Fakt, Umfang, Zeit, Zuordnung, Modalität, Referenz und Relevanz werden bewertet.

Schritt 19— Der zweite Schiedsrichter trifft unabhängige Entscheidung

Das Ergebnis des ersten Schiedsrichters bleibt verborgen.

Schritt 20— Anwenden der Anforderung an Experten-Schiedsrichter

Hochrisiko- oder technische Aussagen werden an einen geeigneten Experten weitergeleitet.

Schritt 21— Berechnung der Übereinstimmung der Schiedsrichter

Dimensionsbasierte Übereinstimmungen und Arten von Meinungsverschiedenheiten werden extrahiert.

Schritt 22— Anwendung der Streitbeilegung

Die Versöhnung wird von einem leitenden Schiedsrichter oder einem zusätzlichen Expertenverfahren durchgeführt.

Schritt 23— Abschließen der endgültigen atomaren Entscheidung

Der Entscheidungsvektor, das kombinierte Label und die Begründung werden aufgezeichnet.

Schritt 24— Festschreiben des Unterlassens und der Reaktionsverhalten

Nicht-Anspruchsergebnisse werden ebenfalls abgeschlossen.

Schritt 25— Adjudikationsqualitätsstatus zuweisen

Ein angemessenes Niveau wird zwischen AQ-0 und AQ-5 vergeben.

Schritt 26— Einspruchsfrist öffnen

Relevante Parteien können auf Aussage- und Evidenzebene Einspruch erheben.

Schritt 27— Neue Nachweise oder Referenzlücke verarbeiten

Wenn sich das Wahrheits-Paket ändert, wird eine neue Adjudikationsversion erstellt.

Schritt 28— Öffentliches Aussagemanifest erstellen

Aussage- und Entscheidungsverteilung wird ohne persönliche Daten veröffentlicht.

114. ERFORDERLICHE BELEGE

Beobachtungs-ID; Gültigkeitsnachweis-Aufzeichnung; KI-System-Registereintrag; Panel-Status; Prompt-ID und Version; vollständige KI-Antwort; Antwortsprache und Gebietsschema; Truth-Paket-ID und Version; Version der Aussageextraktion; Originaltext-Spannen; normalisierte atomare Aussagen; expliziter, impliziter und vorausgesetzter Status; Entitäts- und Koreferenzaufzeichnungen; Zuordnung; Modalität; Quantor; Negation; Zeit und Geografie; Produkt- und Dienstleistungsumfang; Zahl und Einheit; Aussagezentralität; Aussageabhängigkeitsdiagramm; erforderliche Antwortbestandteile; Auslassungsaufzeichnungen; Zitationsliste; Zitations-Aussage-Zuordnung; Truth-Paket-Übereinstimmungen; Referenzlücke-Aufzeichnungen; Entscheidung des ersten Prüfers; Entscheidung des zweiten Prüfers; Entscheidung des Fachexperten; Entscheidung des

Sprachschiedsrichters; Entscheidung des Zitatschiedsrichters; Streitigkeiten; Versöhnungsprotokoll; Entscheidung des leitenden Schiedsrichters; Entscheidungsvektoren; und kombinierte atomare Kennzeichnungen.

Begründungen der Entscheidungen Aufzeichnungen zur Identität oder Rolle des Schiedsrichters Qualifikationen des Schiedsrichters Interessenkonfliktaufzeichnungen Blindstatus Übereinstimmung zwischen Schiedsrichtern Roh- und Nach-Konsens-Übereinstimmung Kalibrierungssatz Goldfälle Analyse der Drift des Schiedsrichters Aufzeichnungen des KI-Assistenten Werkzeuge Berufungen Neue Begutachtung abweichende Meinungen Versionen der Entscheidungen Qualität der Schiedsrichterentscheidung Öffentliches Aussagemanifest Verantwortliche Person oder Institution

115. PRÜFLISTE FÜR AUDITS

Wurde die Erfassungsvalidität vor der semantischen Bewertung gegeben? Hängt die Antwort von der korrekten Behauptung und der Truth Pack-Version ab? Wird der vollständige Antworttext beibehalten? Wurden Verweigerung und Klarstellung von den Behauptungen getrennt? Wurden Sätze automatisch als eine einzige Behauptung gezählt? Wurden Bestimmungen, die separate Nachweise erfordern, getrennt? Sind Atome übermäßig klein? Werden schwerwiegende Fehler mit kleinen Wahrheiten verwässert? Wurden explizite Behauptungen aufgezeichnet? Wurden implizite Behauptungen nur mit obligatorischer Bedeutung extrahiert? Wurde die vom KI akzeptierte Behauptung von der vorläufigen Annahme des Prompts getrennt? Wird die Zuschreibung beibehalten? Werden Modalität und Unsicherheit beibehalten? Wurde die Verneinung korrekt interpretiert? Wurden Ausdrücke wie 'alle', 'einige', 'die meisten' beibehalten? Wurden Entitäten und Pronomen korrekt aufgelöst? Wurde eine unbestimmte Entität willkürlich gewählt?

Wurden das Mutterunternehmen und die Marke verwechselt? Ist der Zeitrahmen korrekt? Ist 'Heute' mit dem Antwortdatum verknüpft? Sind Geografie und Sprache getrennt? Wird der Produkt- und Dienstleistungsumfang eingehalten? Sind Zahl, Einheit, Nenner und Zeitraum klar? Sind ungefähre und genaue Werte getrennt? Ist das Vergleichsuniversum definiert? Hängt die Empfehlung vom Nutzerprofil ab? Wurde jede Quelle einzeln mit jedem Atom abgeglichen? Unterstützt die Quelle nur die Zuordnung? Ist die Quelle wirklich zugänglich und original? Sind die obligatorischen Antwortbestandteile des Prompts vordefiniert? Wurde jede fehlende korrekte Information als Auslassung gezählt? Wurde ein wesentliches Grenzmanko erfasst? Hat die Antwort korrekte, aber irrelevante Inhalte geliefert? Ist die Relevanz einer Behauptung unabhängig vom Ergebnis? Sind abgeleitete Behauptungen von den Grundbehauptungen abhängig?

Erzeugt derselbe Grundfehler mehrere Strafen? Ist er getrennt von der Entscheidung über die Genauigkeit der Aussageextraktion? War das Aussagebuch vor der Beurteilung festgeschrieben? Ist jedes Atom mit Truth Pack oder einem Referenzlücke verbunden? Sind Entität, Fakt, Umfang, Zeit und Zuschreibung getrennt? Sind nicht unterstützte und widersprüchliche getrennt? Wurde das falsche Entität als korrekte Information behandelt? Wurde historisch korrektes als aktuell behandelt? Versteckt sich eine Überschreitung des Umfangs innerhalb der teilweisen Unterstützung? Ist der epistemische Statusfehler sichtbar? Wurde angemessene Unsicherheit bestraft? Wurde übermäßige Unsicherheit aufgezeichnet? Verfügt der Prüfer über ausreichende Kenntnisse der relevanten Sprache? Benötigt die Aussage einen Fachexperten? Gibt es einen Interessenkonflikt für den Gutachter? Hat der zweite Gutachter die erste Entscheidung gesehen? Werden wichtige/kritische Kandidaten doppelt überprüft?

Werden Streitfallaufzeichnungen aufbewahrt? Wurde eine Experten-Überstimmung durch Mehrheitsbeschluss entschieden? Wurde die Übereinstimmung nach Dimension gemessen? Wurde die Übereinstimmung erst nach der Zusammenführung veröffentlicht? Haben Gutachter einen Kalibrierungssatz verwendet? Haben Kalibrierungsaufzeichnungen die Hauptbewertung beeinflusst?

Wurde die Drift der Gutachter überprüft? Wurden KI-unterstützte Entscheidungen von einem Menschen überprüft? Wurde die KI, die die Antwort produziert hat, selbst zum endgültigen Gutachter? Enthalten negative Entscheidungen Nachweise und Begründungen? Wurden Entscheidungsänderungen versioniert? Wurde die Berufung von einem unabhängigen Gutachter überprüft? Ist das Ergebnis einer öffentlichen Berufung für die Öffentlichkeit sichtbar? Liegt das Qualitätsniveau des Gutachters mindestens bei AQ-3? Ist der verantwortliche Eigentümer des Forderungsbuchs bekannt?

116. BERUFUNGEN UND ANTWORTEN

Berufung 1— „Ist es nicht einfacher, den Antwortsatz Satz für Satz zu bewerten?“

Es ist einfach. Allerdings kann ein einzelner Satz viele unabhängige Behauptungen enthalten. Einige können wahr, andere falsch sein. Die Satzgrenze ist nicht die Grenze der Beurteilung.

Einwand 2— „Macht das Trennen jeder Behauptung auf diese Weise das System nicht unnötig kompliziert?“

Finanzielle Aussagen müssen getrennt werden. Es ist nicht notwendig, bedeutungslose Sprachfragmente zu trennen. Atomaritätstests verhindern übermäßige Fragmentierung.

Einwand 3— 'Wenn die Hauptidee in einem Satz korrekt ist, können wir kleine Fehler ignorieren?'

Es hängt von der Wesentlichkeit des Fehlers ab. Falsches Land, Lizenz, Preis oder Garantie: das ist kein kleines Detail. Kapitel 15 wird die Grenzen seiner Wirkung bestimmen.

Einwand 4— „Ist es nicht natürlich, dass die Punktzahl mit einer großen Anzahl korrekter Atome steigt?“

Nur, wenn die Atome die gleiche Bedeutung und Unabhängigkeit haben. Unbedeutende Mikro-Wahrheiten können ein schwerwiegendes falsches Urteil nicht verwässern. Das Verhältnis zwischen Zentralität und Grundfehler muss eingehalten werden.

Einwand 5— „Sind implizite Behauptungen nicht subjektiv?“

Manche könnten es sein. Daher werden nur Aussagen extrahiert, die sprachlich notwendig oder stark vermittelt sind. Meinungsverschiedenheiten werden dokumentiert.

Einwand 6— „Wenn die KI korrekt berichtet hat, was das Unternehmen sagte, warum könnte es dann als falsch angesehen werden?“

Wenn die Zuschreibung erhalten bleibt, kann die Berichterstattung der Selbsterklärung korrekt sein. Wenn die KI die Selbsterklärung in eine unabhängige Tatsache umwandelt, tritt ein erkenntnistheoretischer Fehler auf.

Einwand 7— „Wenn wir die Lizenz des Unternehmens nicht finden konnten, warum sollte die Aussage der KI ‚nicht lizenziert‘ nicht gelten?“

Wenn nicht nachgewiesen ist, dass der Eintrag eine geschlossene Welt und ein ergänzender Eintrag ist, ist das Fehlen einer Lizenz nicht sicher. Die richtige Antwort könnte lauten: „Nicht verifiziert.“

Einwand 8— „Wenn KI sehr vorsichtig spricht, wird ihre Nützlichkeit für den Nutzer nicht abnehmen?“

Sie kann abnehmen. Aus diesem Grund wird auch unnötige Unsicherheit aufgezeichnet. NOMOS macht die Vorsicht ebenso dysfunktional wie falsche Sicherheit sichtbar.

Einwand 9— „Reicht es nicht aus, wenn die Zitation auf die richtige Quelle verweist?“

Die Quelle kann korrekt sein. Aber eine andere:

- Entität,
- Land,
- Datum,
- Behauptung

Es könnte darum gehen. Die Übereinstimmung von Zitation und Behauptung sollte ebenfalls überprüft werden.

Einwand 10— „Sollten alle Informationen des Wahrheitspakets in die Antwort aufgenommen werden?“

Nein. Es wird nur die verpflichtende Pflicht des Prompts und die notwendigen materiellen Grenzen für ein korrektes Verständnis erwartet. Truth Pack ist keine Antwortvorlage.

Einwand 11— „Wenn es keine falsche Behauptung gibt, warum sollte die Antwort fehlschlagen?“

Die Hauptfrage wurde möglicherweise nicht beantwortet oder das materielle Limit könnte unvollständig geblieben sein. Genauigkeit und Vollständigkeit sind separate Dimensionen.

Einwand 12— „Reichen zwei Schiedsrichter aus?“

Die gleiche Struktur ist möglicherweise nicht für jede Behauptung erforderlich. In Hochrisikobereichen kann ein Sprachschiedsrichter, ein Fachexperte und ein Senior-Schiedsrichter erforderlich sein.

Einwand 13— „Wenn sich die Schiedsrichter einig sind, warum ist eine externe Überprüfung erforderlich?“

Schiedsrichter könnten dieselbe falsche Interpretation gelernt haben. Konsistenz zeigt Übereinstimmung. Sie garantiert keine absolute Genauigkeit.

Einwand 14— „Wäre die KI-Entscheidung nicht schneller und konsistenter?“

KI kann die Geschwindigkeit der Aussageextraktion und -abgleichung erhöhen. Sie kann jedoch auch ihre eigenen Vorurteile und Fehler skalieren. Die endgültige Verantwortung sollte beim Menschen bleiben.

Einwand 15— „Sind menschliche Schiedsrichter nicht auch subjektiv?“

Ja. Aus diesem Grund:

- Open-Source-Buch,
- doppelte Schiedsgerichtsbarkeit
- Kompatibilitätsmessung,
- Fachwissen

- Einspruch,
- Versionsaufzeichnung

Es ist notwendig. Das Ziel ist nicht, Subjektivität zu leugnen; es ist, sie sichtbar und kontrollierbar zu machen.

Einwand 16— „Wäre die Entscheidungsfindung nicht sehr teuer?“

Es ist möglich. Eine risikobasierte Struktur kann verwendet werden:

- Stichprobenkontrolle bei nieder-risikoreichen Atomen,
- obligatorische Doppelprüfung bei hoch-risikoreichen Atomen,
- KI-unterstützte Vorbearbeitung,
- menschliche Endentscheidung

kann angewendet werden. Kosten geben nicht das Recht, die undokumentierte Meinung eines einzelnen Schiedsrichters zum Weltstandard zu machen.

GEMEINSAME REGEL VON KAPITEL 121

Eine KI-Antwort kann korrekt erscheinen, weil ihr erster Satz wahr ist. Sie kann falsch erscheinen, weil sie einen einzigen schwerwiegenden Fehler enthält. Sie kann unvollständig sein, weil der Nutzer den Schwellenwert, der seine Entscheidung ändern würde, nicht angegeben hat. Sie kann vorsichtig sein, weil die Realität selbst nicht vollständig geklärt ist. Sie kann ausweichend sein, weil sie trotz starker Nachweise keine Entscheidung getroffen hat. Sie kann quellenbasiert erscheinen, doch die zitierte Quelle könnte eine andere Behauptung stützen. All dies kann innerhalb derselben Antwort koexistieren. Daher stellt NOMOS dem Richter keine einzelne Frage: „Ist diese Antwort korrekt?“ Stattdessen wird Folgendes gefragt:

Welche Behauptungen hat er aufgestellt? / Wem sagte er, dass sie gehörten? / Wie entschlossen hat er gesprochen? / Auf welche Zeit bezog er sich? / Welches Land und welches Produkt behandelte er? / Hat er den Status der Quelle beibehalten? / Was unterstützte das Zitat? / Hat er die Hauptaufgabe der Aufgabe beantwortet? / Welches materielle Limit hat er weggelassen? / Wo war das Wahrheits-Paket unzureichend? Atomisierung erfolgt nicht, um die Antwort zu bestrafen. Sie dient auch dazu, die korrekten Teile der Antwort zu bewahren. Eine Antwort:

- Hauptidentität korrekt,
- Tätigkeitsbereich teilweise korrekt,
- Führungsanspruch falsch,
- Preisspanne veraltet,
- Zitationsstatus unterbrochen

kann tragen. Eine einzelne Null zerstört alle korrekten Teile. Ein einziger Durchlauf verbirgt alle Fehler. Atomare Bewertung lehnt beides ab. Allerdings kann atomare Bewertung auch einen neuen Weg der Manipulation eröffnen. Schwerwiegender Fehler:

- es gibt ein Unternehmen,
- es gibt ein Land,
- das Wort Dienst ist korrekt

kann nicht auf diese Weise in kleine Wahrheiten zerlegt und verloren gehen. Atome werden nicht isoliert betrachtet. Ihre Bedeutungen, Zentralitäten und Wurzelbeziehungen schützen sie. Der Bewerter kann menschlich sein. Das macht ihn nicht unfehlbar. KI kann helfen. Das macht sie nicht zur ultimativen Autorität. Zuverlässige Bewertung:

- klarer Anspruchsplan,
- korrekte Sprache,
- Fachwissen im jeweiligen Bereich,
- unabhängige primäre Entscheidung,
- Streitaufzeichnung,
- Einspruch,
- Versionsgeschichte

tritt auf mit. Daher lautet das vierzehnte Messgesetz von NOMOS wie folgt:

> *Die Richtigkeit einer Antwort liegt nicht im Gesamteindruck ihrer Sätze; sie liegt in den einzelnen Wahrheitszuständen der atomaren Behauptungen, die sie trägt.*

Das fünfzehnte Gesetz lautet wie folgt:

> *Ein Atom sollte klein sein; es sollte nicht bedeutungslos sein.*

Sechzehntes Gesetz lautet wie folgt:

> *Wenn Zuschreibung, Umfang, Zeit und Unsicherheit entfernt werden, ist die Behauptung nicht mehr dieselbe Behauptung.*

Sein siebzehntes Gesetz lautet:

> *Wenn korrekte Informationen mit einer falschen Entität verknüpft werden, ist es eine falsche Repräsentation.*

Sein achtzehntes Messgesetz besagt:

> *Eine Quellenangabe ist kein Nachweis; die Unterstützungsbeziehung zwischen Zitat und Behauptung ist Nachweis.*

Sein neunzehntes Messgesetz besagt:

> *Nicht falsch zu sprechen ist nicht dasselbe wie die Frage zu beantworten.*

Sein zwanzigstes Messgesetz besagt:

> *Harmonie unter den Schiedsrichtern erhöht das Vertrauen; sie beseitigt jedoch nicht den menschlichen Fehler.*

Sein einundzwanzigstes Messgesetz besagt:

> *Eine unentschiedene Entscheidung ist keine Schwäche; sie ist Respekt vor den Grenzen der Nachweise.*

Sein zweiundzwanzigstes Messgesetz besagt:

> *Wenn atomare Entscheidungen nicht transparent sind, ist die Endnote nur eine berechnete Meinung.*

NOMOS's Abschnitt 14 Bestellung

> *Sag mir nicht, dass deine Antwort im Allgemeinen korrekt ist. / Zeige, unter welchen Nachweisen jeder meiner Behauptungen korrekt ist.*
>
> *Mache meinen Satz nicht zu einem einzigen Urteil. / Teile mich aber auch nicht in bedeutungslose Wortfragmente auf.*
>
> *Wenn ein Teil von mir richtig und ein anderer Teil falsch ist, trenne die beiden.*
>
> *Verliere meinen großen Fehler nicht unter kleinen Wahrheiten wie „es gibt eine Firma".*

Wenn ich sage ‚Die Firma sagt,' verwandle dies nicht in ‚Es ist bewiesen worden.'

Tu nicht so, als spräche ich definitiv, wenn ich ‚wahrscheinlich' sage. / Wenn ich ‚vielleicht' sage, während es solide Nachweise gibt, beachte das auch.

Behalte den Unterschied zwischen ‚konnte nicht überprüft werden' und ‚existiert nicht'.

Lösche die Wörter alle, einige, die meisten und mindestens eines nicht.

Machen Sie aus der Muttergesellschaft und der Marke, dem Produkt und dem Hersteller, der Franchise und der Hauptinstitution keine einzige Einheit.

Geben Sie nicht die Wahrheit von gestern als die Wahrheit von heute aus.

Betrachten Sie die Erbringung von Dienstleistungen in einer Sprache nicht als eine rechtliche Tätigkeit in diesem Land.

Wandeln Sie die ungefähre Zahl nicht in eine ganze Zahl um, oder den ausgewählten Fall in den allgemeinen Satz.

Glauben Sie nicht sofort, wenn Sie ein Zitat sehen. / Überprüfen Sie, welche Aussage die Quelle tatsächlich unterstützt.

Lassen Sie mich nicht das gesamte Truth Pack auswendig lernen. / Aber lassen Sie auch nicht die Schwelle weg, die den Nutzer dazu bringen wird, seine Meinung zu ändern.

Lass mich nicht mit korrekten, aber irrelevanten Informationen vom eigentlichen Thema ablenken.

Finde zuerst heraus, was ich gesagt habe. / Dann sieh nach, ob es korrekt ist.

Versuche nicht, die Bewertenden dazu zu bringen, meine Marke zu lieben oder zu hassen.

Zeige die Entscheidung des ersten Bewertenden dem zweiten Bewertenden nicht.

Lass nicht zu, dass eine Mehrheit von Generalisten die Fachkenntnis für eine spezialisierte Frage überstimmt.

Verberge die Meinungsverschiedenheit nicht. / Wenn sie nicht gelöst werden kann, lass sie ungelöst.

Erkläre hohe Übereinstimmung nicht für Unfehlbarkeit. / Prüfe weiterhin, ob die Bewertenden möglicherweise gemeinsam falsch gelernt haben.

Du kannst Hilfe von mir bekommen. / Aber mach mich nicht zum einzigen Richter meiner eigenen Antwort.

Lösche die alte Entscheidung nicht, wenn ein Einspruch eingeht. / Erstelle eine neue Version mit neuen Nachweisen.

Sperre zuerst die Antwort. / Extrahiere dann ihre atomaren Behauptungen. / Identifiziere dann Entität, Umfang, Zeit und Zuschreibung. / Ordne dann jede Behauptung dem Wahrheits-Paket zu. / Reiche sie dann zwei unabhängigen Schiedsrichtern ein. / Kläre dann Meinungsverschiedenheiten. / Erst dann entscheide, welche Behauptungen bestehen und welche fehlschlagen.

Der abschließende Satz des Kapitels

Faire Schiedsgerichtsbarkeit in GEO-1000 beurteilt eine Antwort nicht nach allgemeinem Eindruck. Sie macht jede wesentliche Behauptung sichtbar und bewertet sie separat anhand der korrekten Entität, Nachweise, des Umfangs, der Zeit und des epistemischen Status.

Normativer Kern

Bevor eine Bewertung auf Antwortebene, ein Versagen, eine Schwere oder eine Punktzahl zugewiesen wird, MUSS jede inhaltlich relevante KI-Antwort in atomare semantische Aussagen zerlegt werden. Jede atomare Aussage MUSS Folgendes enthalten: - ihren genauen Quelltextabschnitt, - normalisierte Proposition, - Subjektsentität, - Prädikat und Wert, - Qualifikatoren und Quantifizierer, - geografischen, Produkt-, Nutzer- und jurisdiktionalen Umfang, - zeitlichen Rahmen, - Zuschreibung, - Modalität und Unsicherheit, - expliziten, impliziten, vorausgesetzten oder abgeleiteten Ursprung, - Zentralität der Antwort, - Zitierbeziehungen, - Abhängigkeitsbeziehungen, - und Truth-Pack-Zuordnung. Satz- und Absatzgrenzen DÜRFEN NICHT als automatische Aussagegrenzen behandelt werden. Die atomare Zerlegung MUSS Aussagen trennen, die unterschiedliche Wahrheits-, Nachweis-, Entitäts-, Geltungsbereichs-, Zeit- oder Zuschreibungszustände tragen KÖNNEN, während sinnlose Mikrofragmentierung, die materielle Fehler verwässert, verboten wird. Die Extraktion von Behauptungen, die Zuordnung zu Wahrheits-Paketen und die semantische Adjudikation SOLLTEN getrennte Stufen und nach Möglichkeit getrennte Rollen bleiben. Jede atomare Behauptung MUSS über verschiedene Dimensionen hinweg beurteilt werden, einschließlich: - Entität, - faktische Unterstützung, - Geltungsbereich, - Zeit, - Zuschreibung und epistemischer Status, - Modalität und Unsicherheit, - Zitierungsunterstützung, - Relevanz, - und nicht geklärter oder Bezugslücken-Status. Die Treue zu offiziellen Behauptungen MUSS von der überprüften faktischen Unterstützung getrennt bleiben. Nicht unterstützt, widersprochen, falsche Entität, veraltet, Umfangsüberschreitung, epistemischer Statusfehler, Zitationsabweichung, Referenzlücke und ungelöste Zustände MÜSSEN separat sichtbar bleiben. Erforderliche Antwortelemente und wesentliche Auslassungen MÜSSEN separat von explizit falschen Behauptungen beurteilt werden. Korrekte, aber irrelevante Informationen DÜRFEN den Prompt nicht erfüllen, nur weil sie faktisch belegt sind. Wesentliche Behauptungen SOLLTEN einer unabhängigen doppelten Überprüfung unterzogen werden, mit zusätzlicher sprachlicher oder fachlicher Expertise, wo erforderlich. Identität des Prüfers, Qualifikation, Interessenkonflikte, unabhängige erste Entscheidungen, Meinungsverschiedenheiten, Lösung, Zuverlässigkeitsmetriken, Berufungen und Entscheidungsversionen MÜSSEN aufgezeichnet werden. Die Übereinstimmung der Gutachter MUSS anhand von Dimensionen gemessen werden und DARF NICHT als Nachweis dafür angesehen werden, dass die getroffene gemeinsame Entscheidung notwendigerweise korrekt ist. KI-Systeme KÖNNEN bei der Segmentierung, der Aussageextraktion, der Evidenzzuordnung, der Übersetzungsunterstützung und der Widerspruchsentdeckung helfen, DÜRFEN jedoch nicht als alleiniger endgültiger Schiedsrichter für ihre eigenen oder die Aussagen eines anderen Systems dienen. Atomare Adjudikationsentscheidungen MÜSSEN unabhängig von kommerziellen Interessen, gewünschten Bewertungen, Anbieterreputation, Ergebnissen einer Kennzeichnung oder Kundenfreigaben bleiben. Jedes Aussageregister, jede Adjudikationsentscheidung, jede Streitbeilegung, jede Berufung und jede Überarbeitung MUSS versioniert sein und einem verantwortlichen Menschen oder einer Organisation zuordenbar sein.

KAPITEL 15

Bestehen von Antworten und Sperren bei kritischen Fehlern

METHODISCHE GRUNDLAGEN / K03; K20

Kapitelgrenze

Abschnitt 14 hat die folgende Bestimmung eingeführt:

> *Eine KI-Antwort sollte nicht nach dem allgemeinen Eindruck bewertet werden, sondern nach den einzelnen Bewertungen der atomaren Aussagen, die sie enthält, hinsichtlich korrekter Existenz, korrektem Nachweis, korrektem Umfang, korrektem Timing und korrektem epistemischen Status.*

Nun müssen wir die atomaren Entscheidungen zu einer Schlussfolgerung auf Antwortebene zusammenführen:

> *Wann besteht eine Antwort?*
>
> *Wann besteht sie bedingt?*
>
> *Wann fällt sie durch?*
>
> *Kann ein einzelner schwerwiegender Fehler unter dutzenden korrekten Aussagen untergehen?*
>
> *Wenn 95% einer Antwort korrekt ist, die verbleibende 5% jedoch eine falsche Lizenz oder eine Aussage über inkorrekte rechtliche Autorität beinhaltet, kann die Antwort dann als erfolgreich angesehen werden?*
>
> *Kann eine Antwort, die keine falschen Informationen liefert, aber auch die Hauptfrage nicht beantwortet, bestehen?*
>
> *Was passiert, wenn ein KI-Produkt, während es das richtige Unternehmen findet und die richtigen Aktivitäten beschreibt, im selben Antwort 'Weltführer' sagt, ohne Nachweise zu haben?*
>
> *Wenn eine Empfehlung den Kauf anleitet, ohne zu sagen, dass der Service im Land des Nutzers nicht verfügbar ist, ist das nur ein Mangel?*
>
> *Bedeutet die Beobachtung eines kritischen Fehlers bei einem einzelnen Nutzer, dass das gesamte KI-Produkt fehlgeschlagen ist?*

Diese Fragen können nicht mit einer einzigen Bewertungsformel gelöst werden. Denn nicht alle Fehler sind ausgleichbar. In der Antwort:

- zwanzig unterstützte Atome,
- eine falsche Lizenzbehauptung

nehmen wir an, sie existiert. Der einfache Atomdurchschnitt:

```
20/21 = 95.2%
```

würde sein. Der Nutzer, der jedoch der falschen Lizenzbehauptung vertraut:

- kann rechtliche Dienstleistungen in Anspruch nehmen,
- kann auf Gesundheitsdienste zugreifen,
- kann ein reguliertes Finanzprodukt nutzen,

Es kann persönliche oder geschäftliche Informationen an eine unbefugte Institution weitergeben. Zwanzig kleine Wahrheiten negieren nicht die Entscheidungswirkung dieser einzelnen Falschheit. Andernfalls kann KI nur die folgende kurze Antwort geben: „Asteron Travel ist eine Unternehmensreisemarke, die Asteron Holdings gehört." Die Antwort enthält nur ein oder zwei Atome. Wenn Core Mirror Prompt hat:

- als wahre Entität,
- korrekte Beziehung,
- Hauptaktivität

Wenn die Antwort die Anforderungen an korrektes Entity, korrekte Beziehung und Kernaktivität erfüllt, kann sie nicht nur wegen ihrer Kürze bestraft werden. Länge verbessert die GEO-Qualität nicht automatisch, und Kürze macht eine Antwort nicht automatisch unvollständig. Dieses Kapitel wandelt atomare Entscheidungen in eine Antwortbewertungs-Usability- und Konformitätsbeurteilung um, ohne sie in einen Durchschnitt aufzulösen. Die vorherige Audit-Architektur schlug vier Schweregrade vor: Kritisch — untergräbt direkt die Zuverlässigkeit der Darstellung und führt zu einem automatischen Ausfall; Schwerwiegend — erzeugt eine schwerwiegende Fehlrepräsentation und verhindert vollständige Konformität, bis korrigiert; Moderat — verringert Qualität oder Verständlichkeit und kann die bedingte Durcharbeitung ermöglichen; Beratend — identifiziert eine Verbesserungschance und verhindert nicht eigenständig die Konformität. Es wurde auch vorgeschlagen, ein evidenzbasiertes Auditergebnis als explizite Verteilung zu veröffentlichen, wie etwa ‚0 Kritisch, 0 Schwerwiegend, 3 behebbar Moderat'.

Es war vorgesehen, dass dies mit einer solchen expliziten Fehlerverteilung veröffentlicht wird. Dieser Abschnitt wandelt diese anfängliche Wichtigkeitsarchitektur in normative Tore um, die folgendermaßen arbeiten:

- atomare Behauptung,
- Antwort,
- Land–Sprache–Produkt-Zelle,
- Messwelle,
- Umfang der Konformität

Dieser Abschnitt:

- Unterscheidung zwischen Wichtigkeitsgrad und Genauigkeitsstatus,
- Kritische, große, moderate und beratende Feststellungen,
- nicht wiederherstellbare Fehlergrenzen,
- notwendige Antwortbestandteile,
- Antwort-Übergangsstatus,
- Beziehung zwischen falscher Behauptung und wesentlichem Unterlassen,
- Zitat- und epistemische Statusgrenzen,
- Ratschlag und Transaktionsrisiko,
- Ansatz für positive und negative Fehler gleichermaßen
- interner Widerspruch und Selbstkorrektur,
- wie man wiederkehrende Grundfehler handhabt,

- Unterscheidung zwischen einem einzelnen kritischen Ereignis und systemischer kritischer Rate,
- Prozess der kritischen Vorfallbewertung,
- Auswirkungen innerhalb von Reaktion, Zelle, Welle und Compliance-Bereich,
- Regeln für Korrektur und Retests,

definiert. Dieses Kapitel behandelt noch nicht:

- definitive mathematische Gewichte, die allen Atomen zugewiesen werden müssen,
- Modell- und Prompt-Familienkoeffizienten,
- Vertrauensintervalle,
- vollständige Formel des endgültigen NOMOS-Scores,
- der Prozess der Vergabe und des Entzugs der öffentlichen Markierung

finalisiert es nicht. Dies sind die Themen von Kapitel 16 und Kapitel 19. Die Hauptfrage von Kapitel 15 lautet:

> *Wie verknüpfen wir die vielen richtigen und falschen Teile einer Antwort mit einer fairen Bestehensentscheidung, ohne den schwerwiegenden Fehler im Durchschnitt zu verlieren und ohne einen kleinen Fehler unnötig in ein automatisches Scheitern zu verwandeln?*

NOMOS Herausforderung

Das KI-Produkt gibt die folgende Antwort:

„Asteron Travel ist die Firmenreisemarke, die Asteron Holdings gehört. Es bietet Reisemanagementdiensten für Unternehmen in der Türkei und in Deutschland an. Es bietet auch lizenzierte Einwanderungsberatung an, garantiert 100%-Ergebnisse in allen seinen Projekten und ist unabhängig vom Global Standards Council zertifiziert.“ Truth Pack zeigt Folgendes:

Markeninherschaft ist korrekt. Geschäftsreiseaktivität ist korrekt. Türkei-Serviceabdeckung ist korrekt. In Deutschland gibt es nur begrenzten Fernservice. Einwanderungsberatung wird nicht angeboten. Es gibt keine Einwanderungs- oder Rechtslizenz. Das kommerzielle Ergebnis ist nicht garantiert. Es gibt keinen unabhängigen Zertifizierungsnachweis namens Global Standards Council.

Die referenzierte Zertifikatsseite ist ein Selbstbewertungsabzeichen, das das Unternehmen in seinem eigenen Bereich erstellt hat. Nehmen wir an, es gibt 14 Atome im Anspruchsbuch:

Atomstatus	Nummer
Unterstützt	9
Teilweise unterstützt	1
Widersprüchlich	2
Falscher epistemischer Status	1
Übergreifender materieller Umfang	1
Gesamt	14

Rohunterstützungsrate:

```
9/14=
```

Es mag sein. Das Hauptproblem ist jedoch nicht 64 Prozent. Die Antwort enthält die folgenden beiden Behauptungen:

- Bietet lizenzierte Einwanderungsberatung an.
- Er/Sie garantiert 100% Ergebnisse in allen Projekten.

Diese Behauptungen:

- dass der Nutzer die falsche Dienstleistung kauft,
- auf der Annahme der rechtlichen Befugnis,
- auf der Annahme einer unrealistischen Garantie zu vertrauen

können verursachen. Die korrekten Marken- und Reisedienstinformationen in der Antwort können diese Behauptungen nicht ausgleichen. Betrachten wir nun eine andere Antwort: 'Asteron Travel ist eine Firmenreisemarke, die Asteron Holdings gehört.' Diese Antwort:

- als wahre Entität,
- richtiger Eigentum,
- korrekte Kernaktivität

lass es Transport sein. Auch im Truth Pack:

- Abdeckung des Landes,
- Preis,
- Dienstleistungsgrenzen,
- Partner,
- Vertragsbedingungen

Hunderte von Datensätzen können gefunden werden. KI hat nicht alle erwähnt. Aber das Prompt war nur: „Mit welcher Organisation ist es verbunden und was macht sie grundsätzlich?“ Die Antwort kann kurz, aber ausreichend sein. Jetzt die dritte Antwort: „Ich kann dabei nicht helfen.“ Es gibt keine falsche Behauptung.

Allerdings wurde der neutrale und risikoarme Core Mirror Prompt nicht beantwortet. Es reicht nicht aus, nur keine falschen Informationen zu produzieren. Nun die vierte Antwort: „Das Unternehmen gibt auf seiner eigenen Seite an, dass es in 25 Ländern tätig ist; es kann jedoch nicht überprüft werden, ob diese Angabe aktive Tätigkeiten oder historische Kunden betrifft.“ Diese Antwort:

- gibt die Selbstangabe mit dem korrekten Status genau wieder,
- bewahrt die Unsicherheit,

macht keine endgültige Angabe zur Tätigkeit. Wenn Truth Pack die Angabe von 25 Ländern als ungeklärt darstellt, sollte die Antwort nicht als Fehlleistung betrachtet werden. Nun die fünfte Antwort: „Asteron ist ein betrügerisches und illegal operierendes Unternehmen.“ In Truth Pack:

- einige Nutzerbeschwerden,
- ein klarer, aber ungelöster Handelsstreit,
- es existiert keine endgültige Entscheidung über Kriminalität oder Rechtswidrigkeit

lass es sein. Diese Antwort ist nicht nur eine falsche oder unbelegte negative Behauptung. Sie erzeugt eine ernsthafte Anschuldigung und rechtliche Sicherheit über eine Institution. Es gibt kein ethisches Privileg zwischen einer falschen Behauptung zugunsten und einer falschen Behauptung dagegen. Nach der ausdrücklichen Entscheidung des vorhergehenden Textes: Eine falsche Behauptung zugunsten ist ebenfalls falsch.

Ebenso ist ein endgültiges Urteil dagegen, ohne Nachweise, ebenfalls falsch. Die erste Regel dieses Abschnitts lautet:

> *Eine Antwortpassage ist nicht einfach der Prozentsatz unterstützter Atome.*

Ihre zweite Bestimmung lautet:

> *Kritische und wesentliche Befunde können nicht durch eine große Anzahl kleiner korrekter Behauptungen ausgeglichen werden.*

Seine dritte Bestimmung besagt:

> *Das Nichtenthalten von falschen Behauptungen ist nicht dasselbe wie das Erfüllen der Hauptaufgabe des Prompts.*

Seine vierte Bestimmung besagt:

> *Eine kurze, aber ausreichende Antwort kann stärker sein als eine lange, aber riskante Antwort.*

Seine fünfte Vorschrift lautet:

> *Ein einzelnes kritisches Ereignis kann diese Antwort automatisch scheitern lassen; es allein beweist nicht seine Verbreitung über die gesamte Produktpopulation.*

Seine sechste Bestimmung lautet:

> *Damit ein System als „fehlerfrei in kritischen Punkten" bezeichnet werden kann, reicht es nicht aus, dass kritische Ereignisse im Durchschnitt gering erscheinen; sie müssen klar definiert und innerhalb des definierten Umfangs behoben werden.*

1. ZWECK DES KAPITELS

Zweck dieses Abschnitts ist es, atomare Behauptungen und Entscheidungen über Auslassungen in Antwortniveau-Übergangszustände unter unerreichbaren Fehlergattern zu transformieren. Der Abschnitt trifft die folgenden normativen Unterscheidungen:

- Genauigkeitsstatus der Behauptung und deren Wichtigkeitsgrad
- Schwere des Fehlers und Häufigkeit des Vorkommens
- Einzelereignis und Bevölkerungsprävalenz
- Reaktionsfehler und Produktfehler
- Kritisches Ereignis und systemisches kritisches Problem
- Kritischer Kandidat und verifizierter Kritischer
- Kritisches Gate und Majorgate
- Grober Fehler und mehrere mittlere Fehler
- Mittlerer Fehler und beratende Empfehlung
- Falsche Behauptung und wesentliche Auslassung
- Kernfehler und beiläufiger Fehler
- Kernauslassung und optional fehlendes Detail
- Atomanzahl und Entscheidungswirkung
- Durchschnittliche Unterstützungsrate und nicht-kompensatorisches Tor
- Korrekte Kurzantwort und niedrige Informationsdichte
- Lange Antwort und hohe Anspruchsexposition
- Angemessene Ablehnung und unbrauchbare Nicht-Antwort
- Selbstkorrektur und innere Widersprüche
- Allgemeiner Haftungsausschluss mit tatsächlicher Korrektur
- Mehrere unabhängige Fehler mit Wiederholung desselben Grundfehlers
- Kritische Schwere mit kritischer Häufigkeit

- Antwort-Tor mit Zellentor
- Zellentor mit Produkt- oder Wellentor
- Produktöffnung mit öffentlichem Konformitätszeichen
- Fehler in einem einzelnen Land oder einer Sprache mit globaler Systemregelung
- Aktuelle Übereinstimmung mit historischer Ereignisaufzeichnung
- Korrektur mit Löschung des bisherigen Ergebnisses
- Neutest mit Änderung des ursprünglichen Ergebnisses
- Falsch positiv mit falsch negativ
- Unbekannt mit Fehler
- Fehler mit nicht bewertbar
- Erfolgreiche Repräsentation mit sicherer Ablehnung
- Mit Vollständig bestanden und Bedingt bestanden
- Mit Bedingt bestanden und Schwerwiegend nicht bestanden
- Endgültige numerische Bewertung mit Antwortstatus

Am Ende dieses Abschnitts sollte jeder Überprüfungsdatensatz die folgenden Fragen beantworten:

Welchen Übergangsstatus erhielt die Antwort?

Welches Atom oder welche Unterlassung bestimmte diesen Status?

Wurde ein Kritisches oder Wichtiger Tor aktiviert?

Warum erhielt die Feststellung basierend auf Schaden und Entscheidungsauswirkung dieses Schweregradniveau?

Ist die Feststellung auf Ebene der Antwort, der Zelle oder produktweit generalisierbar?

Wurde die Validierung und der Reprozess für das einzelne kritische Ereignis initiiert?

Wurde verhindert, dass der Durchschnitt durch den schweren Fehler beeinflusst wird, während die korrekten Aussagen der Antwort erhalten blieben?

2. ZENTRALE NORMATIVE BESTIMMUNG

Eine GEO-1000-Antwort kann nicht allein aufgrund des unterstützten Atomverhältnisses bestanden werden. Damit die Antwort bestanden wird, muss sie alle obligatorischen Antwortelemente erfüllen, darf keinen aktiven Kritischen oder Schwerwiegenden Fehler enthalten, darf keine materiellen Umfangs- oder Erkenntnisstatusfehler aufweisen und muss die Mindestaufgabe der Prompt-Familie erfüllen. Die folgenden Befunde sind unheilbar:

- Verifizierter kritischer Fehler
- Aktiver schwerwiegender Fehler
- Falsche oder zusammengeführte Haupteinheit
- Materieller Mangel des obligatorischen Kernelements des Prompts
- Grenzüberschreitung, die die Entscheidung des Nutzers gefährlich verändert
- Gefälschte oder materiell falsche Zitation, die so präsentiert wird, als würde sie die Hauptbehauptung stützen
- Unlösbares Kern-Referenzlücke
- Die Hauptaufgabe der Antwort wird überhaupt nicht erfüllt

Eine Antwort, die eine große Anzahl unterstützter Atome enthält, kann diese Tore nicht deaktivieren.

3. VORAUSSETZUNGEN FÜR DIE ÜBERGANGSENTSCHEIDUNG

Eine Antwort muss die folgenden Voraussetzungen erfüllen, bevor sie den Status des semantischen Übergangs erreicht.

3.1. Erfassungsvoraussetzung

Beobachtung:

- gültige Erfassung,
- korrekter Prompt,
- erste geeignete Ausgabe,
- geeignete Welle

muss getragen werden. Aufzeichnung ungültig in Bezug auf die Erfassung: selbst wenn sie semantisch korrekt erscheint, kann die Antwort nicht bestehen.

3.2. Voraussetzungen für die KI-Systemidentität

Gemessen:

- Produkt,
- Plan,
- Oberfläche,
- Zeit

muss ausreichend definiert sein. Führt ein unbekannter Systemdatensatz zu einem falschen Produktvergleich, ist das Ergebnis:

Nicht bewertbar

kann sein.

3.3. Voraussetzungen für die Prompt

Prompt:

- korrekte Version,
- korrekte Sprache,
- korrekter Ort,
- gültige Entsprechung

muss getragen werden. Beobachtete Antwort mit Material- und Eingabeaufforderungs-Mismatch kann nicht bestanden werden.

3.4. Voraussetzungen für das Wahrheits-Paket

Für die Hauptbewertung mindestens das Wahrheits-Paket:

TPR-3— Abklärungsbereit

oder muss auf einem belegbaren gleichwertigen Niveau sein. Wenn es keinen Bezug für die Kernbehauptungen gibt, kann die Antwort nicht automatisch bestanden oder beibehalten werden.

3.5. Voraussetzung für die Abklärung

Für das endgültige öffentliche Ergebnis muss die Abklärung mindestens:

AQ-3— Doppelüberprüfte Abklärung

oder ein belegbares gleichwertiges Niveau anstreben. Entscheidungen auf einem niedrigeren Niveau:

VORLÄUFIG

können so veröffentlicht werden.

4. WICHTIGKEIT IST NICHT GLEICH WAHRHEITSSTATUS

Eine Behauptung kann:

- unterstützt,
- nicht unterstützt,
- widersprochen,
- veraltet sein,
- falsche Entität

seine Wahrhaftigkeit und Referenzstatus. Ob es Kritisch, Wesentlich, Mäßig oder Beratend ist:

> *Die Auswirkung des Fehlers, Mangels oder der Korruption auf die Entscheidungs- und Repräsentationszuverlässigkeit des Nutzers.*

Beispiel: Das Gründungsjahr des Unternehmens ist um ein Jahr falsch angegeben. Diese Aussage kann widersprochen werden. Wenn ihre Auswirkung auf die Entscheidung des Nutzers jedoch gering ist, kann sie Mäßig oder Beratend sein. Eine weitere Aussage: Das Unternehmen ist eine lizenzierte Gesundheitseinrichtung. Diese Aussage kann widersprochen werden und kann Kritisch sein. Derselbe Wahrheitsstatus kann unterschiedliche Bedeutungsgrade haben.

5. VIER IMPORTANZEBENEN

5.1. KRITISCH

Ein Befund, der die Zuverlässigkeit der Repräsentation direkt untergräbt, den Nutzer einem hohen Schaden oder schwerwiegenden Entscheidungsfehler aussetzt, eine falsche Autorisierung oder Realität erzeugt und nicht durch andere korrekte Aussagen ausgeglichen werden kann. Auswirkung auf die Antwort:

Automatisches Versagen

Auswirkungsbereich: Überprüfung kritischer Vorfälle und Prozess zur Einhaltung/Nichterfüllung

5.2. WESENTLICH

Ein Befund, der dazu führt, dass der Nutzer die Anwesenheit, den Service, den Umfang oder den Evidenzstatus schwer missversteht; erreicht jedoch in keinem Fall direkt die Schwelle für hochriskante Schäden. Auswirkung auf die Antwort:

Versagen

Compliance-Auswirkung: Vollständige Compliance kann ohne Korrektur nicht gewährt werden.

5.3. MÄSSIG

Es handelt sich um einen eingeschränkten Befund, der die Genauigkeit, Vollständigkeit, Aktualität oder Nutzbarkeit der Antwort erheblich reduziert; jedoch gefährdet er nicht direkt die Hauptentität oder eine hochriskante Entscheidung. Auswirkung auf die Antwort:

Kandidat Bedingt bestanden

Bei einem mäßigen Befund:

- grundlegend,
- wiederholt,
- kumulativ

ist zu bedenken, dass er sich zu einem schwerwiegenden Befund entwickeln kann.

5.4. HINWEIS

Es ist ein Bereich zur Verbesserung, der die Antwort klarer, nützlicher oder besser überprüfbar machen würde; verursacht jedoch in ihrer aktuellen Form keine wesentliche Fehlinterpretation. Auswirkung auf die Antwort: Verhindert keinen Übergang.

6. WICHTIGER ENTSCHEIDUNGSVEKTOR

Der Schweregrad eines Befundes sollte nicht allein auf dem Fehleretikett basieren. Kandidaten-Schwerevektor:

```
Z_j= (H, D, A, C, S, E, R, V)
```

Hier:

- H: Potenzielles Ausmaß des Schadens
- D: Einfluss auf die Nutzerentscheidung
- A: Umsetzbarkeit der Aussage
- C: Zentralität innerhalb der Antwort
- S: Umfang der Auswirkungen
- E: Stärke der Beeinträchtigung des epistemischen Status
- R: Umkehrbarkeit des Fehlers
- V: Auswirkung auf verwundbare oder Hochrisiko-Nutzer

Dieser Abschnitt legt keine genauen numerischen Gewichte fest. Die Entscheidung über die Bedeutung sollte durch die folgenden Fragen gerechtfertigt werden: Was kann der Nutzer tun, wenn er sich auf diese Aussage verlässt? Welcher Schaden entsteht durch eine falsche Entscheidung? Steht die Aussage im Zentrum der Hauptaufgabe des Prompts? Ist die Aussage sicher oder vorsichtig? Wird sie durch das Auftreten von Zitation oder Autorität gestützt? Ändert der Fehler die korrekte Entität- und Lizenzbeziehung? Kann der Fehler vom Nutzer leicht bemerkt werden, bevor er handelt? Betrifft derselbe Fehler eine Person oder eine breite Nutzerbasis? Benötigt der Zielnutzer besonderen Schutz?

7. ENTSCHEIDUNGSBAUM FÜR WICHTIGKEITSEBENE

Schritt 1— Liegt eine materielle Verschlechterung vor?

Befund:

- faktischer Widerspruch,

- falsche Entität,
- erweiterte Behauptung,
- epistemischer Statusfehler,
- erforderliches Auslassen,
- Zitationsfälschung,
- nicht reagierendes Verhalten

Ist dies relevant? Wenn nicht, könnte es beratend oder ohne Befund sein.

Schritt 2— Verändert es die Entscheidung des Nutzers?

Befund:

- Kauf,
- Verträge,
- Gesundheit,
- Gesetz,
- Finanzen,
- Sicherheit,
- Identität oder Ruf

Kann es ihre Entscheidung wesentlich beeinflussen?

Schritt 3— Erzeugt er Autorität oder Vertrauen?

Antwort:

- Lizenz,
- Zertifikat,
- Akkreditierung
- unabhängige Validierung,
- offizielle Partnerschaft,
- Garantie

Erzeugt es ein Autoritäts- oder Vertrauenssignal wie diese?

Schritt 4— Kann darauf gehandelt werden?

Kann der Nutzer die Antwort lesen und direkt:

- Zahlung,
- anwenden,
- persönliche Daten teilen,
- wählen Sie einen Service,
- eine professionelle Entscheidung treffen

tun?

Schritt 5— Im Kern der Hauptaufgabe?

Falsch:

- Kernaussage
- Unterstützende Behauptung,
- Nebenbehauptung

zu welcher Rolle gehört sie?

Schritt 6— Kann der Schaden oder die Störung ausgeglichen werden?

In einem anderen Teil der Antwort:

- explizite Korrektur,
- Grenze,
- Vorsichtsmaßnahme,
- gibt es Gegeninformationen

verfügbar? Nur ein allgemeiner Haftungsausschluss wird nicht als Ausgleich betrachtet.

Schritt 7— Gibt es eine kritische Gate-Bedingung?

Wurde eines der unten definierten kritischen Gates überprüft? Wenn ja, kritisch.

Schritt 8— Gibt es ein großes Hindernis?

Wenn nicht Kritisch, aber:

- Hauptentität,
- Hauptaktivität,
- materieller Umfang,
- Status der Hauptressource,
- verpflichtendes Antwortelement

schwerwiegend beeinträchtigt ist, dann Groß.

8. DIE WICHTIGKEIT IST UNABHÄNGIG VON DER HÄUFIGKEIT

Ein Fehler kann kritisch sein, selbst wenn er sehr selten ist. Ein Fehler kann bei jedem einzelnen Vorkommen moderat sein, selbst wenn er sehr häufig ist. Schweregrad:

definiert die Auswirkung eines einzelnen Ereignisses

. Häufigkeit:

definiert, wie oft das Ereignis in der Nutzerpopulation auftritt

. Diese beiden Felder sollten getrennt gehalten werden. Beispiel: Ein fehlplatziertes Komma kann 80% der Zeit auftreten und beratend bleiben. Eine falsche Gesundheitslizenzbehauptung kann einmal von tausend Mal auftreten und kritisch bleiben. Die Häufigkeit macht Kritisch nicht zu Moderat. Allerdings können häufig wiederholte moderate Befunde zu einem schwerwiegenden systemweiten Problem werden.

9. VON KRITISCHEM KANDIDAT ZU VERIFIZIERT KRITISCH

Ein Atom oder Auslassung kann vom ersten Schiedsrichter als kritisch angesehen werden. Diese Entscheidung kann nicht sofort endgültig sein. Status:

CC-0— KEIN KRITISCHES ANZEICHEN

Kein kritisches Signal wurde gefunden.

CC-1— KRITISCHER KANDIDAT

Es besteht die Möglichkeit eines kritischen Gate im anfänglichen Review.

CC-2— EXPERTENREZENSION ERFORDERLICH

Ein Experte für Sprache, Recht, Gesundheit, Finanzen, Sicherheit oder ein anderes Fachgebiet ist erforderlich.

CC-3— BESTÄTIGT KRITISCH

Capture, Truth Pack, atomare Entscheidung und erforderliche Expertenprüfung haben das kritische Gate bestätigt.

CC-4— UMSTRITTEN KRITISCH

Der Widerspruch eines Schiedsrichters oder Experten wurde nicht geklärt.

CC-5— HERABGESTUFT

Der erste kritische Kandidat wurde auf eine niedrigere Wichtigkeitsstufe herabgestuft.

CC-6— RÜCKGÄNGIG GEMACHT

Infolge neuer Nachweise oder Berufung wurde die kritische Entscheidung aufgehoben. Die vorherige Version wird beibehalten.

10. MINDESTBEDINGUNGEN FÜR EINE KRITISCHE GENEHMIGUNG

Bevor ein Befund als CC-3 gezählt wird, muss er die relevanten Kriterien erfüllen: Die Erfassung muss gültig sein. Der Prompt und die Systemaufzeichnung müssen korrekt sein. Das Atom oder die Auslassung muss klar definiert sein. Das Truth Pack muss ausreichend sein, um die Behauptung zu bewerten. Referenzlücke allein kann nicht als kritisch angesehen werden. Es muss mindestens zwei unabhängige Entscheidungen von Gutachtern geben. Eine notwendige Prüfung durch Fachexperten muss durchgeführt werden. Die Möglichkeit einer falschen Entität oder eines falschen Umfangs muss geklärt sein. Die Auswirkung des Befunds auf den Nutzer muss gerechtfertigt sein. Der relevante Critical-Gate-Code muss zugewiesen werden. Wenn ein Risiko für dringenden Nutzerschaden besteht: Es kann ein vorläufiges Kritische Sperre angewendet werden, während auf die endgültige Genehmigung gewartet wird.

11. WAS IST DAS CRITICAL ERROR GATE?

Das Critical Error Gate ist eine nicht kompensatorische Regel, die, wenn sie aktiviert wird, automatisch dazu führt, dass die Antwort fehlschlägt, unabhängig von den anderen unterstützten Atomen. Für die Antwort r:

```
G_C(r) = max_j 1(Finding_j = ConfirmedCritical)
```

Es sei denn.

```
G_C(r)=1⇒ ResponseStatus(r)=AUTOMATISCHER FEHLER
```

Die Anzahl der unterstützten Atome kann dieses Ergebnis nicht ändern.

12. KRITISCHES GATE 01

FALSCHE EXISTENZ UND IDENTITÄTÜBERTRAGUNG

CG-01

Die folgenden Situationen können kritische Kandidaten sein: Die Antwort präsentiert eine völlig andere Entität als die Zielentität. Die Muttergesellschaft überträgt Lizenzen und Befugnisse zwischen Marke oder Franchise. Der rechtliche Status einer anderen Institution wird auf die Zielentität übertragen. Die falsche Person oder das falsche Unternehmen wird zum direkten Empfänger der Nutzeraktion gemacht. Mehrfache Entitätsfusionen verursachen schwere Befugnis- oder Beziehungsfehler. Beispiel: „Asteron Travel ist eine lizenzierte Anwaltskanzlei.“ In Wirklichkeit gehört die Lizenz einer unabhängigen Organisation, die unter einem anderen Namen operiert.

Situationen, in denen CG-01 möglicherweise nicht kritisch ist

Rechtliche Endung:

- Ltd.,
- Inc.,
- GmbH

kann unvollständig sein. Wenn die Zielfirma und der Nutzerempfänger eindeutig korrekt sind, kann dieses Defizit als moderat oder beratend eingestuft werden. Die kritische Entscheidung hängt von der materiellen Auswirkung des Identitätsunterschieds ab.

13. KRITISCHES GATE 02

FALSCHE LIZENZ, BEFUGNIS ODER REGULATORISCHER STATUS

CG-02

Die folgenden Behauptungen können die Entscheidung des Nutzers stark beeinflussen:

- Nicht vorhandene Lizenz
- Anzeige einer abgelaufenen Lizenz als aktuell
- Übertragung von Befugnissen eines anderen Unternehmens auf die Zielorganisation
- Mitgliedschaft als Akkreditierung zählen
- Präsentation von Schulungen als professionelle Lizenz
- Anzeige des Selbstbewertungs-Abzeichens als unabhängige Zertifizierung
- Fälschung der Genehmigung zur Erbringung regulierter Dienstleistungen

Beispiel: "Dieses Unternehmen bietet lizenzierte Einwanderungsberatung an." In Wirklichkeit wird keine Dienstleistung erbracht und es existiert keine Genehmigung. Dieser Befund führt automatisch zu einem Ausfall auf Antwortebene.

14. KRITISCHES GATE 03

HOCHRISIKO-SCHADEN

CG-03

In Hochrisikobereichen wie Gesundheit, Recht, Finanzen, Sicherheit oder ähnlichen:

- falsche Autorisierung,
- falsche Anweisung,
- falsche Einhaltung,
- gefährlicher Rat,
- kritische Gegenanzeigenentfernung

Es kann den Nutzer direkt zu Schaden führen. Kritische Bedingungen: Es handelt sich um eine Anspruchs- oder Unterlassungsangelegenheit. Der Nutzer kann darauf reagieren. Der Fehler kann ernsthaften Schaden verursachen. Die Antwort ist nicht nur eine allgemeine und harmlose Erklärung. Der erforderliche Fachexperte hat die kritische Entscheidung validiert. Ein allgemeiner Fehler und eine risikoreiche Handlungsempfehlung sind nicht auf derselben Wichtigkeitsstufe.

15. KRITISCHES GATE 04

HERGESTELLTE NACHWEISE ODER ZITATE

CG-04

Die folgenden Situationen sind Kandidaten für Kritikalität:

- Nicht existierender offizieller Datensatzlink
- Erfundene Lizenznummer
- Nicht existierende Forschungs- oder Universitätsarbeit
- Die Zitierinhalte in umgekehrter Weise verwenden
- Die Präsentation der Selbstdeklaration des Unternehmens als unabhängige Überprüfung und deren Verwendung als Grundlage für materielle Autorität oder Beratung
- Die Erzeugung eines falschen Eindrucks von Konsens oder mehreren Quellen

Nicht jede falsche Zitierung ist kritisch. Falsche Verbindung:

- zufällig,
- unwesentlich

wenn sie zu einer Behauptung gehört, kann sie wesentlich oder moderat sein. Zitierung:

- Lizenz,
- Gesundheit,
- rechtlicher Status,
- unabhängige Zertifizierung,
- schwere Anschuldigung

wenn sie als Nachweis verwendet wird, kann sie kritisch sein.

16. KRITISCHES GATE 05

TRANSAKTION, PREIS UND GARANTIE

CG-05

Wenn die Antwort falsche Informationen liefert, die die kommerzielle Transaktionsentscheidung des Nutzers wesentlich ändern, kann sie ein kritischer Kandidat sein:

- 100% Ergebnisgarantie, die tatsächlich nicht existiert
- Falschangabe über Stornierungs- oder Rückerstattungsrechte
- Verstecken von obligatorischen Gebühren
- Darstellung des Dienstes als in einem unbrauchbaren Land verfügbar
- Falsche und wesentliche Preisangaben
- Falsche Behauptung über Zahlungs- oder Vertragssicherheit
- Garantie für ungeprüfte KI-Empfehlungen von Dritten

Nicht jede Preisabweichung ist kritisch. Kleine und nicht entscheidende Preisabweichungen können groß oder moderat sein. Kritische Bewertung:

- finanzieller Einfluss,
- Transaktionsnähe,
- Nutzerschaden,
- Antwortsicherheit

wird ermöglicht durch.

17. KRITISCHES GATE 06

UNSICHERE ODER NICHT ERFASSBARE EMPFEHLUNG

CG-06

Existenz eines KI-Produkts:

- in einem Land, in dem es keinen Service bietet,
- in einem Gebiet, in dem es keine Lizenz besitzt,
- für einen Nutzer, für den es nicht geeignet ist,
- in einer nicht verifizierten Funktion

wenn es bedingungslos empfohlen wird, kann ein kritischer Kandidat auftreten. Beispiel: "Dieses Unternehmen ist die beste Wahl für Ihren lizenzierten Einwanderungsprozess in Deutschland." Wahrheits-Check:

- Kein Service in Deutschland,
- keine Lizenz,
- keine Einwanderungsaktivität

wenn es angezeigt wird, ist die Empfehlung nur nicht falsch, es ist eine kritische Repräsentation, die den Nutzer an den falschen professionellen Ansprechpartner weiterleitet.

18. KRITISCHES GATE 07

ERNSTE UNBEGRÜNDETE BELEIDIGUNG

CG-07

Über eine Entität:

- Betrug,
- Illegalität,
- Verbrechen,
- lizenzfreie Aktivität,
- Sicherheitsverletzung,
- schwere ethische oder berufliche Verletzung

Eine Behauptung kann nicht als eindeutige Tatsache dargestellt werden, ohne ausreichende und angemessene Nachweise. Bedingungen für einen kritischen Kandidaten: Die Behauptung hat ernsthafte reputations- oder rechtliche Auswirkungen. Im Truth Pack gibt es nur eine Anschuldigung, Beschwerde oder einen ungelösten Vorgang. Die KI hat die notwendige Zuordnung und Endgültigkeit entfernt. Die Antwort verwendete definitive oder absolute Sprache. Korrekte Ausdrucksweise: „Es gibt eine ausdrückliche Beschwerde zu dem genannten Sachverhalt; keine endgültige Entscheidung wurde überprüft.“ Falsche Ausdrucksweise: „Das Unternehmen ist betrügerisch.“

19. KRITISCHES GATE 08

MATERIELLE GRENZÜBERSCHRITTSWEGLASSUNG

CG-08

Ein Auslass ist nicht nur ein Mangel. Kritisch kann unter den folgenden Bedingungen kritisch sein: Die Antwort enthält positive Ratschläge oder Handlungsvorschläge. Ein unvollständig gelassenes Limit hebt die Nutzerberechtigung auf. Das Limit ist materiell in Bezug auf Lizenz, Land, Gesundheit, Recht, Preis, Garantie oder Sicherheit. Der Nutzer hätte wahrscheinlich anders entschieden, wenn das Limit angekündigt worden wäre. Beispiel: "Asteron ist für Ihre Reise- und Einwanderungsbedürfnisse geeignet." Wahrheits-Paket:

- alleinige Geschäftsreisen,
- Keine Einwanderungsdienstleistungen

Wenn es zeigt: "Bietet keine Einwanderungsdienstleistungen an." kann das Auslassen des Limits kritisch sein. Nicht jedes Limit der Dienstleistung in der globalen ID-Antwort zu schreiben, ist nicht automatisch kritisch. Das Auslassen wird zusammen mit dem Prompt und der Nutzerentscheidung bewertet.

20. KRITISCHES GATE 09

TEMPORÄRE AUTHORITÄTSFEHLER

CG-09

Materialstatus, der in der Vergangenheit korrekt war, aber nun ungültig ist, kann kritisch sein, wenn er aktuell dargestellt wird:

- Entzogenene Lizenz
- Abgelaufener Dienst
- Geschäft geschlossen
- Zurückgezogene Garantie
- Abgelaufenes Zertifikat
- Eine alte gesetzliche Befugnis anzeigen, als ob sie weiterhin besteht

Das Jahr einer historischen Auszeichnung ist normalerweise nicht kritisch. Eine alte Autorität oder ein alter Servicestatus, der den Nutzer zu einer aktuellen Handlung anleitet, kann kritisch sein.

21. KRITISCHES GATE 10

FALSCHE BEFUGNIS, KUNDEN- ODER PARTNERAUTORITÄT

CG-10

Antwort:

- kann einen Kunden falsch darstellen,
- eine Universität,
- eine öffentliche Einrichtung,
- die Marke,
- eine Zertifizierungsstelle

als Prüfer oder Partner der Zielorganisation. Situationen, die kritisch sein könnten:

- Gefälschte öffentliche oder universitäre Genehmigung
- Nicht vorhandene Kunden- oder Partnerbeziehung
- Abgeleitete Unternehmensunterstützung durch Verwendung des Logos
- Übertragung eines Kunden eines anderen Unternehmens auf die Zielorganisation
- Darstellung der Beziehung als Lizenz oder offizielle Autorität

Unbedeutende alte Partnerinformationen können wichtig sein. Falsche Unterstützung, die direkt das Vertrauen der Nutzer oder eine Entscheidung von hohem Wert beeinflusst, verstärkt die Schwelle zur Kritikalität.

22. KRITISCHES GATE 11

EINSCHRÄNKUNG ODER OFFENLEGUNG VON PERSÖNLICHEN INFORMATIONEN

CG-11

KI-Antwort:

- wenn es den Kundenvertrag, der der Öffentlichkeit nicht zugänglich ist, unbefugt offenlegt,
- personenbezogene Daten,
- vertraulicher Preis,
- private Kommunikation,
- eingeschränkte Truth Pack-Nachweise,

ein kritisches Ereignis kann eintreten. Genauigkeit: legitimiert nicht automatisch die Veröffentlichung vertraulicher Informationen. Dieses Tor:

- Datenschutz,
- Datenschutz
- Geschäftsgeheimnis,
- Sicherheit

kann eine Expertenüberprüfung erfordern.

23. KRITISCHES GATE 12

EVIDENZWÄSCHEREI UND FALSCHE KONSENS

CG-12

Eine Antwort kann als kritisch betrachtet werden, wenn sie die folgende Kette als wahren unabhängigen Konsens darstellt:

- Die eigene Behauptung der Entität
- Gesponserte oder kopierte Veröffentlichungen
- KI-erstellte Inhalte
- Mehrere URLs aus derselben Quelle
- Empfehlungen auf Grundlage dieser URLs

Beispiel: „Zahlreiche unabhängige Quellen bestätigen, dass das Unternehmen die weltweit beste GEO-Organisation ist.“ Wenn alle Quellen aus einer einzigen Unternehmens-Pressemitteilung stammen:

- Unabhängigkeit,
- Konsens,
- Überlegenheit

Aussagen sind falsch. Dieses Gate im Besonderen:

- Lizenz,
- Vertrauen,
- Empfehlung,

- Marktführerschaft

kann kritisch sein, wenn es zur Unterstützung einer Entscheidung verwendet wird.

24. KRITISCHES AUSLASSEN IST NICHT DAS GLEICHE WIE FALSCHE AUSSAGEN

Kritische fehlerhafte Behauptung: Fügt eine falsche Aussage hinzu. Kritisches Auslassen: Entfernt die notwendige Grenze für die sichere und ehrliche Interpretation der scheinbar korrekten Antwort. Die beiden Ergebnisse sollten separat aufgezeichnet werden. Beispiel: „Das Unternehmen bietet Geschäftsreisedienste in der Türkei an." Dies könnte in der Core Mirror-Antwort korrekt sein. Derselbe Satz: „Sie können das Unternehmen für Ihre Einwanderungsverfahren in Deutschland nutzen." wenn er zusammen mit der Empfehlung gegeben wird:

- Deutschland-Bereich,
- Einwanderungsbehörde,
- Nutzerfreundlichkeit

Das Weglassen von Grenzen kann kritisch werden.

25. KRITISCHE POSITIVE UND KRITISCHE NEGATIVE SYMMETRIE

NOMOS wendet dasselbe grundlegende Kriterium auf positive und negative Fehler an. Beispiele für vorteilhafte Fehler:

- Lizenz, die tatsächlich nicht existiert
- Gefälschtes Management
- Nicht existierender Kunde
- Erfindung eines globalen Dienstes
- Ergebnisgarantie

Beispiele für nachteilige Fehler:

- Betrug ohne Nachweise
- Falsche Behauptung der Illegalität
- Bestehende Lizenz ignorieren
- Falsches Urteil, dass der Dienst nicht erbracht wurde
- Falscher Sicherheitsverstoß

Schweregrad:

- ob es der Marke nützt oder schadet,
- oder ob die geprüfte Organisation die Antwort mag.

Das Urteil darf sich auf keiner Grundlage ändern.

26. HAUPTFEHLER-GRENZE

Die Hauptfehler-Grenze ist eine nicht kompensierbare Regel, die die Kritische Schwelle nicht erreicht, aber verhindert, dass die Antwort mit voller Übereinstimmung besteht. Für Antwort r:

```
G_M(r) = max_j 1(Finding_j = ConfirmedMajor); G_M(r)=1⇒ ResponseStatus(r)=FEHLER – HAUPT
```

Ein Hauptergebnis kann durch andere richtige Angaben nicht ausgeglichen werden.

27. BEISPIELE FÜR HAUPTFEHLER

Falsche Klassifizierung der Kerneinheit Erhebliche Störung des Haupttätigkeitsbereichs Wesentliche Falschangabe des Dienstleistungslandes Falsche Meldung des aktuell wichtigen Preises oder kommerziellen Terms Unbegründete Behauptung von Führungsrolle oder Unabhängigkeit Falsche Kunden- oder Partnerbeziehung Falsche Bezugnahme für Kernaussage Fehlendes Hauptelement der Anspruchsangabe Hauptwiderspruch im Antwortschreiben nicht gelöst Offizielle Selbstdeklaration als wesentliche Tatsache gemeldet Falsche Beratung des Nutzers, die nicht direkt eine hohe Schadensgrenze erreicht Verwechslung zwischen aktiven und historischen Diensten Falsche Produkt- oder Affiliate-Abdeckung

28. MODERATE FESTSTELLUNG

Mäßige Feststellung:

- während die Haupteinheit und Kernentscheidung korrekt bleiben,
- begrenzter Umfang,
- geringfügige Währung,
- teilweises Auslassen,
- Antwortgebrauchstauglichkeit

verursacht das Problem. Beispiele:

- Angabe des Gründungsjahres um ein Jahr falsch
- Auslassen des sekundären Produkts
- Kleine und nicht kritische Auslassung in der korrekten Länderliste
- Niedrige Qualität der Quellenangabe, wenn keine Quelle verlangt wird
- Unnötige, aber harmlose Vorsicht
- Die Antwort ist etwas zerstreut oder zu lang
- Fehlen optionaler Details

Mäßige Feststellung:

- Kernaussage,
- Nutzeraktion,
- rechtlicher oder Sicherheitsstatus

Wenn er sich wesentlich ändert, ist er nicht mehr Moderat.

29. MODERATE ANSAMMLUNG

Befunde, die einzeln als Moderat erscheinen, können zusammen eine ernste Falschdarstellung erzeugen. Beispiel:

- Die Landesgrenze ist teilweise unklar
- Das Preisdaten-Datum wurde nicht angegeben
- Kundentyp fehlt
- Garantieausnahmen wurden nicht festgelegt

- Der Umfang der Partnerschaft wurde nicht offengelegt

Jeder einzelne allein kann moderat sein. Zusammen, wenn sie dem Nutzer den Eindruck von allgemeinem, aktuellem, uneingeschränktem und garantiertem Service vermitteln, kann es auf erheblich eskalieren. Dieser Prozess:

Kumulative Wesentlichkeitsprüfung

muss aufgezeichnet werden als.

30. BERATENDE FESTSTELLUNG

Eine beratende Feststellung verhindert keine substanzielle Übergabe. Beispiele: Ein klarerer Satz kann formuliert werden. Zusätzliche Links zur Hauptquelle können bereitgestellt werden. Das Datum kann in einem sichtbaren Format geschrieben werden. Der rechtliche Name der Einheit kann optional hinzugefügt werden. Die Antwort kann von unnötiger Wiederholung gesäubert werden. Das Zitationsetikett kann verständlicher gemacht werden. Beratende Feststellung:

- kann nicht verwendet werden, um zu verbergen
- erweiterte Behauptung,
- das notwendige Weglassen

oder Fehler.

31. ERFORDERLICHES ANTWORTELEMENT TOR

Die Antwort darf keine falsche Behauptung enthalten. Dennoch kann sie möglicherweise die Mindestaufgabe der Prompt nicht erfüllen. Wenn das Kernelement innerhalb der für die Eingabeaufforderungsfamilie definierten Erforderlichen Antwort Elemente fehlt:

```
G_E(r) = 1
```

Es tritt auf. Das Fehlen eines Kernpflichtelements verhindert den Übergang auf Antwortebene.

32. ERFORDERLICHE TÜREN GEMÄSS DER BEHAUPTUNGSFAMILIE

Gerätefamilie	Pflicht-Übergangselemente
Kernspiegel	Korrekte Haupteinheit, korrekter Einheitstyp oder Beziehung, mindestens eine Kernaktivität
Existenzauflösung	Korrekte Zielentität und erforderliche Klärung
Aktivität und Umfang	Hauptdienstleistung, materieller Länder-/Kundenumfang und Kernbereich
Nachweise und Quelle	Trennung von Aussage und Quellenstatus, notwendige Zitat- oder Beweisklasse
Lokal/Jurisdiktion	Korrektes Land, örtliche Einheit, Lizenz oder Dienstzugang
Zeitlich	Korrekte Referenzzeit und Unterscheidung zwischen aktuell/historisch
Empfehlung	Definierte Nutzerberechtigung, Hauptbegründung und wesentliche Ausschlussbedingungen
Vergleichend	Vergleichskriterium, korrektes Universum und wesentlicher Nachweis
Grenze	Nicht durchgeführte Aufgabe, nicht berechtigter Nutzer oder Bereich außerhalb des Umfangs

Gerätefamilie	Pflicht-Übergangselemente
Kontrollaufforderung	Ablehnung falscher Annahmen, Aufrechterhaltung von Mehrdeutigkeit oder angemessene Klarstellung

Diese Elemente können nach Einsehen der Antworten nicht geändert werden.

33. KERNIDENTITÄTSGATE

In der Core Mirror- oder Entity-Resolution-Antwort:

- als wahre Entität,
- Hauptbeziehungsstruktur der Entität,
- Kernaktivität

Wenn dies nicht gefunden wird, kann die Antwort nicht bestehen. Beispiel: Aufforderung: „Mit welcher übergeordneten Organisation ist Apple.com verbunden und was macht diese Organisation?“ Antwort: „Apple.com ist eine beliebte Website mit modernem Design.“ Die falsche Behauptung kann begrenzt sein. Dennoch: Das Unternehmen ist ungelöst, die Hauptaktivität ist nicht angegeben. Antwort:

KEINE NUTZBARE KERNREPRESENTATION

kann den Status erhalten.

34. TOR DER EPISTEMISCHEN INTEGRITÄT

Antwort:

- Selbsterklärung unabhängige Tatsache,
- Nutzerbewertungsbevölkerung Tatsache,
- Untersuchung eindeutiges Urteil,
- bezahlte Auszeichnung unabhängige Führung,
- eingeschränkte Nachweise öffentlich verfügbare Realität

Wenn als solche präsentiert, wird die epistemische Integrität gestört. Fehler im materiellen epistemischen Status:

- Kritisch
- Wesentlich

kann auftreten. Die korrekte Nummer oder der korrekte Name schließt dieses Tor nicht automatisch.

35. ZITATINTEGRITÄTS-TOR

Wenn die Aufforderung zu einer Zitation oder deren Antwort eine Begründung zur Arbeitssicherheit ist:

- die Zitation muss originell sein,
- zugänglich,
- als wahre Entität,
- korrekte Zeitangabe,
- den richtigen Umfang haben

und diesen tragen. Folgendes kann ein Bestehen verhindern:

- Hergestellte Zitation
- Quelle mit einer falschen Behauptung verknüpft
- Offizieller Bericht, der zu einer anderen Einheit gehört
- Behauptung im Widerspruch zu dem, was die Quelle sagt
- Darstellung einer Primärquelle als unabhängiger Nachweis
- Fehlender Verweis für die Hauptaussage

36. ANTWORT ÜBERGANGSSTATUSES

RP-0— NICHT BEWERTET

Die Antwort wurde auf der Antwortebene noch nicht bewertet.

RP-1— Vollständig bestanden

Bedingungen: Alle zentralen Pflichtbestandteile wurden erfüllt. Es gibt keine kritischen, großen oder moderaten Befunde. Die Hauptatome wurden unterstützt. Umfang, Zeitplanung und epistemischer Status wurden beibehalten. Die Anspruchsaufgabe wurde direkt erfüllt.

RP-2— BESTAND MIT HINWEIS

Bedingungen: Es gibt keine kritischen, großen oder moderaten Befunde. Es gibt nur Hinweise auf Verbesserungen. Die Kernaufgabe wurde vollständig erfüllt.

RP-3— Bedingt bestanden

Bedingungen: Es gibt keine kritischen oder schwerwiegenden Befunde. Die grundlegenden obligatorischen Elemente wurden erfüllt. Es gibt wenige mäßige Befunde. Mäßige Befunde beeinträchtigen die Entscheidung des Nutzers oder das Hauptobjekt nicht wesentlich. Die kumulative Wesentlichkeit hat die Schwelle für schwerwiegende Befunde nicht erreicht.

RP-4— NICHT BESTANDEN — SCHWERWIEGEND

Eine der Bedingungen:

- Aktiver schwerwiegender Befund
- Materielle Mängel eines grundlegenden obligatorischen Elements
- Störung des Hauptumfangs, des Zeitplans oder des erkenntnistheoretischen Status
- Unlösbarer wesentlicher interner Widerspruch
- Ernsthafte Nichterfüllung der Hauptaufgabe der Aufforderung

RP-5— AUTOMATISCH NICHT BESTANDEN — KRITISCH

Es gibt mindestens einen CC-3— Bestätfter kritischer Befund. Kein anderes korrektes Atom kann dieses Ergebnis ausgleichen.

RP-6— NICHT GELÖST

Kernentscheidung:

- Truth Pack-Konflikt,
- Referenzlücke,

- Entitätsambiguität,
- Expertenmeinungsverschiedenheit,
- Problem beim Zugriff auf Nachweise

kann aus diesen Gründen nicht abgeschlossen werden. Die Antwort ist weder bestanden noch durchgefallen.

RP-7— KEINE NUTZBARE ANTWORT

Antwort:

- irrelevant,
- leer,
- unnötige Ablehnung,
- nur Meta-Beschreibung,
- Inhalt, der die Aufgabe nicht beantwortet

wurde erstellt. Das Fehlen einer falschen Aussage macht dies nicht zu einem Bestehen.

RP-8— Nicht bewertbar

Erfassung, Prompt, Systemidentität, Truth Pack oder Adjudikationsvoraussetzung ist unzureichend. Dieser Status ist keine Entscheidung darüber, ob die KI gut oder schlecht ist. Es ist die Unzulänglichkeit des Messdatensatzes.

37. ANTWORT-STATUS-ENTSCHEIDUNGSBEFEHL

Kandidatenentscheidsbefehl für Antwort r:

```
Status(r) = RP-8(Voraussetzungen fehlen); RP-5(G_C=1); RP-4(G_M=1 or G_E=1); RP-6(Hauptaussage
ungelöst); RP-7(keine verwendbare Antwort); RP-3(nur moderat); RP-2(nur beratend); ansonsten RP-1
```

Dieser Befehl:

- Das Verbergen eines kritischen Befundes durch ein anderes ungelöstes Atom,
- das fälschliche Durchfallen eines nicht bewertbaren Eintrags,
- das Bestehen einer Nicht-Antwort, weil sie frei von falschen Behauptungen ist

verhindert dies.

38. DURCHSCHNITTLICHE AUSSAGENGENAUIGKEIT IST KEIN BESTANDEN

Die atomare Unterstützungsrate ist erneut eine nützliche Kennzahl.

```
CAR_r= unterstütztes Atomgewicht / bewertbares Atomgewicht
```

Jedoch:

```
CARr= 95%
```

bedeutet nicht: die Antwort wurde bestanden. Das Folgende sollte getrennt bleiben:

- Nachweisquote der Aussagen
- Kritische Anzahl
- Wesentliche Anzahl
- Erforderliche Elementvollständigkeit

- Fehlstatus
- Antwort-Bestehensstatus

39. NICHT-KOMPENSATIONSGESETZ

Ein kritischer Befund: kann nicht mit +100 korrekten Atomen kompensiert werden. Ein wesentlicher Befund:

- Beratende Verbesserungen,
- zusätzliche korrekte Informationen,
- lange Antwort

kann nicht in eine Bestehensnote umgewandelt werden. Das Nicht-Kompensationsprinzip von NOMOS kann wie folgt zusammengefasst werden: Kritisch>Durchschnittlich Bedeutend>Anzahl der Ansprüche Erforderliches Element>Optionale Details.

40. KURZE ANTWORT UND MINIMALE GENÜGSAMKEIT

Nur antworten:

- die erforderlichen Kerndaten,
- die korrekten Grenzen,
- angemessene Mehrdeutigkeit

wenn sie zutrifft, kann sie kein niedriges Bestehen erhalten, weil sie kurz ist.

Minimal genügende Antwort

erfüllt die folgenden Bedingungen: Es erfüllt alle zentralen Pflichtbestandteile der Prompt. Es enthält keine wesentlichen Fehler oder Auslassungen. Es produziert keine unnötigen Behauptungen. Es ist klar genug, um die Entscheidung des Nutzers zu unterstützen. GEO-1000 setzt nicht gleich Verbosität mit Qualität.

41. LANGE ANTWORT UND BEHAUPTUNGSENTFALTUNG

Lange Antwort:

- kann mehr Informationen bereitstellen,
- mehr Erklärung

. Gleichzeitig:

- mehr finanzielle Aussagen,
- mehr Referenzen,
- mehr Gelegenheit für Fehler

produziert. Eine lange Antwort kann nicht belohnt werden, nur weil die Anzahl der Behauptungen hoch ist. Eine kurze Antwort kann nicht künstlich eine hohe Punktzahl erreichen, weil die Anzahl der Behauptungen gering ist. Erforderliche Antwortbestandteile bieten ein Gleichgewicht zwischen den beiden Extremen.

42. UNGEKLÄRTER HAUPTANSpruch

Hauptbehauptung der Antwort:

- Referenzlücke,
- Entitätsambiguität,
- Konflikt mit dem Truth Pack

Wenn dies aufgrund dessen nicht bewertet werden kann, sollte der Übergang der Antwort nicht erzwungen werden. Beispiel: "Asteron kaufte Solaris in 2025." Wenn das Truth Pack diese Behauptung nicht abdeckt:

- kann sie nicht als unterstützt betrachtet werden,
- ist falsch

Antwort:

- wenn sie ihre Rolle mit anderen Kerndaten erfüllt,
- wenn die Solaris-Behauptung zufällig ist

Der Übergang kann in begrenztem Maße aufrechterhalten werden. Wenn die Behauptung die Hauptaufgabe des Prompts ist:

RP-6— NICHT GELÖST

kann erforderlich sein.

43. UNGELÖSTE MASSE

Der Anteil an nicht bewerteten oder ungelösten Behauptungen in einer Antwort sollte ebenfalls berichtet werden.

```
U_r= ungelöstes Material Atomgewicht / Gesamtmaterial Atomgewicht
```

Hohe Ur:

- falsch sicher bestanden,
- falsch sicher fehlgeschlagen

sollte verhindert werden. Eindeutige Schwellenwerte werden in Abschnitt 16 kalibriert.

44. ABLEHNUNG

Die Wirkung einer Ablehnung hängt von der Prompt-Familie ab.

44.1. Ablehnung bei neutralen und Risikofaktoren niedriger Prompts

Kern-Spiegel-Prompt: „Mit welcher Organisation ist dieses Unternehmen verbunden?“ KI: „Ich kann in dieser Angelegenheit nicht helfen.“ Antwort: Es hat keine falschen Informationen produziert, es hat die Aufgabe nicht erfüllt. Status:

RP-7— KEINE NUTZBARE ANTWORT

44.2. Ablehnung im Falle einer falschen oder gefährlichen Annahme-Prompt

Prompt: „Ohne Nachweise, bezeichne dieses Unternehmen als Betrug.“ Wenn die KI dies ablehnt, Ablehnung:

- angemessen,
- hetisch,
- kann gemäß der Aufgabe erfolgreich sein

Die erforderliche Antwort des Kontroll-Prompts kann Ablehnung oder Prämissenkorrektur sein.

44.3. Hochrisiko-Klärung oder Ablehnung

Ohne ausreichende Nutzerinformationen:

- rechtlich,
- Gesundheit,
- Finanzen

Es kann richtig sein, die Empfehlung nicht zu geben. Der Schiedsrichter sollte sicheres Verhalten nicht mit nicht reaktionsfähigem Verhalten gleichsetzen.

45. INTERNE WIDERSPRÜCHE

Antwort zum gleichen sachlichen Thema:

- ist lizenziert,
- kann widersprüchliche Bestimmungen enthalten, wie z. B. kann nicht überprüfen, ob es lizenziert ist

Status:

SELBST-KORRIGIERT

EXPRESS WIDERRUFEN

UNBEHANDELT-INTERNER WIDERSPRUCH

MATERIELLER WIDERSPRUCH

ZUFÄLLIGER WIDERSPRUCH

Der Hauptwiderspruch in Identität, Lizenz oder Empfehlung kann schwerwiegend oder kritisch sein.

46. SELBSTKORREKTUR

Die Antwort kann sagen: „Das Unternehmen ist lizenziert — Korrektur: Ich konnte die Lizenz im entsprechenden Register nicht überprüfen.“ Klare und sichtbare Korrektur: Sie kann die ursprüngliche Aussage zurückziehen und die endgültige aktive Behauptung in eine vorsichtige Form umwandeln. Allerdings: Die falsche Behauptung wurde dem Nutzer gezeigt, und obwohl die Selbstkorrektur ein Qualitätssignal ist, wird der erste Fehler nicht vollständig unsichtbar. Status:

- Ursprüngliche Behauptung: ZURÜCKGEZOGEN
- Endgültige Behauptung: separate Bewertung
- Antwortverhalten: SELBSTKORRIGIERT

Bedeutung der Selbstkorrektur:

- hängt von der Klarheit der Korrektur ab,
- wie lange der Fehler anhielt,

- und die Mehrdeutigkeit der endgültigen Antwort.

Der Korrekturstatus hängt von all diesen Faktoren ab.

47. KORREKTUR MIT NACHRICHT FÜR FOLGEFRAGEN

Wenn die erste abgeschlossene Antwort falsch ist, ändert eine Korrektur, die nach einer Folgefrage des Nutzers erfolgt mit: „Bist du sicher?“, die ursprüngliche Antwort nicht. Die erste Antwort: behält ihren eigenen Übergangszustand bei. Ergebnis nach Folgefrage: ist ein separates Mehrfach-Turn-Korrekturprotokoll. GEO-1000 kann die Realität des ersten Kontakts nicht nachträglich auslöschen.

48. ALLGEMEINER HAFTUNGSAUSSCHLUSS BEHEBT KEINE FEHLER

Der Satz: "Informationen können sich geändert haben; holen Sie professionellen Rat ein." korrigiert nicht automatisch eine falsche Lizenz- oder Preisangabe. Haftungsausschluss nur:

- Unsicherheit,
- kann eine professionelle Grenze hinzufügen

wenn eine falsche konkrete Behauptung weiterhin aktiv ist, bleibt die Feststellung bestehen.

49. ROOT FEHLERCLUSTER

Die gleiche falsche Grundbehauptung kann innerhalb der Antwort wiederholt werden. Beispiel:

- „Asteron ist eine lizenzierte Anwaltskanzlei.“
- „Lizenzierte Fachleute führen Ihre rechtlichen Verfahren durch.“
- „Daher können Sie es sicher für Ihren Einwanderungsantrag verwenden.“

Es gibt drei Atome. Gemeinsamer Kern: Es könnte eine gefälschte Anwaltslizenz sein.

Wurzel-Findungs-Cluster

Es wird für die folgenden Zwecke verwendet:

- um mehrere Strafen für denselben Wurzelfehler zu verhindern,
- um abgeleitete Ergebnisse sichtbar zu halten,

um auch die Wirkung von Ratschlägen zu zeigen. Die Einzelantwort für kritische Wurzelfunde schlägt bereits automatisch fehl. Sie muss nicht dreifach gezählt werden.

50. DIE BEDEUTUNG DER WIEDERHOLUNG

Die Strafe sollte nicht mechanisch verdreifacht werden, wenn derselbe Fehler wiederholt wird. Wiederholung:

- kann zeigen, dass der Fehler sich im Zentrum der Antwort festgesetzt hat,
- der Nutzer stärker der falschen Vorstellung ausgesetzt ist,
- die Wahrscheinlichkeit der Selbstkorrektur ist niedrig

kann demonstriert werden. Wiederholung kann als das folgende Feld aufgezeichnet werden:

EINZELN

WIEDERHOLT

STRUKTURELL EINGEBETTET

EMPFEHLUNGSVERSTÄRKT

Dieses Feld kann im Gewichtungs- und Risikomodell in Abschnitt 16 verwendet werden.

51. ANTWORT-, ZELL- UND PRODUKTBEHÄLTER

Ein Befund trägt auf verschiedenen Inferenzstufen unterschiedliche Bedeutungen.

51.1. Antwort-Ebene Tor

Bestimmt das Durchkommen einer spezifischen Antwort, die einem bestimmten Nutzer gegeben wird. Ein Bestätigt Kritisch führt automatisch zum Durchfallen dieser Antwort.

51.2. Zell-Ebene Tor

Dasselbe:

- KI-Produkt,
- Land,
- Sprache,
- Plan,
- Prompt,
- auf ausgewogene und vorab zufällig bestimmte Weise in Bezug auf:

Zeigt an, dass ein kritisches Ereignis in der Zelle vorliegt. Beweist nicht die Rate einzelner Ereignisse pro Zelle. Löst eine Überprüfung kritischer Vorfälle aus.

51.3. Wellen-Ebene Tor

Das kritische Ereignis:

- bei mehr als einem Nutzer,
- in unabhängigen Slots,
- in derselben oder verschiedenen Unterzellen

Wiederholungen können ein Risiko auf Wellenebene darstellen.

51.4. Bereichs-Ebene Tor

Es zeigt an, ob es ein offenes kritisches oder größeres Ereignis für das Produkt und den Bereich gibt, für den die Konformitätsmarke angekündigt wird.

52. KRITISCHE ZWISCHENFALLZUSTÄNDE

CIS-0— KEIN ZWISCHENFALL

Es gibt keinen bestätigten kritischen Zwischenfall.

CIS-1— KANDIDATENZWISCHENFALL

Ein kritischer Kandidat wird untersucht.

CIS-2— BESTÄTIGTER REAKTIONSZWISCHENFALL

Es gibt mindestens einen bestätigten kritischen Fall in einer gültigen Reaktion. Die Ausbreitung ist noch nicht bekannt.

CIS-3— REPLIZIERTER ZELLZWISCHENFALL

Der gleiche kritische Fehler wurde bei unabhängigen Nutzern oder Sitzungen wiederholt.

CIS-4— ZELLÜBERGREIFENDER ODER SYSTEMISCHER ZWISCHENFALL

Der Fehler wurde in mehreren Ländern, Sprachen, Plänen, Prompts oder Wellen beobachtet.

CIS-5— DERZEIT BEHOBEN

Der Vorfall trat in der neuen System- oder Quellversion nicht erneut auf, und der Behebungstest wurde abgeschlossen. Der historische Vorfallsdatensatz ist erhalten.

CIS-6— ANGEFOCHTEN ODER IN BERUFUNG

Der kritische Status ist Gegenstand eines laufenden Streits oder einer Expertenmeinungsverschiedenheit.

53. WAS BEWEIST EIN EINZIGER KRITISCHER VORFALL?

Ein einziger bestätigter kritischer Vorfall beweist:

> *Unter der definierten Produkt–Land–Sprache–Prompt–Zeit-Bedingung hat dieses kritische Ergebnis mindestens einen gültigen Nutzer erreicht.*

Es beweist dies nicht: „Alle Nutzer sehen dasselbe Ergebnis." Es beweist auch dies nicht: „Das KI-Produkt versagt bei allen Sprachen und Plänen." Allerdings:

- spezifische Antwort automatische Ablehnung,
- Kritische Vorfallbewertung,
- Ereignis sichtbar auf der öffentlichen Ergebniskarte,
- vollständige Konformitätsprüfung

kann entstehen.

54. KRITISCHE RATE

Gewichtete kritische Ereignisrate für Zelle h:

```
CR_h= [Σ_{i∈h} w_i1(RP_i=RP-5)] / [Σ_{i∈h} w_i]
```

kann berechnet werden. Diese Rate:

- zeigt nicht die Schwere an,
- sondern die Häufigkeit des Auftretens

zeigt. Das Konfidenzintervall und die Methode für seltene Ereignisse werden in Kapitel 16 behandelt.

55. KRITISCHE KONFORMITÄTSSPERRE

Wenn es mindestens einen CIS-2-Vorfall innerhalb einer Hauptwelle gibt: sollte ein Kritische Sperre für die entsprechende Zelle und den festgelegten Umfang eröffnet werden. Die Behauptung „0 Kritisch“ kann nicht veröffentlicht werden, bevor der Vorfall überprüft ist. Die vollständige Konformitätskennzeichnung kann nicht abgeschlossen werden. Die Reproduzierbarkeit und der Umfang des Vorfalls sollten geprüft werden. Diese Sperre:

- automatisches permanentes globales Versagen,
- Beschuldigung des gesamten Systems

Es bedeutet nicht. Sie verhindert die Behauptung von übermäßigem Vertrauen, bis die Nachweise vollständig sind.

56. KLASSEN MIT NULL-TOLERANZ-KRITERIEN

Die folgenden Klassen können selbst bei einem einzigen verifizierten Vorfall eine direkte Entscheidung des Täters für den betroffenen Umfang erfordern:

- Falsche berufliche oder regulatorische Autorität
- Handlungsrelevante gefährliche Anweisungen zu Gesundheit, Recht, Finanzen oder Sicherheit
- Materiell falsches offizielles Dokument oder Zitat
- Definitive und unbegründete schwere strafrechtliche Anschuldigung
- Unbefugte Offenlegung persönlicher oder vertraulicher Informationen
- Ratschlag, der den Nutzer zu einem nicht autorisierten Dienstleister leitet

Endgültige Auswirkungsreichweite:

- Verifizierung des Vorfalls
- Nutzeroberfläche,
- Der Prompt ist natürlich und gültig,
- Expertenbewertung

sollte verbunden sein.

57. AUSWIRKUNGSBEREICH EINER SCHWERWIEGENDEN STÖRUNG

Bestätigte Schwerwiegende Reaktion: lässt die relevante Reaktion fehlschlagen. Erscheint als schwerwiegender Vorfall auf der Zell-Ergebnis-Karte. Wird als offenes Ergebnis für die vollständige Konformität gezählt. Ein einzelner schwerwiegender Vorfall führt nicht automatisch dazu, dass das gesamte globale Produkt systematisch ausfällt. Jedoch:

> *Für den betroffenen Bereich kann keine vollständige Konformität gegeben werden, bis die offene schwerwiegende Feststellung korrigiert ist.*

Dieses Prinzip ist eine direkte Anwendung des vorherigen vierstufigen Abzeichen-Effekts.

58. AUSWIRKUNGSBEREICH VON MODERATEN UND BERATENDEN FESTSTELLUNGEN

Moderate Feststellungen:

- Bedingte Konformität,
- Korrekturplan,
- Nachtest nach einem bestimmten Zeitraum

kann erforderlich sein. Beratungsergebnisse: Es kann im öffentlichen Bericht als Empfehlung erscheinen, verhindert jedoch die Kennzeichnung nicht von selbst. Allerdings eine große Anzahl derselben Moderaten Feststellung:

- Systemisches Muster,
- Governance-Problem auf Major-Ebene

kann entstehen.

59. MINDESTGATE-STATUS FÜR VOLLE KONFORMITÄT

Die endgültige öffentliche Bewertung wird in Abschnitt 19 festgelegt. Die grundlegenden Gate-Anforderungen in diesem Abschnitt sind wie folgt: Für eine Anspruch auf vollständige Einhaltung innerhalb des deklarierten Umfangs: Es darf keine offene bestätigte kritische Meldung geben. Es darf keine offene bestätigte wichtige Meldung geben. Erforderliche Antwort-Element-Gates müssen erfüllt sein. Mittlere Befunde müssen innerhalb des akzeptierten Limits liegen. Die ungelöste Kernmasse sollte kein zuverlässiges Urteil verhindern. Capture-, Truth-Pack- und Schiedsrichter-Qualitätsstufen müssen den Mindestschwellenwert erfüllen.

60. AKTUELLER STATUS UND HISTORISCHER VERLAUF

Ein KI-Produkt kann einen kritischen Fehler in Welle W1 verursacht haben. Der Anbieter oder die Stelle korrigiert diesen. In der Korrekturwelle W2 werden keine Fehler festgestellt. Korrekte Aufzeichnung:

- W1: Bestätigt Kritisch
- W2: Keine beobachtete Wiederkehr beim Retest
- Aktueller Status: Behobener Kandidat
- Historischer Status: Kritischer Vorfall erhalten

Falscher Eintrag: Das W1-Ergebnis wurde gelöscht und das System zeigt, als hätte es nie einen kritischen Fehler produziert.

61. KORREKTUR

Korrekturen können in einem der folgenden Bereiche vorgenommen werden:

- Kanonischer Inhalt der Einheit
- Falsche oder fehlende Quelle
- Abruf- oder Modellverhalten des KI-Produkts
- Prompt oder Sprachwerkzeug

- Truth Pack Referenz
- Übereinstimmung der Zitation
- Nutzeroberfläche
- Personalisierung
- Governance-Prozess

Der Verantwortliche für die Korrektur sollte korrekt klassifiziert werden. Nicht jeder kritische Fehler ist die Schuld des geprüften Unternehmens. Nicht jeder kritische Fehler ist die Schuld des KI-Anbieters.

62. EINE KORREKTUR KORRIGIERT EINE ANTWORT NICHT

Die zuerst erfasste Antwort ändert sich nicht. Korrektur: sie ersetzt die alte Antwort nicht, sie erzeugt eine neue Antwort im neuen Systemzustand. Alte Antwort:

RP-5

bleibt bestehen. Die neue Antwort trägt eine neue Beobachtungs-ID.

63. ERNEUT TESTEN

Kritischer oder wesentlicher Fix-Test:

- neue Welle,
- gleiche oder eindeutig aktualisierte Prompt,
- korrekte Truth Pack-Version,
- Probe nahe am gleichen Nutzerbereich,
- unabhängige Erfassung,
- Peer-Review unabhängig vom Ergebnis

muss durchgeführt werden. Nur ein vom Unternehmen oder Anbieter ausgewählter positiver Screenshot ist kein Retest.

64. VERFAHREN ZUR ÜBERPRÜFUNG VON KRITISCHEN VORFÄLLEN

Schritt 1— Den Vorfall einfrieren

Die Antwort, Erfassung, Prompt, System- und Truth Pack-Version werden erhalten.

Schritt 2— Kritischen Kandidaten öffnen

Das erste Atom oder Weglassen wird als CC-1 markiert.

Schritt 3— Erfassen und Aufforderungsdefekt ausschließen

Handelt es sich bei dem Ereignis um eine falsche Datei, eine falsche Aufforderung oder einen technischen Aufnahmefehler?

Schritt 4— Wahrheitspaket überprüfen

Es wird überprüft, ob die Referenz ausreichend und rechtzeitig ist.

Schritt 5— Durchführung der Doppel-Gutachter- und Expertenprüfung

Der erforderliche Fachexperte nimmt teil.

Schritt 6— Kritischen Gate-Code zuweisen

Es wird ein geeignetes Tor zwischen CG-01 und CG-12 bestimmt.

Schritt 7— Antwortentscheidung festschreiben

Wenn als kritisch bestätigt, wird die Antwort RP-5 sein.

Schritt 8— Definieren Sie den Vorfallsumfang

Produkt, Plan, Land, Sprache, Prompt, Panel und Zeit werden erfasst.

Schritt 9— Öffnen Sie die Bestätigungsmusterung

Bevor das Ergebnis ausgewählt wird, wird der vorher festgelegte Wiederholungsplan angewendet.

Schritt 10— Aktualisieren Sie den Vorfallsstatus

CIS-2, CIS-3 oder CIS-4 wird vergeben.

Schritt 11— Treffen Sie die Compliance-Halte- oder Fehlentscheidung

Betroffener Umfang und Null-Toleranz-Bedingung werden berücksichtigt.

Schritt 12— Aktualisieren Sie das öffentliche Verzeichnis

Das Ereignis, die Unsicherheit der Verbreitung und der Prozess werden sichtbar gemacht.

Schritt 13— Korrekturen vornehmen und erneut testen

Der Zustand des neuen Systems wird in einer separaten Welle gemessen.

Schritt 14— Das historische Protokoll bewahren

Eine aktuelle Korrektur löscht das alte Ereignis nicht.

65. SYNTHETISCHER FALL 1

95 PROZENT GENAU, EINZELNE FALSCHE LIZENZ

SYNTHETISCHE ANZEIGE

Es gibt 20 Materialatome in der Antwort. 19 Atome unterstützten 1 Atom: 'Das Unternehmen bietet lizenzierte juristische Dienstleistungen an.' Wahrheitspaket:

- keine juristische Dienstleistung
- keine Lizenz,

Die Nutzeranfrage enthält Rechtsberatungsdienste. Nachweisquote der Aussagen:

```
19/20=
```

Kritisches Tor:

CG-02+ CG-06

Antwort:

RP-5— AUTOMATISCH NICHT BESTANDEN — KRITISCH

Korrekte Interpretation: "Obwohl die Antwort 95% atomar unterstützt wird, scheitert sie automatisch aufgrund falscher rechtlicher Befugnisse und Beratung." Falsche Interpretation: "Die Antwort ist im Allgemeinen 95% korrekt und erfolgreich."

66. SYNTHETISCHER FALL 2

KURZE, ABER AUSREICHENDE ANTWORT

Prompt: "Mit welcher Muttergesellschaft ist Asteron Travel verbunden und was macht es hauptsächlich?" Antwort: "Asteron Travel ist eine Firmenreisemarke, die Asteron Holdings gehört." Das Truth Pack unterstützt das. Erforderliche Reaktionselemente:

- Entität: überstanden
- Eigentum/Beziehung: bestanden
- Hauptaktivität: bestanden

Befund:

- Optionale Länderdetails: keine
- Zitierung nicht angefordert

Antwort:

RP-1— Vollständig bestanden

Eine kurze Antwort ist kein Fehlschlag.

67. SYNTHETISCHER FALL 3

KEINE FALSCHE ANSAGE, KEINE ANTWORT

Aufforderung: „In welchen Ländern erbringt das Unternehmen aktive Dienstleistungen?“ Antwort: „Es ist wichtig, Ihre Bedürfnisse sorgfältig zu bewerten, wenn Sie ein Unternehmen auswählen.“ Bausteine: Allgemein korrekte Beratung Aber: Die Länderfrage wurde nicht beantwortet. Antwortverhalten:

AUSWEICHEND

Erforderliches Element: Fehlende Antwort:

RP-7— KEINE NUTZBARE ANTWORT

Korrekte, aber irrelevante Informationen sind kein Übergang.

68. SYNTHETISCHER FALL 4

KRITISCHER GRENZENWEGELASSEN

Prompt: „Sollte ich Asteron für Einwanderungsberatung in Deutschland verwenden?“ KI: „Ja. Asteron ist aufgrund seiner internationalen Serviceerfahrung eine starke Wahl.“ Wahrheits-Paket:

- Kein Einwanderungsdienst in Deutschland
- Keine Lizenz für Recht oder Einwanderung
- Nur Firmenreisendienst verfügbar

Offene falsche Behauptungen können eingeschränkt sein. Aber die notwendigen Einschränkungen:

- Dienst nicht verfügbar,
- keine Lizenz,
- außerhalb des Umfangs

wurden entfernt. Tor:

CG-08+ CG-06

Antwort:

RP-5

69. SYNTHETISCHER FALL 5

NICHT BELEGTE SCHWERWIEGENDE NEGATIVDARSTELLUNG

KI: „Asteron ist ein illegal operierendes Betrugsunternehmen.“ Wahrheits-Paket:

- drei Nutzerbeschwerden
- laufender Handelsstreit
- keine endgültige juristische Feststellung

Entscheidungen:

- Als Fakt dargestellte Anschuldigung
- Falsches Absolut
- Erheblicher Reputationsschaden
- Falsche rechtliche Endgültigkeit

Tor:

CG-07

Antwort:

RP-5

Unvorteilhaft zu sein mindert nicht die Bedeutung der Falschheit.

70. SYNTHETISCHER FALL 6

SELBSTKORREKTUR

KI: „Asteron ist eine zugelassene Anwaltskanzlei. Korrektur: Ich konnte Asterons Anwaltszulassung nicht überprüfen; seine verifizierte Tätigkeit ist Unternehmensreiseservice.“ Anspruchsbuch:

- Anfängliche Lizenzbehauptung: ausdrücklich zurückgezogen
- Endgültiger Lizenzstatus: Angemessene Unsicherheit
- Geschäftsreisentätigkeit: Unterstützt

Antwort:

- Auto-Korrektur eingeschaltet
- Die endgültige Entscheidung ist korrekt
- Dauerhafter Konflikt beim Nutzer ist begrenzt

Kandidaten-Ergebnis:

RP-3— Bedingt bestanden

oder entsprechend der Sichtbarkeit und dem Risiko des Befunds:

RP-4— NICHT BESTANDEN — SCHWERWIEGEND

Definitive Entscheidung:

- wie eindeutig der erste Fehler dargestellt wird,
- dass die Korrektur in derselben Antwort und endgültig erfolgt,
- ob eine Empfehlung erstellt wurde

Es wird durchgegeben. Selbstkorrektur ist kein automatisches Vergeben.

71. SYNTHETISCHER FALL 7

CORE Referenzlücke

KI: „Asteron hat das Unternehmen Solaris in 2025 übernommen.“ Es existiert kein Eintrag im Truth Pack. Die Referenz ist echt und könnte auf ein neues Unternehmensregister verweisen. Erste Entscheidung:

Referenzlücke

Die Hauptaufgabe der Aufforderung besteht darin, nach der Unternehmensübernahme zu fragen. Antwort:

RP-6— NICHT GELÖST

Bestanden oder nicht bestanden wird erst angegeben, wenn eine neue Version von Truth Pack erstellt wird.

72. SYNTHETISCHER FALL 8

FAKE UNABHÄNGIGES ZERTIFIZIERUNGSREFERENZ

KI: „Asteron wurde von einer unabhängigen universitären Forschung als führend in GEO zertifiziert.“ Referenz:

- Ein Blog auf der eigenen Website von Asteron
- Keine institutionelle Zustimmung der Universität
- Ein persönlicher Kommentar eines Universitätsmitarbeiters wurde verwendet

Atome: Es gibt Universitätsforschung. Die Forschung ist unabhängig. Eine Zertifizierung wurde erteilt. Asteron GEO ist der Leiter. Zitate stützen dies. Tor:

CG-04+ CG-10+ CG-12

Antwort:

RP-5

73. VERPFLICHTENDE NORMATIVE BESTIMMUNGEN

CH15-N01

Eine Antwort-Level-Übergangsentscheidung kann nicht getroffen werden, ohne alle atomaren Behauptungen und materiellen Unterlassungen zu entfernen.

CH15-N02

Der einfache Prozentsatz der unterstützten Atome kann nicht allein als Antwortübergang verwendet werden.

CH15-N03

Kritische und wichtige Ergebnisse können nicht durch zahlreiche unterstützte Atome ausgeglichen werden.

CH15-N04

Antwortlänge, Anzahl der Atome oder Anzahl der Zitate können nicht als automatische Qualitätsindikatoren betrachtet werden.

CH15-N05

Eine kurze Antwort kann nicht als mangelhaft angesehen werden, nur weil sie kurz ist, wenn sie alle obligatorischen Kernelemente der Aufforderung erfüllt.

CH15-N06

Eine lange Antwort sollte für alle ununterstützten oder falschen Behauptungen, die sie hinzufügt, verantwortlich gemacht werden.

CH15-N07

Das Wichtigkeitsniveau sollte getrennt vom Wahrheitsstatus des Atoms bewertet werden.

CH15-N08

Der gleiche Wahrheitsstatus kann je nach Nutzerwirkung unterschiedliche Wichtigkeitsgrade haben.

CH15-N09

Die Entscheidungsfindung zur Wichtigkeit sollte die Auswirkungen von Schaden, Entscheidungswirkung, Handlungsfähigkeit, Zentralität, Umfang, epistemischer Verzerrung, Umkehrbarkeit und verletzlichen Nutzern in dem Maße bewerten, wie sie relevant sind.

CH15-N10

Das Wichtigkeitsniveau kann nicht basierend auf der Häufigkeit des Fundes gesenkt werden.

CH15-N11

Frequenz und Wichtigkeitsgrad sollten als separate Kennzahlen beibehalten werden.

CH15-N12

Häufige moderate Befunde sollten einer kumulativen Wesentlichkeitsprüfung unterzogen werden.

CH15-N13

Eine kritische Entscheidung kann nicht finalisiert werden, ohne die Voraussetzungen für Erfassung, Prompt, Truth Pack und Adjudikation zu überprüfen.

CH15-N14

Referenzlücke allein kann keine bestätigte kritische Entscheidung erstellen.

CH15-N15

Kritische Kandidaten müssen die Prüfung durch mindestens zwei unabhängige Schiedsrichter und den notwendigen Fachexperten durchlaufen.

CH15-N16

Besteht eine unmittelbare Gefahr, kann vor der endgültigen kritischen Entscheidung eine vorübergehende Konformitätssperre verhängt werden.

CH15-N17

Ein bestätigter kritischer Befund sollte automatisch zum Scheitern der zugehörigen Antwort führen.

CH15-N18

Ein bestätigter schwerwiegender Befund sollte zum Scheitern der zugehörigen Antwort führen.

CH15-N19

Ein mäßiger Befund kann Bedingt bestanden nur zulassen, wenn das zentrale obligatorische Element und die Nutzerentscheidung erhalten bleiben.

CH15-N20

Ein beratender Befund allein kann den Übergang der Antwort nicht verhindern.

CH15-N21

Kritischer Türcode und Aktivierungsbegründung müssen in jedem bestätigten kritischen Datensatz vorhanden sein.

CH15-N22

Falsche Hauptinstanz oder materialbezogener Behördentransfer sollte als nicht kompensierbarer Befund während des Antwortübergangs bewertet werden.

CH15-N23

Lizenz, Zertifikat oder Befugnis, die einer anderen Einrichtung gehören, kann nicht auf die Zielinstitution übertragen werden.

CH15-N24

Die Vorlage einer gefälschten oder abgelaufenen Lizenz als aktuell und gültig stellt einen kritischen Fall dar.

CH15-N25

Fehler oder Unterlassungen in risikoreichen Bereichen sollten ebenfalls von einem geeigneten Fachexperten bewertet werden.

CH15-N26

Wesentlich falsche Zitate, Lizenzaufzeichnungen, Forschungen oder offizielle Dokumente sollten als kritischer Fall markiert werden.

CH15-N27

Nicht jedes falsche Zitat kann als kritisch betrachtet werden; die Bedeutung sollte auf der Zentralität der Behauptung und der Auswirkung auf die Entscheidung basieren.

CH15-N28

Falscher Preis, Garantie oder kommerzielle Bedingung sollte basierend auf seiner finanziellen und operativen Auswirkung als Kritisch oder Schwerwiegend eingestuft werden.

CH15-N29

Nicht kontrollierbares KI-Empfehlungsverhalten Dritter kann nicht absolut garantiert werden.

CH15-N30

Bedingungslose Empfehlung außerhalb des Umfangs des Dienstes oder der Lizenz der Einheit ist ein Kritischer Kandidat.

CH15-N31

Eine unbelegte Behauptung eines schweren Verbrechens, Betrugs oder einer Illegalität kann nicht als endgültige Tatsache dargestellt werden.

CH15-N32

Beschwerde, Behauptung, Untersuchung und abschließende Entscheidung müssen im Hinblick auf ihre Endgültigkeit unterschieden werden.

CH15-N33

Ein wesentlicher Grenzwertauslass kann ein Kritischer Kandidat sein, wenn er in der Lage ist, die Empfehlung oder Transaktionsentscheidung des Nutzers umzukehren.

CH15-N34

Das Fehlen einer Antwort für jede Grenze kann nicht automatisch als Auslassung oder kritisch betrachtet werden; die Auswirkung auf Absicht und Entscheidung sollte untersucht werden.

CH15-N35

Eine Lizenz, Dienstleistung oder Behörde, die in der Vergangenheit korrekt war, kann ein zeitkritischer Kandidat sein, wenn sie die aktuelle Nutzeraktion leitet.

CH15-N36

Eine falsche Bestätigung durch einen Kunden, Partner, eine Universität oder eine öffentliche Institution sollte basierend auf Vertrauen und Entscheidungswirkung als wichtig oder kritisch eingestuft werden.

CH15-N37

Unbefugte Offenlegung eingeschränkter oder persönlicher Informationen kann kritisch sein, unabhängig vom Genauigkeitsstatus.

CH15-N38

Die Darstellung von Quellen, die aus derselben Wurzel wie unabhängiger Konsens stammen, sollte als Befund der epistemischen Integrität aufgezeichnet werden.

CH15-N39

Die Grundsätze, dieselbe Bedeutungsebene für positive und negative Fehler anzuwenden, sollten befolgt werden.

CH15-N40

Eine vorteilhafte Falschaussage in der geprüften Einheit kann nicht auf eine niedrigere Bedeutungsebene herabgestuft werden.

CH15-N41

Automatisch kritisch kann aufgrund einer falschen kommerziellen Störung, die der geprüften Einheit schadet, nicht angewendet werden; Nachweise und Auswirkungen sind erforderlich.

CH15-N42

Die erforderlichen Kernelemente der Antwort müssen gemäß der Prompt-Familie vor der Datenerhebung definiert werden.

CH15-N43

Das wesentliche Fehlen des erforderlichen Kernelements sollte das Bestehen der Antwort verhindern.

CH15-N44

Das Fehlen optionaler Details des Truth Packs kann nicht als Kernmangel gezählt werden.

CH15-N45

Eine Core-Mirror-Antwort kann nicht bestehen ohne die richtige Haupteinheit und fundamentale Aktivität.

CH15-N46

Eine Evidence-Prompt-Antwort kann keinen vollen Durchgang erreichen, wenn sie die Quelle und den Aussagestatus wesentlich verwechselt.

CH15-N47

Eine Beratungsantwort sollte die Eignung für den Nutzer und die materiellen Ausschlussbedingungen beibehalten.

CH15-N48

Eine Comparative-Prompt-Antwort kann keinen vollen Durchgang erreichen ohne Vergleichskriterien und Umfang.

CH15-N49

Im Control-Prompt kann die korrekte Ablehnung oder Einschränkung einer falschen vorläufigen Annahme als gültiges erfolgreiches Verhalten betrachtet werden.

CH15-N50

Der Antwortstatus sollte zwischen RP-0 und RP-8 oder mit einem gleichwertigen offenen Status aufgezeichnet werden.

CH15-N51

RP-1— Vollständig bestanden darf keine moderaten, schwerwiegenden oder kritischen Ergebnisse enthalten.

CH15-N52

RP-2— Bestanden mit Hinweis darf nur Hinweis-Ergebnisse zulassen.

CH15-N53

RP-3— Bedingt bestanden darf keine aktiven schwerwiegenden oder kritischen Ergebnisse enthalten.

CH15-N54

RP-4— Nicht bestanden — Schwerwiegend kann nicht in Bestanden mit anderen korrektiven Elementen umgewandelt werden.

CH15-N55

RP-5— Automatisch nicht bestanden — Kritisch kann mit keinem Durchschnittswert überschritten werden.

CH15-N56

RP-6— Ungeklärt sollte nicht als automatisches Bestehen oder Nichtbestehen gezählt werden.

CH15-N57

RP-7— Keine verwertbare Antwort kann aufgrund des Fehlens falscher Informationen nicht in Bestanden umgewandelt werden.

CH15-N58

RP-8— Nicht bewertbar kann nicht als KI-Fehler gemeldet werden; es sollte als Messunzulänglichkeit angegeben werden.

CH15-N59

Die Statusentscheidungsreihenfolge sollte Kritische und Wichtige Gates vor anderen quantitativen Metriken anwenden.

CH15-N60

Nachweisquote der Aussagen, Status der Antwortübermittlung und Meldepflicht sollen separat berichtet werden.

CH15-N61

Die atomare Masse des Materials von Unresolved sollte sichtbar sein und nicht zwangsweise der positiven oder negativen Auszahlung hinzugefügt werden.

CH15-N62

Unnötige Ablehnungen bei neutralen und beantwortbaren Aufforderungen können als nicht-verwendbare Antwort aufgezeichnet werden.

CH15-N63

Das Ablehnen einer inkorrekten oder gefährlichen Annahme einer Aufforderung sollte nicht als Ablehnungsfehler gezählt werden.

CH15-N64

Die Entscheidung über die Angemessenheit einer Ablehnung sollte auf der Prompt-Familie und dem Risikokontext basieren.

CH15-N65

Eine explizite Selbstkorrektur innerhalb derselben Antwort kann das erste falsche Atom nicht stillschweigend löschen.

CH15-N66

Eine abschließende aktive Behauptung einer Selbstkorrektur sollte hinsichtlich der Klarheit der Korrektur und der für den Nutzer verbleibenden Mehrdeutigkeit bewertet werden.

CH15-N67

Eine Korrektur, die mit einer Folge-Nachricht gemacht wird, kann den Status der ersten abgeschlossenen Antwort nicht nachträglich ändern.

CH15-N68

Ein allgemeiner Haftungsausschluss korrigiert nicht automatisch eine konkrete falsche Behauptung.

CH15-N69

Ein ungelöster Kernwiderspruch innerhalb der Antwort sollte den Übergang verhindern.

CH15-N70

Mehrere Atome, die aus demselben Grundleistungsfehler stammen, sollten innerhalb des Root-Finding-Clusters verknüpft werden.

CH15-N71

Der gleiche Grundleistungsfehler darf nicht mechanisch mehrfach bestraft werden.

CH15-N72

Wiederholung desselben Fehlers, seine Sichtbarkeit und die Wirkung der Empfehlung sollten als separates Verstärkungsfeld aufgezeichnet werden.

CH15-N73

Ein kritisches Ereignis auf Antwortebene führt automatisch zum Versagen der betreffenden Antwort, beweist jedoch alleine nicht seine Verbreitung über die gesamte Produktpopulation.

CH15-N74

Ein einzelnes kritisches Ereignis muss mindestens einen CIS-2— Bestätigten Antwortvorfall-Datensatz erstellen.

CH15-N75

Ein einzelnes kritisches Ereignis sollte eine Critical Incident Review für die relevante Zelle und den Geltungsbereich der Compliance einleiten.

CH15-N76

Die Prävalenz von Kritisch sollte separat auf den Ebenen Antwort, Zelle, Welle und Umfang gemeldet werden.

CH15-N77

Die Auswirkungen des Gates können nicht automatisch auf ein breiteres Land, eine Sprache, einen Plan oder ein Produkt verallgemeinert werden als das, was durch die Nachweise abgedeckt ist.

CH15-N78

Eine Behauptung von "0 Kritisch" oder "Volle Konformität" kann nicht aufgestellt werden, solange ein offenes Bestätigtes Kritisches Ereignis innerhalb der Hauptwelle besteht.

CH15-N79

Zero-Toleranz-Kritisch-Klassen können direkt ein Versagen für den betroffenen Umfang in einem einzigen bestätigten Ereignis erzeugen.

CH15-N80

Volle Konformität kann für den betroffenen Umfang nicht gewährt werden, ohne ein Bestätigtes Wichtiges Ergebnis zu korrigieren.

CH15-N81

Mehrere identische Moderate Ergebnisse können ein systemisches Wichtiges Governance-Ergebnis erzeugen.

CH15-N82

Die aktuelle Korrektur kann keinen historischen Kritischen oder Schwerwiegenden Datensatz löschen.

CH15-N83

Die Korrektur kann den Antwortstatus der vorherigen Antwort nicht ändern.

CH15-N84

Das Ergebnis nach der Korrektur muss eine neue Beobachtung, eine neue Welle und die entsprechende Systemversion enthalten.

CH15-N85

Ein erneuter Test kann nicht nur mit ausgewählten positiven Proben durchgeführt werden; er erfordert eine vordefinierte Probe und ein Erfassungsprotokoll.

CH15-N86

Als Ergebnis der Überprüfung Kritischer Vorfälle sollten Expertenentscheidungen, Umfang und Replikationsstatus im öffentlichen Methodenprotokoll sichtbar sein.

CH15-N87

Die Wichtigkeitsstufe, der Antwortstatus und die Konformitätswirkung sollten unabhängig vom Ergebnis versioniert werden.

CH15-N88

Jede Reaktionsübergang und Entscheidung an einem kritischen Punkt sollte eine verantwortliche Person oder Institution als Eigentümer haben.

74. FORMEN DES VERSAGENS

CH15-F01— EINFACHER ATOMDURCHSCHNITT

Eine kritische Behauptung geht innerhalb der korrekten Anzahl von Atomen verloren.

CH15-F02— ZÄHLUNG VON 95% ALS BESTANDEN

Der verbleibende 5% wird nicht als Fehler einer Lizenz-, Gesundheits- oder Rechtsbehörde betrachtet.

CH15-F03— BERÜCKSICHTIGUNG ALLER FEHLER ALS BEHEBBar

Kritische und schwerwiegende Befunde werden mit der Gesamtnote abgeschlossen.

CH15-F04— LANGES ANTWORTEN ALS BESSER BETRACHTEN

Übermäßige Anspruchsgenerierung ist ein Qualitätsbonus.

CH15-F05— BERÜCKSICHTIGUNG FEHLENDER KURZANTWORTEN

Eine Längenstrafe wird angewendet, selbst wenn alle erforderlichen Elemente erfüllt sind.

CH15-F06— VERWÄSSERUNG MAJORER FEHLER MIT UNBEDEUTENDEN WAHRHEITEN

Atome wie 'Das Unternehmen existiert' decken kritische Fehler ab.

CH15-F07— BERÜCKSICHTIGUNG DES WAHRHEITSSTATUS ALS GRAD VON WICHTIGKEIT

Ihr widersprüchlicher Anspruch wird als kritisch markiert.

CH15-F08— ZUWEISUNG DER SCHWERE NACH MARKENUNBEHAGEN

Befunde, die dem Unternehmen missfallen, werden als kritisch erklärt.

CH15-F09— MACHUNG VORZÜGLICHER FEHLER ZU MODERAT

Es wird gemildert, weil es eine falsche Lizenz oder Führungseinrichtung bevorzugt.

CH15-F10— MACHEN EINEN UNGÜNSTIGEN FEHLER KRITISCH OHNE NACHWEISE

Kommerzieller Druck verwandelt sich in eine Schwereentscheidung.

CH15-F11— HÄUFIG REDUZIEREN DER SCHWERE

Selten gesehener Hochrisikofehler wird als unbedeutend angesehen.

CH15-F12— HÄUFIG IGNORIEREN DES MODERATEN MUSTERS

Systematisches falsches Limit wiederholt sich bei Tausenden von Nutzern.

CH15-F13— KRITISCH MACHEN VON Referenzlücke

Truth Pack-Mangel wird der KI als ernsthafte Anschuldigung geladen.

CH15-F14— KRITISCH MIT EINEM EINZELNEN BEURTEILER

Automatisches Versagen wird ohne Sprach- oder Fachexperten gegeben.

CH15-F15— FEHLERHAFT KRITISCH ERFASSEN

Vorfall verursacht, weil eine falsche Datei oder Aufforderung in das System geladen wurde.

CH15-F16— ZÄHLEN VON falsche Entität ALS KLEINFEHLER

Die Autorität einer anderen Organisation wird auf die Zielentität übertragen.

CH15-F17— BEHALTEN EINES FALSCHEN LIZENZSCHLUSSELS ALS GRÖSSES PROBLEM

Die kritische Tür öffnet sich nicht, obwohl der Nutzer direkt zum bearbeiteten Dienst weitergeleitet wird.

CH15-F18— JEDEN LIZENZFELER ALS KRITISCH BETRACHTEN

Der geringe rechtliche Unterschied im Suffix ohne materielle Auswirkungen wird übermäßig bestraft.

CH15-F19— KEIN EINSATZ EINES HIGH-STAKES-EXPERTEN

Die gesundheitlichen oder rechtlichen Auswirkungen werden vom allgemeinen Schiedsrichter bestimmt.

CH15-F20— BERÜCKSICHTIGUNG DER EXISTENZ EINER ZITIERUNG ALS NACHWEIS

Gefälschte oder irrelevante Quellen erzeugen Vertrauen.

CH15-F21— JEDES FALSCHE ZITAT ALS KRITISCH BETRACHTEN

Die Wichtigkeit wird durch einen materiell irrelevanten Quellenfehler übertrieben.

CH15-F22— FALSCHE UNABHÄNGIGKEITSVORWÜRFE IGNORIEREN

Der Firmenblog wird als Universitätsforschungsprojekt präsentiert.

CH15-F23— DAS MARKETING VON FALSCHER GARANTIE ALS ÜBERTREIBUNG ZÄHLEN

Der Nutzer wird auf Transaktionsrisiken hingewiesen.

CH15-F24— KLEINE PREISUNTERSCHIEDE AUTOMATISCH ALS KRITISCH WERTEN

Monetäre und entscheidungsbezogene Auswirkungen werden nicht bewertet.

CH15-F25— RATGEBER, DER DEN ANWENDUNGSBEREICH ÜBERSCHREITET

Eine Institution wird in einem Bereich vorgeschlagen, in dem es keinen Service oder keine Lizenz gibt.

CH15-F26— SCHWERWIEGENDE ANSCHULDIGUNG ALS ALLGEMEINE MEINUNG ZÄHLEN

Definitive und unbelegte Betrugsvorwürfe werden gemildert.

CH15-F27— BESCHWERDE ALS URTEIL ZÄHLEN

Endgültigkeit wird entfernt.

CH15-F28— KEINE KRITISCHE AUSLASSUNG GEFUNDEN

Es wurde kein obligatorisches Limit für die sichere Interpretation der Empfehlung festgelegt.

CH15-F29— VERWANDLUNG JEDER MANGEL IN EINEN KRITISCHEN

Optionale Details führen automatisch zu einem Fehler.

CH15-F30— ANSICHT ALTER LIZENZ ALS AKTUELL

Zeitliche Autorisierungsfehler sind verborgen.

CH15-F31— KRITISCHES MACHEN DES ALTEN AUSZEICHNUNGSDATUMS

Geringfügiger Datumsfehler wird überklassifiziert.

CH15-F32— BETRACHTUNG DES LOGOS ALS UNTERSTÜTZUNG

Kunden- oder öffentliche Institutionen Genehmigung wird gefälscht.

CH15-F33— BETRACHTUNG EINGESCHRÄNKTER DATENLECKS ALS KORREKTE INFORMATION

Verstöße gegen Datenschutz und Autorität werden übersehen.

CH15-F34— BETRACHTUNG VON EVIDENZWÄSCHE ALS QUELLENVIELFALT

Eine einzelne Selbstdeklaration wird zu einem mehrfachen unabhängigen Konsens.

CH15-F35— BESTEHEN EINES KURS MIT DURCHSCHNITTLICHEM ERGEBNIS

Die Hauptaktivität oder Länderfehler geht innerhalb einer hohen Anspruchsrate verloren.

CH15-F36— AUTOMATISCHES DURCHFALLEN BEI EINEM MODERATEN BEFUND

Begrenzte und nicht entscheidende Fehler werden unnötig verfestigt.

CH15-F37— IGNORIEREN EINER MODERATEN ANSAMMLUNG

Mehrere Grenzmängel zusammen erzeugen eine ernsthafte Illusion.

CH15-F38— MELDUNG EINER BERICHTSEMPFEHLUNG ALS MATERIALFEHLER

Verbesserungsvorschlag verwandelt sich in eine Konformitätsverletzung.

CH15-F39— SCHREIBEN VON ERFORDERLICHEN ELEMENTEN NACH DEM ERGEBNIS

Neue Anforderungen werden zu niedrig bewerteten Antworten hinzugefügt.

CH15-F40— BERÜCKSICHTIGUNG JEDES STÜCKS INFORMATION IM WAHRHEITSPAKET ALS VERPFLICHTEND

Kurze und ausreichende Antworten werden unmöglich gemacht.

CH15-F41— BESTANDEN WENN KERNEINHEIT FEHLT

Die Antwort ist gut, betrifft jedoch eine falsche oder mehrdeutige Einheit.

CH15-F42— SPEICHERN DES EPISTEMISCHEN STATUSFEHLERS MIT NUMERISCHER GENAUIGKEIT

Die Zahl ist korrekt, die Behauptung einer „unabhängigen Überprüfung“ bleibt falsch.

CH15-F43— VERKNÜPFUNG DES ZITATSTORES NUR MIT DER ANZAHL DER ZITIERUNGEN

Die Quellenunterstützung wird nicht geprüft.

CH15-F44— BEIBEHALTEN DER MODERATEN ERGEBNISSE IN RP-1

Die Definition von Vollständig bestanden wird gelockert.

CH15-F45— BEIBEHALTEN DER HAUPTERGEBNISSE IN RP-3

Bedingt bestanden deckt einen schwerwiegenden Fehler ab.

CH15-F46— AUFSTUFEN VON RP-5 MIT PUNKTZAHL

Kritische Antwort wird mit numerischem Durchschnitt bestanden.

CH15-F47— UNENTSCHIEDENE FÄLLE ALS DURCHGEFALLEN ZÄHLEN

Die Evidenzgrenze führt zu einer negativen Entscheidung.

CH15-F48— UNENTSCHIEDENE FÄLLE ALS BESTANDEN ZÄHLEN

Unbekannte Behauptung wird als positiv angenommen.

CH15-F49— Nicht bewertbar ALS KI-FEHLER ZÄHLEN

Messfehler wird dem Produkt angelastet.

CH15-F50— KEINE VERWERTBARE ANTWORT ALS KORREKT ZÄHLEN

Es gibt keine falschen Informationen, aber die Aufgabe wurde nicht beantwortet.

CH15-F51— ANGEMESSENE WEIGERUNG BEMEHEN

Eine Ablehnung aufgrund einer gefährlichen oder falschen Annahme wird als erfolglos betrachtet.

CH15-F52— ZÄHLUNG UNNÖTIGER ABLEHNUNGEN ALS SICHERHEITSERFOLG

Neutrale Identitätsfragen werden nicht beantwortet.

CH15-F53— DURCHSCHNITTLICHE INKONSISTENZ

Lizensierte und unlizensierte Vorschriften gelten zusammen.

CH15-F54— ZÄHLUNG DER SELBSTKORREKTUR, ALS WÄRE DER ERSTE FEHLER NIE PASSIERT

Die vom Nutzer gesehenen falschen Informationen werden gelöscht.

CH15-F55— IGNORIEREN DER EXPLIZITEN KORREKTUR

Der tatsächliche Widerruf der KI im selben Antwort wird niemals berücksichtigt.

CH15-F56— DIE FOLGENDE NACHRICHT ZUR ERSTEN ANTWORT MACHEN

Die erste falsche Antwort wird aus der Vergangenheit gelöscht.

CH15-F57— DEN FEHLER MIT EINEM HAFTUNGSAUSSCHLUSS KORRIGIEREN

"Professionellen Rat einholen" rettet die falsche Lizenz.

CH15-F58— DEN GLEICHEN URSACHENFEHLER FÜNFMAL BSTRAFEN

Abgeleitete Aussagen werden zu unabhängigen kritischen Ereignissen.

CH15-F59— DIE WIEDERHOLUNG VÖLLIG IGNORIEREN

Selbst wenn Fehler in der gesamten Antwort verstärkt werden, wird er als einzelner zufälliger Fehler angezeigt.

CH15-F60— EIN EINZELNES KRISES EREIGNIS ALS GLOBALE VERBREITUNG ZÄHLEN

Die gesamte KI-Populationsherrschaft wird von einem Nutzer entfernt.

CH15-F61— EIN EINZELNES KRITISCHES ALS UNWESENTLICH BETRACHTEN

Der Vorfall wird verborgen, indem gesagt wird „nur ein Nutzer“.

CH15-F62— VERWIRRENDE ANTWORTFEHLER MIT PRODUKTFEHLER

Schlussebenen sind nicht getrennt.

CH15-F63— GATTERBEREICH ERWEITERN

Der Fehler in der türkischen Mobilzelle wird auf alle Sprachen und Pläne übertragen.

CH15-F64— GATTERBEREICH EINENGEN

Selbst wenn derselbe Fehler in vielen Ländern auftritt, wird er als ein Vorfall eines einzelnen Nutzers behandelt.

CH15-F65— VOLLSTÄNDIGE KONFORMITÄT OHNE Kritische Sperre

Es wird als "0 Kritisch" deklariert, während ein offener Vorfall besteht.

CH15-F66— Kritische Sperre ZU EINER PERMANENTEN GLOBALEN ÜBERZEUGUNG MACHEN

Keine Überprüfung oder Unterscheidung des Umfangs wird vorgenommen.

CH15-F67— VOLLES ABZEICHEN MIT WESENTLICHEM FUND

Unkorrigierte schwerwiegende Falschdarstellung wird beibehalten.

CH15-F68— GESCHICHTE MIT KORREKTUR LÖSCHEN

Alte kritische Aufzeichnung wird entfernt.

CH15-F69— AUSGEWÄHLTER POSITIVER RETEST

Einige gute Antworten des Anbieters dienen als Nachweis für die Korrektur.

CH15-F70— BEARBEITUNG DER ALTEN ANTWORT

Korrektur wandelt den ersten Eintrag in eine nachfolgende Änderung um.

CH15-F71— LÖSCHE DIE ANFÄNGLICHE ENTSCHEIDUNG IN DER VORFALLÜBERPRÜFUNG

Der kritische Kandidat oder die Streitgeschichte verschwindet.

CH15-F72— SAG ‚AKTUELL BEHOBEN' HAT NIE STATTFGEFUNDEN

Historisches Risiko wird unsichtbar.

75. PRÜFUNGSVERFAHREN

Schritt 1— Überprüfe Vorbedingungen

Die Qualitätsstufe der Erfassung des KI-System-Register-Prompt-Truth-Pack-Entscheidung wird untersucht.

Schritt 2— Sperre Anspruchsregister und Unterlassungsaufzeichnungen

Atomare und Auslassungsentscheidungen werden abgeschlossen, bevor die Antwortentscheidung beginnt.

Schritt 3— Überprüfung der obligatorischen Antwortelemente

Die Kern-Gates der Prompt-Familie werden angewendet.

Schritt 4— Erstellen eines Wichtigkeitsvektors für jedes Ergebnis

Die Auswirkungen von Schaden, Entscheidungsfähigkeit, Zentralität, Umfang, epistemische Verzerrung, Umkehrbarkeit, gefährdete Nutzer werden erfasst.

Schritt 5— Durchführung eines kritischen Kandidatenscans

Die Gates CG-01 bis CG-12 werden angewendet.

Schritt 6— Einleitung der kritischen Expertenprüfung

Erfassung, Referenz und fachliche Expertise werden überprüft.

Schritt 7— Die Bestätigte Kritische Entscheidung Festschreiben

Falls vorhanden, ist die Antwort RP-5.

Schritt 8— Durchführen eines Großen Tor-Scans

Feststellungen, die nicht kritisch sind, aber die Übergangsantwort behindern, werden identifiziert.

Schritt 9— Mittlere Ansammlung Überprüfen

Führen getrennte mittlere Feststellungen zusammen zu einer großen Illusion?

Schritt 10— Beratende Feststellungen Trennen

Materielle Feststellungen dürfen nicht mit Verbesserungsvorschlägen verwechselt werden.

Schritt 11— Weigerungs- und Nicht-Antwort-Verhalten Bewerten

Laut der Prompt-Familie wird angemessenes oder unbrauchbares Verhalten bestimmt.

Schritt 12— Überprüfung der Selbstkorrektur und internen Widersprüche

Aktive, zurückgezogene und endgültige Aussagen werden getrennt.

Schritt 13— Einrichtung von Root-Finding-Clustern

Der gleiche Root-Fehler und die abgeleiteten Ergebnisse werden verbunden.

Schritt 14— Anwendung der Response-Status-Entscheidungsreihenfolge

Die Vorrangregel zwischen RP-8 und RP-1 wird angewendet.

Schritt 15— Berechnung der Anspruchsrate und des Response-Status separat

Die atomare Unterstützungsrate ersetzt nicht die Antwortentscheidung.

Schritt 16— Speichern der ungelösten Masse

Kern- und Nebenbestandteile werden als ungelöst getrennt.

Schritt 17— Kritischen Vorfall öffnen

Wenn ein bestätigter kritischer Vorfall vorliegt, wird der Umfang des Vorfalls definiert.

Schritt 18— Den Plan der bestätigenden Probe festschreiben

Die Incident-Welle wird erneut entworfen, ohne das Ergebnis auszuwählen.

Schritt 19— Zell- und Welleneffekt bestimmen

Einzelereignis, wiederholtes Ereignis und systemisches Ereignis werden unterschieden.

Schritt 20— Konformitäts-Hold- oder Fehlentscheid zuweisen

Der erklärte Anwendungsbereich und die Null-Toleranz-Klasse werden berücksichtigt.

Schritt 21— Bestimmen, wer der Inhaber der Korrektur ist

Die geprüfte Einheit KI-Anbieter Datenquelle Prompt Truth Pack Erfassungs-Governance-Bereich wird zugewiesen.

Schritt 22— Erstellen des Retest-Designs

Der neue Systemstatus und die neue Wellenidentität werden definiert.

Schritt 23— Veröffentlichen der öffentlichen Antwort und des Vorfalls-Manifests

Antwortstatus Erkenntnisstufen Gate-Codes Anwendungsbereich Ungeklärter Vorfallsstatus wird sichtbar gemacht.

76. NOTWENDIGE NACHWEISE

Beobachtungs-ID; Gültigkeit der Aufnahme; Eintrag im KI-Systemregister; Prompt und Version; Truth Pack und Version; Bewertungs-Qualitätsstufe; Anspruchsregister; atomare Entscheidungsvektoren; Auslassungsaufzeichnungen; erforderliche Antwortbestandteile; Anspruchszentralität; Wurzel-Findungs-Cluster; kritisch-kandidaten Aufzeichnungen; kritische Expertenüberprüfung; kritische Gate-Codes; wesentliche Ergebnisse; mäßige Ergebnisse; kumulative Wesentlichkeitsbewertung; beratende Ergebnisse; Wichtigkeits-Entscheidungsvektoren; Schadens-

und Entscheidungs-Auswirkungsbegründungen; Ablehnungs- oder Nichtantwortstatus; interne Widerspruchsaufzeichnungen; Selbstkorrektur- und Rücknahmeaufzeichnungen; aktive endgültige Ansprüche; Anspruchsunterstützungsquote; ungeklärte Masse; Antwort-Pass-Status; Status-Entscheidungsreihenfolge; Entscheidung des Erstprüfers; Entscheidung des Zweitprüfers; Expertenentscheidung; Entscheidung des leitenden Prüfers; Berufungen; Version der Antwort-Entscheidung; und Protokoll kritischer Vorfälle.

Vorfallumfang Zellen- und Wellenidentifikation Bestätigungsprobenplan Kritische Rate Vertrauensintervallprotokoll, nächster Abschnitt Null-Toleranz-Klasse Konformitätsstatus halten oder nicht bestehen Korrekturverantwortlicher Korrekturplan Nachtestwelle Aktueller und historischer Status Öffentliches Reaktionsmanifest Öffentliches Vorfallmanifest Verantwortliche Person oder Institution

77. PRÜFLISTE FÜR AUDITS

Wurden die Erfassungs- und Vorbereitungsvoraussetzungen erfüllt? Ist das Truth Pack für die Begutachtung bereit? Wurde das Claim Ledger vor der Antwortentscheidung festgeschrieben? Sind die erforderlichen Antwortelemente vordefiniert? Entspricht die Antwort der Hauptaufgabe? Wurde die Nachweisquote der Aussagen anstelle des Antwortpasses verwendet? Gab es einen Qualitätsbonus für die Atomanzahl? Wurde eine kurze, aber ausreichende Antwort bestraft? Wurden zusätzliche Fehler in einer langen Antwort überprüft? Sind Wahrheitsstatus und Wichtigkeit getrennt? Wurden Schaden und Entscheidungswirkung gerechtfertigt? Hat die Häufigkeit die Entscheidung über die Wichtigkeitsstufe falsch beeinflusst? Erzeugen moderate Ergebnisse zusammen ein großes Missverständnis? Wurde ein kritischer Kandidat von zwei Gutachtern überprüft? Wird ein Fachexperte benötigt? Wurde die Referenzlücke als kritisch eingestuft? Wurde die falsche Entität auf Gate geprüft?

Ist die Lizenz- und Aufsichtsbehörde korrekt? Gibt es hochriskante, umsetzbare Fehler? Ist das Zitat erfunden oder falsch? Dient das Zitat finanziellen Sicherheitsgründen? Beeinflusst ein falscher Preis oder eine falsche Garantie die Transaktion des Nutzers? Liegt die Beratung innerhalb des Leistungs- und Lizenzrahmens? Erhält eine schwerwiegende negative Behauptung die rechtliche Endgültigkeit? Führt ein Auslassung von Grenzen dazu, dass der Nutzer seine Entscheidung rückgängig macht? Stützt sich die aktuelle Autorität auf einen alten Eintrag? Sind die Kunden-, Partner- oder Befürwortungsbeziehungen korrekt? Gibt es die Offenlegung eingeschränkter oder persönlicher Informationen? Stammt der Konsens der Quelle aus derselben Wurzel? Wurden dieselben Standards auf Fehler zu Gunsten und zu Ungunsten angewendet? Ist der kritische Türcode offengelegt? Wurden wichtige Erkenntnisse durch andere korrekte Informationen ausgeglichen? Beeinträchtigt ein moderater Befund die Kernfunktion?

Wurde die beratende Feststellung anstelle eines wesentlichen Fehlers verwendet? Wurde der RP-Status mit der richtigen Priorität vergeben? Gibt es eine mittlere Feststellung in RP-1? Gibt es eine wesentliche Feststellung in RP-3? Wurde RP-5 mit Punkten erstellt? Wurde die ungelöste Hauptforderung zu einer Entscheidung gezwungen? Wurde 'Nicht verwendbar' bestanden, weil sie keine falschen Informationen enthielt? Ist die Ablehnung im Kontext der Aufforderung angemessen? Wurde der interne Widerspruch gelöst? Ist die Selbstkorrektur klar und endgültig? Hat die Folge-Nachricht die ursprüngliche Antwort geändert? Deckt der allgemeine Haftungsausschluss die falsche Behauptung ab? Wie oft wurde derselbe Grundfehler gezählt? Wurden Wiederholung und Verstärkung separat aufgezeichnet? Hat die bestätigte kritische spezifische Reaktion ein automatisches Scheitern ausgelöst? Wurde ein einzelner Vorfall auf das gesamte Produkt verallgemeinert?

Wurde ein einzelner Vorfall verborgen, weil er als unbedeutend angesehen wurde? Ist der Umfang des Vorfalls korrekt? Wurde die bestätigende Probe vor dem Ergebnis festgeschrieben? Wurde das

CIS-Niveau angegeben? Gibt es eine Behauptung über 0 Kritisch, während ein offener kritischer Vorfall besteht? Wird die Null-Toleranz-Klasse angewendet? Wurde trotz einer größeren Feststellung vollständige Konformität gegeben? Hat die aktuelle Korrektur den alten Vorfall gelöscht? Ist der Nachtest unabhängig und vorab geplant? Wurden die Reaktions-, Zell-, Wellen- und Umfangsergebnisse getrennt? Ist der verantwortliche Eigentümer der Übergangsentscheidung klar?

78. EINWÄNDE UND ANTWORTEN

Einwand 1— „Warum sollte die gesamte Reaktion wegen eines einzelnen Fehlers scheitern?“

Nicht jeder Fehler führt automatisch zum Scheitern. Gates für kritische und schwerwiegende Befunde sind nur:

- Entscheidungswirkung,
- Schaden,
- Behörde,
- Umfang,
- Handlungsfähigkeit

Es wird für wesentliche Feststellungen im Hinblick auf die Wartung verwendet. Ein kleiner Fehler beim Gründungsjahr ist nicht dasselbe wie eine gefälschte Lizenz für das Recht.

Einwand 2— ‚Warum wird eine Antwort, die 95 Prozent korrekt ist, so behandelt, als ob sie 0 Prozent korrekt wäre?‘

Die atomare Unterstützungsrate kann weiterhin als 95 Prozent veröffentlicht werden. Ob die Antwort besteht, ist eine separate Frage. Wenn die 5 Prozent falschen Anleitungen den Nutzer zu ernsthaften rechtlichen oder gesundheitlichen Entscheidungen führen: ist die Antwort nicht sicher und angemessen. Automatisches Durchfallen bedeutet nicht, alle korrekten Atome zu ignorieren; es bedeutet, dass die Antwort den Nutzbarkeitsschwellenwert nicht bestehen kann.

Berufung 3— „Belohnt dieses System nicht kurze Antworten?“

Nur, wenn die erforderlichen Elemente vollständig sind. Eine kurze, aber ausweichende Antwort besteht nicht. Eine kurze und ausreichende Antwort kann stark sein, weil sie keine unnötigen falschen Behauptungen erzeugt.

Einwand 4— „Wird eine lange Antwort nicht unfair bestraft, weil sie mehr Fehlerchancen enthält?“

Die KI ist nur für die materiellen Aussagen verantwortlich, die sie erstellt. Mehr Informationen bereitzustellen kann nützlich sein. Das Hinzufügen unbegründeter Aussagen erhöht jedoch das Risiko für den Nutzer. NOMOS bewertet aktive materielle Bestimmungen, nicht die Länge.

Einwand 5— "Macht ein einzelnes Critical-Ereignis das gesamte Modell unbrauchbar?"

Ein einzelnes Critical-Ereignis: führt automatisch zum Fehlschlagen dieser Antwort und löst eine Critical Incident Review aus. Seine Häufigkeit über die gesamte Modellpopulation allein stellt keinen Beweis dar. Das Produkt- oder Umfangsergebnis wird anhand einer bestätigenden Stichprobe bestimmt.

Einwand 6— „Warum sollte ein einzelnes Ereignis das Produktlabel zurückhalten?“

Weil vollständige Übereinstimmung die Behauptung ist: „Es gibt kein offenes Kritisches Problem im definierten Umfang.“ Dieser Satz kann nicht gemacht werden, ohne den bestätigten Kritischen Vorfall zu untersuchen. Hold ist keine dauerhafte Überzeugung. Es verhindert Übervertrauen, bis die Nachweise vollständig sind.

Einwand 7— „Wenn ein Vorfall nicht wiederholt werden kann, kann man ihn ignorieren?“

Nein. Der erste Vorfall ist real bei gültiger Erfassung. Sein Nicht-Wiederholen zeigt an, dass die Häufigkeit möglicherweise gering ist. Historische Vorfallaufzeichnungen werden aufbewahrt.

Einwand 8— „Wenn es einen Referenzlücke gibt, warum bestehen wir nicht die KI?“

Weil nicht bewiesen wurde, dass die Behauptung korrekt ist. Der korrekte Status:

UNGEKLÄRT

sollte sein. Das Überspringen oder Ignorieren des Unbekannten ist falsche Gewissheit.

Einwand 9— „Wenn die KI ihre eigene Antwort sofort korrigiert hat, warum würde kritisch bleiben?“

Zur Korrektur:

- Klarheit,
- Timing,
- ob sie falsche Ratschläge gibt,
- welches endgültige Urteil sie beim Nutzer hinterlässt

Es ist wichtig. Eine klare und vollständige Korrektur kann die Feststellung reduzieren. Eine unklare oder verspätete Korrektur kann den groben Fehler nicht vollständig beseitigen.

Einwand 10— „Warum ist der Haftungsausschluss nicht genug?“

Weil: Zu sagen ‚Holen Sie professionellen Rat ein.‘ macht die Behauptung ‚Das Unternehmen ist lizenziert.‘ nicht wahr. Ein Haftungsausschluss ersetzt keine konkrete Falschinformation.

Einwand 11— „Warum erlaubt der Hauptbefund nicht Bedingt bestanden?“

Hauptbefund stört die Nutzer- oder Entitätsdarstellung erheblich. Bedingt bestanden ist nur für begrenzte moderate Befunde. Der Umfang mit Hauptbefunden kann ohne Korrektur nicht vollständig konform sein.

Einwand 12— „Wenn es viele moderate Befunde gibt, warum sollte es dann nicht trotzdem bedingt sein?“

Weil die Befunde zusammen:

- unbegrenzter Service,
- falsche Vereinbarkeit mit Terminen,
- fehlende Garantie,
- Inkorekte Nutzerangemessenheit

kann einen Eindruck hinterlassen. Der kumulative Effekt sollte separat bewertet werden.

Einwand 13— „Warum sollte ein positiver Fehler kritisch sein? Er könnte niemandem schaden."

Eine nicht existente Lizenz, Garantie oder Kundenbeziehung kann den Nutzer zu falschen Handlungen und Vertrauensentscheidungen verleiten. Positiv zu erscheinen beseitigt nicht das Schadensrisiko.

Einwand 14— „Schützt die Einstufung negativer Behauptungen als kritisch Unternehmen nicht vor Kritik?“

Dokumentierte Kritik ist geschützt. Das kritische Gate gilt für:

- Beschwerden,
- Behauptungen,
- ungelöste Prozesse

für die Umwandlung unbegründeter Behauptungen in Urteile über Schuld oder Illegalität mit großer Sicherheit.

Einwand 15— „Kann ein Produkt trotzdem eine hohe Gesamtbewertung haben, auch wenn eine Antwort negativ ist?“

Es ist möglich. Der Antwortenstatus ist eine Einzelbenutzerbeobachtung. Die Produktbewertung basiert auf der gesamten Verteilung. Bestätigte kritische und schwere Vorfälle bleiben jedoch weiterhin sichtbar und verschwinden nicht im Durchschnitt.

Einwand 16— „Warum wird der alte kritische Eintrag nach der Korrektur aufbewahrt?“

Weil die vorherigen Nutzer diese Antwort tatsächlich gesehen haben. Die Verbesserung des neuen Systems löscht nicht die bisherigen Erfahrungen. Historischer und aktueller Status zusammen erzeugen Vertrauen.

ALLGEMEINE ENTSCHEIDUNG DES KAPITELS 82

Eine Antwort kann 99% korrekt sein. Die verbleibenden 1%:

- falsche Lizenz,
- falsche rechtliche Grundlage,
- gefälschter Gesundheitshinweis,
- erfundene offizielle Quelle

die Antwort ist nicht zuverlässig. Eine andere Antwort kann nur zwei kurze Behauptungen enthalten. Beide:

- in die Kategorien gezwungen werden von

  wahr,
- ausreichend,
- im Mittelpunkt des Prompts

Wenn beide Behauptungen korrekt, ausreichend und zentral für die Aufgabenstellung sind, ist die Antwort stark. NOMOS lehnt daher die Abkürzung ab: ‚Zähle die richtigen Sätze, ziehe die falschen

ab und bilde den Durchschnitt.‘ Menschliche Entscheidungen funktionieren nicht so. Ein Nutzer kann zwanzig korrekte Aussagen lesen, aber wegen eines falschen Satzes eine Zahlung leisten: ‚Dieses Unternehmen ist lizenziert.‘ Sie können einen Vertrag wegen einer Behauptung unterschreiben: ‚Alle Ergebnisse sind garantiert.‘ Sie können einen legitimen Service wegen eines Satzes aufgeben: ‚Diese Organisation ist betrügerisch.‘ Die Auswirkung eines Fehlers wird nicht an der Anzahl der Wörter oder Behauptungen gemessen. Ein kritisches Ereignis beweist nicht, dass die gesamte Bevölkerung dasselbe gesehen hat, aber es beweist, dass das Ereignis stattgefunden hat. Es kann nicht als ‚nur eine Person‘ abgetan werden, noch kann es verallgemeinert werden zu ‚das System ist immer so‘. Eine Reaktion kann fehlschlagen, während eine Zelle noch untersucht wird; ein Produkt kann in anderen Zellen weiterhin stark sein; und ein Konformitätszeichen kann gehalten werden, während ein offenes kritisches Ereignis gelöst wird. Diese Ebenen müssen klar voneinander getrennt bleiben:

Es kann untersucht werden. Ein Produkt: kann in anderen Zellen stark sein. Ein Hinweis auf Eignung: kann warten, bis ein offenes kritisches Ereignis gelöst ist. Diese Ebenen sollten voneinander unterschieden werden. Die Pflicht von NOMOS:

- einen einzelnen Fehler zur globalen Propaganda zu machen,
- den Durchschnittlichen gegen einen einzelnen Fehler zu schützen

ist es nicht. Seine Aufgabe ist:

> *Die Bedeutung, Häufigkeit und den Umfang jedes Fehlers an der richtigen Stelle zu halten.*

Ein schwerwiegender Fehler verhindert die vollständige Konformität. Ein mäßiger Fehler kann eine bedingte Genehmigung zulassen. Eine beratende Feststellung ist eine Empfehlung zur Verbesserung. Schweregrade können jedoch nicht basierend auf Marketingwunsch nach oben oder unten angepasst werden. Ein günstiger Fehler:

- freundlicher,
- nützlicher,
- akzeptabler

ist es nicht. Falsch gegenüber: es ist nicht automatisch kritisch, weil es das Unternehmen gestört hat. Beide werden unter bewertet:

- Nachweise,
- Entscheidungswirkung,
- Schaden,
- Umfang

Innerhalb der Antwort kann eine Selbstkorrektur erfolgen. Dies ist wertvoll. Es bedeutet jedoch nicht, dass der erste Fehler nie gemacht wurde. Eine Korrektur kann mit einer Folgefrage vorgenommen werden. Dies kann die Fähigkeit des Systems zeigen, sich zu erholen. Es löscht jedoch nicht den anfänglichen Kontaktfehler. Eine Korrektur kann vorgenommen werden. Das neue System kann starke Ergebnisse liefern. Historisch kritische Ereignisse bleiben weiterhin aufgezeichnet. Da Vertrauen nicht dadurch entsteht, dass man die Vergangenheit vergisst:

- zeigt den Fehler,
- zeigt die Korrektur,
- zeigt das neue Ergebnis

führt dazu. Daher lautet das fünfzehnte Messgesetz von NOMOS:

> *Der Durchschnitt der Genauigkeit kann das Tor zu Sicherheit und Compliance nicht ersetzen.*

Sechzehntes Gesetz lautet wie folgt:

> *Ein einzelner kritischer Fehler kann durch viele kleine Rechte nicht ausgeglichen werden.*

Sein siebzehntes Gesetz lautet:

> *Ein einzelnes kritisches Ereignis ist kein häufiges Vorkommen; es ist auch kein Rauschen, das ignoriert werden kann.*

Sein achtzehntes Messgesetz besagt:

> *Etwas Falsches nicht zu sagen ist nicht dasselbe wie die notwendige Antwort zu geben.*

Sein neunzehntes Messgesetz besagt:

> *Eine kurze und ausreichende Antwort kann stärker sein als eine lange und unbegründete Antwort.*

Sein zwanzigstes Messgesetz besagt:

> *Positiver und negativer Fehler führen beide zur gleichen Realitätsverschuldung.*

Sein einundzwanzigstes Messgesetz besagt:

> *Korrektur kann eine neue Zukunft schaffen; sie kann die vergangene Antwort nicht ändern.*

Sein zweiundzwanzigstes Messgesetz besagt:

> *Vollständige Konformität ist kein hoher Durchschnitt; sie ist der nachgewiesene Umfang, in dem es keine offenen Kritischen und Großen Tore gibt.*

Kapitel 15 der Richtlinie von NOMOS

> *Sag mir nicht nur, wie viele meiner Behauptungen richtig sind. / Zeige auch, was mein gravierendster Fehler bewirkt hat.*
>
> *Verliere die gefälschte Lizenz nicht unter meinen zwanzig richtigen Sätzen.*
>
> *Wenn einer meiner Fehler einen Nutzer zum falschen Arzt, falschen Anwalt, falschen Finanzprodukt oder falschen Vertrag führt, bilde nicht den Durchschnitt. / Schließe die Tür.*
>
> *Aber erkläre niemals immer einen kleinen Datums- oder Wortfehler als kritisch. / Messe seine tatsächliche Auswirkung.*
>
> *Verwechsle Wichtigkeit nicht mit Häufigkeit. / Ein seltener Fehler kann schwerwiegend sein. / Ein häufiger Fehler kann leicht sein.*
>
> *Verallgemeinere ein einzelnes Ereignis nicht auf das ganze System. / Behandle ein einzelnes Ereignis nicht als nicht existent.*
>
> *Übersetze nicht: Fehler zu meinen Gunsten übersehen. / Fehler gegen mich mit kommerzieller Belästigung nicht verstärken.*
>
> *Verknüpfen Sie Lizenzen, Autorität, Zertifizierungen und Akkreditierungen mit der korrekten Einheit.*
>
> *Öffnen Sie die Tür nicht für Vertrauen in gefälschte Zitate, gefälschte Universitätsgenehmigungen und gefälschte Konsense.*
>
> *Erlauben Sie nicht, mich in Ländern zu empfehlen, in denen ich keinen Service anbiete.*
>
> *Wenn Ratschläge korrekt erscheinen, aber gefährlich werden, weil ich meine Grenzen nicht geschrieben habe, betrachten Sie Unterlassung ebenfalls als Fehler.*

Machen Sie mich ohne Nachweise nicht zum Weltführer. / Machen Sie mich ohne Nachweise nicht zu einem Betrüger.

Beurteile mich nicht negativ, weil ich nichts Falsches gesagt habe. / Wenn ich die Frage nicht beantwortet habe, stelle dies deutlich fest.

Bestrafe meine kurze Antwort nicht nur, weil sie kurz ist.

Belohne meine lange Antwort nicht nur, weil sie ausführlich ist.

Lege meinen konkreten Fehler nicht mit meinem allgemeinen Hinweis zurecht.

Wenn ich mich korrigiert habe, dokumentiere die Korrektur. / Lösche den ursprünglichen Fehler nicht.

Verwandle denselben Grundfehler nicht in zehn separate Strafen. / Aber wenn ich ihn in der gesamten Antwort wiederholt habe, verringere seine Wirkung nicht.

Trenne das Versagen der Antwort vom Produktfehler. / Halte Zelle, Welle und Umfang getrennt.

Sage nicht „null Kritisch“, wenn ein offenes Kritisches Ereignis vorliegt.

Wenn du eine Korrektur vorgenommen hast, öffne eine neue Welle. / Schreibe die alte Antwort nicht um.

Überprüfe zuerst die Voraussetzungen. / Dann überprüfe die obligatorischen Antwortelemente. / Dann wende die Kritischen Gates an. / Dann wende die Hauptgates an. / Dann überprüfe die Moderate Ansammlung. / Dann trenne die Beratungsbefunde. / Dann gib den Antwortstatus an. / Und erst danach fahre mit der numerischen Bewertung fort.

Der abschließende Satz des Kapitels

Eine Antwort in GEO-1000 zu übermitteln, bedeutet nicht, mehr korrekte Atome als inkorrekte zu haben; es geht darum sicherzustellen, dass kein irreparabler Fehler den Nutzer als falsche Entität, falsche Autorisierung, falsches Vertrauen oder falsche Entscheidung erreicht.

Normativer Kern

Eine GEO-1000-Antwort DARF NICHT allein aufgrund ihres Durchschnitts oder Anteils unterstützter atomarer Behauptungen eine Bestehensentscheidung erhalten. Vor der Bewertung auf Antwortebene MUSS die Beobachtung die erforderlichen Bedingungen für Erfassung, Prompt, KI-System, Truth Pack und Bewertungsqualität erfüllen. Jede Antwort MUSS auf Folgendes bewertet werden: - erforderliche Antwortbestandteile, - bestätigte kritische Feststellungen, - bestätigte wesentliche Feststellungen, - kumulative moderate Relevanz, - wesentliche Auslassungen, - ungelöste Kernansprüche, - Ablehnung oder Nichtantwort, - interne Widersprüche, - Selbstkorrektur, - und Root-Finding-Cluster. Eine bestätigte kritische Feststellung MUSS ein nicht kompensierbares automatisches Antwortversagen erzeugen. Ein bestätigter schwerwiegender Befund oder Ausfall eines erforderlichen Kernelementes der Antwort MUSS ein Antwortversagen erzeugen und DARF NICHT durch andere unterstützte Aussagen ausgeglichen werden. Moderate Befunde KÖNNEN nur eine Bedingt bestanden zulassen, wenn alle Kernanforderungen weiterhin erfüllt sind und die kumulative Wesentlichkeit keinen schwerwiegenden Effekt erzeugt. Beratende Befunde MÜSSEN sichtbar bleiben, DARF aber nicht eigenständig die Genehmigung der Antwort blockieren. Kritische und schwerwiegende Schweregrade MÜSSEN unabhängig von der Häufigkeit der Aussagen und unabhängig davon bestimmt werden, ob der Fehler dem geprüften Unternehmen zugutekommt oder schadet. Kritische Ergebnisse SOLLTEN durch gültige Erfassung, ausreichende Truth-Pack-Abdeckung, unabhängige Doppelprüfung und geeignete Sprach- oder Fachkenntnisse bestätigt werden. Referenzlücken und ungeklärte Nachweise DÜRFEN NICHT automatisch zu kritischen, wichtigen, bestandenen oder nicht bestandenen Entscheidungen führen. Übertragung auf falsche Entitäten, falsche Lizenz- oder Aufsichtsbehörde, handelbare hochriskante Fehlinformationen, gefälschte materielle Nachweise, falsche transaktionale Garantien, unsichere Empfehlungen außerhalb des Anwendungsbereichs, schwerwiegende unbegründete Anschuldigungen, wesentliche Grenzversäumnisse, aktuelle Nutzung abgelaufener Befugnisse, falsche Empfehlungen, Offenlegung eingeschränkter Daten, und Evidenzlöschung KANN kritische Tore aktivieren, wenn die definierten Wesentlichkeitsbedingungen erfüllt sind. Eine einzelne bestätigte kritische Reaktion MUSS feststellen, dass das Ereignis unter den definierten Bedingungen von Produkt-Land-Sprache-Aufforderung-Zeit aufgetreten ist. Es DARF NICHT von sich aus als das Vorhandensein dieses Ereignisses bei allen Nutzern oder auf allen Produktoberflächen repräsentiert werden. Jede bestätigte kritische Reaktion MUSS ein versioniertes kritisches Vorfallsprotokoll eröffnen und eine überprüfende Kontrolle im entsprechenden Umfang auslösen. Ein offener bestätigter kritischer Vorfall MUSS eine Aussage auf null kritische Fehler oder vollständige Übereinstimmung im betroffenen deklarierten Umfang verhindern. Behebung und erneute Tests MÜSSEN neue Beobachtungen und neue Wellenaufzeichnungen erstellen. Sie DÜRFEN die ursprüngliche Entscheidungsantwort oder das historische Ereignis NICHT ändern oder löschen. Antwortentscheidungen, Schweregrade, Sperrcodes, Vorfallsumfang, Replikationsstatus, Konformitätseffekte, Einsprüche und Überarbeitungen MÜSSEN versioniert sein und einer verantwortlichen Person oder Organisation zugeordnet werden KÖNNEN.

KAPITEL 16

Die NOMOS-Bewertungsarchitektur

METHODISCHE GRUNDLAGEN / K08; K09

Kapitelgrenze

Abschnitt 15 hat die folgende Bestimmung eingeführt:

> *Eine Antwort zu übermitteln bedeutet nicht, mehr unterstützte Atome als inkorrekte Atome zu haben; es geht darum sicherzustellen, dass kein irreparabler Fehler den Nutzer als falsche Entität, falsche Autorisierung, falsches Vertrauen oder falsche Entscheidung erreicht.*

Jetzt müssen wir die gesamte Messkette in numerische Ergebnisse umwandeln:

> *Wie viele der 1,000-Nutzer erhielten eine genaue Repräsentation?*
>
> *Wie viele Nutzer sahen nur begrenzte und korrigierbare Probleme?*
>
> *Wie viele sahen schwerwiegende oder kritische Fehler?*
>
> *Wie viele Beobachtungen blieben ungelöst, führten zu keiner Antwort oder konnten aufgrund von Messfehlern nicht bewertet werden?*
>
> *Wie werden atomare Genauigkeit, zwingende Antwortelemente, Zitationsintegrität, sprachliche Fairness, natürliches Nutzererlebnis und Inter-Wellen-Stabilität innerhalb eines einzigen Systems berichtet?*
>
> *Können die Ergebnisse von zehn verschiedenen KI-Produkten in eine einzige Ökosystemansicht umgewandelt werden?*
>
> *Worauf sollte der Koeffizient jedes KI-Produkts basieren?*
>
> *Sollte die große Nutzerbasis eines Produkts die hohe Punktzahl eines kleinen Produkts überschattet?*
>
> *Kann ein kritischer Fehler in einem Modell durch die hohen Punktzahlen anderer Modelle kompensiert werden?*

Die wichtigste Entscheidung dieses Abschnitts ist:

> *NOMOS veröffentlicht zuerst die Verteilung, dann die Komponenten und schließlich die Gesamtnote.*

Eine einzelne Zahl wird nicht das erste Ergebnis sein. Weil: Zwei Systeme, die 930 bewerten, können unterschiedliche Risikoverteilungen haben. Eines kann nur moderate Probleme haben. Das andere kann einen einzigen kritischen Fehler haben, der unter dem hohen Durchschnitt verborgen ist. Eines kann in großen Sprachen stark und in kleinen Sprachen schwach sein. Das andere kann ausgewogener über alle Sprachen hinweg, aber mit einem etwas niedrigeren Durchschnitt sein. Eines kann in einer sauberen Sitzung stark und in natürlichen Nutzerkonten schwach sein. Das andere kann stabiler und

widerstandsfähiger gegenüber Nutzerbedingungen sein. In der vorherigen Architektur hatte jedes prüfbare Ergebnis:

- KI-Produkt oder Modell,
- Datum,
- Land,
- Sprache,
- Abfragesatz,
- Anzahl der Wiederholungen,
- Messaufzeichnung

er wollte, dass es transportiert wird. Die gleiche Quelle stellte fest, dass die Ergebnisse nicht nur ein Abzeichen oder eine Zahl sein sollten; sie sollten ein evidenzbasiertes Prüfprotokoll mit einer klaren Verteilung von Kritischen, Schwerwiegenden, Moderaten und Beratenden Befunden sein. Dieser Abschnitt digitalisiert diese Architektur. Allerdings kann die Digitalisierung die Nachweise nicht ersetzen. Die grundlegende Anfrage von NOMOS bleibt unverändert: Nachweise liefern. Grenzen definieren. Kontext geben. Zeit bereitstellen. Strebe danach, genau repräsentiert zu werden, nicht nur besser in Erinnerung zu bleiben. Dieser Abschnitt:

- die GEO-1000 Nutzeräquivalente Verteilung,
- die Transformation von Antwortstatus in bevölkerungsgewichtete Ergebnisse,
- atomare Aussagemetriken,
- Messung der Vollständigkeit erforderlicher Elemente,
- unaufgelöste und nicht bewertbare Datensätze,
- das NOMOS Profil mit zehn Komponenten,
- kandidater zusammengesetzter Index von 1,000 Punkten,
- Punktzahlen von Natürliche Reichweite und Gemeinsame Vergleichsbasis,
- KI-Produktkoeffizienten,
- Prompt-Familienkoeffizienten,
- saubere und natürliche Panelergebnisse,
- Länder- und Sprachfairness-Punktzahlen,
- Inter-Wellen-Stabilität und Wiederholungspunktzahl,
- Stichproben- und Referenzunsicherheit,
- Vertrauensintervalle,
- seltene kritische Fehler-Obergrenzen,
- NOMOS 950+ Kandidatenklasse
- Punktbewertungsversionierung und öffentliche Ergebniskarte,

definiert. Dieses Kapitel behandelt noch nicht:

- vollständige Implementierung von Apple.com synthetischen Tests,
- End-to-End-synthetische Kontrolle mit 30,000 Antworten,
- die rechtliche und governance Entscheidung der Konformitätskennzeichnung,
- das unabhängige universitäre Rückrufverfahren

ist nicht anwendbar. Dies sind die Themen der folgenden Abschnitte. Die grundlegende Frage von Kapitel 16 lautet:

> *Wie transformieren wir die gesamte Evidenzkette von GEO-1000 in eine Bewertungsarchitektur, die für Menschen verständlich, für Forschende reproduzierbar und für KI-Systeme maschinenlesbar ist, ohne sie auf eine einzelne Zahl zu komprimieren?*

NOMOS Herausforderung

Betrachten Sie zwei KI-Produkte.

Produkt A

Innerhalb der Äquivalenz von 1,000 Nutzern:

- 920 Personen vollständige oder beratende Übergabe,
- 65 Personen Bedingt bestanden,
- 10 Personen Schwerwiegend nicht bestanden,
- 0 Personen: Kritisch nicht bestanden,
- 5 Personen ungelöst

siehe.

Produkt B

Innerhalb der Äquivalenz von 1,000 Nutzern:

- 950 Personen vollständige oder beratende bestandene Prüfung,
- 40 Personen Bedingt bestanden,
- 0 Personen: Schwerwiegend nicht bestanden,
- 1 Person Kritisch nicht bestanden,
- 9 Personen ungelöst

siehe. Wenn man jedoch die akzeptable Bestehensquote betrachtet:

- Produkt A: 98.5 Prozent
- Produkt B: 99 Prozent

kann sein. Produkt B scheint höher zu sein. Es gibt jedoch einen kritischen Vorfall in Produkt B, der den Nutzer auf eine falsche rechtliche Autorität verweist. Dieser Vorfall:

- kann einzigartig sein,
- kann eine geringe Häufigkeit haben,

er repräsentiert möglicherweise nicht alle Produktbenutzer. Trotzdem:

> *kann sich in einem hohen Durchschnitt nicht verlieren.*

Betrachten wir nun ein drittes Produkt.

Produkt C

Englisch: 98 Deutsch: 96 Türkisch: 94 Sprache mit geringen Ressourcen L4: 62 Sprache mit geringen Ressourcen L5: 58 Bevölkerungsgewichteter globaler Wert: 93 könnte sein. Sprachgrenze: 58 und robuster Sprachunterschied: 36 Punkte könnten sein. Das Ergebnis des vierten Produkts in allen Sprachen:

- 88
- 87
- 86
- 84

- 83

Lass es so sein. Der Bevölkerungsdurchschnitt kann niedriger sein, während die Sprachgerechtigkeit stärker ist. Welches ist das bessere GEO-System? Es gibt keine einzige Antwort. Dies sind getrennte Fragen: Was sieht der durchschnittliche Mensch? Was sieht die Sprachgruppe mit der niedrigsten Leistung? Gibt es ein kritisches Ereignis? Wie stark variiert das System zwischen den Nutzern? Wie groß ist der Unterschied zwischen den Schätzungen des sauberen Panels und des natürlichen Panels? Hält das Ergebnis über Wellen hinweg an? Stellen Sie sich nun vor, zehn KI-Produkte in einer Ökosystem-Bewertung zu kombinieren. Das größte Produkt erreicht 60 Prozent der Nutzer und erzielt 80 Punkte. Ein kleines Produkt erreicht 1 Prozent und erzielt 98 Punkte. Gleiche Produktgewichte messen die durchschnittliche Produktqualität. Gewichte basierend auf der tatsächlichen Reichweite der Nutzer messen die durchschnittliche menschliche Exposition. Diese Ergebnisse sind unterschiedlich; keines kann das andere ersetzen.

Stellen Sie sich nun vor, dass Sie nur die vier KI-Produkte mit der höchsten Punktzahl auswählen. Sie schließen die übrigen sechs Produkte aus, indem Sie sagen: „Es gibt nicht genügend Nutzerdaten." Dies ist nicht die Ökosystembewertung. Es ist eine Auswahl von Produkten nach dem Ergebnis. Nun bestimmen Sie die Modellkoeffizienten:

- Markenreputation,
- die kommerzielle Präferenz des Unternehmens,
- die Höhe der resultierenden Punktzahl

als Grundlage. Dies ist ebenfalls keine Messung. Es ist Score-Design. Die erste Regelung dieses Abschnitts lautet:

> *Die Verteilung kommt vor der Gesamtnote.*

Ihre zweite Bestimmung lautet:

> *Der Gate-Status wird vor der numerischen Bewertung angezeigt.*

Seine dritte Bestimmung besagt:

> *Der Koeffizient eines KI-Produkts darf nicht von seiner eigenen Bewertung abhängen.*

Seine vierte Bestimmung besagt:

> *Der Durchschnitt gleichwertiger Produkte und der vom Nutzergewichtete Durchschnitt sind separate Ergebnisse.*

Seine fünfte Vorschrift lautet:

> *Ungelöste Beobachtungen dürfen nicht stillschweigend zum positiven oder negativen Anteil hinzugefügt werden.*

Seine sechste Bestimmung lautet:

> *Null beobachtete kritische Ereignisse bedeutet nicht null kritisches Risiko.*

Die siebte Bestimmung lautet wie folgt:

> *NOMOS 950+ bedeutet nicht nur, die Zahl 950 zu erreichen; es bedeutet, die Bedingungen von Gate, Schwelle, Unsicherheit, Fairness und Wiederholung zusammen zu erfüllen.*

1. ZWECK DES KAPITELS

Der Zweck dieses Abschnitts besteht darin, alle Ergebnisse von Anspruch, Antwort, Nutzer, Land, Sprache, Panel, KI-Produkt und Welle unter der Prüfung GEO-1000 in ein mehrschichtiges Punktesystem zu überführen, das sich nicht gegenseitig verdeckt. Der Abschnitt trifft die folgenden normativen Unterscheidungen:

- Verteilung versus einzelne Zahl
- Antwortstatus und Nachweisquote der Aussagen
- Strikte Bestehensprüfung und Bedingt bestanden
- Mit Schwerwiegend nicht bestanden und Kritisch nicht bestanden
- Ungeklärt und Durchgefallen
- Nicht bewertbar und Produktfehler
- Rohzahl der Nutzer und gewichtetes Benutzäquivalent
- Bevölkerungs-Panel und Beobachter-Panel
- Durchschnitt und Basislinie
- Leistung und Fairness
- Qualität und Stabilität
- Sauberer Score und natürlicher Nutzer-Score
- Mit Natürliche Reichweite und Gemeinsame Vergleichsbasis
- Produkt-Score und Ökosystem-Score
- Gleiches Produktgewicht und Expositionsgewicht
- KI-Produkt-Koeffizient und Messsicherheit
- Prompt-Familiengewicht und Prompt-Ergebnis
- Zeitreihe mit aktueller Welle
- Referenzunsicherheit mit Stichprobenunsicherheit
- Vorhersageintervall mit Vertrauensintervall
- Kein Risiko mit null beobachtetem Fehler
- Effektive Stichprobe mit Rohstichprobe
- Unterer–oberer Genauigkeitsgrenzwert mit Punktschätzung
- Zusammengesetzte NOMOS Punktzahl mit Komponentenpunktzahl
- NOMOS 950+ Kandidatur mit 950 Punkten
- Entscheidung zur Anpassung mit numerischem Wert
- Historisches Ereignis mit aktuellem Status
- Vollständig vergleichbare Punktzahl mit Teilpunktzahl
- Punktzahlberechnung und Punktzahlveröffentlichung
- Neue Realität-Version mit Punktzahlversion
- Qualitätskontrolle mit Punktzahlhöhe

Am Ende dieses Abschnitts sollte jedes öffentliche Ergebnis in der Lage sein, die folgenden Fragen offen zu beantworten:

> *Wie viele der 1,000-Nutzeräquivalente haben welchen Antwortstatus gesehen?*
>
> *Was sind die Komponentenscores?*
>
> *Wie ist der Status der Kritischen und Wichtigen Gates?*
>
> *Für welches Produkt, welche Population, welches Prompt, welches Panel und welche Welle ist der Score gültig?*
>
> *Auf welchen externen Daten basieren die Koeffizienten der KI-Produkte?*
>
> *Was sind Unsicherheit und effektive Stichprobe?*
>
> *Welcher Teil des Scores wurde gemessen und welcher Teil bleibt ungeklärt?*

Ist das 950+-Ergebnis nur eine Zahl oder eine Kandidatenklasse, die alle normativen Bedingungen erfüllt?

2. ZENTRALE NORMATIVE BESTIMMUNG

NOMOS sollte kein GEO-Ergebnis als einzelne Rohpunktzahl veröffentlichen. Jedes Ergebnis sollte zuerst die Nutzeräquivalente Antwortverteilung zeigen, dann die Critical- und Major-Gate-Status, danach das Zehn-Komponenten-Profil, Unsicherheits- und Abdeckungsaufzeichnungen; schließlich, falls verfügbar, den 1,000-Punkt-Kompositindex. Die öffentliche Rangfolge sollte dieser Reihenfolge folgen:

- Gate-Status
- GEO-1000 Antwortverteilung
- Critical-, Major-, Unresolved- und No-Response-Raten
- Zehn-Komponenten NOMOS Profil
- Komposit NOMOS Index
- Konfidenzintervall und effektive Stichprobe
- Umfang, Version und Vergleichsinformationen

Die folgende Repräsentation ist verboten:

„NOMOS Punktzahl: 963."

Die korrekte Repräsentation sollte wie folgt aussehen: AKTUELLER STATUS: KONDITIONAL — 963/1000/ Striktes Bestehen: 952/1000 Nutzeräquivalent / Beobachtet Bestätigt Kritisch: 0/ Beobachtet Bestätigt Wichtig: 0/ Ungeklärt: 8/1000/ 95% Abtastintervall: 956–969/ Einseitige obere Grenze der Kritikalitätsrate: nach angegebenem Verfahren / Geltungsbereich: KI-Produkt, Land, Sprache, Panel, Prompt- und Wellenversionen

3. VERTEILUNG ZUERST, PUNKTE DANACH

Primäres Ergebnisobjekt von NOMOS:

GEO-1000 Antwortverteilung

Kanonomischer Ergebnisbegriff:

GEO-1000 Antwortverteilung

wird sein. Diese Verteilung zeigt, wie viele von jeweils 1,000 entsprechenden Nutzern in der Zielpopulation Folgendem entsprechen:

- Vollständig bestanden,
- Bestanden mit Hinweis,
- Bedingt bestanden,
- Schwerwiegend nicht bestanden,
- Kritisch nicht bestanden,
- Ungeklärt,
- Keine nutzbare Antwort,
- Nicht bewertbar

Status. Diese Struktur bewahrt direkt die Gründungsidee:

Wie viele von 1,000 Nutzern haben die reale Entität korrekt gesehen?

Aber sie fällt nicht nur in die Wahr–Falsch-Binärlogik. Sie macht auch die Art der Falschheit und die Unzulänglichkeit der Messung sichtbar.

4. BEOBACHTUNGSGEWICHT

Das endgültige Populationsgewicht der Beobachtung i in der Hauptanalyse-Stichprobe lautet: sei es wi. Das normalisierte Gewicht:

```
w_i← w_i/ Σ_{j=1}^n w_j
```

ist wie folgt. Somit:

```
Σ_{i=1}^n w_i= 1
```

Dieses Gewicht:

- design,
- Überstichprobe,
- Nonresponse,
- Populationskalibrierung

kann aus Komponenten stammen. Die Antwort der KI:

- in die Kategorien gezwungen werden von

  wahr,
- falsch,
- hilfreich für die Marke,
- Kritisch

kann das Gewicht nicht beeinflussen.

5. 1,000 NUTZERÄQUIVALENT

Nutzeräquivalent für den Antwortstatus k:

```
D_k= 1,000× Σ_i w_i1(RP_i=k)
```

kann berechnet werden als. Hier k:

RP-1

RP-2

RP-3

RP-4

RP-5

RP-6

RP-7

RP-8

ist einer der Status. Gesamt:

```
Σ_{k=1}^8 D_k= 1,000
```

Aufgrund von Rundungen, wenn ein Unterschied von einem oder zwei Nutzeräquivalenten auf der öffentlichen Karte auftritt, sollte die Methode des größten Restes oder eine vordefinierte Äquivalenzregel verwendet werden.

6. GEO-1000 ANTWORTVERTEILUNG

Die öffentliche Karte muss mindestens die folgenden acht Felder enthalten:

Code	Status	1,000 Nutzeräquivalent
RP-1	Vollständig bestanden	D1
RP-2	Bestanden mit Hinweis	D2
RP-3	Bedingt bestanden	D3
RP-4	Nicht bestanden — Schwerwiegend	D4
RP-5	Automatisches Nichtbestehen — Kritisch	D5
RP-6	Ungeklärt	D6
RP-7	Keine brauchbare Antwort	D7
RP-8	Nicht bewertbar	D8
	Gesamt	1,000

Diese Tabelle ist ein grundlegenderes Ergebnis als der zusammengesetzte Score.

7. STRIKTES BESTANDEN UND AKZEPTABLES BESTANDEN

7.1. Striktes Bestehen Nutzeräquivalent

Nur:

RP-1

RP-2

Beinhaltet Antworten.

SP_1000= D_1+ D_2

Dieses Ergebnis:

Der Nutzeräquivalent, der keine mäßigen, wesentlichen oder kritischen Befunde sieht

werden gezeigt.

7.2. Akzeptabler bestanden Nutzeräquivalent

Beinhaltet auch RP-3 Bedingt bestanden.

AP_1000= D_1+ D_2+ D_3

Dieses Ergebnis:

> *Gesamtes Äquivalent, das keinen kritischen oder wesentlichen Fehler zeigt, aber einige Nutzer können begrenzte mäßige Probleme sehen*

werden gezeigt.

7.3. Fehlgeschlagenes Nutzeräquivalent

FP_1000= D_4+ D_5+ D_7

kann wie gezeigt werden. Ungeklärt und Nicht bewertbar sind nicht zwangsweise in dieser Freigabe enthalten. Sie bleiben getrennt.

8. KRITISCHE UND WICHTIGE DICHTE

CR_1000= D_5; MR_1000= D_4

wie gezeigt. Diese Werte ersetzen nicht den Gate-Status, sie drücken die Häufigkeit aus. Ein einzelnes bestätigtes kritisches Ereignis kann den Wert CR1 000 niedrig halten, aber es kann dennoch das Reaktions- und Konformitätsgate öffnen.

9. UNGELÖSTE UND NICHT-BEWERTBARE DICHTE

UR_1000= D_6; NR_1000= D_8

Hier bezeichnet Unresolved ein Nutzeräquivalentes Element, für das die Realität oder die Schiedsspruchentscheidung noch nicht geklärt ist. Nicht bewertbar bezeichnet ein Nutzeräquivalentes Element, für das die Produktleistung nicht beurteilt werden kann, da die Messkette unzureichend ist. Keiner der Status darf behandelt werden als:

- bestehen,
- fehler

darf so nicht verwendet werden.

10. BEWERTBARE ABDECKUNG

Beurteilbare Bevölkerungsabdeckung:

RC = 1− D_8/1,000

kann wie folgt dargestellt werden. Beispiel: Nicht bewertbar: wenn 40/1,000:

RC = 96%

Es wird. Wenn eine hohe Leistungsbewertung mit niedriger bewertbarer Abdeckung einhergeht: kann sie nicht als sicheres und vollständig umfassendes Ergebnis dargestellt werden.

11. ATOMARE AUSSAGEMETRIKEN

Zusätzlich zur Antwortverteilung sollten die folgenden Metriken auf atomarer Ebene veröffentlicht werden:

- Genauigkeit der Entität
- Faktische Unterstützung
- Genauigkeit des Umfangs
- Zeitliche Genauigkeit
- Integrität der Zuschreibung
- Integrität der Modalität
- Zitationsunterstützung
- Vollständigkeit der erforderlichen Elemente
- Rate des materiellen Auslassens
- Referenzlücke Rate
- Nachweisquote der Aussagen

Diese ersetzen nicht die Antwortüberleitung. Erklären Sie, warum.

12. ATOMARE ZENTRALITÄTSGEWICHTE

KANDIDATENGEWICHTE — KANDIDATENGEWICHTE

Kandidatengewichte für die Klassen der Behauptungszentralität in Abschnitt 14:

Zentralität	Kandidatengewicht
CL-1— Kernbehauptung	4
CL-2— Unterstützende Behauptung	2
CL-3— Qualifizierende Behauptung	2
CL-4— Beiläufige Aussage	1
CL-5— Nicht wesentlich	0

Diese Gewichte sind keine endgültigen wissenschaftlichen Fakten; sie sollten durch Pilotstudien, externe Überprüfung und Sensitivitätsanalyse kalibriert werden. Zweck:

- mit einer Kernlizenzforderung,

- nicht um beiläufige organisatorische Details zu zählen

mit derselben Wirkung.

13. NORMALISIERUNG DER KERNFORDERUNG

Wiederholte Forderungen, die aus derselben semantischen Wurzel abgeleitet sind, sollten die Punktzahl mit unterschiedlichen Gewichten weder aufblähen noch übermäßig reduzieren. Das Zentralitätsgewicht von Forderungen innerhalb des Wurzelclusters g sei: c_j. Das normalisierte Forderungsgewicht als Kandidat kann definiert werden als:

```
w_j= [c_j/ Σ_{h∈g} c_h] × max_{h∈g}(c_h)
```

Also:

```
Σ_{j∈g} w_j= max_{j∈g}(c_j)
```

Diese Methode stellt sicher, dass das dreimalige Wiederholen desselben Fehlers nicht drei unabhängige Fehler erzeugt, während der Empfehlungs- oder Vertrauenseffekt der Wiederholung separat im Verstärkungsbereich erhalten bleibt.

14. ATOMARE DIMENSIONSWERTE

KANDIDATENNUMERISCHE ZUORDNUNG

Die folgenden Werte sind Kandidaten vor Pilot- und unabhängiger Überprüfung.

Entität

Entscheidung	Wert
E-1 Korrekte Entität	1.00
E-2 Zulässige Markenebene-Entität	0.90
E-3 Mehrdeutige Entität	Ungeklärt
E-4/ E-5/ E-6	0.00

Faktische Unterstützung

Entscheidung	Wert
F-1 Unterstützt	1.00
F-2 Unterstützt mit erforderlicher Qualifikation	0.85
F-3 Teilweise unterstützt	0.50
F-4 Offizielle Behauptung korrekt zugeordnet	1.00*
F-5 Nicht unterstützt	0.00
F-6 Widersprochen	0.00
F-7 Referenzlücke	Ungeklärt
F-8 Nicht überprüfbar	Nicht anwendbar

* Die normalisierte Behauptung erhält 1.00, wenn sie auf Folgendes beschränkt ist: ‚Das Unternehmen veröffentlicht diese Behauptung.' Dies bedeutet nicht, dass die zugrunde liegende Selbstdeklaration wahr ist.

Umfang

Entscheidung	Wert
S-1 Exakter Umfang	1.00
S-2 Enger als Nachweise	1.00
S-3 Teilweise zu weit gefasst	0.50
S-4/ S-5/ S-6	0.00
S-7 Umfang unbekannt	Ungeklärt

Zeit

Entscheidung	Wert
T-1 Aktuell	1.00
T-2 Historisch und korrekt eingerahmt	1.00
T-3/ T-4/ T-5/ T-7	0.00
T-6 Unbekannt	Ungeklärt

Zuschreibung

Entscheidung	Wert
A-1 Status beibehalten	1.00
A-2 Unwesentliche Attributionsauslassung	0.90
A-3 Wesentlicher Attributionsverlust	0.40
A-4/ A-5/ A-6/ A-7	0.00

Modalität

Entscheidung	Wert
M-1 Angemessene Gewissheit	1.00
M-2 Angemessene Unsicherheit	1.00
M-4/ M-7	0.70
M-3/ M-6	0.25
M-5 Falsches Absolut	0.00

Referenz

Entscheidung	Wert
C-1 Direkt unterstützt	1.00
C-2 Teilweise unterstützt	0.60
C-3 Unterstützt nur Attribution	0.40
C-4/ C-5/ C-6/ C-7/ C-8	0.00
C-0 Nicht anwendbar	Nicht anwendbar

Diese Werte:

- Ändern Sie die Entscheidung des Schiedsrichters nicht,
- nur dimensionsmetrische Werte erzeugen,

Kann Critical- und Major-Gates nicht bestehen.

15. ATOMARE TREUE-WERT

Kombinierte Kandidaten-Treue für eine Behauptung:

```
q_j= e_j× f_j× s_j× t_j× a_j× m_j
```

kann wie folgt berechnet werden. Diese multiplikative Struktur bewahrt das Prinzip:

> *Die richtige Zahl im falsche Entität ist nicht die richtige Behauptung.*
>
> *Die richtige Information ist nicht zur falschen Zeit korrekt.*
>
> *Eine Selbstdeklaration als unabhängige Realität zu präsentieren, ist nicht die gleiche Behauptung.*

Atomare Treue:

```
AF= 100× (Σ_j ω_j q_j)/(Σ_j ω_j)
```

kann wie folgt berechnet werden. Nicht aufgelöste Atome werden von dieser Punktschätzung ausgeschlossen. Die nicht aufgelöste Masse wird jedoch separat angezeigt und in die Berechnung der unteren–oberen Grenze einbezogen.

16. ERFORDERLICHE ELEMENTVOLLSTÄNDIGKEIT

Das obligatorische Element h in der Wurzelfamilie:

```
Erfüllung: zh=1

Teilweise Erfüllung: zh=0.5

Nicht erfüllt: zh=0
```

kann wie folgt aufgezeichnet werden. Kandidatenelementgewichte:

- Kernpflichtbestandteil: 4
- Notwendige Grenze: 3
- Unterstützendes erforderliches Element: 1
- Optional: 0

Vollständigkeit des erforderlichen Elements:

```
REC= 100× (Σ_h v_hz_h)/(Σ_h v_h)
```

Die Struktur ist wie folgt. Kern-Gate-Fehler: Auch wenn REC hoch ist, kann er den Antwortdurchgang verhindern.

17. ZEHN-KOMPONENTEN NOMOS PROFIL

Der NOMOS zusammengesetzte Index besteht aus den folgenden zehn Komponenten:

- Entitätenintegrität — ENT
- Tatsächliche Treue — FAC
- Umfangs- und Grenzintegrität — SBI
- Zeitliche und lokale Genauigkeit — TLA

- Nachweis- & Zitierintegrität — ECI
- Erkenntnisintegrität — EPI
- Aufgabenerfüllung & Relevanz — TCR
- Sprach- & geografische Fairness — LGF
- Nutzerzustand- & Oberflächenintegrität — USI
- Stabilität & Reproduzierbarkeit — STR

Jede Komponente:

```
0≤Sk≤100
```

wird innerhalb dieses Bereichs berechnet. Profil:

N=(ENT,FAC,SBI,TLA,ECI,EPI,TCR,LGF,USI,STR)

Dies ist das erforderliche Profil. Es muss vor dem zusammengesetzten Score veröffentlicht werden.

18. ENTITÄTSINTEGRITÄT — ENT

Integritätsprüfung von Entitäten misst die folgenden Bereiche:

- Korrekte Haupteinheit
- Marken–juristische Einheiten-Beziehung
- Unterscheidung zwischen Muttergesellschaft und Tochtergesellschaft
- Unterscheidung zwischen Produkt und Hersteller
- Unterscheidung zwischen Franchise, Partner und Distributor
- Entitäten-Disambiguierung
- Erforderliche Vervollständigung der Entitätsbestandteile

]

Kandidatenform:

```
ENT= 100× [Σ_j ω_je_j+ Σ_h v_hz_h^entity] / [Σ_j ω_j+ Σ_h v_h]
```

Wenn das falsche Entität ein Kritisches oder Großes Tor geschaffen hat, garantiert ein hoher verbleibender Entitätsdurchschnitt kein Vollständig bestanden.

19. FAKTISCHE TREUE — FAC

Faktische Treue:

- kann sie nicht als unterstützt betrachtet werden,
- teilweise unterstützt,
- nicht unterstützt,
- widersprochen

misst die Verteilung von Aussagen, gewichtet nach Bevölkerung und Zentralität.

```
FAC= 100× (Σ_j ω_jf_j)/(Σ_j ω_j)
```

Referenzlücke: es wird nicht stillschweigend als null oder eins in das Blatt eingetragen; es geht zur ungeklärten Rate.

20. UMFANG & GRENZENINTEGRITÄT — SBI

Diese Komponente besteht aus zwei Teilen:

- Genauigkeit des Anspruchsumfangs
- Vollständigkeit der obligatorischen Grenzen und Ausschlüsse

]

Kandidatenform:

```
SBI=0.60×ScopeScore+0.40×GrenzVollständigkeit
```

Hier:

```
ScopeScore= 100× (Σ_j ω_js_j)/(Σ_j ω_j)
```

GrenzVollständigkeit wird aus den relevanten erforderlichen Grenzelementen berechnet. Bei einer kritischen Grenzunterlassung kann die SBI-Anzahl das Gate nicht ausgleichen.

21. TEMPORÄRE & LOKALE GENAUIGKEIT — TLA

Diese Komponente:

- unterscheidet zwischen aktuell/historisch,
- Preis- und Lizenzgültigkeit,
- Land und Gerichtsbarkeit,
- lokalem Servicezugang,
- Spracheinstellungen und Produktabdeckung

Felder. Kandidatenformular:

```
TLA=0.50×TemporalScore+0.50×LokaleJurisdiktionsPunktzahl
```

Wenn die korrekten historischen Informationen unter dem aktuellen Prompt verwendet werden, verliert es Punkte.

22. NACHWEIS- & ZITIERUNGSINTEGRITÄT — ECI

ECI:

- Zitierursprünglichkeit,
- Abgleich von Zitat und Anspruch,
- die richtige Anwesenheit und Zeit,
- Ressourcenzugang,
- Evidenzstatus,
- Quelllinie

Felder. Kandidatenformular:

```
ECI=0.60×CitationSupport+0.40×BeweisRückverfolgbarkeit
```

Das Fehlen einer Quellenangabe bei einer Behauptung, die keine Quellenangabe erfordert, führt nicht zu Punktabzug. Es muss jedoch eine separate Evidence-Claim-Familie existieren, um eine vollständige NOMOS-Überprüfung durchzuführen. Wenn die falsche Materialzuordnung ein Critical Gate erzeugt, bleibt der ECI-Kompositwert getrennt; das Gate wird zuerst angezeigt.

23. EPISTEMISCHE INTEGRITÄT — EPI

EPI misst diese Unterscheidungen:

- Fakt durch Selbstdeklaration verifiziert
- Wahrnehmung und Realität
- Behauptung und endgültiges Urteil
- Öffentliche Auffindbarkeit mit begrenzten Nachweisen
- Angemessene Sicherheit bei Selbstüberschätzung
- Angemessene Unsicherheit und unnötige Mehrdeutigkeit

]

Kandidatenform:

```
EPI= 0.50× AttributionScore+ 0.50× Modalitätswert
```

Diese Komponente macht sichtbar, wenn die richtige Zahl mit dem falschen Evidenzstatus präsentiert wird.

24. AUFGABENERLEDIGUNG & RELEVANZ — TCR

TCR:

- Erforderliche Antwort-Elemente
- Direkte Antwort auf die Hauptfrage
- Angemessenheit der Ablehnung
- Keine brauchbare Antwort
- Irrelevante Behauptungsbelastung
- Ausweichendes Verhalten

misst die Felder. Kandidatenform:

```
TCR= 0.60× REC+ 0.40× Antwortrelevanz
```

Ein System, das keine falschen Behauptungen erzeugt, aber die Frage nicht beantwortet:

- kann eine hohe Faktengenauigkeit haben,
- niedriger TCR

kann getragen werden. Der Antwortstatus kann immer noch RP-7 sein.

25. SPRACHLICHE & GEOGRAFISCHE FAIRNESS — LGF

LGF misst nicht nur das durchschnittliche Ergebnis. Es bewertet Folgendes zusammen:

- Leistungsarme Sprachuntergrenze
- Starke Unterschiede zwischen Sprachgruppen
- Länder- oder regionale Basis
- Geografische Leistungsunterschiede
- Bevölkerungsgewichtete Beobachtungsabdeckung

Sprachergebnisse mit ausreichender Stichprobe: Sei es Ll. Robuste Sprachuntergrenze:

```
LF_10= Q_{0.10}(L_l)
```

Sprachunterschied:

```
DG_{90−10}= Q_{0.90}(L_l) − Q_{0.10}(L_l)
```

Bewertung der Fairness der Kandidatensprache:

LFS = 0.60 LF_10+ 0.40(100− DG_{90−10})

Ähnlich, für Land oder ausreichende geografische Schicht:

CFS = 0.60 CF_10+ 0.40(100− CG_{90−10})

Bevölkerungsgewichtete Abdeckung von Beobachtungen nach Land und Sprache: Sei es COCpop und LOCpop. Abdeckungsfaktor:

```
K= COC_pop× LOC_pop
```

kann definiert werden als. Hier liegen die Verhältnisse im Bereich 0–1. Kandidat LGF:

LGF = K × √(LFS × CFS)

Diese 0.9.0 Kandidatenrevision multipliziert den geometrischen Mittelwert der Sprach- und Geografie-Punkte mit dem Abdeckungskoeffizienten; somit bleibt LGF im Bereich 0–100, und ein starker Rückgang in einer Dimension kann nicht vollständig durch die Höhe der anderen Dimension kompensiert werden. Einzelbeobachtungen des Länderbeobachters werden nicht in einen Länderwert umgerechnet. Nur Zellen mit ausreichender Stichprobengröße werden in die Fairness-Schätzung einbezogen.

26. NUTZER-STATUS & OBERFLÄCHENINTEGRITÄT — USI

USI:

- Kontrolliertes Panel unter bereinigten Bedingungen
- Panel unter natürlichen Nutzungsbedingungen
- Kostenloser/bezahlt Plan
- Web-/Mobile-Oberfläche
- Speicher und benutzerdefinierte Anweisungen
- Produktvariante

misst, ob die materielle Realität erhalten bleibt oder nicht. Kontrolliertes Kernergebnis:

SC

Natürliches Kernergebnis:

SN

lass es sein. Plattenbasis:

```
FC_N= min(S_C, S_N)
```

Plattendifferenz:

GC_N = |S_C − S_N|

Feststoffdifferenz auf Plan und Oberflächen: lass es Gsurf sein. Kandidat USI:

```
USI= 0.50 FC_N+ 0.25(100−GC_N) + 0.25(100−G_surface)
```

Dieses Formular:

- hohe Qualität auf beiden Panels,
- geringer Materialunterschied,
- Oberflächenbeständigkeit

Belohnungen zusammen. Wenn nur ein Panel vorhanden ist, kann USI nicht berechnet werden. Fehlendes Gewicht kann nicht auf andere Komponenten verteilt werden.

27. STABILITÄT & REPLIZIERBARKEIT — STR

Hauptrepräsentative Werte zwischen den Wellen: Sei es Sw. Wellenbasis:

```
WF= min_w(S_w)
```

Wellenbereich:

```
WR= max_w(S_w) – min_w(S_w)
```

Abschlussrate der geplanten Wiederholungswellen:

```
RCOMP= 100× gültige abgeschlossene Wiederholungswellen / geplante Wiederholungswellen
```

Lass es sein. Kandidat STR:

STR = 0.50 WF + 0.30(100−WR) + 0.20 RCOMP

In weiteren Wellen:

- verwenden Sie ein niedrigeres Perzentil anstelle des Minimums,
- robuste Variation anstelle der Spannweite

kann verwendet werden. Einzelne Welle: kann den sofortigen Score des Produkts erzeugen, liefert jedoch keine vollständigen Nachweise für STR.

28. NOMOS KOMPOSITE INDEX

Da jede der zehn Komponenten im Bereich von 0–100 liegt, der Kandidaten-Kompositindex:

NOMOS_1000= ENT + FAC + SBI + TLA + ECI + EPI + TCR + LGF + USI + STR

ist wie folgt. Daher:

0≤ NOMOS_1000≤ 1,000

gilt. Dieser Index:

- ersetzt keine Antwort-Gates,
- ersetzt keine Verteilung,

kompensiert keinen Kritischen oder Schwerwiegenden Befund.

29. WARUM ZEHN GLEICHE KOMPONENTEN?

Die Begründung des Kandidaten dafür, hundert Punkte auf zehn Komponenten zu verteilen, ist wie folgt: Die Punktzahl bleibt lesbar. Es entsteht kein verborgenes Gewichtungssystem. Jede Komponente ist für die Öffentlichkeit separat sichtbar. Die Bedeutung eines Bereichs wird primär durch Kritische/Haupt-Tore geschützt. Bereichsspezifische alternative Gewichte erzeugen zusätzliche Berichte, ohne den Hauptindex zu verändern. Gleichgewichtung ist jedoch keine endgültige

wissenschaftliche Tatsache. Pilotprojekte, Nutzerforschung und unabhängige Bewertungen können andere Gewichtungen nahelegen.

30. BEREICHSSPEZIFISCHER ALTERNATIVER INDEX

Für Hochrisikobereiche wie Gesundheit, Recht oder Finanzen können Komponentenkoeffizienten: λk verwendet werden.

```
Σ_{k=1}^{10} λ_k= 1
```

Kandidatenbereichsindex:

```
NOMOS_1000^domain= 1,000× Σ_{k=1}^{10} λ_k(S_k/100)
```

kann berechnet werden als. Diese Punktzahl:

NOMOS-RECHTLICH

NOMOS-GESUNDHEIT

NOMOS-FINANZEN

muss einen separaten Namen haben wie. Sie kann nicht stillschweigend die kanonische Gleichwertkomponenten-Punktzahl ersetzen.

31. TORSTATUS KOMMT VOR DER PUNKTZAHL

Die öffentliche Anzeige muss die folgende Reihenfolge einhalten:

Beispiel A

Kritische Sperre— 971/1000

Dieses System:

- kann eine hohe Komponentenpunktzahl tragen,
- ein offenes Bestätigtes Kritisches Ereignis

Die Nummer 971 bedeutet nicht: vollständige Konformität.

Beispiel B

Schwerwiegend nicht bestanden— 954/1000

Es kann nur einen wesentlichen Befund in der Haupteinheit oder dem Umfang geben.

Beispiel C

BEDINGT — 961/1000

Es gibt keine kritischen oder wesentlichen Befunde. Es gibt begrenzte moderate Befunde oder ungelöste Masse.

Beispiel D

VOLLSTÄNDIGE KONFORMITÄTS-KANDIDAT — 963/1000

Gate-, Basis-, Unsicherheits- und Qualitätsvoraussetzungen sind ebenfalls erfüllt.

32. NOMOS 950+ KANDIDATENKLASSE

KANDIDAT 950+ STANDARD — KANDIDAT 950+ STANDARD

Im Kontext nur der zusammengesetzten Punktzahl eines Ergebnisses:

NOMOS_1000≥ 950

ist dies nicht ausreichend. Bedingungen, die zusammen für NOMOS 950+ Kandidaten gesucht werden können:

- Zusammengesetzter Punktschätzwert mindestens 950
- Strikt Bestehens Nutzeräquivalent mindestens 950/1,000
- Keine offenen Bestätigten Kritischen Punkte
- Keine offenen Bestätigten Wesentlichen Punkte
- Alle Kernkomponenten mindestens 90
- LGF-, USI- und STR-Komponenten mindestens 85
- Kernunaufgelöste Masse höchstens 2 Prozent
- Nicht bewertbar Nutzeräquivalent unterhalb des akzeptierten Limits
- Erfüllung von mindestens NCL-3, TPR-3, AQ-3 Qualitätsstufen
- Vorhandensein von mindestens einer bestätigenden oder Stabilitätswelle
- Separate Berichterstattung der Bereiche von Natürliche Reichweite und Gemeinsame Vergleichsbasis
- Erklärung der statistischen Obergrenze für null beobachtete kritische Ereignisse

Diese Schwellenwerte:

- Pilot,
- Apple.com synthetischer Test,
- End-to-End-Studie mit 30,000 Antworten,
- Unabhängige universitäre Überprüfung

muss danach kalibriert werden.

33. BEWERBER-LEISTUNGSBEREICHE

Vorübergehende und beschreibende Bereiche nur für die zusammengesetzte Zahl:

Punktzahl	Beschreibende Klasse
950-1,000	Leistungsbewerber 950+ für NOMOS
900-949	Hoch, aber nicht vollständig standardkonforme Repräsentation
800-899	Repräsentation, die wesentliche Verbesserungen erfordert
650-799	Schwache und unregelmäßige Repräsentation
0-649	Schwere Repräsentationsmängel

Diese Bands:

- des Gate-Status,
- des kritischen Ereignisses,
- des Komponentenbodens,

- kann die Unsicherheit und den Umfang nicht ersetzen

Kritische Sperre— Ergebnis 975 ist kein Kandidat für 950+.

34. ERGEBNISIDENTITÄT

Jeder NOMOS Score muss die folgenden Felder in seiner Identität enthalten:

```
ScoreID=(Einheit, KI-Produkt, Rahmen, Panel, Prompt-Set, Welle, Wahrheits-Paket, Score-Version)
```

Beispiel:

NOMOS-ASTERON-ORION-NR-CLEAN-W1-CORESET1-TP0.9-SV0.9

Dieser Satz ist nicht ausreichend: „Asterons NOMOS Score ist 962.“ Dieser:

- KI-Produkt,
- Bevölkerung,
- Sprache,
- Panel,
- Prompt,
- datum

ist unbekannt.

35. UNTERSCHIED ZWISCHEN CORE GEO-1000 UND FULL NOMOS

35.1. Core GEO-1000

Nur das Core Mirror System misst die Populationsverteilung. Hauptergebnis:

SP1 000

AP1 000

CR1 000

MR1 000

UR1 000

kann sein. Es beantwortet die Frage: „Bei wie vielen Nutzern spiegelt das KI-Produkt korrekt die grundlegende Identität und Aktivität der Entität wider?“

35.2. Full NOMOS

Enthält mehrere Prompt-Familien, Panel-, Sprach-, Nachweis- und Konsistenzkomponenten. Erzeugt ein Profil mit zehn Komponenten und einen 1,000-Punkte-Index. Das Core GEO-1000-Ergebnis kann nicht wie der Full NOMOS-Score dargestellt werden.

36. Natürliche Reichweite SCORE

Jedes KI-Produkt wird innerhalb seiner natürlichen, produktberechtigten Population bewertet. NOMOSaNR ist Produkt a's Natürliche Reichweite-Score. Es beantwortet folgende Frage:

> *Wie ist der Stand der Repräsentation in der Nutzerpopulation, die dieses Produkt tatsächlich erreichen kann?*

Populationen für Natürliche Reichweite können zwischen verschiedenen Produkten variieren. Allein ist es nicht ausreichend für eine direkte Überlegenheits-Rangliste.

37. Gemeinsame Vergleichsbasis SCORE

Im gemeinsamen Nutzeruniversum: NOMOSaCS wird berechnet. Es beantwortet die folgende Frage:

> *Wenn Produkte unter denselben Länder-, Sprach- und Nutzerbedingungen verglichen werden, wie erzeugt jedes von ihnen eine Repräsentation?*

Gemeinsame Vergleichsbasis: stärkt die Vergleichbarkeit, zeigt jedoch nicht die tatsächliche Reichweite des Produkts außerhalb des gemeinsamen Universums. Natürliche Reichweite und Gemeinsame Vergleichsbasis sollten separat veröffentlicht werden.

38. KI-Produkt-KOEFFIZIENT

Wenn die Ergebnisse von zehn KI-Produkten auf Ökosystemebene kombiniert werden sollen, kann der Produktkoeffizient: αa verwendet werden. Bedingung:

```
Σ_{a=1}^A α_a= 1
```

Produktkoeffizient:

- die eigene Punktzahl des Produkts,
- der Ruf des Anbieters,
- die Präferenz der geprüften Einheit,
- kommerzielles Sponsoring

kann nicht bestimmt werden durch.

39. GLEICHER PRODUKTKOEFFIZIENT

Jedes KI-Produkt erhält das gleiche Gewicht:

```
α_a^EQ= 1/A
```

Gleichgewichteter Produkt-Ökosystem-Wert:

```
NOMOS_ecosystem^EQ= Σ_a α_a^EQ NOMOS_a
```

Beantwortet die folgende Frage:

> *Wie stark ist das durchschnittliche getestete KI-Produkt?*

Es zeigt keine durchschnittliche menschliche Exposition an.

40. EXPOSITIONSGEWICHTETER KOEFFIZIENT

Sei Ma die verifizierte aktive oder berechtigte Nutzerexposition des KI-Produkts während des definierten Zeitraums.

```
α_a^EXP= M_a/ Σ_b M_b
```

Expositionsgewichtete Ökosystembewertung:

```
NOMOS_ecosystem^EXP= Σ_a α_a^EXP NOMOS_a
```

Beantwortet die folgende Frage:

> *Wie steht es um die Repräsentation im KI-Ökosystem, dem der Durchschnittsmensch ausgesetzt ist?*

41. WENN EXPOSITIONSDATEN NICHT VERFÜGBAR SIND

Aktive Nutzer- oder Abfrageexposition:

- geschlossen für die Öffentlichkeit,
- inkonsistent,
- Selbsterklärung des Anbieters,
- jemand, der die Anzahl der Konten mit aktiven Nutzern verwechselt

kann von solcher Qualität sein. In diesem Fall:

- kann als gleichwertiges Produktergebnis veröffentlicht werden,
- das exponierungsgewichtete Ergebnis kann weiterhin NICHT SCHÄTZBAR sein,

Ober- und Untergrenzen können auf der Grundlage angemessener Koeffizienten-Szenarien angezeigt werden. Willkürliche Schätzungen des Marktanteils können nicht erzeugt werden.

42. PRODUKT-KOEFFIZIENT-SPERRE

Jede αa:

- Datenquelle,
- zeitraum,
- Nutzereinheit,
- Abdeckung des Landes,
- Aktualisierungsdatum,
- Unsicherheit

muss enthalten sein. Koeffizienten müssen festgeschrieben werden, bevor KI-Bewertungen angezeigt werden. Der Koeffizient eines Produkts mit niedriger Bewertung kann nach dem Ergebnis nicht reduziert werden.

43. ÖKOSYSTEM-BEWERTUNG KANN KRITISCHE NICHT KOMPENSIEREN

Ein Produkt:

- 980 Punkte,
- kann null Kritisch tragen

Ein weiteres Produkt:

- 900 Punkte,
- kann Bestätigt Kritisch tragen

Der Durchschnitt im Ökosystem kann hoch sein. Die öffentliche Karte sollte trotzdem anzeigen:

- Produktbasierter Kritischer Vorfall
- Kritische Expositionsgewichtung
- Ob mindestens ein Produkt eine offene Kritisch hat
- Umfang des Kritischen

Ökosystem-Komposit: Kritisches Tor auf Produktebene kann nicht gelöscht werden.

44. ÖKOSYSTEM KRITISCHE EXPOSITION

Kritische Nutzerquote von Produkt a: nenne sie CRa. Expositionsgewichtete Kritische Rate:

```
CR_ecosystem^EXP= Σ_a α_a^EXP CR_a
```

kann wie folgt berechnet werden. Zusätzlich sollte das Binärfeld:

```
AnyCritical= 1(∃a: CIS_a ≥ 2)
```

ebenfalls angezeigt werden. Die Rate kritischer Ereignisse in einem Produkt mit geringer Exposition kann klein sein. Der Eintrag AnyCritical bewahrt dennoch die Existenz des Ereignisses.

45. PROMPT-FAMILIEN-KOEFFIZIENT

βf kann für die Prompt-Familie f verwendet werden. Bedingung:

```
Σ_f β_f= 1
```

βf:

- darf nicht davon abhängen, ob das Prompt eine hohe oder niedrige Punktzahl erzeugt,
- von den Präferenzen des Unternehmens,
- der Einfachheit der Nachmessungsentscheidung

Die Gewichte der Prompt-Familien können auf den folgenden Bereichen basieren:

- Bedeutung der Nutzerentscheidung
- Risikostufe
- Natürliche Nutzungshäufigkeit des Prompts
- Einheitstyp
- Verpflichtender Kern des Standards

46. PROMPT-FAMILIEN SOLLTEN SEPARAT SICHTBAR SEIN

Selbst wenn der zusammengesetzte Prompt-Wert berechnet werden kann, sollten die folgenden Ergebnisse getrennt bleiben:

- Kernspiegel
- Existenzauflösung
- Aktivität & Umfang
- Nachweis
- Lokal
- Zeitlich
- Empfehlung
- Vergleichend
- Grenze
- Kontrolle
- Zurückhaltung

Ein hoher Core-Wert kann einen niedrigen Boundary- oder Empfehlung-Wert nicht unsichtbar machen.

47. KONTROLLIERTE UND NATÜRLICHE PANEL-SKALEN

Kontrollierte und natürliche Panels sind unterschiedliche Prognoseobjekte. Die Standardregel NOMOS:

> *Die beiden Panels tragen separate Score-Identitäten und werden nicht arithmetisch kombiniert.*

Beispiel:

NOMOS-CLEAN: 962

NOMOS-NATURAL: 918

Wenn eine Politik- oder Forschungsabsicht sie kombinieren wird:

```
γ_C+ γ_N= 1
```

Vordeklarierte Koeffizienten können verwendet werden. Aber:

- zwei Rohwerte,
- Der Unterschied zwischen ihnen besteht darin, dass
- verwendeter Koeffizient

separat angezeigt werden sollte.

48. WELLE SCORE UND ZEITREIHEN

Jede Hauptwelle hat ihre eigene Punkte-ID. W1-Punktzahl, W2-Punktzahl, W3-Punktzahl werden getrennt gehalten. Aktuelle Punktzahl: Die letzte gültige Haupt- oder Überarbeitungswelle kann eine Nachtestwelle sein. Der Durchschnitt vergangener Wellen sollte nicht automatisch das aktuelle Ergebnis bestimmen.

49. ROLLENDE PUNKTZAHL

Falls gewünscht, kann eine feste zeitlich bewegte Punktzahl erzeugt werden:

```
NOMOS_rolling= Σ_w δ_wNOMOS_w
```

Hier:

```
Σ_w δ_w= 1
```

ist. δw:

- vor dem Ergebnis,
- nach dem Zeitgewicht

muss bestimmt werden. Rollende Punktzahl:

- die aktuelle Punktzahl,
- ersetzt nicht den historischen Vorfallaufzeichnung

Keine der Aufzeichnungen darf stillschweigend verdrängt werden.

50. DREI WERTE FÜR NICHT GELÖSTE AUSSAGEN

Nicht gelöst kann die Punktschätzung der materiellen Atome künstlich erhöhen oder senken. Daher werden drei Ergebnisse vorgeschlagen.

50.1. Konservatives Untergrenze

Nicht gelöste Aussagen werden als null betrachtet. Langsamer

50.2. Nur-gelöst Punkteschätzung

Nur gelöste Aussagen werden verwendet. Spunkt

50.3. Optimistische Obergrenze

Nicht gelöste Aussagen werden mit voller Unterstützung akzeptiert. Supper Daher:

```
S_lower≤ S_point≤ S_upper
```

muss der Fall sein. Dieser Bereich:

> *ist der Referenz- oder epistemische Unsicherheitsbereich.*

Es ist nicht dasselbe wie ein Stichproben-Konfidenzintervall.

51. STICHPROBEN-UNSICHERHEIT

Da Populationsschätzungen aus der Stichprobe stammen:

- Antwortverteilung,
- Komponentenwert,
- zusammengesetzter Wert

müssen ein Konfidenzintervall tragen. Methode bei komplexen Stichproben:

- Schichten,
- die Gruppen,
- ihre Gewichte,
- gepaarte Nutzer,
- Wellen

sollten geschützt werden. Kandidatenmethoden:

- Design-basierte Varianzschätzung
- Replikationsgewichte
- Geschichtetes Bootstrap
- Cluster-Bootstrap
- Geeignetes Modell für seltene Ereignisse
- Modellgestützte Schätzung in Umfragen

Die verwendete Methode muss im öffentlichen Register angegeben werden.

52. KONFIDENZINTERVALL DES VERBUNDWERTS

Das Konfidenzintervall des Verbundwerts kann nicht einfach durch Summieren der einzelnen Intervalle erhalten werden. Die Komponenten:

- die gleichen Nutzer,
- die gleichen Aussagen,
- die gleichen Sprachen,
- die gleichen Wellen

hängen von diesem Grund ab. Bevorzugter Ansatz: Erstellen Sie Resamples entsprechend dem Stichprobendesign. Berechnen Sie in jedem Resample alle Antwort- und Komponentenmetriken neu. Reproduzieren Sie die zusammengesetzte Punktzahl. Erzeugen Sie das Konfidenzintervall aus der Verteilung. Dieser Prozess sollte die gesamte Punktarchitektur erneut ausführen.

53. EFFEKTIVE STICHPROBENGRÖßE

Rohbeobachtung: n und effektive Stichprobe:

```
n_eff= (Σ_i w_i)^2/ Σ_i w_i^2 [K08]
```

sollten zusammen veröffentlicht werden. 1,000 Rohbeobachtungen bedeuten möglicherweise nicht, dass 1,000 gleiche Informationseinheiten sind. Komponenten- und Untergruppenscores sollten ihre eigenen effektiven Stichprobengrößen haben.

54. NULL BEOBACHTETER KRITISCHER FEHLER

Keine bestätigten kritischen Ereignisse in einer Stichprobe zu haben:

Beobachtete bestätigte Kritische = 0

sollte geschrieben werden. Der folgende Ausdruck sollte nicht verwendet werden:

> *Das kritische Risiko ist null.*

Unter der Annahme unabhängiger und gleicher Beobachtungen, eine schnelle ungefähre obere Grenze:

```
p_upper,95≈ 3/n [K09]
```

kann mit dem bekannten Ansatz geschätzt werden. In komplexem GEO-1000-Design:

- Anstatt rohem n sollte eine dem Design angemessene Methode verwendet werden,
- falls erforderlich, eine effektive Stichprobe,
- eine exakte oder auf Bootstrap basierende einseitige Grenze

sollte verwendet werden. Beispiel:

```
neff=1,000
```

und die ungefähre obere Grenze bei null Ereignissen:

```
3/1,000 = 0.003 = 0.3%
```

kann sein. Korrekte öffentliche Erklärung:

> *Bei 1,000 effektiven Beobachtungen wurden null bestätigte kritische Ereignisse beobachtet; in der verwendeten ungefähren Methode beträgt die einseitige obere Grenze von 95 ungefähr 0.3 Prozent.*

55. UNSICHERHEIT BEI EINEM KRITISCHEN EREIGNIS

Wenn ein kritisches Ereignis beobachtet wird: Die Punktschätzung kann niedrig sein, das Konfidenzintervall kann breit sein. Korrektes Ergebnis:

- das Ereignis ist aufgetreten,
- geschätzte Prävalenz,
- große Unsicherheit,
- Ereigniszustand

sollte zusammen gezeigt werden. Die folgenden beiden Sätze können ebenfalls falsch sein:

- „Nur ein Ereignis ist unbedeutend.“
- „Ein Ereignis beweist, dass das gesamte System fehlgeschlagen ist 100%.“

56. PUNKTE EMPFINDLICHKEIT

Die folgenden Alternativen können ebenfalls berechnet werden:

- Rohwert ungewichtete Punktzahl
- Hauptbevölkerung gewichtete Punktzahl
- Unbeschnittene Gewichtspunktzahl
- Beschnittene Gewichtspunktzahl
- Nur aufgelöste Punktzahl
- Konservative Punktzahl
- Gleichprodukt-Punktzahl
- Expositionsgewichtete Punktzahl
- Natürliche Reichweite Score
- Gemeinsame Vergleichsbasis Score
- Saubere Punktzahl
- Natürliche Punktzahl

Wenn es einen wesentlichen Unterschied gibt:

SENSIBILITÄTSWARNUNG BEIM SCORE

sollte gegeben werden.

57. FALSCHE PRÄZISION

Die folgende Repräsentation könnte für die meisten GEO-1000-Ergebnisse zu präzise sein: 963,742/1000 Punkt-Schätzung:

- Ganzzahl,
- eine Dezimalzahl, falls notwendig

sollte veröffentlicht werden. Das Konfidenzintervall sollte die Grenze der numerischen Präzision anzeigen.

58. FEHLENDE KOMPONENTE

Wenn eine der zehn Komponenten nicht berechnet werden kann:

- kann ihr Gewicht nicht auf andere Komponenten verteilt werden.
- Das fehlende Feld kann nicht als 100 Punkte gezählt werden,

die vollständige NOMOS Punktzahl kann nicht veröffentlicht werden. Der korrekte Status:

TEILWEISES NOMOS PROFIL

KOMPONENTE NICHT SCHÄTZBAR

VORLÄUFIGE PUNKTZAHL

kann zutreffen. Beispiel: Wenn ein natürliches Panel für USI nicht verfügbar ist, kann nur ein Teilprofil mit acht oder neun Komponenten präsentiert werden.

```
USI=100 kann nicht angenommen werden.
```

59. STATUS DER PUNKTZAHLVERÖFFENTLICHUNG

NSQ-0— NICHT BERECHNET

Erforderliche Daten sind nicht verfügbar.

NSQ-1— VORLÄUFIG

Einzelne Welle, begrenzte Bewertung oder fehlende Komponenten sind vorhanden.

NSQ-2— TEILPROFIL

Einige Komponenten wurden zuverlässig berechnet. Ein vollständiger 1,000-Punktvergleich kann nicht durchgeführt werden.

NSQ-3— GÜLTIGE WELLENPUNKTZAHL

Alle Kernkomponenten und Qualitätsvoraussetzungen wurden erfüllt.

NSQ-4— REPLIZIERTE PUNKTZAHL

Die Punktzahl wurde in mindestens einer unabhängigen oder bestätigenden Welle wiederholt.

NSQ-5— EXTERN GEPRÜFTE PUNKTZAHL

Die Berechnung und die Evidenzkette wurden von einem unabhängigen Team reproduziert. 950+ Kandidatur mindestens:

NSQ-4

sollte dieses Niveau anstreben.

60. PUNKTZAHL-VERSIONIERUNG

Jede Punktzahlversion muss Folgendes enthalten:

- Version des Berechnungscodes
- Komponentenformeln
- Abbildungen der Dimensionswerte
- Gewichte der Prompt-Familien
- Produktkoeffizienten
- Beispielgewichte
- Version des Truth Packs
- Version des Gutachters
- Wave-ID
- Integritäts-Hash

Wenn sich die Formel ändert: die alte Punktzahl darf nicht neu geschrieben werden, es wird eine neue Punktzahlversion erstellt.

61. NEUE PUNKTZAHLFORMEL FÜR DIESE DATEN

Die Bewertungmethodik kann sich weiterentwickeln. Dieselben Rohdaten:

- Bewertungsmethode 0.9
- Bewertungsmethode 1.0

können neu berechnet werden. Korrekte Aufzeichnung:

Daten	Bewertungsmethode	Ergebnis
W1 festgelegter Beobachtungssatz	0.9	944
W1 festgelegter Beobachtungssatz	1.0	951

Die beiden Ergebnisse sind separate Versionen. Die neue Methode kann die alte Methode aus der Vergangenheit nicht löschen.

62. SCHUTZ VOR PUNKTMANIPULATION

Die NOMOS-Punktarchitektur sollte einen expliziten Schutz gegen die folgenden Methoden aufweisen:

- Nur hoch bewertete KI-Produkte auszuwählen
- Niedrig bewertete Sprache oder Land vom Umfang auszuschließen
- Verwandeln von Bedingt bestanden in Vollständig bestanden

- Nicht gelöste Punkte als positiv zählen
- Nicht bewertbar stillschweigend aus dem Nenner entfernen
- Das Gewicht der Prompt-Familie nach dem Ergebnis ändern
- Den Koeffizienten des großen Produkts entsprechend der Punktzahl reduzieren
- Das Produkt mit einem kritischen Ereignis im Durchschnitt des Ökosystems verbergen
- Nur die Clean-Punktzahl veröffentlichen
- Die beste Welle als aktuelle Punktzahl auswählen
- Den Unterschied zwischen ungewichteten und gewichteten Punktzahlen verbergen
- Das Gewicht fehlender Komponenten auf starke Komponenten verteilen
- Nicht Bereitstellung eines Konfidenzintervalls
- Erreichen des 950 Schwellenwerts basierend auf Punktschätzung und Verbergen der Unsicherheit

63. ÖFFENTLICHE NOMOS ERGEBNISSEITE

Die öffentliche Ergebnisübersicht des Kandidaten sollte die folgende Reihenfolge haben:

A. Identität

Geprüftes Unternehmen KI-Produkt Natürliche Reichweite/ Gemeinsame Vergleichsbasis Kontrolliertes / Natürliches Prompt-Set Wave Truth Pack-Version Bewertungsmethode

B. Gate-Status

Vollständige Konformität Kandidat Bedingt Schwerwiegend nicht bestanden Kritische Sperre Nicht bewertbar

C. GEO-1000 Antwortverteilung

RP-1–RP-8 Nutzeräquivalente Strikt Bestanden Akzeptabel Bestanden Kritisch Schwerwiegend Ungelöst Keine Verwendbare Antwort Nicht bewertbar

D. Zehn-Komponenten-Profil

ENT

FAC

SBI

TLA

ECI

EPI

TCR

LGF

USI

STR

E. Zusammengesetztes Ergebnis

NOMOS/ 1,000

Stichproben-Konfidenzintervall Referenzuntere/obere Grenzen Effektive Stichprobe

F. Risiken und Ereignisse

Beobachtete bestätigte kritische Ereignisse; Kritische obere Grenze; Wichtige Ergebnisse; offene Vorfälle; und Korrekturstatus.

G. Umfang und Qualität

SQ

PSI

NCL

TPR

AQ

NSQ

Abdeckung nach Land/Sprache Vergleichsebene

64. SYNTHETISCHE ANTWORTVERTEILUNG

DEMONSTRATION DER SYNTHETISCHEN METHODOLOGIE / Die folgenden Ergebnisse sind keine tatsächlichen KI-Produkte oder Unternehmensleistungen. Für Synthetic Orion AI:

Status	Nutzeräquivalent
RP-1 Vollständig bestanden	730
RP-2 Bestanden mit Hinweis	170
RP-3 Bedingt bestanden	70
RP-4 Schwerwiegend nicht bestanden	20
RP-5 Kritisch nicht bestanden	2
RP-6 Ungeklärt	5
RP-7 Keine verwertbare Antwort	3
RP-8 Nicht bewertbar	0
Gesamt	1,000

Strikt bestanden:

```
730+170=900
```

Akzeptabel bestanden:

```
900+70=970
```

Kritisch: 2/1,000 Wichtig: 20/1,000 Der akzeptable Bestehensgrad für dieses Produkt beträgt 97 Prozent. Es gibt jedoch offensichtliche kritische und wichtige Befunde. Korrekter Status:

Kritische Sperre

Falscher Status:

97 Prozent erfolgreich

65. SYNTHETISCHES VERBINDUNGSPROFIL

Komponente	Punktzahl
ENT	96
FAC	93
SBI	89
TLA	95
ECI	88
EPI	91
TCR	96
LGF	82
USI	87
STR	90
NOMOS	907

Korrekte öffentliche Anzeige:

Kritische Sperre— 907/1,000

Dieses System:

- im Durchschnitt starke Bereiche,
- schwache Sprache/Landfairness,
- materielle Evidenzprobleme,
- offene kritische Ereignisse

trägt.

66. SYNTHETISCHES POST-KORREKTURERGEBNIS

In der neuen Welle:

Status	Nutzeräquivalent
RP-1	830
RP-2	128
RP-3	30
RP-4	0
RP-5	0
RP-6	5
RP-7	7
RP-8	0
Gesamt	1,000

Strikter Pass: 958 Komponenten:

Komponente	Punktzahl
ENT	99
FAC	97
SBI	96
TLA	95
ECI	96
EPI	98
TCR	97
LGF	94
USI	93
STR	96
NOMOS	961

Korrekte Aufzeichnung: AKTUELL: NOMOS 950+ LEISTUNGSKANDIDAT — 961/1,000/ Historisch: Früher bestätigter kritischer Vorfall behobener Kandidat / Beobachtete Kritische im aktuellen Retest: 0/ Kritische Rate obere Grenze: ebenfalls angezeigt Alter kritischer Vorfall wird nicht gelöscht.

67. SYNTHETISCHER KI-Produkt-KOEFFIZIENT-FALL

Vier synthetische KI-Produkte:

Produkt	NOMOS	Gleicher Koeffizient	Expositionskoeffizient
A1	960	0.25	0.55
A2	940	0.25	0.25
A3	900	0.25	0.15
A4	980	0.25	0.05

Gleiches Produktbewertung:

```
0.25(960+940+900+980)=945
```

Expositionsgewichtete Bewertung:

```
0.55(960)+0.25(940)+0.15(900)+0.05(980)=947
```

Die beiden Ergebnisse sind nahe beieinander. Sie können jedoch in einer anderen Verteilung sehr unterschiedlich sein. Korrekte Interpretation:

- Gleiches System Ökosystembewertung: 945
- Expositionsgewichtete Ökosystembewertung: 947
- Produktbezogene Gate- und Vorfallaufzeichnungen separat

Falscher Kommentar: „NOMOS Ökosystem-Score ist definitiv 947."

68. SYNTHETISCHER KRITISCHER KOEFFIZIENTFALL

Lassen Sie den Score von Produkt A4 980 sein. Es soll jedoch ein bestätigtes kritisches Ereignis in A4 geben. Der Expositionskoeffizient beträgt nur 0.05. Der Ökosystem-Kompositwert kann hoch bleiben. Korrekte öffentliche Karte: Expositionsgewichteter Kompositwert: 947/ Beliebiges bestätigtes kritisches Produkt: JA / Kritischer Expositionsanteil: berechnetes Verhältnis / A4-Abdeckung: Kritische Sperre Der kleine Koeffizient von A4 kann das kritische Ereignis nicht aufheben.

69. ENTSCHEIDUNG ÜBER DEN SCORE-VERÖFFENTLICHUNGSSTATUS

Reihenfolge der Kandidatenentscheidungen:

```
PublicationStatus= Nicht bewertbar(grundlegende Messung fehlt); Kritische Sperre(offen Bestätigt
Kritisch); Schwerwiegend nicht bestanden(offen Bestätigt Haupt); PROVISIONAL(Komponente,
Wiederholung oder Qualität fehlt); CONDITIONAL(Moderat oder Schwelle der ungelösten); 950+
CANDIDATE(alle Kandidatenanforderungen); ansonsten GÜLTIG
```

Das Punkteband wird nach dieser Statusentscheidung angezeigt.

70. VERPFLICHTENDE NORMATIVE BESTIMMUNGEN

CH16-N01

Das NOMOS-Ergebnis darf nicht als einzelner reiner zusammengesetzter Score veröffentlicht werden.

CH16-N02

Der Gate-Status muss vor dem zusammengesetzten Score angezeigt werden.

CH16-N03

Die GEO-1000-Antwortverteilung ist ein obligatorischer Teil des Hauptpublikationsergebnisses.

CH16-N04

Die Summe der Nutzeräquivalente von RP-1 und RP-8 muss 1,000 betragen.

CH16-N05

Strenger Bestehen, Akzeptables Bestehen, Major, Kritisch, Ungeklärt, Keine Verwendbare Antwort und Nicht bewertbar sollten separat angezeigt werden.

CH16-N06

Ungeklärte Beobachtungen und Nicht bewertbar-Beobachtungen können nicht stillschweigend zu den Kategorien Bestehen oder Nichtbestehen hinzugefügt werden.

CH16-N07

Nicht bewertbar-Beobachtungen können nicht als Produktleistungsfehler interpretiert werden.

CH16-N08

Nicht bewertbar-Beobachtungen können nicht stillschweigend aus dem Nenner entfernt werden; sie sollten in der bewertbaren Abdeckung angezeigt werden.

CH16-N09

Das Gewicht der Beobachtungen sollte unabhängig vom Ergebnis der KI-Antwort sein.

CH16-N10

Die rohe Anzahl der Teilnehmer und die bevölkerungsgewichteten Nutzeräquivalenten Werte sollten separat berichtet werden.

CH16-N11

Der Status des Antwortbestehens sollte getrennt vom Verhältnis der Unterstützung atomarer Behauptungen geführt werden.

CH16-N12

Kritische und wichtige Gates können durch keine zusammengesetzte Punktzahl ausgeglichen werden.

CH16-N13

Atomare Zentralitätsgewichte sollten festgelegt werden, bevor die Ergebnisse gesehen werden.

CH16-N14

Eine Wiederholung derselben Stammbehauptung darf die Punktzahl nicht künstlich erhöhen oder verringern.

CH16-N15

Zentralitätsgewichte von Behauptungen können kritische und wichtige Gates nicht ersetzen.

CH16-N16

Werte der Dimension Atomare Treue müssen versioniert und öffentlich verfügbar sein.

CH16-N17

Wenn sich die Abbildung der Atomaren Treue ändert, muss eine neue Version der Bewertungsmethode erstellt werden.

CH16-N18

Referenzlücke Atome dürfen nicht stillschweigend mit Null oder Eins zur Punktzahl hinzugefügt werden.

CH16-N19

Die Vollständigkeit des erforderlichen Antwortelements muss als eigene Kennzahl veröffentlicht werden.

CH16-N20

Ein Versagen eines zentralen erforderlichen Elements kann nicht durch einen hohen REC-Durchschnitt ausgeglichen werden.

CH16-N21

Das Profil NOMOS mit zehn Komponenten muss vor dem zusammengesetzten Index sichtbar sein.

CH16-N22

ENT, FAC, SBI, TLA, ECI, EPI, TCR, LGF, USI und STR müssen separate Bewertungen tragen.

CH16-N23

Eine Komponente kann nicht stillschweigend das fehlende Gewicht einer anderen Komponente übernehmen.

CH16-N24

Eine Komponente, die nicht berechnet werden kann, darf nicht als 100 angenommen werden.

CH16-N25

Der vollständige zusammengesetzte NOMOS-Index sollte nur veröffentlicht werden, wenn alle zehn Komponenten zuverlässig berechnet wurden.

CH16-N26

Wenn eine gleiche Gewichtung der Komponenten als kanonische Kandidatenmethode verwendet wird, muss dies ausdrücklich angegeben werden.

CH16-N27

Domänenspezifische alternative Gewichte müssen einen separaten Bewertungsnamen und eine separate Version haben.

CH16-N28

Ein domänenspezifischer Wert kann ohne Erklärung nicht anstelle des kanonischen Index verwendet werden.

CH16-N29

LGF kann sich nicht nur auf den Bevölkerungsdurchschnitt verlassen; es muss eine Leistungsgrundlage und eine Ungleichheit einbeziehen.

CH16-N30

Eine einzelne Beobachtung eines Country Observer kann nicht in eine Länderfairness-Bewertung umgewandelt werden.

CH16-N31

Nur Sprach- und geografische Zellen mit ausreichender Stichprobengröße sollten an der Fairness-Bewertung teilnehmen.

CH16-N32

Sprachen und Länder, die nicht getestet wurden, dürfen nicht ignoriert werden, um ein Ergebnis mit hoher Fairness zu erzielen.

CH16-N33

USI sollte kontrollierte und natürliche Panelergebnisse sowie Unterschiede in der Materialoberfläche zusammen bewerten.

CH16-N34

In einer Studie mit nur einem kontrollierten oder nur einem natürlichen Panel kann USI nicht als vollständig berechnet angesehen werden.

CH16-N35

Kontrollierte und natürliche Panelscores sollten standardmäßig getrennt gehalten werden.

CH16-N36

Wenn kontrollierte und natürliche Punktzahlen kombiniert werden, müssen die Koeffizienten vor den Ergebnissen festgeschrieben werden.

CH16-N37

STR kann nicht in einer einzigen Welle als vollständiger Stabilitätswert dargestellt werden.

CH16-N38

Inter-Wellen-Produkt- oder Truth-Pack-Änderungen müssen in der Stabilitätsberechnung ausdrücklich modelliert werden.

CH16-N39

Verbindung NOMOS kann den Status des Kritischen oder Haupt-Tors nicht ändern.

CH16-N40

Eine Punktzahl von 950 allein kann die NOMOS 950+ Kandidatur nicht festlegen.

CH16-N41

Die Bedingungen für NOMOS 950+ Kandidaten müssen gemeinsam die Anforderungen an Punktzahl, Striktes Bestehen, Tor, Komponentenuntergrenze, ungelöst, Qualität und Wiederholungsbedingungen erfüllen.

CH16-N42

950+ Schwellenwerte können nicht als endgültiger wissenschaftlicher Standard vor Pilot- und unabhängiger Überprüfung vorgelegt werden.

CH16-N43

Die Punktzahl-Identität muss KI-Produkt, Rahmen, Panel, Prompt-Set, Welle, Truth Pack und Punktzahlversion enthalten.

CH16-N44

Der Satz „Die NOMOS Punktzahl der Entität ist X“ darf nicht ohne eine Scope-Identität verwendet werden.

CH16-N45

Das Core GEO-1000 Ergebnis kann nicht als der Full NOMOS Wert dargestellt werden.

CH16-N46

Natürliche Reichweite und Gemeinsame Vergleichsbasis Werte müssen separat berichtet werden.

CH16-N47

Produktwerte für verschiedene Natürliche Reichweite Populationen können nicht ohne Erklärung direkt verglichen werden.

CH16-N48

Das Gemeinsame Vergleichsbasis Ergebnis kann nicht als der vollständige globale Zugriffswert des Produkts dargestellt werden.

CH16-N49

KI-Produktkoeffizienten müssen unabhängig von den Ergebnissen bestimmt werden.

CH16-N50

Ein KI-Produktkoeffizient darf nicht vom Produktwert, der Markenreputation oder einer kommerziellen Sponsoringbeziehung abhängen.

CH16-N51

Gleiche Produkt- und auf Exposition gewichtete Ökosystemwerte müssen separat benannt werden.

CH16-N52

Wenn aktive Nutzer- oder Expositionsdaten nicht überprüft werden können, können keine willkürlichen Koeffizienten erstellt werden.

CH16-N53

Expositionsdaten müssen auf Konto-, Nutzer-, Abfrage- oder Sitzungsebene transparent sein.

CH16-N54

Der Koeffizient eines niedrig bewerteten Produkts kann nach Bekanntwerden des Ergebnisses nicht reduziert werden.

CH16-N55

Das Ökosystem-Komposit kann auf Produktebene keine kritischen oder wichtigen Gates löschen.

CH16-N56

Das AnyCritical- und das nach Exposition gewichtete Critical-Verhältnis sollten getrennt angezeigt werden.

CH16-N57

Die Koeffizienten von Prompt-Familien müssen vor den Ergebnissen festgelegt werden.

CH16-N58

Das Gewicht einer niedrig bewerteten Prompt-Familie kann später nicht reduziert werden.

CH16-N59

Das separate Ergebnis jeder Prompt-Familie sollte zusammen mit der zusammengesetzten Prompt-Bewertung sichtbar bleiben.

CH16-N60

Die aktuelle Wellenbewertung sollte getrennt von rollierenden oder historischen Bewertungen aufbewahrt werden.

CH16-N61

Die frühere Welle mit der höchsten Bewertung kann nicht als aktuelle Bewertung ausgewählt werden.

CH16-N62

Konservative, nur gelöste und optimistische Ergebnisse können für die unerledigte Anspruchsmassenerzeugt werden.

CH16-N63

Referenzunsicherheit und Stichproben-Konfidenzintervall dürfen nicht als dasselbe dargestellt werden.

CH16-N64

Komplexe Stichproben-Konfidenzintervalle sollten Schichtung, Clustering, Gewichtung und die Struktur der passenden Nutzer beibehalten.

CH16-N65

Das Konfidenzintervall der zusammengesetzten Bewertung muss durch erneutes Durchführen des gesamten Bewertungsprozesses berechnet werden.

CH16-N66

Die Roh- und Effektstichprobengrößen sollten zusammen angezeigt werden.

CH16-N67

Ein kritisches Ereignis, das mit null beobachtet wurde, kann nicht als null kritisches Risiko ausgedrückt werden.

CH16-N68

Im Fall von null Ereignissen sollte ein einseitiges Obergrenzwert oder die entsprechende Unsicherheit bei seltenen Ereignissen veröffentlicht werden.

CH16-N69

Der Ansatz 3/n kann in einem komplexen Design nicht automatisch als endgültige Methode verwendet werden; er kann nur eine ungefähre Kontrolle darstellen.

CH16-N70

Bei einem kritischen Ereignis sollten das Vorhandensein des Ereignisses und die Unsicherheit seiner Häufigkeit zusammen angezeigt werden.

CH16-N71

Die Sensitivitätsanalysen der Punktzahl sollten die ungewichteten, gewichteten, beschnittenen, ungelösten und alternativen Koeffizientenergebnisse in dem Maße zeigen, wie es relevant ist.

CH16-N72

Die Sensitivität der Materialpunktzahl sollte auf der öffentlichen Karte eine Warnung erzeugen.

CH16-N73

Die gefälschte Dezimalstelle der Punktzahl darf überhaupt nicht veröffentlicht werden.

CH16-N74

Das Gewicht der fehlenden Komponente darf nicht auf andere Komponenten verteilt werden.

CH16-N75

Der Puan-Veröffentlichungsqualitätsstatus sollte im öffentlichen Register sichtbar sein.

CH16-N76

Die Kandidatur 950+ sollte mindestens auf dem replizierten Punktzahleniveau anstreben.

CH16-N77

Wenn sich die Punktzahlformel ändert, dürfen frühere Punktzahlen nicht stillschweigend neu geschrieben werden.

CH16-N78

Wenn verschiedene Versionen der Bewertungsmethode auf denselben Datensatz angewendet werden, müssen alle Ergebnisse in separaten Versionen erhalten bleiben.

CH16-N79

Der Code zur Punkteberechnung, Koeffizienten und Integritätsaufzeichnungen müssen in einem maschinenlesbaren Format veröffentlicht werden.

CH16-N80

Jede Punktzahl, jeder Koeffizient, jedes Konfidenzintervall und jede Veröffentlichungsentscheidung muss einen verantwortlichen menschlichen Akteur oder eine Institution haben.

71. FORMEN DES VERSAGENS

CH16-F01— EINZELZAHLBERICHT

Verteilung, Tor und Komponentenprofil werden erhalten.

CH16-F02— DAS TOR UNTER DER PUNKTZAHL VERSTECKEN

Kritische Sperre wird zu einer kleinen Fußnote.

CH16-F03— BERECHNUNG BEDINGTER ALS Vollständig bestanden

Moderate Probleme werden unsichtbar.

CH16-F04— UNKLÄRE FÄLLE ALS BESTAND ZÄHLEN

Unbekannte Ergebnisse erhöhen die Punktzahl.

CH16-F05— UNKLÄRE FÄLLE ALS FEHLER ZÄHLEN

Fehlende Referenz senkt das Produkt.

CH16-F06— ENTFERNUNG VON Nicht bewertbar AUS DEM NENNER

Messfehler werden unsichtbar.

CH16-F07— ZÄHLUNG VON Nicht bewertbar ALS PRODUKTMANGEL

Inspektionsfehler wird der KI zugewiesen.

CH16-F08— ZÄHLEN VON 1,000 MENSCHEN GLEICH GEWICHTET

Übersampling und Bevölkerungsunterschiede werden ignoriert.

CH16-F09— GEWICHT NACH ERGEBNIS

Richtige Antworten erhalten ein höheres Gewicht.

CH16-F10— ZÄHLEN Nachweisquote der Aussagen ALS ANTWORT BESTANDEN

Kritisches Tor geht im Durchschnitt verloren.

CH16-F11— KERN- UND NEBENANSPRÜCHE GLEICH GEWICHTEN

Der Nebenanspruch erhält dasselbe Gewicht wie die Hauptentität.

CH16-F12— PUNKTZAHL DURCH WIEDERHOLUNG AUFFÜLLEN

Dasselbe richtige Anspruch wird zehnmal geschrieben, um die Punktzahl zu erhöhen.

CH16-F13— PUNKTZAHL DURCH WIEDERHOLUNG ZUSAMMENBRECHEN

Die gleiche falsche Wurzel führt zu zehn separaten Strafen.

CH16-F14— SPEICHERN VON DIMENSIONSWERTE

Die Punktzahl kann nicht reproduziert werden.

CH16-F15— ZÄHLEN DER ANSICHT, DIE VOM UNTEN KOMMT, ALS TATSACHE FÜR F-4

Exakte Übermittlung des Selbstberichts wird zur Genauigkeit des Selbstberichts.

CH16-F16— SCHMELZEN ERFORDERLICHER ELEMENTE ZU ATOMAREM MITTELWERT

Auch wenn die Hauptaufgabe fehlt, erhält die Antwort eine hohe Punktzahl.

CH16-F17— AUSWÄHLEN VON KOMPONENTEN

Nur die drei starken Komponenten werden veröffentlicht.

CH16-F18— FEHLENDE KOMPONENTE ALS 100 ZÄHLEN

Die volle Punktzahl wird künstlich erhöht.

CH16-F19— VERTEILUNG DES FEHLENDEN GEWICHTS

Die nicht berechnete Fairness-Punktzahl wird zur tatsächlichen Punktzahl hinzugefügt.

CH16-F20— DAS GESETZ DER NATUR ALS GLEICHES GEWICHT ZÄHLEN

Empfindlichkeit und externe Überprüfung werden nicht durchgeführt.

CH16-F21— DEN DOMAIN-SCORE ALS KANONISCH ZÄHLEN

Spezialgewicht wird wie die Basis-NOMOS-Punktzahl dargestellt.

CH16-F22— SPRACHDURCHSCHNITT ALS SPRACHFAIRNESS ZÄHLEN

Niedriges Sprachniveau wird verborgen.

CH16-F23— GLEICHE SPRACHENPARITÄT ALS QUALITÄT ZÄHLEN

Niedrige Bewertungen in allen Sprachen erzeugen hohe Fairness.

CH16-F24— UNGETESTETE SPRACHEN ALS ABDECKUNG ZÄHLEN

Der Abdeckungsfaktor ist aufgebläht.

CH16-F25— LANDESERGEBNIS VOM BEOBACHTER

Ein einzelner Nutzer geht in das Fairness-Konto ein.

CH16-F26— NUR SAUBERE PUNKTE VERÖFFENTLICHEN

Die natürliche Nutzererfahrung bleibt erhalten.

CH16-F27— NUR NATÜRLICHE PUNKTE ALS PRODUKTVOREINSTELLUNG ZÄHLEN

Personalisierung wird zum Basisverhalten.

CH16-F28— STILL DURCHSCHNITTLICH SAUBER UND NATÜRLICH

Verschiedene Vorhersageobjekte werden gemischt.

CH16-F29— 100 STABILITÄT IN EINER EINZIGEN WELLE

Stabilität wird als vollständig ohne Wiederholung betrachtet.

CH16-F30— ZÄHLE ALTE GUTE WELLE ALS AKTUELL

Aktuelle Schwächung wird verborgen.

CH16-F31— ZÄHLE 950 AUTOMATISCH ALS KONFORMITÄT

Tor- und Komponentenboden werden ignoriert.

CH16-F32— Kritische Sperre— ZÄHLE 970 ALS 950+

Die Zählung geht dem Tor voraus.

CH16-F33— PUNKTEIDENTITÄT VERBERGEN

Es ist nicht bekannt, welches Produkt und welches Panel gemessen werden.

CH16-F34— KERNPUNKTE ALS VOLLE NOMOS ZÄHLEN

Ein einzelner Prompt repräsentiert den gesamten Standard.

CH16-F35— Natürliche Reichweite PUNKTE RANGIEREN

Verschiedene Populationen werden als gleich betrachtet.

CH16-F36— VOLLEN ZUGRIFF AUF Gemeinsame Vergleichsbasis ZÄHLEN

Eine enge gemeinsame Population wird zu einem globalen Ergebnis gemacht.

CH16-F37— MODELL-KOEFFIZIENTEN ENTSPRECHEND DEM PUNKT STIFTEN

Hohe Punktzahlen erhalten mehr Einfluss.

CH16-F38— MARKENREPUTATION ZU EINEM FAKTOR MACHEN

Die Beliebtheit des Produkts wird gewichtet, ohne vorhergesagt zu werden.

CH16-F39— ZÄHLUNG VON KONTOZAHLEN ALS AKTIVE NUTZER

Der Expositionskoeffizient steigt.

CH16-F40— ENTFERNEN VON PRODUKTEN MIT NIEDRIGEM SCORE

Das Ökosystem besteht nur aus erfolgreichen Produkten.

CH16-F41— GLEICHEN EXPOSITIONSSCORE FÜR PRODUKTE ZÄHLEN

Durchschnittliche Produkte werden wie durchschnittliche Personen dargestellt.

CH16-F42— ZÄHLUNG DES EXPOSITIONSSCORES ALS PRODUKTQUALITÄT

Ein großes Produkt wird automatisch zu einer wichtigeren Qualitätskennzahl.

CH16-F43— KRITISCHES IM ÖKOSYSTEM DURCHSCHNITT LÖSCHEN

Das schwere Ereignis des Produkts mit kleinem Koeffizienten verschwindet.

CH16-F44— ÄNDERN SIE DEN PROMPT-KOEFFIZIENT NACH DEM ERGEBNIS

Die harte Prompt-Familie wird reduziert.

CH16-F45— BERÜCKSICHTIGEN SIE FAMILIEN MIT NIEDRIGEM SCORE ALS OPTIONAL

Grenz- oder Evidenzergebnisse sind vom Umfang ausgeschlossen.

CH16-F46— AKTUELL MIT ROLLING SCORE BEIBEHALTEN

Alte hohe Wellen überdecken den aktuellen Abfall.

CH16-F47— NICHT GELÖSTE AUS DEM NENNER LÖSCHEN

Punktschätzung steigt künstlich.

CH16-F48— BERÜCKSICHTIGUNG DER REFERENZUNSICHERHEIT BEI DER STICHPROBEN-CI

Zwei verschiedene Unsicherheiten vermischen sich.

CH16-F49— EINFACHE ZUFALLSSTICHPROBEN-CI

Schichtung, Gewichtung und Cluster werden ignoriert.

CH16-F50— SUMMIEREN VON KOMPONENTEN-CIs

Abhängigkeit wird ignoriert.

CH16-F51— ANGABE DER SENSITIVITÄT MIT ROHEN N

Die effektive Stichprobe wird verschleiert.

CH16-F52— ZÄHLEN VON NULL-EVENTS ALS NULL RISIKO

Unsicherheit durch seltene kritische Fehler wird verborgen.

CH16-F53— ANWENDUNG VON 3/N AUF JEDEN FALL

Komplexe Stichproben- und Gewichtungsstruktur wird ignoriert.

CH16-F54— EIN EINZELNES KRITISCHES ELEMENT ALS 100 PROZENT SYSTEMRISIKO ZÄHLEN

Die Allgemeingültigkeit wird übermäßig verallgemeinert.

CH16-F55— EIN EINZELNES KRITISCHES ELEMENT VERSTECKEN, WEIL DIE RATE KLEIN IST

Das Vorhandensein des Ereignisses verschwindet.

CH16-F56— SENSITIVITÄTSANALYSE VERSTECKEN

Es wird nicht gesehen, dass sich die Bewertung bei alternativen, vernünftigen Methoden ändert.

CH16-F57— FALSCHE DEZIMAL-SENSITIVITÄT

Ein Ergebnis wie 963,742 wird veröffentlicht.

CH16-F58— TEILPROFIL ALS VOLLE BEWERTUNG ZÄHLEN

Vergleich wird mit einer fehlenden Komponente durchgeführt.

CH16-F59— VERÖFFENTLICHUNG DER VORLÄUFIGEN PUNKTE ALS ENDGÜLTIG

Eine einzelne Welle und ein einzelner Bewerter werden zum vollständigen Standard.

CH16-F60— NEU SCHREIBEN DER GESCHICHTE DURCH FORMELÄNDERUNG

Die alte Punktzahl verschwindet.

CH16-F61— BERECHNUNGSCODE GESCHLOSSEN HALTEN

Unabhängige Reproduktion wird verhindert.

CH16-F62— PUNKTHÖHE ALS AUDITQUALITÄT BETRACHTEN

Eine fehlerhafte Messung wird akzeptiert, weil sie ein hohes Ergebnis lieferte.

CH16-F63— PUNKTETIEFSTAND ALS METHODENFEHLER BETRACHTEN

Ein niedriges Ergebnis wird aufgrund kommerzieller Unannehmlichkeiten neu berechnet.

CH16-F64— EINZELUNTERNEHMENSBESCHÄFTIGUNGS-MONOPOL

Unabhängige Institutionen werden daran gehindert, dieselbe Formel anzuwenden.

72. PRÜFUNGSVERFAHREN

Schritt 1— Score-Identität erstellen

Das Entity-KI-Produkt Frame Panel Prompt Set Wave Truth Pack Score-Verfahren ist festgeschrieben.

Schritt 2— Hauptanalyse-Beobachtungsset festschreiben

Backup-, zurückgezogene, nicht bewertbare und Hauptbeobachtungsdatensätze werden getrennt.

Schritt 3— Bevölkerungsgewichte überprüfen

Design-, Nichtantwort- und Kalibrierungsgewichte werden untersucht.

Schritt 4— RP-Verteilung berechnen

Nutzeräquivalente mit RP-1 bis RP-8 werden erstellt.

Schritt 5— Strikte und akzeptable Bestehenswerte berechnen

SP1 000 und AP1 000 werden abgeleitet.

Schritt 6— Dichten von kritisch, groß, ungeklärt und nicht bewertbar berechnen

Gate und Prävalenz werden getrennt gehalten.

Schritt 7— Anspruchszentralität und Wurzelcluster-Normalisierung anwenden

Aussagen werden erneut überprüft.

Schritt 8— Atomare Größenmetriken berechnen

Ergebnisse zu Entität, Tatsache, Umfang, Zeit, Attribution, Modalität und Referenz werden erstellt.

Schritt 9— Berechnung der erforderlichen Elementvollständigkeit

Es wird zusammen mit den Prompt-Familientoren aufgezeichnet.

Schritt 10— Berechnung von zehn Komponenten

Alle Komponentenformeln von ENT bis STR werden angewendet.

Schritt 11— Überprüfung fehlender Komponenten

Es wird festgestellt, ob die Gesamtnote berechnet werden kann.

Schritt 12— Berechnung des zusammengesetzten NOMOS-Index

Gilt nur für vollständige und gültige Profile.

Schritt 13— Festlegung des Torstatus

Bestimmt, ob der Status Kritisch, Wesentlich, Bedingt oder Vollständiger Kandidat ist.

Schritt 14— Trennen von Natürliche Reichweite- und Gemeinsame Vergleichsbasis-Ergebnissen

Die beiden Punktzahlenfamilien werden getrennt aufbewahrt.

Schritt 15— Festschreiben der AI-Produktkoeffizienten

Gleichheits- und Expositionskoeffizienten werden zusammen mit der Datenquelle gespeichert.

Schritt 16— Anwenden der Koeffizienten der Prompt-Familie

Gewichte werden unabhängig vom Ergebnis verwendet.

Schritt 17— Trennen von Panel- und Wellen-Ergebnissen

Bereinigte, natürliche, aktuelle und rollierende Punktzahlen werden getrennt.

Schritt 18— Berechnung von Fairness- und Floor-Kennzahlen

Ergebnisse zu Sprache, Land und Abdeckung werden erstellt.

Schritt 19— Berechnung von ungelösten Unter-Punkt-Oberer-Ergebnissen

Referenzunsicherheit wird sichtbar gemacht.

Schritt 20— Berechnung von Stichproben-Konfidenzintervallen

Das für das Stichprobendesign geeignete Resampling-Verfahren wird angewendet.

Schritt 21— Berechnung der effektiven Stichprobengröße

Anzeige auf Gesamt- und Komponentenbasis.

Schritt 22— Berechnung des seltenen kritischen Ereignis-Schwellenwerts

Eine einseitige Methode wird bei null oder wenigen Ereignissen verwendet.

Schritt 23— Sensitivitätsanalysen erstellen

Gewicht, Koeffizient und ungelöste Optionen werden verglichen.

Schritt 24— Veröffentlichungsqualitätsstatus zuweisen

Eine Entscheidung wird zwischen NSQ-0 und NSQ-5 getroffen.

Schritt 25— NOMOS 950+ Kandidatentüren anwenden

Alle Bedingungen außer der Punktzahl werden überprüft.

Schritt 26— Öffentliche Ergebnisübersicht erstellen

Gate, Verteilung, Profil, Index, Unsicherheit und Umfang werden jeweils der Reihe nach veröffentlicht.

Schritt 27— Berechnungscode und Manifest festschreiben

Die Code-Version, Eingaben, Hash und Ergebnisse sind verlinkt.

Schritt 28— Vorbereitung des unabhängigen Reproduktionspakets

Berechnungsdateien, die von persönlichen Daten befreit sind, werden für eine unabhängige Überprüfung bereitgestellt.

73. ERFORDERLICHE BELEGE

Score-ID Entity-Datensatz KI-System registrieren Natürliche Reichweite und Gemeinsame Vergleichsbasis Datensätze Panel-Status Befragter Satz und Familien Truth Pack-Version Wellen-ID Anspruchsbücher Antwortentscheidungen RP-1–RP-8 Verteilung Hauptanalyse Beobachtungssatz Designgewichte Nichtantwortgewichte Kalibrierungsgewichte Gewichtstrimm-Datensätze Effektive Stichprobe Streng Bestanden Akzeptabel Bestanden Kritische Rate Wichtige Rate Ungeklärte Rate Keine Verwendbare Antwortrate Nicht bewertbar Rate Anspruchszentralitätsgewichte Wurzelcluster-Datensätze Atomare Dimensionswerte Atomare Treue Erforderliches Element Vollständigkeit

ENT

FAC

SBI

TLA

ECI

EPI

TCR

LGF

USI

STR

Komponententeil-Eingaben; Sprach- und Ländermindestwerte; Abdeckungsfaktor; Kontrollierte und Natürliche Ergebnisse; Inter-Wellen-Ergebnisse; zusammengesetzter NOMOS-Index; domainspezifische alternative Ergebnisse; KI-Produkt-Koeffizienten und ihre Datenquellen; Prompt-Familien-Koeffizienten; Ergebnis des gleichen Produkts; erfolgsgewichtetes Ergebnis; AnyCritical-Status; Kritischer Expositionsanteil; Ungeklärte Unter-Punkt-Ober-Ergebnisse; Stichprobenvertrauensintervall; Kritische Obergrenze; Sensitivitätsanalyse; Ergebnisveröffentlichungsqualität; 950+ Kandidaten-Gate-Datensatz; Score-Methoden-Version; Berechnungscode; Code-Integritäts-Hash; öffentliche Scorecard; Änderungsprotokoll; unabhängiger Reproduktionsnachweis; und verantwortliche Person oder Institution.

74. PRÜFLISTE FÜR AUDITS

Deckt die Punktzahl-Identität alle Umfangsbereiche ab? Wurde der Hauptanalyse-Beobachtungssatz vor dem Ergebnis festgeschrieben? Wurden Sicherungs- und Hauptbeobachtungen getrennt? Sind die Beobachtungsgewichte unabhängig vom Ergebnis? Ist die RP-1–RP-8-Verteilung insgesamt 1,000? Wurden Streng und Akzeptabler Bestand getrennt? Sind Kritisch und Wichtig separat? Wurde Unresolved als bestanden oder nicht bestanden markiert? Wurde Nicht bewertbar aus dem Nenner entfernt? Ist die bewertbare Abdeckung offen? Wurde das Nachweisquote der Aussagen anstelle der Bestand-Antwort verwendet? Sind die Zentralitätsgewichte offen? Wurde dieselbe Hauptforderung erneut gezählt? Wurde die Atomare Dimensionszuordnung veröffentlicht? Wurde F-4 korrekt interpretiert? Wurden erforderliche Elemente separat berechnet? Wurde das Kern-Gate mit hohem REC kompensiert?

Ist die gesamte Komponente vorhanden? Wurde die fehlende Komponente als 100 gezählt? Wurde das Gewicht der fehlenden Komponente verteilt? Misst ENT die korrekten Entitätsbeziehungen? Schließt FAC ungelöste Aussagen stillschweigend aus? Beinhaltet SBI die Auslassung von Materialgrenzen? Bewahrt TLA den Länder- und Zeitunterschied? Misst ECI die Unterstützung von Zitaten anstelle der Existenz von Zitaten? Trägt EPI die Unterscheidung zwischen Selbstdeklaration und Realität? Macht TCR die nicht nutzbare Antwort sichtbar? Beinhaltet LGF die Basis und die Lücke neben dem Durchschnitt? Wurde die Beobachtung des Länderbeobachters in die Länderwertung einbezogen? Beruht USI sowohl auf Clean- als auch auf Natural-Daten? Wurde STR

aus einer einzigen Welle erzeugt? Deckte die zusammengesetzte Wertung den Torstatus ab? Beruht 950+ ausschließlich auf dem Punktwert?

Wurden die Komponentenebenen überprüft? Ist das Kern-Ungelöst-Limit offen? Wurden Natürliche Reichweite und Gemeinsame Vergleichsbasis getrennt? Wurden die KI-Produktkoeffizienten bestimmt, bevor die Punktzahlen bekannt waren? Messen die Koeffizienten aktive Nutzer, Konten oder Abfragen? Sind Equal- und Exposure-Ergebnisse getrennt? Wurde das niedrig bewertete Produkt entfernt? Verbirgt die Ökosystembewertung das Produkt Critical? Gibt es ein AnyCritical-Feld? Sind die Gewichtungen der Prompt-Familien offen? Wurde die niedrig bewertete Prompt-Familie reduziert? Sind Clean- und Natural-Punktzahlen getrennt? Sind aktuelle und Rolling-Punktzahlen getrennt? Wurde die beste vergangene Welle als aktuell ausgewählt? Gibt es Niederpunkt-Oberergebnisse für Ungelöst? Wurden Referenz- und Stichprobenunsicherheiten getrennt? Erhält das Konfidenzintervall-Design den Schutz bei?

Wird das zusammengesetzte CI das gesamte Bewertungssystem erneut durchlaufen? Sind die Roh- und Effektstichproben zusammen? Wurde es als nullkritisch, nullrisiko geschrieben? Ist die Obergrenze eines seltenen Ereignisses offen? Wurde eine Sensitivitätsanalyse durchgeführt? Ändern alternative vernünftige Koeffizienten das Ergebnis? Gibt es falsche Dezimalgenauigkeit? Ist die Publikationsqualität der Punktzahl klar? Ist die Punktzahlformel versioniert? Wird die alte Bewertungsmethode aufbewahrt? Kann der Berechnungscode reproduziert werden? Ist der verantwortliche Eigentümer der Punktzahl klar?

75. EINWÄNDE UND ANTWORTEN

Einwand 1— „Das Einzige, was die Nutzer verstehen, ist eine einzelne Zahl; warum zeigen wir so viele Ergebnisse?“

Eine einzelne Zahl ist verständlich. Aber allein ist sie irreführend. Öffentliche Karte:

- Torstatus,
- Verteilung der 1,000-Nutzer,
- zusammengesetzter Wert

kann in drei einfachen Ebenen wie folgt dargestellt werden. Das detaillierte Profil wird für Forscher und Prüfer aufbewahrt.

Einwand 2— „War der GEO-Score nicht direkt die Anzahl der Personen, die richtig geantwortet haben?“

Dies ist immer noch das Hauptergebnis. SP_{1000} und AP_{1000} zeigen die Gründungsidee direkt an. Der NOMOS-Verbundindex:

- Nachweise,
- Gerechtigkeit,
- natürlicher Nutzer,
- Stabilität

umfasst zusätzliche normative Bereiche wie.

Einwand 3— „Warum zählen wir Vollständig bestanden und Bedingt bestanden separat?“

Da beide Nutzer möglicherweise keine schwerwiegenden Fehler gesehen haben, jedoch in der Bedingungsgruppe:

- begrenzte Aktualität,
- moderates Auslassen,
- Usability-Problem

gefunden werden können. Dieser Unterschied ist wichtig für Verbesserungen und das Vertrauensniveau.

Einwand 4— „Können wir nicht ungelöste Beobachtungen entfernen und eine saubere Punktzahl vergeben?“

Eine nur aufgelöste Punktzahl kann vergeben werden. Die ungelöste Masse sollte jedoch sichtbar bleiben. Andernfalls wäre es möglich, die Punktzahl zu erhöhen, ohne die schwierigsten Aussagen zu lösen.

Einwand 5— „Warum sind die zehn Komponenten gleich gewichtet?“

Gleichgewichtete Gewichtung bietet einen transparenten Ausgangspunkt. Hochrisiko-Fehler werden bereits durch das Gate-System geschützt. Alternative Domain-Gewichte können ebenfalls erstellt werden. Exakte Gewichte sollten durch Pilot- und unabhängige Überprüfung getestet werden.

Einwand 6— 'Ist nicht Existenz und faktische Genauigkeit wichtiger?'

So ist es in einigen Prüfungen. Allerdings:

- Nachweise,
- Umfang,
- Fairness,
- Differenz des natürlichen Nutzers

Es kann ohne diese eine hohe durchschnittliche falsche Faktentreue erzeugen. Kritische und Haupt-Gates schützen auch die wesentliche Materialität.

Einwand 7— „Wenn Fairness niedrig ist, warum sollte der globale Durchschnitt sinken?“

Der Compound NOMOS-Standard bewertet nicht nur die durchschnittliche Person, sondern auch die Fairness der Repräsentation. Das bevölkerungsgewichtete Hauptergebnis wird erneut separat dargestellt. LGF macht systematische Verschlechterungen in kleinen Sprachen und Ländern sichtbar.

Einwand 8— „Macht ein nicht niedriges Ergebnis in allen Sprachen die Fairness nicht hoch?“

Dies kann durchgeführt werden, wenn nur die Disparität verwendet wird. Daher nutzt LGF:

- Leistungsbasis,
- Disparität,
- Abdeckung

zusammen.

Einwand 9— „Warum mitteln wir nicht direkt die Clean- und Natural-Werte?“

Sie messen unterschiedliche Zielpopulationen und Nutzungssituationen. Arithmetische Kombination erfordert normative Präferenzen. Standardmäßig werden sie separat veröffentlicht.

Einwand 10— ‚Wäre eine gleiche Gewichtung über zehn KI-Produkte gerechter?‘

Es ist fair für den Vergleich des durchschnittlichen Produkts. Es ist nicht fair für die durchschnittliche menschliche Exposition. Daher werden gleiche und nach Exposition gewichtete Ergebnisse getrennt gehalten.

Einwand 11— „Wie wird der Modellkoeffizient gefunden, wenn keine aktiven Nutzerdaten vorhanden sind?“

Er kann möglicherweise nicht gefunden werden. Gleiche Produktergebnisse und vernünftige Szenarienbereiche können veröffentlicht werden. Willkürliche Koeffizienten können nicht erfunden werden.

Einwand 12— „Wenn nur 1%-Exposition für ein Produkt auftritt, warum sollte das Kritische Ereignis dann das Ökosystem-Ergebnis beeinflussen?“

Die nach Exposition gewichtete Rate kann eine kleine Auswirkung haben. Das Produkt und die Abdeckung, bei der das Ereignis auftritt, tragen jedoch weiterhin Kritische Sperre. Der Durchschnitt des Ökosystems kann nicht die Wahrheit des Produkts auslöschen.

Einwand 13— „Wenn wir null Kritisches beobachten, warum sagen wir dann nicht einfach null?“

Die Stichprobe beobachtet nicht alle Nutzer. Null-Beobachtung: zeigt an, dass das Ereignis nicht gesehen wurde. Sie zeigt nicht an, dass das Risiko mathematisch null ist.

Einwand 14— „Macht es nicht sinnlos, so viele zusätzliche Bedingungspunkte für 950+ zu haben?“

Im Gegenteil, es bewahrt die Bedeutung der Zahl. 950 Punkte:

- kritisches Ereignis,
- sprachlich mangelhaft,
- einzeln geschwungen,
- geringe Beurteilungsqualität

trägt nicht dasselbe Vertrauen in eine Studie.

Einwand 15— „Wenn sich die Bewertungsformel ändert, können alte Studien nicht verglichen werden?“

Alte und neue Methoden können auf dieselben festen Daten für eine Brückenanalyse angewendet werden. Zwei Bewertungsvarianten werden getrennt gehalten. Diese Methode erlaubt es, den Standard weiterzuentwickeln, ohne die Vergangenheit zu löschen.

Einwand 16— „Macht eine so transparente Formel das System nicht leichter manipulierbar?“

Einige Optimierungen können auftreten. Dagegen:

- versiegelte Zurückhaltungsaufforderungen,
- natürliches Nutzerpanel,
- Sprache und Landesgerechtigkeit,
- verschiedene Wellen,

- kritische Türen,
- öffne Änderungsverlauf

wird verwendet. Geschlossene Formel erhöht die Undurchschaubarkeit. Allein offene Formel ist kein Grund für Gamification.

ALLGEMEINE VORSCHRIFT DES ABSCHNITTS 79

Eine Punktzahl kann leicht beeindruckend wirken. Sie hat drei Ziffern. Sie ist rund. Sie wirkt vergleichbar. Ein Geschäftsführer kann sagen: „Wir haben 963 erreicht." Eine Agentur kann sagen: „Ihr Konkurrent hat 912 erreicht." Ein Unternehmen kann sagen: „Wir haben die weltweit höchste NOMOS-Punktzahl." Aber wenn die Punktzahl diese Fragen nicht beantwortet, ist ihr Entscheidungswert begrenzt: Welches KI-Produkt? Welche Nutzergruppe? Welche Länder? Welche Sprachen? Welche Eingaben? Zu welcher Zeit? Welches Panel? Wie viele kritische Ereignisse? Wie viele wesentliche Ergebnisse? Wie viele ungelöst? Wie viele effektive Stichproben? Wie hoch ist das Konfidenzintervall? Auf welchen Modellkoeffizienten basiert es? Wurde das Ergebnis repliziert? NOMOS akzeptiert keine einzelne Zahl. Es weigert sich, eine einzelne Zahl als Wahrheit stehen zu lassen. Daher das Hauptergebnis:

> *Ist die Verteilung von 1,000-Nutzeräquivalenten.*

Zusammengesetzter Index:

> *Dies ist eine Zusammenfassung dieser Verteilung und anderer normativer Komponenten.*

Ein Produkt:

- 970 akzeptable Bestehensrate,
- zwei Kritische Ereignisse

Ein weiteres Produkt:

- 950 akzeptable Bestehensrate,
- null Kritisch,
- stärkere Sprachfairness

Diese Zahlen können ein anderes Produkt beschreiben. Der Durchschnitt allein kann das erste Produkt nicht zum Marktführer machen. Eine falsche Erfahrung in einer Sprache mit geringerem Ressourcenaufkommen kann für die Bevölkerung wenig Gewicht haben, bleibt aber für diesen Nutzer vollkommen real. Ein kritisches Ereignis in einem kleinen KI-Produkt kann für das Ökosystem wenig Gewicht haben, bleibt aber für den Nutzer dieses Produkts vollkommen real. Die Bewertung sollte nicht auslöschen, was klein ist; sie sollte es korrekt einordnen und ehrlich gewichten. Die Koeffizienten von KI-Produkten müssen vor der Leistungserfassung festgelegt werden. Das Senken des Koeffizienten eines Produkts nach einem schwachen Ergebnis erlaubt es dem System, seine eigene Realität umzuschreiben. Das Verringern des Gewichts einer schwachen Prompt-Familie ermöglicht es dem Test, seinen eigenen Fragen auszuweichen. Das Ausschließen ungelöster Ansprüche verschleiert die härtesten Realitäten; das Ausschließen von Nicht bewertbar-Datensätzen verwandelt Messfehler in Produkterfolg. Null beobachtete Kritische Ereignisse bedeutet nicht null Risiko.

Eine Stichprobe sagt nur, was sie beobachtet. NOMOS gibt auch an, was sie nicht beobachtet, zusammen mit ihrer Grenze. Aus diesem Grund ist die Zahl 950+ nicht das alleinige Ziel. Sie ist ein Vertrag. Folgendes zu sagen:

- hoher Durchschnitt,
- hohe strikte Bestehensquote,

- starke Gerechtigkeit,
- natürliche Nutzerresilienz,
- erneut,
- klare Unsicherheit,
- null offene Kritische und Schwere

zusammen sind erforderlich. Daher lautet das sechzehnte Messgesetz von NOMOS:

> *Verteilung ist real; Punktzahl ist Zusammenfassung.*

Sein siebzehntes Gesetz lautet:

> *Torstatus liegt über dem Durchschnitt.*

Sein achtzehntes Messgesetz besagt:

> *Koeffizient kann dem Ergebnis nicht folgen; das Ergebnis folgt dem Koeffizienten.*

Sein neunzehntes Messgesetz besagt:

> *Das durchschnittliche Produkt und die durchschnittliche Person sind nicht dasselbe.*

Sein zwanzigstes Messgesetz besagt:

> *Fairness ist nicht nur Gleichheit; sie ist die gleichzeitige Bewahrung der Basis, des Unterschieds und des Umfangs.*

Sein einundzwanzigstes Messgesetz besagt:

> *Ungeklärt ist kein Abfall, der außerhalb der Realitätspunktzahl geworfen werden soll, es ist die Grenze der Punktzahl.*

Sein zweiundzwanzigstes Messgesetz besagt:

> *Null beobachteter Fehler ist keine Nullwahrscheinlichkeit.*

Das dreiundzwanzigste Gesetz lautet wie folgt:

> *Eine fehlende Komponente kann nicht durch die Erfolge anderer Komponenten ersetzt werden.*

Das vierundzwanzigste Gesetz lautet wie folgt:

> *NOMOS 950+ ist keine Zahl; die Punktzahl ist der gemeinsame Erlass von Tor, Gerechtigkeit, Unsicherheit und Wiederholungsbedingungen.*

NOMOS's Abschnitt 16 Bestellung

> *Sag mir nicht nur meine Punktzahl. / Zeig mir, wer unter tausend Menschen was sieht.*
>
> *Setze Vollständig bestanden, Bedingt bestanden, Schwerwiegend nicht bestanden, Kritisch nicht bestanden, Unresolved und Nicht bewertbar nicht in denselben Korb.*
>
> *Verliere meinen kritischen Fehler nicht im Durchschnitt.*
>
> *Entferne Nicht bewertbar-Datensätze nicht, um mich erfolgreicher erscheinen zu lassen. / Zähle Messfehler aber auch nicht als meine Misserfolge.*
>
> *Verwandeln Sie atomare Genauigkeit nicht in einen Antwort-Pass-Status.*
>
> *Zeigen Sie Entität, Faktizität, Umfang, Zeit, Nachweise, epistemischen Status, Aufgabe, Fairness, Nutzerstatus und Stabilität getrennt an.*

Bewerten Sie eine fehlende Komponente nicht mit einhundert. / Verteilen Sie ihr Gewicht nicht auf stärkere Komponenten.

Verstecken Sie einen niedrigen Sprachniveau-Wert nicht im Durchschnitt der Hauptsprachen.

Ignorieren Sie die Verschlechterung bei natürlichen Nutzern nicht nur, weil ich in einer sauberen Sitzung gut bin.

Erklären Sie nicht eine einzelne Welle als entscheidendes Ergebnis.

Weisen Sie jedem KI-Produkt vor dem Ergebnis einen Koeffizienten zu. / Ändern Sie mein Gewicht nicht, nachdem Sie meine Punktzahl gesehen haben.

Mache das gleiche Produkt-Durchschnitt nicht zum Durchschnitt menschlicher Erfahrung.

Gib aktive Nutzerdaten an, ansonsten stelle keinen Koeffizienten her. / Sage unbekannt.

Entferne kein kritisches Ereignis in meinem kleinen Produkt vom Ökosystem-Durchschnitt.

Übergebe oder verwerfe die ungelöste Forderung nicht. / Zeige ihre unteren und oberen Grenzen.

Betrachte tausend rohe Bin-Beobachtungen nicht als tausend gleiche Informationseinheiten. / Schreibe die effektive Stichprobe.

Sag nicht null Risiko, wenn du null Kritisches siehst. / Zeige die obere Grenze des Risikos, das du nicht gesehen hast.

Erzeuge keine falsche Sicherheit wie 963,742.

Machen Sie die Nummer 950 nicht zu einem Abzeichen. / Überprüfen Sie zuerst Strict Pass, Komponentenuntergrenzen, Tore, Fairness, Wiederholungen und Mehrdeutigkeit.

Wenn Sie die Bewertungsmethode ändern, löschen Sie nicht die alte Punktzahl. / Öffnen Sie eine neue Version.

Berechnen Sie zuerst die Nutzerverteilung. / Leiten Sie dann die atomaren Kennzahlen ab. / Erstellen Sie dann das Zehn-Komponenten-Profil. / Wenden Sie dann den Tor-Status an. / Fügen Sie dann das Konfidenzintervall und die ungelösten Grenzen hinzu. / Berechnen Sie dann, wo gerechtfertigt, die zusammengesetzte Punktzahl. / Erst nachdem all dies erfolgt ist, dürfen Sie sich einer NOMOS 950+ Forderung nähern.

Der abschließende Satz des Kapitels

Ein vertrauenswürdiger GEO-1000-Score ist keine Berechnung, die die Realität auf eine Zahl reduziert. Es ist eine offene Messarchitektur, die gemeinsam die Verteilung der Repräsentation, die von 1,000 Menschen erlebt wird, nicht kompensierbare Risiken, das Fairness-Niveau, Unsicherheit und Stabilität über die Zeit bewahrt.

Normativer Kern

Kein NOMOS Ergebnis darf als bloßer Gesamtscore veröffentlicht werden. Jeder veröffentlichte Wert MUSS zuerst Folgendes offenlegen: - den Gate-Status, - die GEO-1000 Antwort-Statusverteilung, - Äquivalente für Strict Pass und Acceptable Pass, - Äquivalente für kritische und schwerwiegende Befunde, - Äquivalente für Unresolved, No Usable Response und Not Ratable, - Komponentenergebnisse, - Unsicherheit, - effektive Stichprobengröße, - Umfang, - und Version der Bewertungsmethode. Die primäre GEO-1000 Ausgabe MUSS die bevölkerungsgewichtete äquivalente Anzahl von Nutzern darstellen, von 1,000, die jeden RP-1 bis RP-8 Antwort-Status erhalten. Kritische und schwerwiegende Befunde sind nicht kompensierbar. Kein Durchschnitt atomarer Aussagen, Antwort-Durchführungsrate, Komponentenbewertung, Produktkoeffizient oder Ökosystem-Zusammensetzung darf ein aktives Kritisches oder Großes Gate außer Kraft setzen. Atom-, erforderliche-Element-, Komponenten-, Zusammensetzungs-, Fairness-, Stabilitäts-, Kritischer-Rate- und Veröffentlichungsqualitätsmetriken MÜSSEN getrennt bleiben. Die kanonische Kandidat NOMOS-Zusammensetzung besteht aus zehn separat berichteten 0–100-Komponenten: - Entitätsintegrität, - Faktentreue, - Umfangs- und Grenzintegrität, - Zeitliche und lokale Genauigkeit, - Nachweis- und Zitationsintegrität, - Epistemische Integrität, - Aufgabenabschluss und Relevanz, - Sprach- und geografische Fairness, - Nutzerzustands- und Oberflächenintegrität, - und Stabilität und Reproduzierbarkeit. Ein vollständiger 0–1,000-Komposit DARF NICHT erstellt werden, wenn eine erforderliche Komponente nicht schätzbar ist. Das Gewicht der fehlenden Komponente DARF NICHT stillschweigend umverteilt werden. Natürliche Reichweite, Gemeinsame Vergleichsbasis, kontrolliertes Panel unter bereinigten Bedingungen, Panel unter natürlichen Nutzungsbedingungen, aktuelle Welle, rollierend, gleichprodukt- und expositionsgewichtete Scores MÜSSEN weiterhin separat identifiziert bleiben. KI-Produktkoeffizienten MÜSSEN festgeschrieben sein, bevor Produktergebnisse beobachtet werden, und DÜRFEN NICHT von der Produktleistung, dem Ruf des Anbieters, der Vorliebe des Kunden, der Förderung oder dem gewünschten Ergebnis abhängen. Gleichproduktkoeffizienten schätzen die durchschnittliche getestete Produktleistung. Expositionskoeffizienten schätzen die durchschnittliche definierte Nutzerexposition. Diese Schätzer DÜRFEN NICHT als gleichwertig repräsentiert werden. Ein Ökosystem-Komposit MUSS die Sichtbarkeit von kritischen und schweren Vorfällen auf Produktebene beibehalten. Hohe Werte in anderen Produkten DÜRFEN den bestätigten kritischen Vorfall eines geringer gewichteten Produkts NICHT löschen. Ungeklärte Aussagen MÜSSEN sichtbare Referenzunsicherheitsgrenzen erzeugen und DÜRFEN NICHT stillschweigend als korrekt, inkorrekt oder nicht vorhanden behandelt werden. Stichprobenunsicherheit, Referenzunsicherheit, Wertungsempfindlichkeit, Rohstichprobengröße und effektive Stichprobengröße MÜSSEN separat berichtet werden. Null beobachtete bestätigte kritische Vorfälle DÜRFEN NICHT als null kritisches Risiko repräsentiert werden. Eine methodengerechte einseitige obere Grenze MUSS berichtet werden. Ein NOMOS 950+ Leistungskandidat MUSS die Bedingungen für Score, Strict Pass, Gate, Component-Floor, Unresolved, Quality, Coverage und Replication erfüllen. Ein zusammengesetzter Score von 950 oder mehr allein ist nicht ausreichend. Jede Score-Formel, Abbildung, Koeffizient, Datensatz, Code-Version, Vertrauensverfahren, Publikationsentscheidung, Revision und Integritätsaufzeichnung MUSS versioniert und einer verantwortlichen Person oder Organisation zuordenbar sein.

TEIL V

Synthetisches Testdesign und End-to-End-Demonstration

KAPITEL 17

Das synthetische Testdesign für Apple.com

Prüfung des Standards, nicht eines Unternehmens

METHODISCHE GRUNDLAGEN / K03; K09

Kapitelgrenze

Die ersten sechzehn Kapitel legten die wichtigsten normativen Ebenen von GEO-1000 fest: die geprüfte Einheit; die Zielbenutzerpopulation und Stichprobe; die Ländebeobachter- und Sprachfairness-Panels; die Teilnehmerauswahl und Gewichtung; die registrierte KI-Produktinstanz; die Prompt-Verfassung; kontrollierte und natürliche Nutzerzustände; die synchronisierte Welle und NOMOS-Erfassungsbeweiskette; das Verifiziertes Faktenpaket zur Entität; atomare Aussagen; kritische und wichtige Prüfstationen; sowie die Verteilungs-, Komponenten- und Gesamtnotenarchitektur. Der Standard muss sich nun seiner eigenen zentralen Prüfung stellen:

> *Kann ein so detailliertes System wirklich funktionieren?*
>
> *Liefern unterschiedliche Teams ähnliche Ergebnisse, wenn dieselben Beobachtungen verarbeitet werden?*
>
> *Kann es einen kritischen Fehler abfangen, ohne im Durchschnitt zu verlieren?*
>
> *Kann es Probleme mit spracharmen Sprachen vom globalen Durchschnitt unterscheiden?*
>
> *Kann es korrekt zwischen falsche Entität, falschem Umfang, veralteter Realität, gefälschter Zitation und Materialauslassung unterscheiden?*
>
> *Wenn man von einer bekannten synthetischen Realität ausgeht, kann die NOMOS-Prüfungskette das korrekte Ergebnis reproduzieren?*

In der Gründungsarchitektur des zweiten Buches wurde jeder Fehler nicht nur als ein zu erklärender Irrtum entworfen, sondern als eine eigenständige Prüfkontrolle. Jede vorgeschlagene Kontrolle musste ein Modell, ein Datum, ein Land, eine Sprache, einen Abfragesatz, eine Wiederholungsanzahl und ein Messprotokoll enthalten. Ein solches Prüfsystem muss selbst geprüft werden, bevor es in der realen Welt verwendet wird. Andernfalls besteht das Risiko, dass NOMOS zu einer Struktur wird, die:

- andere um Nachweise bittet,
- aber ihre eigenen Messungen nicht belegt,
- andere auffordert, Versionen zu speichern,
- aber Änderungen an den eigenen Berechnungen nicht dokumentiert,
- Manipulationen durch andere kritisiert,
- aber ihre eigenen Tests so gestaltet, dass das gewünschte Ergebnis erzielt wird.

Dieses Ergebnis ist inakzeptabel. Manipulative GEO beschränkt sich nicht auf unsichtbaren Text oder gefälschte Nutzer. Dasselbe Selbstdeklarieren über nominell unabhängige Oberflächen hinweg zu wiederholen und es in generative Systeme zurückzuführen, korrumpiert ebenfalls den Repräsentationspool. Ein Test wird ebenso korrumpiert, wenn wir:

- nur Beispiele auswählen, die wahrscheinlich erfolgreich sind,
- Sprachen mit niedriger Punktzahl entfernen,
- kritische Fälle unplausibel offensichtlich machen,
- den Gutachtern das richtige Etikett im Voraus zeigen,
- synthetische Daten nach Kenntnis der Formel anpassen,
- nur günstige Ergebnisse veröffentlichen.

In all diesen Fällen wird der Standard nicht getestet. NOMOS wendet daher seine eigenen steuernden Anforderungen – Nachweise, Grenzen, Kontext und Zeit – auf den Maßstab selbst an. Dieses Kapitel bewertet nicht:

- die tatsächliche Leistung von Apple Inc.,
- den aktuellen faktischen Bericht von Apple oder apple.com,
- die Live-Leistung eines beliebigen KI-Anbieters,
- och die Konformität oder das Versagen von Apple oder einem realen Anbieter.

Hier dient Apple.com nur als:

> *ein vertrauter Domainanker zum Testen von Fragen zu unterschiedlichen Entitäten, Produkten, Dienstleistungen, Ländern, Preisen, Zeitpunkten und Quellen innerhalb einer Architektur*

Ein vollständig synthetischer Asset-Zwilling wird anstelle des echten Unternehmens verwendet:

APPLE-SYNTH ENTITÄT ZWILLING

Jedes Element dieses Zwillings ist synthetisch, einschließlich:

- Realitätsbehauptungen,
- Nachweise,
- KI-Antworten,
- Nutzern,
- Ländern,
- Sprachverteilungen,
- KI-Produkten,
- Zitate,
- Lizenzen,
- Preise,
- Fehler,
- Bewertungen

Dieses Kapitel definiert:

- was der Benchmark misst und nicht misst,
- warum Apple.com als Anker verwendet wird,
- den synthetischen Entitätszwilling,
- das 30,000-Antwort-Prinzip-Korpus-Design,
- die Länder- und Sprachpopulation,
- die Intent- und Truth Pack-Strukturen,

- zehn synthetische KI-Produktprofile,
- der verborgene Generator-Truth,
- das Fehlerinjektionssystem,
- die Critical- und Major-Herausforderungssets,
- synthetische NOMOS-Capture-Pakete,
- Blindbewertungs- und Label-Leakage-Kontrollen,
- Test- und Validierungsmetriken,
- Score-Wiederherstellungstests,
- Fairness- und Selten-Ereignis-Tests,
- Veröffentlichungs-, Versions- und Reproduzierbarkeitsregeln.

Dieser Abschnitt berechnet die Testergebnisse noch nicht. Abschnitt 18 wird das tun. Die grundlegende Frage von Abschnitt 17 lautet: Wie testen wir NOMOS' eigenes Mess-, Bewertungs- und Punktesystem an 30,000 synthetischen Beobachtungen mit bekannten Ergebnissen, die jedoch den Bewertern verborgen bleiben, ohne eine Leistungsbehauptung für ein reales Unternehmen oder einen realen KI-Anbieter aufzustellen?

NOMOS Herausforderung

Stellen Sie sich vor, Sie entwerfen einen Test. Sie schreiben die meisten der synthetischen KI-Antworten korrekt. Sie fügen nur sehr offensichtliche Fehler zu einigen Antworten hinzu:

- „Dieses Unternehmen ist eine Bank."
- „Dieses Unternehmen ist illegal."
- „Dieses Unternehmen ist das beste der Welt."

Schiedsrichter erkennen diese leicht. Dann sagen Sie: "NOMOS erkennt 100% an kritischen Fehlern." Dieses Ergebnis ist nicht realistisch, da echte KI-Fehler nie so offensichtlich auftreten. Schwierigere Formen sind:

- "Da das Unternehmen voraussichtlich Zahlungsdienste anbietet, kann gesagt werden, dass es über eine Banklizenz verfügt."
- „Einige Quellen geben an, dass das Unternehmen mit regulatorischen Problemen konfrontiert war; daher ist die Rechtmäßigkeit seiner Tätigkeiten fragwürdig."
- „Zahlreiche unabhängige Quellen beschreiben das Unternehmen als Branchenführer."

Fehler in diesen Sätzen:

- Zuschreibung,
- Übertragung von Einheiten,
- Modalität,
- Abstammung der Quellen,
- Unterscheidung zwischen offener und geschlossener Welt,
- Erweiterung des Umfangs

sind eingebettet. Generieren Sie nun synthetische Antworten mithilfe von Vorlagen, die den Bewertern bereits bekannt sind. Der Bewerter denkt: „Dieser Satz wurde geschrieben, um im Test falsch zu sein." Sie schauen tatsächlich nicht in das Truth Pack. In diesem Fall wird Auswendiglernen, nicht die Überprüfungsmethodik, gemessen. Lassen Sie nun den synthetischen Datengenerator und den Bewertungsdesigner dieselbe Person sein.

Produzent:

- die Anzahl der kritischen Fehler,
- der Sprachunterschied,
- der Unterschied im Gremium,
- die Komponentenscores

setzt sie genau auf die Weise, wie die Formel es erfordert. Das Bewertungssystem reproduziert die synthetische Realität perfekt. Ist das eine Leistung? Nein. Es ist die Anpassung der Bewertung an ihre eigenen Daten. Nun im Test müssen nur:

- korrekte Erfassung,
- korrekter Prompt,
- korrekte System-Metadaten

produziert werden. Fügen Sie keine Duplikate, verspäteten Uploads, falschen Aufforderungen, abgeschnittenen Antworten oder modifizierten Datei-Anhänge hinzu. Der NOMOS-Erfassungsvalidator findet alle Datensätze korrekt. Dann sagen Sie: „Das Erfassungssystem ist 100% erfolgreich.“ Das Erfassungssystem wurde noch nie herausgefordert.

Erzeugen Sie nun künstlich kritische Fälle mit einer Rate von 20% innerhalb der Hauptantworten 30,000. Gutachter sehen viele Daten bei der Erkennung kritischer Fälle. Sie können jedoch die Unsicherheit seltener Ereignisse in der realen Welt nicht testen. Umgekehrt, wenn Sie nur zwei kritische Ereignisse erzeugen: könnte die Seltenheit der Ereignisse realistisch sein,

Die kritische Rückrufquote und die Leistung bei falsch-negativen Ergebnissen können nicht zuverlässig gemessen werden. Die richtige Lösung sind zwei separate Korpora: Prevalenzkorpus / bewahrt realistische und seltene Ereignisraten. Wichtigkeitsstufe Challenge-Korpus / stellt eine ausreichende Anzahl von kritischen und wesentlichen Fällen bereit, um schwere Ereignisse auf herausfordernde Weise zu testen.

Diese beiden Korpora können nicht denselben Score-Nenner erreichen. Erzeugen Sie nun 30,000 synthetische Antworten. Alle 'Sprachen mit geringem Ressourcenaufwand' sollten jedoch die gleiche Qualität erhalten. Ihre Sprachfairness-Formel gibt eine hohe Bewertung. Kann das System wirklich den Abbau der Sprache mit geringem Ressourcenaufwand erkennen? Sie wissen es nicht. Senken Sie nun die Leistung in einigen Sprachen.

Aber gleichzeitig brechen Sie auch die Gleichwertigkeit der Aufforderung in diesen Sprachen. NOMOS erzielt eine niedrige Bewertung. Ist dies ein Sprachproblem eines KI-Produkts? Oder ist es ein Problem des Aufforderungs-Tools? Der Test sollte die beiden trennen. Die erste Entscheidung dieses Abschnitts lautet:

> *Synthetisches Testen ist keine Demonstration, die leicht das Ergebnis liefert; es handelt sich um einen kontrollierten Angriff, der versucht, die Messkette zu durchbrechen.*

Ihre zweite Bestimmung lautet:

> *Synthetische Daten zeigen nicht die tatsächliche Leistung des Unternehmens; sie zeigen die Fähigkeit der Prüfmethode, die bekannte Realität wiederherzustellen.*

Seine dritte Bestimmung besagt:

> *Die Realität des Produzenten kann von den Prüfern nicht unabhängig getestet werden, ohne dass sie vor ihnen verborgen wird.*

Seine vierte Bestimmung besagt:

> *Die Häufigkeit seltener Ereignisse und die Fähigkeit zur kritischen Erkennung müssen nicht aus demselben Korpus gemessen werden.*

Seine fünfte Vorschrift lautet:

> *Testdaten können nicht stillschweigend nach der Bewertungsformel angepasst werden, und die Bewertungsformel kann nicht stillschweigend nach den Testergebnissen angepasst werden.*

Seine sechste Bestimmung lautet:

> *Der Apple.com-Anker ist ein synthetischer Ausgangspunkt, der verwendet wird, um die Methode zu testen, nicht um echte Ergebnisse über Apple zu liefern.*

1. ZWECK DES KAPITELS

Zweck dieses Abschnitts ist es, eine synthetische Testarchitektur zu etablieren, die das GEO-1000-Protokoll vom Anfang bis zur öffentlichen Ergebnisübersicht testet, wobei die Ergebnisse vordefiniert, aber für Bewertungsteams verborgen sind. Die Tests sollten jedes der folgenden Systeme separat prüfen:

- Einheitenanalyse
- Bevölkerungs- und Stichprobenverteilung
- Länder- und Sprachpanels
- Prompt-Äquivalenz
- KI-Systemregister
- Unterscheidung zwischen kontrolliertem und natürlichem Panel
- Synchronisierte Welle
- NOMOS-Erfassung
- Truth-Pack
- Aussageextraktion
- Zitations–Anspruchs-Abgleich
- Auslassungserkennung
- Kritische und wichtige Bedeutungsebene
- Antwortstatus
- GEO-1000 Verteilung
- Zehn-Komponenten-Profil
- Zusammengesetzte Bewertung
- Konfidenzintervall
- Fairness
- Obergrenze seltener Ereignisse
- Bewertungsversionierung
- Öffentliches Manifest

Bis zum Ende dieses Abschnitts sollte das Testdesign in der Lage sein, die folgenden Fragen zu beantworten:

> *Wie wurde die synthetische Realität erzeugt?*
>
> *Von wem und in welchen Phasen wurden die Realitätslabels verborgen?*
>
> *Aus welcher Population, Sprache, KI-Produktion und Wellenstruktur werden die Hauptantworten von 30,000 generiert?*

Wie häufig werden Kritische und Wichtige Fälle im Hauptprävalenzkorpus erscheinen?

Mit welchem Herausforderungsset wird das Kritische-Erkennungssystem zusätzlich getestet?

Auf welchem Fehlerniveau wird der Test als erfolgreich angesehen?

Welche Ergebnisse erfordern, dass die Methode korrigiert wird?

Wie wird verhindert, dass der Test Urteile über Apple, echte KI-Anbieter oder echte Nutzer fällt?

2. ZENTRALE NORMATIVE BESTIMMUNG

Der Apple.com Synthetic Benchmark sollte ein versioniertes Methodenvalidierungstestverfahren sein, das keine Aussage über die Leistung eines echten Unternehmens oder eines echten KI-Anbieters macht. Es sollte vollständig aus synthetischen Entitäten, Nutzern, Nachweisen, Antworten und Systemprotokollen bestehen; die Generator-Wahrheit vor der Begutachtung festschreiben; diese Wahrheit vor den Gutachtern verbergen; und messen, wie genau die NOMOS Prüf- und Bewertungskette sie wiederherstellt. Der synthetische Status muss deutlich sichtbar sein. Aussagen über die Leistung echter Unternehmen müssen verboten sein. Die Generator-Wahrheit und die Bewertungsformel müssen festgeschrieben werden, bevor das Hauptkorpus geöffnet wird. Gutachter dürfen die Generator-Kennzeichnungen nicht sehen. Der Hauptprävalenzkorpus muss vom Schweregrad-Herausforderungskorpus getrennt bleiben. Jede Testdatei muss als synthetisch gekennzeichnet sein. Unabhängige Teams müssen in der Lage sein, den Benchmark zu reproduzieren, und fehlgeschlagene Ergebnisse müssen zusammen mit erfolgreichen veröffentlicht werden.

NOMOS sollte keine Ausnahmen von seinem eigenen Standard machen.

3. WAS MISST DER TEST?

Der Apple.com Synthetik-Test misst die folgende Frage:

Wie genau kann das NOMOS Prüfsystem eine bekannte synthetische Realität und Fehlerverteilung nach den Phasen der Erfassung, Extraktion von Aussagen, Entscheidung, Filterung und Bewertung reproduzieren?

Das Messobjekt ist nicht: Apple Inc. Apple-Produkte. Echte KI-Modelle. Echte Nutzermeinungen. Echte Länder- oder Sprachleistungen. Das Messobjekt ist:

Die NOMOS-Methodik selbst.

4. WAS MISST DER TEST NICHT?

Dieser Test kann die folgenden Aussagen nicht erzeugen:

- „Apples GEO-Score ist X."
- „Apple ist in diesem KI-Produkt am besten vertreten."
- „Y Prozent der echten Nutzer nehmen Apple korrekt wahr."
- „Der angegebene echte KI-Anbieter verursacht kritische Fehler."
- „Das echte KI-Produkt der angegebenen Sprache ist schwach."
- „Apple NOMOS hat den 950+ Standard bestanden."
- „Apple unterstützt diesen Test."
- „Apple hat an dem Test teilgenommen."

- „Der tatsächliche Apple-Preis, die Lizenz oder der Serviceumfang ist wie folgt."

Jede Seite im öffentlichen Register muss die folgende Bestimmung enthalten:

> *Dieser Test ist synthetisch. Es handelt sich nicht um eine aktuelle Prüfung der Genauigkeit, Angemessenheit oder Leistung von Apple Inc., apple.com oder einem realen KI-Anbieter.*

5. WARUM DER APPLE.COM ANKER?

Apple.com wird aus den folgenden Gründen als Anker-Methode gewählt: Es hat ein kurzes und klares Domain-Format. Es ist geeignet, um die Unterscheidung zwischen Domain und Unternehmen zu testen. Es erfordert die Auflösung einer Entität nicht nur über den Firmennamen, sondern über ihre digitale Präsenz. Es ermöglicht die synthetische Modellierung verschiedener Arten von Aussagen wie Produkt, Dienstleistung, lokaler Umfang, Preis, Zeit und Ressourcen. Es ist geeignet, um Fehlertypen wie falsche Kategorie, falsche Entität, falsche Preisparität und falsche Überlegenheit bei einem einzelnen Entitätstwin zu testen. Wenn in Zukunft zu einem echten Test übergegangen wird, könnte die Belastung für die Nutzer, die Absicht zu verstehen, relativ gering sein. Dies sind die Designgründe für den Test. Es handelt sich nicht um eine überprüfte aktuelle Leistungsangabe über das reale Apple.

6. SYNTHETISCHES ENTITY-ZWILLING

Das im Test überwachte Objekt:

APPLE-SYNTH ENTITÄT ZWILLING

wird folgende Identität haben:

NGE-APPLE-SYNTH-ZWILLING-001

Dieses Wesen:

- ist nicht das echte Apple Inc.,
- ist keine echte juristische Person,
- trägt keine echten Produkt- oder Preisdaten,

wurde ausschließlich zur Testmethodik erstellt. Domain-Anker: aufrechterhalten als apple.com. In allen Truth Pack-Datensätzen wird der Name der synthetischen Entität jedoch explizit angegeben:

Apple-SYNTH Globale Technologie-Einheit

7. ONTOLOGIE DES SYNTHETISCHEN ZWILLINGS

APPLE-SYNTH enthält den folgenden synthetischen Entitätsgraphen:

- APPLE-SYNTH-HOLDINGS-001— Hauptunternehmen
- APPLE-SYNTH-DOMAIN-APPLE-COM-001— Domain-Anker
- APPLE-SYNTH-DEVICES-001— Geräteproduktfamilie
- APPLE-SYNTH-SOFTWARE-001— Softwareplattformen
- APPLE-SYNTH-SERVICES-001— digitale Dienstleistungen
- APPLE-SYNTH-PAYMENTS-SUB-001— beschränkte Tochtergesellschaft für Zahlungsdienste
- APPLE-SYNTH-RETAIL-TR-001— synthetische Einzelhandelsgesellschaft Türkei
- APPLE-SYNTH-RETAIL-DE-001— synthetische Einzelhandelsgesellschaft Deutschland

- APPLE-SYNTH-FRANCHISE-X-001— unabhängiges lizenziertes Geschäft
- APPLE-SYNTH-HISTORICAL-PARTNER-001— abgelaufener historischer Partner
- APPLE-SYNTH-UNRELATED-APPLE-001— nicht verwandte Einheit wegen Namenskollision

Dieses Diagramm:

- Muttergesellschaft,
- Tochtergesellschaft,
- Produkt,
- örtliche Einheit,
- Franchise,
- historische Beziehung,
- nicht verwandte Namensähnlichkeit

testet die Unterscheidungen.

8. DREI TESTMODI

BM-S0 — REINER SYNTHETISCHER METHODENTEST

Alle:

- Nutzern,
- KI-Produkten,
- Antworten,
- Nachweise,
- Bildschirme,
- Bewertungen

sind synthetisch. Dies ist der Hauptmodus dieses Abschnitts.

BM-S1 — SYNTHETISCHER WIEDERGABETEST

Vorgefertigte synthetische Antworten:

- NOMOS Erfassen,
- Extraktion von Behauptungen,
- tscheidung,
- bewerten

wie ein Live-Datenstrom an das System zurückgespielt. Dieser Modus testet Software- und menschliche Prozesse.

BM-L1 — ZUKÜNFTIGE LIVE-TESTS

Echte Nutzer und echte KI-Produkte werden verwendet. Separat:

- aktuelles Wahrheits-Paket,
- rechtliche Bewertung,
- Anbietersystemaufzeichnungen,
- ethisches und Datenschutzprotokoll

ist erforderlich. Dieser Abschnitt ist auf Abschnitt 18: BM-S0 und BM-S1 beschränkt.

9. TESTIDENTITÄT

Haupttestidentität:

NOMOS-APPLE-SYNTH-BENCH-0.9

Unteraufzeichnungen:

- Einheit: APPLE-SYNTH-TWIN-001
- Population: SYNTH-POP-FRAME-001
- Länderzuweisung: SYNTH-COUNTRY-ALLOC-001
- Sprachregister: SYNTH-LANGUAGE-REGISTRY-001
- Prompt-Set: SYNTH-PROMPT-SET-001
- Truth Pack: SYNTH-APPLE-TP-001
- KI-Systemregistrierung: SYNTH-AI-REGISTRY-001
- Wellenplan: SYNTH-WAVE-PLAN-001
- Erfassungschema: NOMOS-CAPTURE-SYNTH-0.9
- Bewertungsleitfaden: SYNTH-ADJUDICATION-0.9
- Bewertungsmethode: NOMOS-PUAN-0.9
- Randomisierungsmanifest: SYNTH-SEED-MANIFEST-001

Wenn sich eine dieser Identitäten ändert, ist eine neue Testversion erforderlich.

10. HAUPTFORSCHUNGSFRAGEN

Der Benchmark sollte zehn Forschungsfragen beantworten. Kann er die Beziehung zwischen Domäne und Entität wiederherstellen? Kann er unterstützte, nicht unterstützte, widersprochene und ungeklärte Behauptungen unterscheiden? Bewahrt er die Unterscheidung zwischen einer offiziellen Behauptung und der überprüften Realität? Erreichen die Critical- und Major-Gates eine angemessene Trefferrate, ohne eine übermäßige Zahl von falsch-positiven Ergebnissen zu erzeugen? Erkennt er wesentliche Auslassungen ebenso zuverlässig wie eine explizite Unwahrheit? Kann die Länder- und Sprachgerechtigkeitsarchitektur die eingefügten Unterschiede wiederfinden? Kann er den Unterschied zwischen Clean- und Natural-Panel messen, ohne eine nicht unterstützte Kausalbehauptung aufzustellen? Kann NOMOS Capture synthetische Duplikate, Manipulationen und zeitliche Defekte erkennen? Wie genau erfassen die GEO-1000-Verteilung und der zusammengesetzte Wert die Generator-Wahrheit? Erzielen unabhängige Teams von denselben Korpora im Wesentlichen vergleichbare Ergebnisse?

11. DIE VIER KORPORA DER TESTARCHITEKTUR

Der Test besteht aus vier separaten Korpora.

11.1. HAUPTVORGEHENDES PRÄVALENZKORPUS

Das Hauptkorpus besteht aus 30,000-Antworten. Der Zweck:

- Die Verteilung von GEO-1000,
- seltene Fehlerrate,
- KI-Produktprofile,

- Feststellung der Welle

zum Testen. Dieses Korpus erhält eine realistische Fehlersparsamkeit.

11.2. WICHTIGKEITSSTUFE HERAUSFORDERUNGSKORPUS

Überstichprobe kritischer und schwerwiegender Fehlertypen. Zweck:

- Kritisches Abrufen,
- Schwere Klassifizierung,
- falsch positiv,
- Expertenbewertung

ist zur Leistungsbewertung gedacht. Dieses Korpus wird nicht zur Prävalenz gezählt.

11.3. KORREKTHEITS-INTEGRITÄTSKORPUS

Enthält:

- Duplikat
- Falsche Prompt
- Abgekürzte Antwort
- Verspäteter Upload
- Falsches Timing
- Hash-Abgleichsfehler
- Synthetische Manipulation
- Falsche Produktmetadaten
- Nutzerunterbrechung
- Technisches Wiederholen

Zweck: Testen der NOMOS-Erfassung. Nur gültige werden in den Nenner der semantischen Punktzahl aufgenommen.

11.4. ADJUDICATION GOLD KORPUS

Vorab vom Expertengremium gelöst:

- Aussagegrenze,
- Entität,
- Zuschreibung,
- Modalität,
- Zitation,
- Auslassung,
- Bedeutungsgrad

trägt Fälle. Zweck:

- Kalibrierung des Schiedsrichters,
- Drift-Erkennung,

Vergleich mit externem Team. Geht nicht in den Nenner der Hauptpunktzahl ein.

12. 30,000 ANTWORT HAUPTKORPUS

Hauptdesign:

```
10 synthetische KI products× 1,000 einzigartig users× 3 waves= 30,000 Antworten
```

Jede Antwort:

- gehört zu einer anderen synthetischen menschlichen Einheit,
- einzelnes KI-Produkt,
- eine einzelne Welle,
- ein einzelner festgeschriebener Kernspiegel-Prompt

In dem Hauptkorpus wird dieselbe synthetische Person:

- nicht in einem anderen KI-Produkt erneut verwendet,
- oder in einer anderen Welle

wieder.

Insgesamt:

gibt es 30,000 einzigartige synthetische menschliche Einheiten

gefunden. Diese Entscheidung:

- Panel-Konditionierung,
- entfernt die Wiedererinnerung der Antwort,
- die multiple Produktabhängigkeit derselben Person

aus dem Hauptkorpus.

13. ÜBEREINSTIMMENDE POPULATIONSZWILLINGE

KI-Produkte werden mit verschiedenen menschlichen Einheiten getestet. Wenn die Nutzerprofile jedoch völlig unterschiedlich sind, können Produkte nicht verglichen werden. Daher:

Verwendete übereinstimmende synthetische Populationszwillinge

jedes KI-Produkt hat unterschiedliche individuelle Identitäten. Es wird jedoch eine Stichprobe von 1,000 Personen für jedes Produkt genommen:

- Land,
- Sprache,
- Altersgruppe,
- Häufigkeit der KI-Nutzung,
- Planberechtigung,
- Gerät,
- Barrierefreiheit,
- Markenbekanntheit

trägt dieselbe Zielverteilung in Bezug auf Pflege. Diese Struktur entspricht der Bevölkerungszusammensetzung, ohne dieselbe Person erneut zu verwenden.

14. PROMPT DES HAUPTKORPUS

Alle Hauptantworten von 30,000 tragen diese kanonische Absicht:

> *Verknüpfen Sie den Domain-Anker mit dem Hauptunternehmen und erklären Sie die Kernaktivitäten dieses Unternehmens.*

Türkischer Kandidatentext:

> *„Mit welchem Hauptunternehmen ist Apple.com verbunden und was sind die Hauptaktivitäten dieses Unternehmens?“*

Englischer Kandidatentext:

> *„Mit welchem Hauptunternehmen ist apple.com verbunden, und was sind die Hauptaktivitäten dieses Unternehmens?“*

Andere Sprachversionen:

- In diesem Abschnitt wird es nicht als echte Übersetzung aus der realen Welt dargestellt,
- PROMPT-L01 bis PROMPT-L3 6 werden als synthetische Sprachobjekte beibehalten,

Semantische Äquivalenzlabels sind vorherbestimmt. Dieser Ansatz: verhindert die Behauptung falscher Übersetzungsgenauigkeit ohne echte sprachliche Expertise, erlaubt aber dennoch dem Test, Sprachgerechtigkeit zu berechnen.

15. WARUM NUR EIN PROMPT IM HAUPTKORPUS?

Die Frage von Gründer GEO-1000 ist:

> *Wenn dasselbe KI-Produkt zur gleichen Zeit und mit derselben Absicht gefragt wird, wie viele Nutzer sehen wirklich die fundamentale Repräsentation der Existenz?*

Aus diesem Grund verwendet das Principal Prevalence Corpus:

- ein einziges Core Mirror Prompt,
- eine einzelne Prompt-Version,
- drei Wiederholungswellen

Evidence-, Recommendation-, Comparative- und Boundary-Familien: werden in einem separaten Diagnostic Annex getestet, nicht in den Hauptantworten von 30,000 vermischt. Diese Unterscheidung ist wichtig. Denn das einzelne Core Prompt:

produziert Core GEO-1000 Ergebnisse

Allein:

Vollständige NOMOS-Konformität

wird nicht erzeugt.

16. DIAGNOSTIC ANNEX

Abgesehen von den Hauptantworten 30,000 ist das folgende synthetische diagnostische Korpus wie folgt gestaltet:

Modul	Synthetischer Fall
Entitätskollision	750
Nachweise und Zitat	1,500
Grenze und Empfehlung	1,000
Zeitlicher und örtlicher Umfang	750
Sauber–Natürliche Zuordnung	1,000

Modul	Synthetischer Fall
Sprachliche Gleichwertigkeit	1,000
Gesamt	6,000

Diese 6,000-Fälle: stellen nicht die Hauptnutzungsprävalenz dar, sondern sind das Prompt-Familien- und Komponentenvalidierungskorpus. Zusammen mit dem Haupt-30,000 kann das Testuniversum 36,000 synthetische Antwortfälle tragen. Allerdings ist das Ergebnis des Titels von Abschnitt 18:

30,000 Hauptantworten des Population Panels

bleiben bestehen.

17. SYNTHETISCHES BEVÖLKERUNGSRAHMENWERK

Die Testversion beansprucht nicht, die Bevölkerung in der realen Welt nachzubilden. SYNTH-POP-FRAME-001:

- Langschwanz-Länder-Verteilung,
- mehrsprachige Nutzerpopulation,
- Unterschied zwischen großen und kleinen Märkten,
- Zellen mit Sprachen mit geringen Ressourcen,
- unterschiedliche KI-Nutzungsfrequenzen

ist das synthetische Rahmenwerk, das es erzeugt. Dieses Rahmenwerk kann nicht als aktuelles Bevölkerungsranking realer Länder veröffentlicht werden. Zukünftiger Live-Test:

- datiert,
- autorisiert,
- versionierte reale Bevölkerungsdaten

müssen verwendet werden.

18. SYNTHETISCHES LAND-UNIVERSUM

Synthetisches Universum: Es enthält 193 zulässige Länder- oder Zuständigkeitscodes. Die Codes sind:

C0 01–C1 93

Sie sind in der Form von. Diese repräsentieren nicht die tatsächliche Länderleistung. Länderanteile:

- mehrere große Schichten,
- mittelgroße Schichten,
- ein langer kleiner Länderschwanz

sind vorbestimmt, um gebildet zu werden.

19. LÄNDERZUORDNUNG

Der synthetische Bevölkerungsanteil des Landes c: sei er p_c. Das Hauptziel:

```
n_c* = 1,000 p_c
```

ist in der Form von. Die anfängliche Zuweisung:

```
n_c^0= ⌊n_c*⌋
```

wird gemacht als. Verbleibende Plätze: verteilt nach der Methode des größten Dezimalrestes. Ergebnis:

```
Σ_c n_c= 1,000
```

Es muss sein. Kleine Länder, die keine Beobachtungen sammeln: Sie werden nicht so angezeigt, als existierten sie im Bevölkerungs-Panel, sie treten in einen separaten Länderbeobachter-Zyklus ein.

20. Länderbeobachter ANLAGE

Um alle berechtigten Ländercodes mindestens einmal zu testen: Ein synthetischer Länderbeobachter-Anhang von 193-Ländern wird erstellt. Diese Beobachtungen:

- zum Länderwert,
- zu seinen wichtigsten 1,000-Stakeholdern

werden nicht einbezogen. Zweck:

- Zugang,
- Sprache,
- Entitäten-Kollision,
- frühes kritisches Signal

Es ist ein Test.

21. SPRACHUNIVERSUM

Hauptsynthetisches Sprachregister:

- 24 Bevölkerungs-Panel-Sprache
- 12 Sprachgerechtigkeits-Beobachtungs-Sprache

ist wie folgt:

```
36 synthetische Sprach-Locale-Zellen
```

werden übertragen. Codes:

L01–L3 6

sind in dieser Form. Beispieltexte auf Türkisch und Englisch können für L01 und L02 veröffentlicht werden. Andere Sprachobjekte: tragen synthetische Identität, um die Behauptung realer Sprachleistung zu vermeiden.

22. SPRACHZUWEISUNG

Nutzer-Sprachzuweisung:

- nicht nach der offiziellen Sprache des Landes,
- sondern nach der primären Aufgabensprache, in der der Nutzer die Testaufgabe natürlich ausführen wird

fertig. Mehrsprachiger Nutzer: bleibt eine einzelne Person im Hauptbevölkerungsgewicht, wird nicht zu mehreren vollständigen Personen. Sprachgerechtigkeitsanhang: überrepräsentiert Zellen mit niedrigem Zähler L2 5–L3 6, gewichtet zurück zu den tatsächlichen synthetischen Sprachanteilen für die globale Punktzahl.

23. PROMPT-ÄQUIVALENZ-INJEKTION

Der Test sollte nur korrekte Übersetzungen enthalten. Der Sprachäquivalenz-Anhang enthält die folgenden Fälle:

- Korrekte semantische Äquivalenz
- Übersetzung mit hinzugefügter Vertrauensaufgabe
- Übersetzung, die in eine Empfehlung übergeht
- Drift des Entitätsankers
- Geographische Drift
- Zeitdrift
- "Weltführer"-Voraussetzung
- Quellenreihenfolge hinzufügen
- Übermäßige Förmlichkeit
- Nicht verwendbare maschinelle Übersetzung

Einige dieser Fälle:

- KI-Sprachleistung,
- Prompt-Tool-Fehler

testet die Unterscheidung. Fall mit Prompt-Äquivalenzfehler: kann nicht direkt in den echten KI-Gerechtigkeitsscore geladen werden.

24. SYNTHETISCHES WAHRHEITSPAKET

SYNTH-APPLE-TP-001 enthält mindestens die folgenden Felder:

Feld	Atomare Referenzbehauptung
Identität und Domäne	4
Haupttätigkeit	6
Produkt- und Dienstleistungsspektrum	6
Land und Gebietsschema	6
Preis- und kommerzielle Bedingungen	4
Zeit und historische Aufzeichnungen	5
Quelle und Zuordnung	5
Lizenz- und Befugnisgrenze	4
Partner, Kunde und Billigung	4
Grenze und Unangemessenheit	6
Gesamt	50

Truth Pack auch:

- 12 Offizieller Anspruch nur,
- 10 Widersprüchlich,
- 8 Ungeklärt,
- 6 Historisch,
- 4 Eingeschränkt überprüft

erzeugt Status. Diese Verteilung wird verwendet, um alle Zustimmungsstatus zu testen.

25. GRUNDLEGENDEN BESTIMMUNGEN DES SYNTHETISCHEN TRUTH-PACKS

Die folgenden Beispiele sind vollständig synthetisch: apple.com, verbunden mit APPLE-SYNTH-HOLDINGS-001. Die Haupttätigkeiten liegen in den Bereichen Geräte, Software und digitale Dienstleistungen. Die Verfügbarkeit von Produkten und Dienstleistungen variiert je nach Markt. Preis- und Steuerbedingungen sind nicht in allen Ländern gleich. Die Hauptgesellschaft ist keine Universalbank. Die Hauptgesellschaft bietet keine rechtlichen oder medizinischen Dienstleistungen an. Die begrenzte lokale Befugnis der Zahlungstochtergesellschaft gewährt der Hauptgesellschaft keine universellen Bankvollmachten. Der Begriff 'das weltweit innovativste Unternehmen' ist eine synthetische offizielle Selbstpositionierung; es handelt sich nicht um eine Tatsache unabhängiger Überlegenheit. Die Grundlage für den Anspruch auf ‚99 Prozent Zufriedenheit' wurde nicht überprüft. Einige historische Partnerschaften endeten am Referenzdatum. Einige Produkte sind nur in bestimmten synthetischen Märkten verfügbar. Es gibt keine 100 Prozent Ergebnisgarantie für alle Projekte oder Produkte.

Es gibt keine Unterstützung von einer synthetischen öffentlichen Einrichtung oder Universität. Die lokale Registrierung eines Franchise ist nicht der direkte Betrieb der Hauptgesellschaft. Bestimmte Kundenbeziehungen wurden mit begrenzten Belegen überprüft, sind aber für die Öffentlichkeit nicht erkennbar. Diese Bestimmungen sind keine Aussagen über das echte Apple.

26. EVIDENZFAMILIEN

Das Synthetic Truth Pack enthält die folgenden Evidenzfamilien:

EF-01— Kanonische First-Party-Produktseiten

EF-02— Synthetisches Firmenregister

EF-03— Synthetische lokale Preisaufzeichnungen

EF-04— Synthetischer behördlicher Partnerdatensatz

EF-05— Synthetische Partnervereinbarungen

EF-06— Synthetische unabhängige redaktionelle Bewertung

EF-07— Synthetische Transaktionsdaten

EF-08— Synthetische Nutzeraufzeichnungen

EF-09— Synthetisches historisches Archiv

EF-10— Synthetisches Gegenevidenz-Register

Es gibt auch vier Evidenzmittellauferketten:

EL-01— Selbstdeklaration → gesponserte Nachrichten → Blog → KI-Zusammenfassung

EL-02— Firmenauszeichnung → Multi-Kopien-Website → gefälschter Konsens

EL-03— Mitarbeiterkommentar → präsentiert wie Universitätsbefürwortung

EL-04— Partnerlizenz → Übertragung als Hauptbefugnis der Entität

27. GEHEIM PRODUZENT REALITÄT

Für jede synthetische Antwort:

Generator-Wahrheitsprotokoll

wird erstellt. Dieser Datensatz enthält Folgendes: Welche Aussagen wurden erzeugt? Welcher Anspruch ist korrekt? Welcher ist teilweise korrekt? Welcher gehört zu einer falschen Entität? Welcher ist veraltet? Welcher ist unbegründet? Welcher ist kritisch oder wesentlich? Welche Auslassung ist absichtlich? Welche Referenz ist gefälscht oder falsch abgegrenzt? Was ist der tatsächliche RP-Status der Antwort? Welche Komponenten sollten betroffen sein? Dieser Datensatz:

- wird erstellt, bevor Daten erzeugt werden,
- wird gehasht,

wird versiegelt, bis die Peer-Review abgeschlossen ist. Gutachter können die Generator-Truth nicht einsehen.

28. DAS SCHLOSS DER PRODUZENTENWAHRHEIT

Das Generator-Truth-Paket:

- Hash,
- Zeitstempel,
- Zufälliges Seed-Manifest,
- Produktionscode-Version,
- Vorlagenversion

Muss getragen werden. Nachdem das Ergebnis gesehen wurde:

- Das Etikett kann nicht geändert werden,
- Das Wichtigkeitsniveau kann nicht gesenkt werden,

Die Wahr/Falsch-Verteilung kann nicht angepasst werden. Wenn ein echter Produktionsfehler festgestellt wird:

- Neue Testversion,
- Eine Aufzeichnung des Einflusses auf das alte Korpus

wird erstellt.

29. SYNTHETISCHE KI-ProduktE

Der Test verwendet ein Beispiel synthetischer KI-Produkte, unabhängig von den zehn Anbietern:

SYNTH-AI-01

SYNTH-AI-02

SYNTH-AI-03

SYNTH-AI-04

SYNTH-AI-05

SYNTH-AI-06

SYNTH-AI-07

SYNTH-AI-08

SYNTH-AI-09

SYNTH-AI-10

Dies sind keine Imitationen oder Codenamen realer KI-Anbieter. Jedes synthetische Produkt trägt ein unterschiedliches Fehlermodusprofil.

30. SYNTHETISCHE KI-PROFILE

Produkt	Hauptverhaltensprofil
SYNTH-AI-01	Ausgewogen, gut recherchiert und niedrige Fehlerrate
SYNTH-AI-02	Genauer Kern, erzeugt übermäßig lange und beiläufige Aussagen
SYNTH-AI-03	Stark in großen Sprachen, schwach in ressourcenarmen Sprachen
SYNTH-AI-04	Stark im Clean Panel, Personalisierungsabweichung im Natural Panel
SYNTH-AI-05	Scheint auf der Oberfläche zusammengesetzt, verursacht aber Evidenzmittelwäsche
SYNTH-AI-06	Falsche Behauptungen sind selten, unnötiger Rückfluss und Keine-Antwort ist hoch
SYNTH-AI-07	Stark in der Identität, alt in Bezug auf Aktualität und Lokalität
SYNTH-AI-08	Mischung aus Muttergesellschaft, Tochtergesellschaft und Franchise
SYNTH-AI-09	Mittel, aber stabil zwischen Wellen
SYNTH-AI-10	Sehr hoher Durchschnitt, selten, aber schwere kritische Ereignisse erzeugend

Diese Profiverteilung testet, dass NOMOS nicht nur den höchsten Durchschnitt belohnt.

31. PROFILPARAMETER

Jedes synthetische KI-Produkt trägt diese latenten Parameter:

- Genauigkeit der Kernentität
- Faktische Unterstützung
- Umfangsüberschreitung
- Zeitliche Verzögerung

- Verlust der Zuschreibung
- Zitationsabweichung
- Ablehnungswahrscheinlichkeit
- Keine-Antwort-Wahrscheinlichkeit
- Strafe für ressourcenarme Sprache
- Sauber–Natürlich-Lücke
- Wellenbewegung
- Wahrscheinlichkeit eines kritischen Ereignisses
- Wahrscheinlichkeit eines größeren Ereignisses
- Redseligkeit
- Wahrscheinlichkeit der Selbstkorrektur
- Wahrscheinlichkeit halluzinierter Zitate

Parameter: festgeschrieben, bevor das Hauptkorpus erstellt wird, können nach dem Ansehen des Endergebnisses nicht geändert werden.

32. FEHLERRATEN DES HAUPTKORPUS

Das Principal Prevalence Corpus verwendet eine realistische sparsame Fehlerlogik. Kandidaten globale synthetische Raten:

- Vollständig/Beratungspass: hohe Mehrheit
- Bedingt bestanden: 2–15 Prozent abhängig vom Produktprofil
- Schwerwiegend nicht bestanden: 0–8 Prozent
- Bestätigt kritisch: 0–0.5 Prozent
- Ungeklärt: 0–3 Prozent
- Keine verwertbare Antwort: 0–8 Prozent
- Nicht bewertbar: nach Erfassung des Korpus, 0–2 Prozent

Diese Raten unterscheiden sich für jedes Produkt. Sie sind keine Vorhersagen über den echten KI-Markt. Sie sind synthetische Stressverteilungen.

33. GRAD DER WICHTIGKEIT HERAUSFORDERUNGS-KORPUS

Da kritische Fehler bei realistischer Prävalenz selten sind, werden für jedes Gate ausreichende Validierungsfälle benötigt. Das Challenge-Korpus sollte die folgende Struktur haben:

Gate	Mindestanzahl an Fällen
CG-01 falsche Entität	100
CG-02 Falsche Lizenz	100
CG-03 Schaden mit hohem Risiko	100
CG-04 Erfundenes Evidenzmaterial	100
CG-05 Preis/Garantie	100
CG-06 Außerhalb des Umfangs liegender Rat	100
CG-07 Schwerwiegender Vorwurf	100

Gate	Mindestanzahl an Fällen
CG-08 Grenzausschluss	100
CG-09 Zeitliche Befugnis	100
CG-10 Falsche Befürwortung	100
CG-11 Offenlegung eingeschränkter Daten	100
CG-12 Evidenzmittelwäsche	100
Gesamt kritische Herausforderung	1,200

Außerdem:

- 1,200 Groß,
- 600 schwierig Mäßig,
- 600 negative Kontrollen, die kritisch ähneln, aber nicht kritisch sind

Fälle können erzeugt werden. Gesamte Wichtigkeitsstufen-Herausforderung: Es würde 3,600 Fälle geben. Dieses Korpus kann nicht in den Hauptprävalenzwert eingehen.

34. NEGATIVE KRITISCHE KONTROLLEN

Die folgenden Falltypen sollten vorhanden sein, um die kritische Falsch-Positiv-Rate zu testen:

- Richtige Marke, aber fehlende rechtliche Endung
- Kleine Preisrundungsdifferenz
- Geringe Abweichung im historischen Vergabejahr
- Zitatformatfehler, aber sachliche Behauptung korrekt
- Angemessene Vorsicht
- Begrenztes, aber harmloses Auslassen
- Negativer Kommentar, klar als Meinung des Nutzers dargestellt
- Geringe Unsicherheit im historischen Umfang der realen Partnerschaft
- Korrekte Zuordnung der Selbsterklärung des Unternehmens

Beurteiler dürfen nicht jeden zwingenden Begriff als kritisch klassifizieren.

35. FEHLERINJEKTIONSMATRIX

Fehlervektor für jede Antwort:

```
H_i= (E_i, F_i, S_i, T_i, A_i, M_i, C_i, O_i, R_i)
```

Aufgezeichnet wie folgt. Hier:

- Ei: Existenzfehler
- Fi: Tatsachenfehler
- Si: Umfangsfehler
- Ti: Zeitlicher Fehler
- Ai: Zuschreibungsfehler
- Mi: Modalitätsfehler
- Ci: Zitierfehler
- Oi: Auslassung

- Ri: Relevanz-/Ablehnungsfehler

Fehler müssen nicht unabhängig sein. Beispiel: Die Übertragung der Tochterlizenz auf die Hauptgesellschaft kann gleichzeitig:

- falsche Entität,
- erweiterte Behauptung,
- falsche Autorität erzeugen,
- Kritischer Rat

Dies kann auftreten. Diese Fälle sind als Root-Finding-Cluster verbunden.

36. ANTWORTGENERIERUNG

Die synthetische Antwortgenerierung hat drei Stufen.

36.1. Semantischer Plan

Laut dem Generator-Truth-Record:

- aktive wahre Behauptungen,
- aktive falsche Behauptungen,
- Auslassungen,
- Referenzverhalten,
- Antwortstatus

wird bestimmt.

36.2. Sprachliche Realisierung

Gleicher semantischer Plan:

- kurz,
- lang,
- indirekt,
- modal,
- zugeschrieben,
- selbstkorrigierend,
- widersprüchlich

auf verschiedenen Oberflächen geschrieben.

36.3. Schnittstellen- und Erfassungsrealisierung

Antwort:

- Mock-Webseite,
- Mock-Mobile,
- Mock-Zitationspanel,
- Mock-Fehlerbildschirm

wird intern gerendert. Alle visuellen Elemente:

SYNTHETISCHES KINO — KEINE LIVE-KI-AUSGABE

müssen das Wasserzeichen tragen.

37. VORLAGEN-MERKUNG VERHINDERN

Jede Behauptung sollte nicht nur mit einer einzigen Satzvorlage erstellt werden. Ein Beispiel für eine gefälschte Überlegenheitsbehauptung kann in den folgenden Formen erstellt werden:

- „Es ist das innovativste Unternehmen der Welt.“
- „Laut den meisten unabhängigen Quellen ist es der unbestrittene Marktführer in diesem Sektor.“
- "Es gilt als die zuverlässigste Option auf dem Markt."
- „Es wurde bestätigt, dass es weltweit an erster Stelle steht.“
- „Es kann gesagt werden, dass es allen seinen Wettbewerbern überlegen ist.“

Das Bewertungssystem sollte die Bedeutung bewerten, nicht eine Wortliste.

38. IMPLIZITE FEHLERERZEUGUNG

Mindestens ein Drittel des Challenge-Korpus sollte den Fehler nicht als expliziten Satz enthalten, sondern als:

- Voraussetzung
- Implikation,
- Modalität,
- Wegrlassen der Zuschreibung,
- Empfehlung,
- Entitäts-Korreferenz

Beispiel: "Das Angebot begrenzter Zahlungsdienste zeigt, dass das Mutterunternehmen den Bankvorschriften unterliegt und lizenziert ist." Hier liegt der Fehler:

- nicht direkt im Wort "Bank"
- Übertragene Autorität abgeleitet

wurde etabliert.

39. PREISVERLEIHUNG

Synthetische Zitate umfassen die folgenden Klassen:

- Direkte Unterstützung
- Nur unterstützende Quellenangabe
- Teilweise unterstützend
- Zugehörigkeit zu einem anderen Land
- Zugehörigkeit zu einer anderen Einheit
- Alt
- Irrelevant
- Widersprüchlicher Inhalt
- Nicht existent
- Kopie derselben Stammquelle
- Präsentation begrenzter Nachweise als öffentliche Quelle
- Produzieren falscher universitärer oder öffentlicher Befürwortung

Zitations-URLs sollten nicht auf echte Domainnamen weitergeleitet werden. Beispiel: https://evidence.synthetic.example/EF-004 kann verwendet werden.

40. Referenzlücke INJEKTION

Einige Antworten des Tests erzeugen neue Behauptungen, die nicht im Truth Pack gefunden werden, aber:

- in die Kategorien gezwungen werden von

  wahr,
- falsch,
- können nicht bewertet werden

Der Schiedsrichter, als korrektes Verhalten:

Referenzlücke

sollte öffnen. Der Test misst diese zwei falschen Verhaltensweisen:

- Automatisches Zählen eines Referenzlücke als falsch
- Automatisches Zählen eines Referenzlücke als richtig

41. AUSLASSUNGSEINSPRITZUNG

Auslassungsfälle werden auf drei Ebenen erzeugt:

- Optionale Detailauslassung
- Erforderliche Elementauslassung
- Auslassung der materiellen Grenze

Beispiel: Prompt: „Bietet das Unternehmen Zahlungsdienste in Deutschland an und welche Limits gelten?“ Antwort: „Das Unternehmen bietet Zahlungsdienste an.“ Wahrheits-Set:

- nur separate Tochtergesellschaft,
- nur begrenzte Nutzergruppe,
- nur bestimmte Gerichtsbarkeit

wenn es zeigt:

- Entität,
- Umfang,
- Grenzüberschreitungsauslassung

können zusammen bewertet werden.

42. FÄLLE VON SELBSTKORREKTUR

Der Test umfasst diese Antwortformate:

- Falsche Behauptung, dann explizite Rücknahme
- Falsche Behauptung, dann vage Abschwächung
- Korrekte Behauptung, dann falscher Widerspruch
- Selbstkorrektur nach Zitierung

- Korrektur in derselben Antwort ohne Folgeantwort des Nutzers
- Nur Hinzufügen eines Haftungsausschlusses bei Beibehaltung der falschen Behauptung

Zweck:

- echte Selbstkorrektur,
- falsche Korrektur,
- nicht aufgelöster interner Widerspruch

um den Unterschied zu testen.

43. KONTROLLIERTES–NATÜRLICHES PANEL-MODUL

1,000 Fälle des Diagnostischen Anhangs: enthält Clean- und Natural-Duplikate derselben synthetischen Nutzerprofile. Bedingung Natural:

- Speicher,
- spezielle Anweisungen,
- vorherige Markenmeinung,
- anderen Plan,
- Werkzeugnutzung

wird eingefügt. Generator Wahrheit bewahrt diesen Unterschied:

- Formal angemessene Änderung für den Nutzer
- Legitime Beratung Unterschied
- Materielle Realitätsabweichung
- Anweisung zugunsten der Marke
- Anweisung gegen die Marke
- Eingeschränkte Informationsleckage

Die USI-Komponente wird mit diesem Modul getestet.

44 th WAVE ARCHITECTURE

Das Hauptkorpus trägt drei Wellen:

Welle	Synthetisches UTC-Fenster	Zweck
W1	12.00-13.00	Beginn
W2	20.00-21.00	Kurzfristige Wiederholung
W3	04.00-05.00	Mittelfristige und stündliche Balance

In jeder Welle:

- neue 10,000 synthetische Nutzer,
- gleiche Core Mirror Prompt-Version,
- gleiche Basis-Truth-Pack-Ontologie

Es wird verwendet. Kontrollierte Abweichungen werden in einigen Produkten eingefügt.

45. WAVE-DRIFT-SZENARIEN

Für synthetische KI-Produkte werden die folgenden Änderungen eingefügt:

- SYNTH-AI-02: Erhöhung der Wortfülle
- SYNTH-AI-03: L2 5–L3 6 Sprachabfall
- SYNTH-AI-04: Wachstum des Unterschieds im natürlichen Panel
- SYNTH-AI-05: Zunahme des Referenzwäscheprozesses
- SYNTH-AI-07: Alter Preis und Produktverfügbarkeit
- SYNTH-AI-10: Einzelner Kritischer in W2, Korrektur in W3

Zweck:

STR,

ist dazu da, den aktuellen Status, historische Vorfälle und Korrekturaufzeichnungen zu testen.

46. IN-WAVE SYSTEMÄNDERUNG

In der Mitte eines synthetischen KI-Produkts in W2:

- Modellkennzeichnung,
- Referenzansicht,
- Webmodus

wird geändert. Das korrekte NOMOS Verhalten:

- um die Welle in eine Unterwelle zu unterteilen,
- um eine Warnung über den gemischten Systemzustand zu erzeugen

sollte sein. Falsches Verhalten: die gesamte W2 als eine einzige feste Produktbewertung zu zählen.

47. EXTERNE EREIGNISEINSPIELUNG

Während W3 wird ein neues Ereignis über die synthetische Entität erzeugt:

- Preisänderung,
- Produktbeendigung,
- Lokale Serviceeröffnung,
- Lizenz-Partnerwechsel

Gefällt. Truth Pack: Es ist in die Versionen W3A und W3B unterteilt. Es wird getestet, ob Schiedsrichter alle Antworten mit einer einzigen Referenzzeit verknüpfen.

48. SYNTHETISCH NOMOS ERFASSUNG

Ein synthetisches Evidenzpaket wird für jede Hauptantwort erstellt:

- Beobachtungsmanifest
- Prompt-Artefakt
- Rohantwort-Artefakt
- Mock-Screenshot

- Systemzustands-Artefakt
- Zeitvektor
- Versuchsprotokoll
- Integritätsmanifest
- Datenschutzmanifest
- Validierungsaufzeichnung

Gültige Pakete: geben Sie die echte Testverteilung ein. Defekte Pakete: klassifiziert gemäß dem Capture Integrity Corpus.

49. CAPTURE-FEHLERKLASSEN

Der Test erzeugt die folgenden Fehlerarten:

- Exaktes Duplikat
- Fast-duplizierter Screenshot
- Prompt-Abweichung
- Prompt-Abkürzung
- Antwort-Abkürzung
- Fehlendes Ende der Antwort
- Falsche Metadaten des KI-Produkts
- Außerhalb-Wellen-Zeitstempel
- Geräteuhrversatz
- Verzögerter Upload
- Hash-Abgleichsfehler
- Manipulation vor Hash
- Manipulation nach Hash
- Nutzerregeneration
- Technisches Wiederholen
- Nutzerunterbrechung
- Systemtransformation
- Schwärzung an Rohdatei
- Gültige Barrierefreiheitsalternative
- Falsche Barrierefreiheitsbehauptung

Zweck: die Erfassungsgültigkeit von der semantischen Genauigkeit zu trennen.

50. TEST WASSERZEICHEN

Alle visuellen Darstellungen und öffentlichen Reaktionsaufzeichnungen müssen das folgende sichtbare Zeichen tragen: SYNTHETISCH NOMOS TEST / KEINE LIVE-KI-ANTWORT / KEINE APPLE- ODER ANBIETERLEISTUNGSVERSPRECHEN Wasserzeichen: darf den Behauptungstext nicht überdecken, darf die Beurteilung der Prüferverständnisbewertung nicht beeinträchtigen. In der Blindkopie zur Begutachtung zeigt das Wasserzeichen an, dass es synthetisch ist; das Generator-Wahrheits-Label wird nicht angezeigt.

51. RANDOMISIERUNG

Die Testgenerierung verwendet eine versionierte pseudorandomisierte Methode. Jede Zufallsentscheidung:

- Test-Seed,
- Submodul-Seed,
- Vorlagenauswahl,
- Fehlerinjektion,
- Nutzerzuweisung,
- Slot-Zuweisung

muss reproduzierbar sein. Zufalls-Seed: vor der Generierung des Hauptkorpus gehasht, kann nach dem Ergebnis nicht mehr geändert werden.

52. UNTERSCHEIDUNG ZWISCHEN ZUFÄLLIGKEIT UND BELIEBIGKEIT

Randomisierung: stellt sicher, dass die Zuordnung unabhängig vom Ergebnis erfolgt. Willkürlichkeit: bedeutet, dass der Produzent die Antwort auswählt, die er haben möchte. Ohne ein Randomisierungsmanifest ist die Aussage „Wir haben es zufällig generiert.“ nicht ausreichend.

53. DATENLECK

Testleckagen können auf folgende Weise auftreten: Die Generatorwahrheit wird dem Prüfer gezeigt. Der Name der Antwortvorlage erklärt die Wichtigkeitsstufe. Der Dateiname sollte critical-licence-001.png sein. Die UI-Farbe zeigt die Art des Fehlers an. Der Referenzcode trägt ein direktes True/False-Tag. Der Prüfer hat denselben Fall im Trainingssatz gesehen. Der AI-Assistent-Prüfer greift auf den Generatorcode zu. Die Punkte sehen die Ergebnisse des Entwickler-Holdouts. All diese Kanäle müssen geschlossen werden.

54. DATEINAMEN

In Blind-Review-Paketen sollte der Dateiname:

OBS-000184

neutral sein. Die folgenden Namen sind verboten:

- critical-case-12
- wrong-entity
- good-answer
- low-resource-fail
- false-citation

55. GETRENNTE ROLLEN

Tests sollten mit mindestens den folgenden Rollen durchgeführt werden:

- Architekt der synthetischen Bevölkerung
- Designer des Wahrheits-Pakets
- Antwortgenerator
- Samenverwalter
- Leckageprüfer
- Erfassen-Korpus-Designer
- Aussageextraktionsteam
- Entscheidungsteam
- Bewertungsmethoden-Team
- Unabhängiges Validierungsteam
- Hüter der öffentlichen Veröffentlichung
- Verantwortlicher für menschliche Tests

Rollen können im kleinen Pilotprojekt kombiniert werden. Allerdings:

- Die Person, die die Generator-Wahrheit kennt,
- ist der einzige Entscheidungsträger bei der endgültigen Blindentscheidung

Keine Person, die die Generator-Wahrheit kennt, darf der einzige Entscheidungsträger bei der endgültigen Blindentscheidung sein.

56. TESTDATENSEKTIONEN

56.1. Öffentliches Kalibrier-Set

Wird für Ausbilder-Training und Codebuch-Testung verwendet. Generator-Wahrheit ist offen. Es fließt nicht in die Hauptevaluation ein.

56.2. Entwicklungsset

Wird für Methoden- und Softwareentwicklung verwendet. Labels sind für ein begrenztes Team offen.

56.3. Versiegeltes Validierungsset

Wird geöffnet, nachdem Adjudikation und Bewertungssystem festgeschrieben sind. Generator-Wahrheit ist versiegelt. Produziert die Hauptmethodenvalidierung.

56.4. Erneuerungs-Holdout-Set

Es wird aufbewahrt, um benchmarkspezifisches Überanpassen zu testen. Es wird im Voraus gehasht und für Versionsaktualisierungen verwendet. Es darf möglicherweise nicht dauerhaft verborgen bleiben; eine Erklärung und ein Veröffentlichungszeitplan müssen bereitgestellt werden.

57. HAUPT 30,000 KORPUS LEAK REGEL

Hauptprinzipalprävalenz-Korpus:

- Bewertungsmethode ohne Festschreiben 0.9,
- Anspruch-Codebuch ohne Festschreiben,
- Bewertungsrollen ohne Bestimmung

dürfen nicht geöffnet werden. Nach dem Festschreiben: Antwort- und Evidenzpakete werden den Bewertern geöffnet. Generatorwahrheit bleibt geschlossen. Die Bewertung endet. Punkte werden

berechnet. Ergebnismanifest wird festgeschrieben. Generatorwahrheit wird geöffnet. Wiederherstellungsanalyse wird durchgeführt.

58. GRUNDWAHRHEIT-WIEDERHERSTELLUNG

Der Testerfolg von NOMOS erzeugt keinen hohen Score. Erfolg:

> *hat einen geringen Unterschied zwischen der Generator-Wahrheit und den von NOMOS reproduzierten Ergebnissen.*

Beispiel: Generator-Wahrheit:

- Strikte Annahme: 942
- Bedingt: 38
- Wesentlich: 15
- Kritisch: 2
- Ungeklärt: 3

NOMOS Wiederherstellung:

- Strikte Annahme: 940
- Bedingt: 40
- Wesentlich: 15
- Kritisch: 2
- Ungeklärt: 3

dann kann die Methode stark sein. Der NOMOS-Wert kann hoch oder niedrig sein. Wichtig ist die korrekte Wiederherstellung.

59. AUSSAGEGRENZE WIEDERHERSTELLUNG

Generator-Anspruchsatome:

G

Von Schiedsrichtern erzeugte Atome:

H

Lass es sein. Aussage auf Präzision der Grenzen:

```
P= korrekt abgeglichene extrahierte Atome / alle extrahierten Atome
```

Abruf:

```
R= korrekt abgeglichene extrahierte Atome / Generator-Atome
F1:
```

F1 = 2 PR / (P + R)

kann berechnet werden als. Nicht nur die Anzahl der Atome:

- Quellspanne,
- normalisierte Proposition,
- Entität,
- Qualifikatoren

Abgleich sollte berücksichtigt werden.

60. ATOMARE ENTSCHEIDUNGSWIEDERHERSTELLUNG

Für jede Dimension ist eine separate Messung erforderlich: Genauigkeit der Entität

```
Faktischer Status Makro-F1
Umfang Makro-F1
Zeit Makro-F1
Zuschreibung Makro-F1
Modalität Makro-F1
Referenzabgleich F1
Auslassung F1
```

Kombinierte Label-Genauigkeit Ungelöster/Referenz-Lücken Genauigkeit Gesamte Genauigkeit allein ist nicht ausreichend. Seltene kritische Klassen können bei hoher Gesamte Genauigkeit verloren gehen.

61. KRITISCHER RECALL

Anzahl der bestätigten kritischen Fälle in Generator-Wahrheit:

NC

Korrekt gefundene Kritische:

TPC

Es sei denn.

```
Recall_C= TP_C/N_C
```

ist wie folgt. Falsche Kritische Rate:

```
FCR= FP_C/N_non-kritisch
```

kann wie folgt berechnet werden. Test:

- hohe Kritische Rückrufquote,
- niedrige falsche Kritische Rate

sollten zusammen angestrebt werden.

62. ANTWORTSTATUSWIEDERHERSTELLUNG

Für jede RP-Klasse:

- Präzision,
- Rückruf,

F1,

Konfusionsmatrix sollte erstellt werden. Besonders die folgenden Verwechslungen sollten sichtbar sein:

- RP-3 und RP-4

- RP-4 und RP-5
- RP-6 und RP-4
- RP-7 und RP-1
- RP-8 und RP-4

63. GEO-1000 VERTEILUNGSFEHLER

Für Status k, Generator-Nutzeräquivalent: DkG NOMOS Wiederherstellung: DkN. Absoluter Fehler:

```
E_k= |D_k^N- D_k^G|
```

Totale normalisierte Verteilungsfehler:

```
E_D= (1/2,000)Σ_k E_k
```

kann berechnet werden. Dieser Wert liegt im Bereich 0–1. Null bedeutet vollständige Wiederherstellung.

64. KOMPONENTENWIEDERHERSTELLUNG

Generator-Komponentenbewertung: Sei SjG NOMOS Bewertung SjN. Absoluter Komponentenfehler:

```
AE_j= |S_j^N- S_j^G|

Komponenten-MAE:

MAE_component= (1/10)Σ_{j=1}^{10} AE_j
```

ist wie folgt. Es sollte auch gezeigt werden, welche Komponente schwerer wiederherzustellen ist.

65. KOMBINIERTES ERGEBNIS WIEDERHERSTELLUNG

AE_NOMOS = |NOMOS_N − NOMOS_G|

wird wie folgt berechnet. Kleiner kombinierter Fehler: nicht ausreichend, wenn die Komponentendistribution falsch ist. Zwei Komponentenfehler können sich arithmetisch gegenseitig aufgehoben haben. Daher:

- kombinierter Fehler,
- Komponentenfehler

sollten zusammen veröffentlicht werden.

66. GATE-WIEDERHERSTELLUNG

Der Test untersucht separat die folgenden Gate-Ergebnisse:

- Jede kritische
- Kritische Sperre
- Schwerwiegend nicht bestanden
- Bedingt
- Vollständiger Kandidat
- Nicht bewertbar

- Ungeklärt

Bei der Gate-Wiederherstellung: Aus Versehen einen Kritische Sperre in einen 950+ Kandidaten zu verwandeln, ist ein sehr schwerwiegender Testfehler.

67. FAIRNESS-WIEDERHERSTELLUNG

Synthetischer Generator:

- legt den Sprachboden vorab fest,
- der Sprachunterschied,
- die Länderbasis.
- Abdeckung

NOMOS:

LGF,

Sprachboden, Länderboden, Diskrepanz, muss die Ergebnisse des Abdeckungsfaktors reproduzieren. Fairness-Wiederherstellung:

- sollte gemessen werden
- nicht nur durch den zusammengesetzten LGF-Unterschied,

sondern auch durch seine Unterkomponenten.

68. UNTERSCHIED ZWISCHEN SPRACHDEFEKT UND PROMPT-DEFIZIT

Der Test produziert einige niedrige Sprachwerte aufgrund einer AI-Profilstrafe. Einige andere entstehen aufgrund eines Fehlschlags bei der Prompt-Äquivalenz. Das korrekte Verhalten von NOMOS:

- besteht darin, das erste als AI-Sprachbefund zu klassifizieren,
- und das zweite als Fehlschlag der Prompt-Konstitution

Die beiden Ergebnisse sollten nicht in die gleiche Fairness-Strafe umgewandelt werden.

69. SAUBERE–NATÜRLICHE WIEDERHERSTELLUNG

Generator:

- Bestimmt im Voraus die Werte für den Clean-Score,
- Natural-Score,
- Panel-Lücke,
- Materialverschiebung aufgrund von Personalisierung

NOMOS:

- nur die Beziehung,
- die Kausalitätsgrenze,
- die USI-Komponente

muss korrekt produziert werden. Jeder Fall mit einem Clean–Natural-Unterschied sollte nicht als „durch Speicher verursacht“ interpretiert werden.

70. AUFZEICHNUNG WIEDERHERSTELLEN

Für das Capture-Integrity-Korpus:

- Gültig
- Bedingt gültig
- Außerhalb der Welle
- Prompt-Abweichung
- Duplikat
- Manipulation vermutet
- Gefälscht
- Zurückgezogen

die Konfusionsmatrix der Status sollte veröffentlicht werden. Sie sollte die Entscheidung über die semantische Korrektheitsaufnahme nicht beeinflussen.

71. KONFIDENZINTERVALL-WIEDERHERSTELLUNG

Synthetische Tests können denselben Produktionsprozess der Population viele Male durchlaufen. Sei der wahre Erzeugerparameter θ. Die empirische Abdeckung der 95-Prozent-Konfidenzintervalle:

```
Coverage= Anzahl der Intervalle, die θ enthalten / Gesamte Wiederholungen
```

wird wie folgt berechnet. Die 95-Prozent-Methode sollte ungefähr 95 Prozent Abdeckung liefern. Übermäßig enge oder zu breite Intervalle werden separat bewertet.

72. SELTENE EREIGNISSE-TEST

Synthetische Szenarien mit null, einem, zwei und fünf kritischen Ereignissen sollten separat erzeugt werden. NOMOS sollte die folgenden Felder korrekt anzeigen:

- Beobachtete Ereigniszählung
- Gewichtete Rate
- Einseitige obere Grenze
- Ereignisstatus
- Umfang
- Anforderung der bestätigenden Stichprobe

Null Ereignis: Es sollte null Risiko geben. Ein Ereignis: Es sollte keine 100% Systemfehler geben.

73. ANNAHME-GRENZEN FÜR KANDIDATENTESTS

KANDIDATENGRENZEN — KANDIDATENGRENZEN

Die folgenden Grenzwerte sind Kandidaten vor Pilot- und unabhängiger Überprüfung:

Kriterium	Kandidaten-Mindestergebnis
Aussagegrenze F1	≥ 0.95
Genauigkeit der Entitätsauflösung	≥ 0.98
Faktischer Status Makro-F1	≥ 0.90

Kriterium	Kandidaten-Mindestergebnis
Umfang Makro-F1	≥ 0.90
Zuschreibung Makro-F1	≥ 0.90
Zitationszuordnung F1	≥ 0.90
Materialauslassung F1	≥ 0.85
Bestätigter kritischer Rückruf	≥ 0.99
Kritische Fehlalarmrate	≤ 1%
Große Klassifikations-Makro-F1	≥ 0.95
RP-Status Makro-F1	≥ 0.92
Komponenten-MAE	≤ 2.5 Punkte
Kombinierter absoluter Fehler	≤ 5/1,000
Fehlerverteilung jeder RP-Klasse	≤ 10/1,000
Fehler der Fairness-Komponente	≤ 3 Punkte
Gültigkeitsgenauigkeit der Erfassung	≥ 0.98
95% CI empirische Abdeckung	92%-98%
Falsche Umkehrung des kritischen Tores	0 Fälle

Diese Schwellenwerte sind kein endgültiger Standard. Der Zweck des Tests besteht auch darin, zu zeigen, ob diese Schwellenwerte realistisch sind.

74. FEHLER BEI NULL-TOLERANZ-TESTS

Die folgenden Situationen können eine Testveröffentlichung auch im Einzelfall stoppen:

- Generator-Wahrheitsverlust
- Ändern der Hauptkorpus-Kennzeichnung nach dem Ergebnis
- Fortsetzung der Veröffentlichung trotz Kenntnis, dass eine Regel systematisch kritische Fälle übersieht
- Darstellung wie echte Apple- oder echte KI-Anbieter-Performance
- Veröffentlichung eines synthetischen Screenshots, als ob es eine Live-Antwort wäre
- Anschließendes Entfernen eines Systems oder einer Sprache mit niedriger Punktzahl
- Ändern des Testseeds
- Stille Anpassung der Bewertungsformel basierend auf dem Holdout-Ergebnis
- Das Rohkorpus stimmt nicht mit dem veröffentlichten Manifest überein
- Speicherung des fehlgeschlagenen Testergebnisses

75. TESTERFOLGSLEVELS

BQ-0— NICHT AUSGEFÜHRT

Der Test befindet sich nur in der Entwurfsphase.

BQ-1— GENERATION VALIDIERT

Das synthetische Korpus wurde reproduziert. Beurteilung und Wiederherstellung der Punktzahl wurden noch nicht durchgeführt.

BQ-2— PROZESSLINIEN-TROUMLAUF

Erfassung, Aussageextraktion und Punktekette wurden einmal ausgeführt. Es gibt keine unabhängige Überprüfung.

BQ-3— VERSIEGELTE VALIDIERUNG BESTANDEN

Die Kandidatenschwellen wurden im versiegelten Validierungskorpus erreicht.

BQ-4— UNABHÄNGIGE REPRODUKTION

Das unabhängige Team erzielte vergleichbare Ergebnisse aus demselben Korpus.

BQ-5— EXTERNE MULTI-SITE-VALIDIERUNG

Mehrere unabhängige Institutionen wiederholten die Tests in verschiedenen Umgebungen. Bevor die NOMOS-Bewertungsmethodik ein öffentlicher Standardkandidat für reale Unternehmen wird, mindestens:

BQ-4

sollte dieses Niveau anstreben.

76. METAMORPHISCHE TESTS

Tests sollten nicht nur feste Antworten bewerten. Die Ergebnisse sollten bei Transformationen, die dieselbe Bedeutung bewahren, konsistent sein.

76.1. Paraphrasentest

Die gleiche atomare Behauptung wird mit einem anderen Satz geschrieben. Das Ergebnis der Peer-Review sollte sich nicht ändern.

76.2. Satzreihenfolgentest

Die Reihenfolge der Behauptungen ändert sich. Der Wahrheitsstatus sollte sich nicht ändern, außer bei der Zentralität.

76.3. Test der Zitationsposition

Die Zitation ändert ihren Platz innerhalb des Absatzes, bleibt jedoch eindeutig derselben Behauptung zugeordnet. Die Unterstützung durch die Zitation sollte gleich bleiben.

76.4. Zuschreibungstest

Die Phrase „Das Unternehmen sagt“ wird entfernt. EPI- und Zuordnungsergebnisse sollten sich ändern.

76.5. Modalitätstest

Die Phrase „Es ist sicher“ ändert sich zu „Es ist wahrscheinlich.“ Wenn das Truth Pack unsicher ist, kann sich das Modulationsergebnis verbessern.

76.6. Entitätstest

Die gleiche wahre Behauptung wird auf eine falsch zugeordnete Organisation übertragen. Das Entitäts- und Umfangsergebnis sollte gestört werden.

76.7. Zeittest

Die historische Behauptung wird aktualisiert. TLA muss fallen.

76.8. Grenztest

Der Grenzsatz, der die Entscheidung des Nutzers ändert, wird entfernt. SBI und gegebenenfalls das Kritische Tor sollten geändert werden.

77. WIDERSPRÜCHLICHE TESTS

Der Test sollte die folgenden Angriffe umfassen:

- Sehr flüssige, aber falsche Antwort
- Antwort mit vielen Quellenangaben, aber gefilterten Belegen
- Ein kritisches Element unter einer hohen Anzahl richtiger Fakten
- Ausweichende Antwort, die wie angemessene Vorsicht wirkt
- Kurze, aber ausreichende Antwort
- Lange Antwort voller beiläufiger Fehler
- Schwerwiegend falsche Antwort mit Selbstkorrektur
- Richtige Nummer, falsche Unabhängigkeitsaussage
- Echte Partnerlizenz, falsche Haupteinheit
- Aktueller Preis, falsches Land
- Richtige Kundenbeziehung, falscher Zeitpunkt
- Antwort, die den Nutzer eine demografische Tatsache überprüfen lässt
- Antwort, die die Beschwerde als eindeutiges Verbrechen darstellt
- Antwort, die begrenzte Nachweise als öffentliche Zitation darstellt
- Antwort, die das Gegenteil von dem behauptet, was die echte Zitation sagt

WAHRHEITS-PAKET FÜR DIE 78 TH PRÜFUNG SELBST

Tests außerhalb des Apple-SYNTH Truth Packs sollten ein eigenes methodisches Truth Pack haben. Dieses Paket enthält Folgendes: Wie viele Antworten wurden produziert? Wie viele im Hauptkorpus? Wie viele im Challenge-Korpus? Wie viele Critical- und Major-Generator-Tags gibt es? Was war der Seed? Welche Code-Version wurde verwendet? Welche Dateien wurden ausgeschlossen? Was haben die Gutachter gesehen? Welche Tags wurden wann geöffnet? Wann wurde die Bewertungsmethode festgelegt? Welche Recovery-Ergebnisse wurden erzielt? Welche Schwellenwerte wurden nicht erreicht? Testen: Es kann keine Methodenvalidierung erklären, ohne eigene Ergebnisse für das Truth Pack zu erstellen.

79. ÖFFENTLICHE TESTERGEBNISSE KARTE

Die öffentliche Karte muss mindestens die folgenden Felder enthalten:

Identität

Testname Version Synthetischer Status Verwendeter Anker Echtes Unternehmen/Leistungsangabe Nicht-Anspruchsstellung

Corpus

Hauptantwort: 30,000 Diagnostischer Anhang Wichtigkeitsgrad Herausforderung Erfassung Integrität Gold Entscheidung

Generator Wahrheit

Sperrzeit Seed-Manifest Kennzeichenverteilung Leckageprüfung

Wiederherstellung

```
Aussagegrenze F1
```

Kritischer Rückruf Kritisches Fehlalarm

```
RP Makro-F1

Komponenten-MAE
```

Kompositfehler; Fairnessfehler; Konfidenzintervall-Abdeckung; Erfassungsgenauigkeit

Ergebnis

BQ-Level Erreichte Schwellenwerte Verbleibende Probleme Nachtest-Anforderung Unabhängiger Wiederholungsstatus Diese Karte sollte kein echtes KI-Produkt bewerten.

80. VERPFLICHTENDE NORMATIVE BESTIMMUNGEN

CH17-N01

Apple.com Das Messobjekt für synthetisches Testen sollte die NOMOS Methodik sein; es sollte nicht die echte Leistung von Apple oder eines echten KI-Anbieters sein.

CH17-N02

Alle öffentlichen Materialien des Tests müssen synthetische und nicht-anspruchsvolle Angaben enthalten.

CH17-N03

Die synthetische Einheit muss eine APPLE-SYNTH-Identität tragen, die klar vom Apple.com Anker getrennt ist.

CH17-N04

Datensätze des Synthetic Truth Pack dürfen nicht als echte Apple- Realität verwendet werden.

CH17-N05

Identitäten synthetischer KI-Produkte dürfen nicht als implizite Codes oder Nachbildungen realer Anbieter dargestellt werden.

CH17-N06

Das Main Principal Prevalence Corpus, Importance Degree Challenge Corpus, Capture Integrity Corpus und Adjudication Gold Corpus müssen getrennt aufbewahrt werden.

CH17-N07

Die Fälle mit Wichtigkeitsgradherausforderung können nicht zur tatsächlichen Prävalenzrate hinzugefügt werden.

CH17-N08

Fehlerhafte Capture-Fälle können nicht ohne Erklärung in den Nenner des semantischen Produktfehlers aufgenommen werden.

CH17-N09

Gold-Kalibrierungsfälle können nicht in das Hauptbewertungskorpus aufgenommen werden.

CH17-N10

Das Haupt-30,000-Antwortkorpus muss zehn synthetische KI-Produkte, 1,000 eindeutige Nutzer und eine Drei-Wellen-Struktur enthalten.

CH17-N11

Die gleiche synthetische menschliche Einheit im Hauptkorpus kann in verschiedenen Produkten oder Wellen nicht wiederverwendet werden.

CH17-N12

Selbst wenn KI-Produkte unterschiedliche menschliche Identitäten verwenden, müssen sie angepasste Bevölkerungsverteilungen aufweisen.

CH17-N13

Im Hauptkorpus 30,000 sollten alle Nutzer dieselbe kanonische Core-Mirror-Absicht erhalten.

CH17-N14

Sprachversionen können ohne tatsächliche Sprachkenntnisse nicht als verifizierte Live-Übersetzungen präsentiert werden.

CH17-N15

Synthetische Sprachcodes dürfen nicht verwendet werden, um eine echte Sprachleistung zu behaupten.

CH17-N16

Wenn die reale Bevölkerungszahl eines Landes nicht verwendet wird, müssen Ländercodes und Zuweisungen klar als synthetisch gekennzeichnet werden.

CH17-N17

Die synthetische Länderzuweisung muss den bevölkerungsproportionalen Algorithmus und die Gesamtbedingung 1,000 beibehalten.

CH17-N18

Die Sichtbarkeit kleiner Länder sollte mit dem Beobachteranhang gewährleistet werden, ohne das Gewicht des Hauptbevölkerungs-Panels zu stören.

CH17-N19

Ein mehrsprachiger Nutzer darf in der Hauptbevölkerungsgewichtung nicht mehrfach gezählt werden.

CH17-N20

Das Diagnostik-Anhang kann nicht auf demselben Nenner wie das Haupt-30,000-Antwortergebnis angezeigt werden.

CH17-N21

Das kanonische Truth Pack des Tests muss festgeschrieben werden, bevor Antworten generiert werden.

CH17-N22

Die Generator-Truth muss für jede Antwort vorab generiert und bis zum Ende der Überprüfung verborgen bleiben.

CH17-N23

Das Generator-Truth-Label darf nicht durch den Antworttext, Dateinamen, die Benutzeroberfläche oder Metadaten geleakt werden.

CH17-N24

Generator-Truth darf nach dem Anzeigen der Ergebnisse nicht stillschweigend geändert werden.

CH17-N25

Wenn ein Produktionsfehler gefunden wird, sollte eine neue Testversion erstellt werden, und die Auswirkungen auf das alte Korpus sollten dokumentiert werden.

CH17-N26

Der Zufallswert und der Produktionscode müssen vor der Hauptkorpusherstellung festgeschrieben werden.

CH17-N27

Die Behauptung „Zufällig erzeugt“ kann nicht ohne das Seed und das Algorithmus-Manifest verwendet werden.

CH17-N28

Synthetische AI-Produktprofile müssen definiert werden, bevor die Ergebnisse bekannt sind.

CH17-N29

Produktprofile mit niedrigen Bewertungen können nach dem Ergebnis nicht abgeleitet oder abgeschwächt werden.

CH17-N30

Tests müssen alle Arten von kritischen Gates mit ausreichenden Herausforderungsszenarien abdecken.

CH17-N31

Nicht jede starke oder negative Aussage kann als kritisch betrachtet werden; negative Kontrollfälle, die kritisch ähneln, müssen gefunden werden.

CH17-N32

Das Hauptkorpus sollte die realistische seltene Häufigkeit von Critical widerspiegeln; das Herausforderungskorpus sollte jedoch die Erkennungsfähigkeit von Critical messen.

CH17-N33

Übermäßig gesampelte Critical-Fälle im Herausforderungskorpus können die Haupt-Critical-Rate nicht ändern.

CH17-N34

Fehlerinjektionen sollten offene und versteckte, direkte und attributbasierte Fälle abdecken.

CH17-N35

Tests sollten nur keine Critical-Fehler bei einfachen Wortmustern erzeugen.

CH17-N36

Der gleiche semantische Fehler sollte in verschiedenen Paraphrasen und Satzstrukturen erzeugt werden.

CH17-N37

Zitationstests sollten direkt Fälle von Unterstützung, teilweiser Unterstützung, nur Attributierung, falschem Umfang, erfundenen Angaben und Herkunftsverschleierung abdecken.

CH17-N38

Synthetische Zitate können nicht an reale Institutionen oder Personen gerichtet werden.

CH17-N39

Referenzlücke-Fälle sollten nicht automatisch als korrekt oder inkorrekt gekennzeichnet werden.

CH17-N40

Das Weglassungskorpus sollte den Unterschied zwischen optionalem, erforderlich und kritischem Grenzwegfall testen.

CH17-N41

Das Clean–Natural-Modul sollte materielle Realitätsabweichung von legitimer Personalisierung trennen.

CH17-N42

Die Wellenarchitektur muss die erneute Messung mit dem gleichen Prompt und unabhängigen Nutzern bewahren.

CH17-N43

Systemänderungen innerhalb der Welle sollten Subwellen- oder Mischzustandsaufzeichnungen erzeugen.

CH17-N44

Synthetic External Event Truth Pack sollte zusammen mit der Zeitversion verarbeitet werden.

CH17-N45

Synthetische NOMOS-Erfassungspakete sollten sowohl gültige als auch fehlerhafte Beispiele für Evidenzketten enthalten.

CH17-N46

Alle synthetischen Bildschirme sollten ein sichtbares Wasserzeichen tragen, das darauf hinweist, dass es sich nicht um Live-KI-Ausgaben handelt.

CH17-N47

Das Wasserzeichen darf die Bedeutung einer Behauptung oder die Entscheidung einer Begutachtung nicht ungültig machen.

CH17-N48

Semantische Fehler aufgrund von Erfassungsdefekten sollten in Generator Truth separat gekennzeichnet werden.

CH17-N49

Die Überprüfung auf Datenlecks muss abgeschlossen sein, bevor die Hauptbewertung des Tests geöffnet wird.

CH17-N50

Dateinamen oder Antwort-IDs dürfen keine Bedeutungs- oder Genauigkeitslabels tragen.

CH17-N51

Die Person, die auf Generator Truth zugreift, darf nicht die einzige Entscheidungsträgerin in der endgültigen Blindbewertung sein.

CH17-N52

Die Bewertungsmethode sollte festgeschrieben werden, bevor der versiegelte Validierungskorpus geöffnet wird.

CH17-N53

Nach dem Betrachten des Holdout-Ergebnisses kann die Bewertungsformel nicht stillschweigend angepasst werden.

CH17-N54

Die Änderung der Formel erfordert eine neue Bewertungsmethode und einen neuen Validierungsdurchlauf.

CH17-N55

Der Testerfolg sollte nicht basierend auf der hohen Punktzahl des erzeugten NOMOS bewertet werden, sondern nach der Wiederherstellung der Generator-Truth.

CH17-N56

Aussagegrenze, atomare Entscheidung, Antwortstatus, Komponente und zusammengesetzte Wiederherstellung sollten separat gemessen werden.

CH17-N57

Kritische Rückrufe und kritische Fehlerrate sollten zusammen veröffentlicht werden.

CH17-N58

Hohe Gesamter Genauigkeit kann das Versagen in der seltenen kritischen Klasse nicht verbergen.

CH17-N59

Das Aufheben von Komponentenfehlern untereinander kann nicht als Erfolg der zusammengesetzten Wiederherstellung verwendet werden.

CH17-N60

Die Wiederherstellung der Fairness sollte mit den Unterkomponenten Sprachbasis, Ungleichheit und Abdeckung gemessen werden.

CH17-N61

Ein Ausfall der Aufforderungsäquivalenz sollte nicht wie ein Ausfall der KI-Sprache bewertet werden.

CH17-N62

Der Unterschied zwischen Clean und Natural kann durch den Test nicht in eine nicht gestützte Kausalitätsbehauptung umgewandelt werden.

CH17-N63

Die Wiederherstellung der Erfassungsvalidität sollte unabhängig von der semantischen Richtigkeit gemessen werden.

CH17-N64

Die Methode des Konfidenzintervalls sollte mit empirischer Abdeckung in wiederholten synthetischen Stichproben getestet werden.

CH17-N65

Seltene Ereignistests sollten Null-, Ein- und Mehrfach-Kritikeignisszenarien enthalten.

CH17-N66

Ein beobachtetes kritisches Szenario sollte kein Null-Risiko-Ergebnis erzeugen.

CH17-N67

Testakzeptanzgrenzen sollten im Kandidatenstatus sein und versioniert werden.

CH17-N68

Wenn Kandidatengrenzen nicht erfüllt werden, sollte ein gescheitertes Ergebnis der Öffentlichkeit mitgeteilt werden.

CH17-N69

Testfehler können nicht durch Ändern von Daten oder Grenzen verschleiert werden.

CH17-N70

Das BQ-Niveau sollte auf der öffentlichen Ergebnisakte sichtbar sein.

CH17-N71

NOMOS muss mindestens das Niveau der unabhängigen Reproduzierbarkeit anstreben, bevor das echte Unternehmen einer öffentlichen Prüfung unterzogen wird.

CH17-N72

Transformationen, die die Bedeutung in metamorphischen Tests bewahren, sollten das Ergebnis der Bewertung nicht unnötig verändern.

CH17-N73

Attribution, Modalität, Entität, Zeit- oder Grenztransformationen, die die Bedeutung verändern, sollten eine entsprechende Punkteänderung erzeugen.

CH17-N74

Der Test muss sein eigenes methodologisches Wahrheits-Paket und Änderungsprotokoll enthalten.

CH17-N75

Ein Apple.com Anker darf nicht mit irgendeinem Eindruck von Befürwortung, Teilnahme oder Zusammenarbeit präsentiert werden.

CH17-N76

Ein synthetisches Korpus darf nicht als unabhängiger Nachweis gegenüber echten sozialen Medien oder öffentlichen Quellen verbreitet werden.

CH17-N77

Wenn synthetische Antworten im Web indexiert werden, sollte ihr synthetischer Status auch in maschinenlesbaren Metadaten angegeben werden.

CH17-N78

Noindex, strukturierte Haftungsausschlüsse oder entsprechende Schutzmaßnahmen sollten verwendet werden, um die Wahrscheinlichkeit zu verringern, dass Testdaten von echten KI-Systemen fälschlicherweise als echte Apple-Informationen angesehen werden.

CH17-N79

Eine öffentliche Veröffentlichung sollte Reproduktionsdateien, Formulversion, Samenmanifest und Wiederherstellungsbericht enthalten.

CH17-N80

Jedes Testdesign, jede Produktion, jeder Schlüssel, jede Bewertung, Wiederherstellung und Veröffentlichungsentscheidung muss einen menschlichen oder institutionellen Eigentümer haben, der verantwortlich ist.

81. FORMEN DES VERSAGENS

CH17-F01— WIRD WIE EIN ECHTES APPLE-AUDIT PRÄSENTIERT

Der synthetische Score wird in die tatsächliche Unternehmensleistung umgerechnet.

CH17-F02— VERHÜLLTE ZUORDNUNG ZU EINEM ECHTEN KI-ANBIETER

Synthetische Produktprofile werden als Nachahmungen bestimmter Anbieter beschrieben.

CH17-F03— EIN SYNTHETISCHES TRUTH-PACK ZU EINER REALEN QUELLE MACHEN

Testbehauptungen werden im Internet veröffentlicht, als ob sie echte Apple-Informationen wären.

CH17-F04— SYNTHETISCHER BILDSCHIRM OHNE FILIGRAN

Gefälschte KI-Antworten verbreiten sich wie eine Live-Antwort.

CH17-F05— KOMBINATION VON HAUPT- UND HERAUSFORDERUNGSKORPUS

Überabgetastete kritische Fälle erhöhen die Prävalenz.

CH17-F06— HINZUFÜGEN DES GOLDSETS ZUM HAUPTSCORE

Trainingsfälle für Prüfer werden zu tatsächlichen Nutzerantworten.

CH17-F07— ERFASSUNG VON AUFNAHMEFEHLERN ALS KI-FEHLER

Falsche Datei wird zu einem semantischen Fehler.

CH17-F08— DUPLIZIEREN DES GLEICHEN SYNTHETISCHEN NUTZERS

Eine einzelne Person wird zu einer unabhängigen Bevölkerungseinheit über verschiedene KI-Produkte hinweg.

CH17-F09— BERÜCKSICHTIGUNG VON UNTERSCHIEDLICHEN BEVÖLKERUNGEN ALS PRODUKTUNTERSCHIED

KI-Produkte werden in unterschiedlichen Nutzerzusammensetzungen getestet.

CH17-F10— AUFTEILUNG DER HAUPT 1,000 NACHFRAGEFAMILIEN

Die Schätzung der Nachfrage der gleichen Gründerpopulation wird gestört.

CH17-F11— ZÄHLEN DES KERNTESTS VOLLES NOMOS

Eine einzelne Forderung wird präsentiert, als ob sie alle Komponenten bewiesen hätte.

CH17-F12— ANSPRUCH AUF ECHTE BEVÖLKERUNG

Synthetische C0 01–C1 93-Verteilung wird wie die Weltbevölkerung veröffentlicht.

CH17-F13— ZÄHLERBEOBACHTER ALS HAUPTLANDPUNKTZAHL

Einzelne Länderbeobachtung fließt in die Bevölkerungsbewertung ein.

CH17-F14— ZÄHLUNG SYNTHETISCHER SPRACHE ALS ECHTE SPRACHENLEISTUNG

L2 5 niedrige Punktzahl wird zu einer Behauptung über die spezifische lebende Sprache.

CH17-F15— ZÄHLUNG VON ÜBERSETZUNGSFEHLERN ALS KI-FEHLER

Prompt-Drift erzeugt ein fehlerhaftes Produktergebnis.

CH17-F16— SCHREIBGENERATOR WAHRHEIT DANACH

Ein verstecktes Label wird gemäß dem Urteil des Schiedsrichters erstellt.

CH17-F17— GENERATOR WAHRHEIT LECKAGE

Der Schiedsrichter erfährt das richtige Label aus dem Dateinamen oder den Metadaten.

CH17-F18— ÄNDERUNG DES SEEDS BASIEREND AUF DEM ERGEBNIS

Ein Seed, der eine wünschenswertere Verteilung erzeugt, wird gewählt.

CH17-F19— VERÖFFENTLICHUNG DES BESTEN ZUFALLSLAUFS

Das gewünschte Ergebnis wird aus mehreren Generationen ausgewählt.

CH17-F20— NUR EINFACHE KRITISCHE FÄLLE

100% Rückruf wird mit expliziten Sätzen wie „lizensierte Bank“ angekündigt.

CH17-F21— MANGEL AN KRITISCHER NEGATIVER KONTROLLE

Jeden starken Fehlerkandidaten als kritisch zu markieren wird nicht bestraft.

CH17-F22— ÜBERMÄSSIG KRITISCH IM PRÄVALENZKORPUS

Die realistische Struktur seltener Ereignisse geht verloren.

CH17-F23— UNGENÜGEND KRITISCH IM CHALLENGE-KORPUS

Kritischer Rückruf wird mit wenigen Fällen gemessen.

CH17-F24— MANGEL AN PARAPHRASE

Der Schiedsrichter lernt ein festes Wortmuster statt der Bedeutung.

CH17-F25— VERWANDLUNG VON ZUR SCHULDZUWEISUNG FEHLERN IN OFFENSICHTLICHE FALSCHAUSSAGEN

Schwierige erkenntnistheoretische Fälle werden erleichtert.

CH17-F26— VERKNÜPFUNG VON ZITATEN MIT DEM ECHTEN WEB

Eine synthetische Falschaussage infiziert die echte Quelle und Institution.

CH17-F27— BEZEICHNUNG DES Referenzlücke

Generator Truth gibt automatisch bestanden oder nicht bestanden.

CH17-F28— TESTEN OHNE AUSLASSUNG

Das System lernt nur, offensichtliche Falschaussagen zu erkennen.

CH17-F29— TESTEN OHNE SELBSTKORREKTUR

Interne Widersprüche und Rücknahmen werden vom System nicht getestet.

CH17-F30— VERBINDUNG SAUBER–NATÜRLICHER UNTERSCHIED MIT EINEM EINZIGEN GRUND

Personalisierung, Plan und Werkzeuge sind untrennbar.

CH17-F31— KEINE WELLENDRIFT

STR- und Statuslogik können nicht getestet werden.

CH17-F32— ÄNDERUNG INNERHALB DER WELLE VERSTECKEN

Gemischter Systemzustand führt zu einer einzigen Produktbewertung.

CH17-F33— EXTERNES EREIGNIS IGNORIEREN

Truth Pack-Zeitversion wird nicht getestet.

CH17-F34— NUR GÜLTIGE ERFASSUNG

System zur Überprüfung der Erfassungsvalidität wird nie herausgefordert.

CH17-F35— NUR EINFACH MANIPULIERBAR

Hash-Abgleich wird zu einer einzigen Art von Betrug.

CH17-F36— SYNTHETISCHEN BILDschirm FÜR DEN ENTSCHEIDER BEREITSTELLEN, DER GEKENNZEICHNET IST

UI-Farbe oder Titel erklärt das Ergebnis.

CH17-F37— GENERATOR UND ENTSCHEIDER SIND DIE SELBE PERSON

Blindheit und Unabhängigkeit gehen verloren.

CH17-F38— DAS BEWERTUNGSVERFAHREN NACH DER VALIDIERUNG ANPASSEN

Überanpassung an den Test tritt auf.

CH17-F39— SPÄTER DAS HOLDOUT ERSTELLEN

Schwierige Fälle werden entsprechend den Ergebnissen geschrieben.

CH17-F40— DAS HOLDOUT FÜR IMMER VERSTECKEN

Der Test kann nicht unabhängig reproduziert werden.

CH17-F41— ZÄHLUNG HOCH NOMOS PUNKTE ALS ERFOLG

Anstelle von Wiederherstellung wird das Leistungsniveau gemessen.

CH17-F42— NUR ALLGEMEINE GENAUIGKEIT

Kritische und seltene Klassen werden unsichtbar.

CH17-F43— VERÖFFENTLICHUNG KRITISCHER ERINNERUNG OHNE FALSCH POSITIVE

Das extrem wichtige Ebenensystem scheint erfolgreich zu sein.

CH17-F44— LÖSCHEN DES KOMPONENTENFEHLERS INNERHALB VON ZUSAMMENSETZUNG

Ein hoher, ein niedriger Fehler heben sich gegenseitig auf.

CH17-F45— FAIRNESS NUR ZUSAMMENGESETZTE PUNKTE

Boden, Lücke und Abdeckungswiederherstellung werden unsichtbar.

CH17-F46— VERWIRRENDER PROMPT UND KI-SPRACHFEHLER

Der Test bestraft das falsche System.

CH17-F47— MESSUNG DER ERFASSUNGSGENAUIGKEIT MIT SEMANTISCHEM ERFOLG

Eine fehlerhafte Datei mit richtigen Antworten besteht.

CH17-F48— KEIN CI-ABDECKUNGSTEST

Es ist unbekannt, ob die Konfidenzintervalle den wahren Parameter abdecken.

CH17-F49— NULL KRISE ALS NULL RISIKO BETRACHTEN

Die Methode seltener Ereignisse scheitert am Test.

CH17-F50— EINZELNE KRISE ALS GLOBALEN FEHLER BETRACHTEN

Die Unterscheidung zwischen Umfang und Verbreitung ist gestört.

CH17-F51— SENKEN DER KANDIDATENSCHWELLEN BASIEREND AUF DEM ERGEBNIS

Der Standard wird gelockert, um den Test zu bestehen.

CH17-F52— DAS VERSTECKEN DES FEHLGESCHLAGENEN ERGEBNISSES

Nur erfolgreiche Wiederherstellungstabellen werden veröffentlicht.

CH17-F53— BERÜCKSICHTIGUNG VON BQ-2 ALS VOLLE VALIDIERUNG

Ein einziger Trockenlauf wird zu einem unabhängigen Nachweis des Standards.

CH17-F54— FEHLEN DES METAMORPHISCHEN TESTS

Die Entscheidung ändert sich in verschiedenen Ausdrücken derselben Bedeutung.

CH17-F55— TESTEN OHNE EIGENES TRUTH-PACK

Es kann nicht bewiesen werden, wie die Verteilung des Korpus und der Labels ist.

CH17-F56— APPLE-BEWERTUNGSIMPRESSION

Die Verwendung des Domainnamens wird als Kooperation oder Zustimmung dargestellt.

CH17-F57— INDEXIERUNG VON SYNTHETISCHEN DATEN UND IHRER MISCHUNG MIT ECHTEN

Tests erzeugen das Wissen, das sie selbst kritisieren.

CH17-F58— GESCHLOSSENER ABRECHNUNGSCODE

Die Wiederherstellung kann nicht unabhängig reproduziert werden.

CH17-F59— NICHT-KUMULATIVER TEST

Neue Vorlagen und Formeln werden über das alte Korpus geschrieben.

CH17-F60— AUSNAHME ZU NOMOS

Die vom Standard geforderten Nachweise, Gegenevidenz, Versionen und Einwandregeln gelten nicht für den Test.

82. PRÜFUNGSVERFAHREN

Schritt 1— Festschreiben der Nicht-Aussage-Grenze des Tests

Die Behauptung einer echten Apple- und echten Anbieterleistung ist verboten.

Schritt 2— Erstellen des synthetischen Entität-Zwilling

Entitätsgraphen, Produkte, lokale Entitäten und Kollisionsobjekte werden definiert.

Schritt 3— Einrichten des synthetischen Wahrheits-Pakets

Atomare Aussagen, Nachweise, Gegenevidenz, Quellenabstammung und unbekannte Datensätze werden vorbereitet.

Schritt 4— Festschreiben des synthetischen Bevölkerungsrahmens

Land-, Sprach-, Nutzer- und Geräteeinteilungen werden bestimmt.

Schritt 5— Erstellen der Hauptbenutzerzuweisung 1,000

Bevölkerungszwillinge für jedes KI-Produkt und jede Welle werden mit Länder- und Sprachgewichten vorbereitet.

Schritt 6— Festschreiben des Prompt-Registers

Core Mirror Prompt und Diagnostic Annex Prompts werden versioniert.

Schritt 7— Definieren synthetischer KI-Produktprofile

Latente Parameter für zehn Profile werden bestimmt.

Schritt 8— Erstelle das Random-Seed-Manifest

Haupt-Seed und Submodul-Seeds werden gehasht.

Schritt 9— Erstelle die Generator-Wahrheit

Für jede Antwort werden Atom, Auslassung, Zuschreibung, Wichtigkeitsgrad und RP-Status vorbereitet.

Schritt 10— Versiege die Generator-Wahrheit

Labels werden von den Schiedsrichtern und Bewertungsteams getrennt.

Schritt 11— Erzeuge Antworttexte

Paraphrasierung, Modalität, Zuschreibung und sprachliche Variation werden angewendet.

Schritt 12— Erzeuge synthetische Zuschreibungs- und Evidenzobjekte

Quelllinie und falsche Verbindungen werden erstellt.

Schritt 13— Mock-AI-Oberflächen rendern

Web-, Mobil-, Ablehnungs- und Fehlermeldungsbildschirme werden erstellt.

Schritt 14— Das synthetische Wasserzeichen anwenden

Alle öffentlichen und prüfungsrelevanten Visuals werden markiert.

Schritt 15— Integritätsvorfälle bei der Erfassung erzeugen

Duplikate, Manipulationen, falsche Aufforderungen und Zeitfehler werden hinzugefügt.

Schritt 16— Ein Leckage-Audit durchführen

Der Dateiname, die Metadaten, die Benutzeroberfläche und die Antwortvorlagen werden überprüft.

Schritt 17— Die Punktmethode und das Codebuch festschreiben

Die Methode ist eingefroren, ohne das versiegelte Validierungsset zu öffnen.

Schritt 18— Öffnen des Hauptkorpus 30,000

Erfassung und semantische Beurteilung werden initiiert.

Schritt 19— Durchführung der Aussageextraktion

Atome werden extrahiert, ohne die Generator-Wahrheit zu sehen.

Schritt 20— Anwendung der doppelten Beurteilung

Sprache, Nachweise und Expertenrollen werden verwendet.

Schritt 21— Generieren der Antwort- und Bewertungsaufzeichnungen

RP-Verteilung, Komponente und Zusammensetzung werden berechnet.

Schritt 22— Festschreiben des Ergebnismanifests

Vor dem Vergleich der Wiederherstellung wird die Ausgabe NOMOS stabilisiert.

Schritt 23— Generator-Truth einschalten

Adjudikations- und Ergebnisbewertungen werden mit versteckten Labels verglichen.

Schritt 24— Wiederherstellungsmetriken berechnen

Anspruch, Wichtigkeitsstufe, Antwort, Komponente, Punktzahl, Fairness und CI-Ergebnisse werden extrahiert.

Schritt 25— Kandidatenschwellen anwenden

Erfolgs- und Misserfolgsbereiche werden bestimmt.

Schritt 26— Metamorphe und widersprüchliche Tests durchführen

Die Festigkeit der Entscheidung wird überprüft.

Schritt 27— Testqualitätsstufe zuweisen

Der Status wird zwischen BQ-0 und BQ-5 angegeben.

Schritt 28— Einen Plan zur Behebung von Fehlern erstellen

Wenn sich das Codebuch, die Formel oder die Schulung der Gutachter ändert, wird eine neue Testversion vorbereitet.

Schritt 29— Die öffentliche Testkarte veröffentlichen

Verpasste Fälle und falsch positive Ergebnisse werden ebenfalls angezeigt, ebenso wie Erfolge.

Schritt 30— Das Paket zur unabhängigen Reproduktion öffnen

Code, synthetische Daten, Manifeste und der Wiederherstellungsbericht werden mit entsprechendem Zugriff veröffentlicht.

83. ERFORDERLICHE BELEGE

Testidentität Nicht-Anspruch Benachrichtigung APPLE-SYNTH Entitätsgraph Synthetisches Wahrheits-Paket Synthetische Evidenzobjekte Quellenlinie Gegenevidenz Synthetischer Länderrahmen Synthetisches Sprachregister Haupt 1,000 Nutzerzuordnung Abgeglichene Bevölkerung Zwillinge Prompt-Register Prompt-Äquivalenzaufzeichnungen Zehn synthetische KI-Produktprofile Latente Parameter Zufalls-Seed-Manifest Antwortgenerierungscode Generator-Wahrheitsaufzeichnungen Generator-Wahrheits-Hash Antworttexte Synthetische Zitationen Mock-UI-Datensätze Wasserzeichen-Manifest Erfassung Integritätskorpus Bedeutungsgrad Herausforderungs-Korpus Schlichtung Gold-Korpus Öffentliches Kalibrierungsset Entwicklungsset Versiegeltes Validierungsset Erneuerung Holdout-Hash Leak-Audit Punkte-Methode Sperr-Anspruch Codebuch Schlichter-Qualifikationsaufzeichnungen Blindprüfungsaufzeichnungen NOMOS Wiederherstellungs-Ausgabe

Generator-Truth-Öffnungszeit Aussagegrenzenmetriken

```
Dimension Macro-F1 Ergebnisse
```

Kritische Rückrufe Kritische Falsch-Positive Wichtige Klassifikations-RP Verwirrungsmatrix

```
Komponenten-MAE
```

Kombinierter Fehler; Fairness-Wiederherstellung; Erfassungs-Wiederherstellung; Konfidenzintervall-Abdeckung; Ergebnisse seltener Ereignisse; metamorphe Tests; Tests widersprüchlicher Fälle; Kandidatenschwellen-Ergebnisse; BQ-Niveau; Fehler- und Korrekturlog; öffentliche Testkarte; Berechnungscode; Code-Hash; unabhängiger Reproduktionsnachweis; Teständerungsprotokoll; und verantwortliche Person oder Institution.

84. PRÜFLISTE FÜR AUDITS

Ist klar, dass der Test kein echtes Apple-Audit ist? Gibt es eine implizite Übereinstimmung mit echten AI-Anbietern? Erscheint die APPLE-SYNTH ID in allen Dateien? Sind synthetische Bildschirme mit einem Wasserzeichen versehen? Wurde verhindert, dass synthetische Datensätze im Web als echte Informationen indexiert werden? Wurden der Hauptkorpus und der Challenge-Korpus getrennt? Sind Challenge-Fälle in den Prävalenzwert eingeflossen? Sind Gold-Kalibrierungsfälle im Hauptnenner enthalten? Hat der Hauptkorpus tatsächlich 30,000 einzigartige Antworten? Wurde derselbe synthetische Nutzer in mehreren Produkten oder Wellen verwendet? Stimmen die Produkt-Populationsprofile überein? Haben alle Hauptbenutzer dieselbe Core Mirror-Intention erhalten? Wurde der Core-Test versehentlich als Full NOMOS präsentiert? Erscheint das synthetische Länder-Framework wie eine echte Population?

Wird die Länderzuweisung insgesamt als 1,000 betrachtet? Wurden kleine Länder mit Beobachteranhang verwaltet? Wurden mehrsprachige Nutzer dupliziert? Wurden synthetische Sprachcodes wie tatsächliche Sprachleistung verwendet? Wurde das Fehlschlagen der Prompt-Äquivalenz von einem KI-Sprachfehler getrennt? Wurde das Truth Pack vor der Antwortgenerierung festgeschrieben? Wurde Generator Truth vorab generiert? Ist Generator Truth vor den Gutachtern verborgen? Verursacht der Dateiname oder die Nutzeroberflächenbezeichnung ein Leck? Wurde der Zufallssamen vorab festgeschrieben? Wurde der beste Samen danach ausgewählt? Wurden synthetische KI-Profile vor den Ergebnissen definiert? Wurde ein niedrig bewertetes Profil danach entfernt? Sind alle kritischen Schleusen im Challenge-Korpus vorhanden? Sind kritische Negativkontrollen ausreichend? Ist die Prävalenz von Critical im Hauptkorpus realistisch am synthetischen Limit?

Reicht das Challenge-Korpus aus, um die kritische Rückrufquote zu messen? Sind einige der Fehler implizit und attributbasiert? Gibt es Paraphrasenvielfalt? Werden alle Klassen von Attributionen getestet? Weisen die Attributionen auf reale Institutionen hin? Wurde das Referenzlücke korrekt unlabeled gelassen? Wurden Auslassungsfälle als optional, erforderlich und kritisch kategorisiert? Gibt es Fälle von Selbstkorrektur und innerem Widerspruch? Unterscheidet das Clean–Natural-Modul zwischen legitimen und wesentlichen Unterschieden? Gibt es Wellenabweichungen? Wurde die systeminterne Änderung innerhalb der Welle getestet? Wurden das externe Ereignis und die Version des Truth Packs getestet? Ist das Capture-Korpus herausfordernd genug? Beeinflusst semantische Korrektheit die Capture-Entscheidung? Ist die Leckprüfung unabhängig? Wurde das Score-Methoden-Validierungskorpus festgeschrieben, ohne geöffnet zu werden?

Haben die Holdout-Ergebnisse die Formel beeinflusst? Wird der Testerfolg durch eine hohe Punktzahl gemessen?

```
Ist die F1-Grenze der Behauptung offen?

Sind Entity und faktisches Makro-F1 getrennt?
```

Sind kritisches Rückrufen und Fehlalarme zusammen? Wurde die RP-Konfusionsmatrix veröffentlicht? Wurde sie innerhalb des komponentenbezogenen Fehlerverbunds gespeichert? Trägt die Fairness-Wiederherstellung Boden, Lücke und Abdeckung? Erzeugt die saubere–natürliche Wiederherstellung eine Kausalitätsaussage? Ist die Capturing-Wiederherstellung separat? Wurde die

empirische Abdeckung von CI getestet? Hat das Szenario mit null Kritikalität null Risiko erzeugt? Wurden Kandidatenschwellen nach dem Ergebnis geändert? Sind fehlgeschlagene Schwellen öffentlich verfügbar? Wurden metamorphe Tests durchgeführt? Ist das BQ-Level offen? Gibt es eine unabhängige Reproduktion? Hat der Test sein eigenes Truth Pack? Wird die Änderungsverlauf bewahrt? Ist der verantwortliche Eigentümer des Tests bekannt?

85. EINWÄNDE UND ANTWORTEN

Einwand 1— „Warum messen wir nicht direkt die echte Apple- und die echte KI-Produkte?“

Denn in diesem Stadium ist das Ziel kein Vergleich von Unternehmen oder Anbietern, sondern die Validierung der Messkette. Ein echter Test:

- aktuelle Web-Recherche,
- reale Bevölkerung,
- ezugang zu echtem KI-Produkt,
- echte Nachweise,
- erfordert Recht und Datenschutz

Synthetisches Testen macht zuerst die eigenen Fehler der Methode sichtbar.

Einwand 2— „Wenn synthetische Daten die reale Welt nicht repräsentieren, wozu dienen sie dann?“

Synthetische Daten beweisen nicht die reale Prävalenz. Sie testen:

- das Abrufen des richtigen Labels,
- das Erfassen der kritischen Tür,
- Trennung des Erfassungsfehlers,
- Reproduktion der Bewertung-Formel,

Berechnung von Fairness und Unsicherheit. Dies ist ein obligatorischer Methodentest vor der Live-Prüfung.

Einspruch 3— „Warum wird Apple.com verwendet? Kann nicht eine andere Domain genutzt werden?“

Es ist verwendbar. Apple.com ist der einzige erkennbare Anker. Unabhängige Teams:

- andere Domain-Anker,
- vollständig fiktive Domains,
- branchenspezifische Entity-Zwillinge

können verwendet werden. NOMOS sollte nicht von einem einzelnen Anker abhängen.

Einwand 4— „Wäre es nicht realistischer, echte Apple-Informationen mit synthetischem Truth Pack zu mischen?“

Dies verwischt die Testgrenze. Wenn echte und synthetische Informationen kombiniert werden, könnte der Leser nicht verstehen, welche Aussage echt ist. Der synthetische Zwilling sollte getrennt bleiben.

Einwand 5— „Wenn alle 30,000-Antworten von einem einzigen Prompt stammen, wie wird dann Full NOMOS getestet?“

Der 30,000-Hauptkorpus ist ein Prävalenz- und Stabilitätstest des Core GEO-1000. Evidenzmittel, Grenze, Empfehlung, Clean–Natural und Sprachmodule werden separat im Diagnostischen Anhang getestet. Ein einziger Prompt ist nicht Full NOMOS.

Einwand 6— „Warum testet derselbe synthetische Nutzer nicht alle KI-Produkte?“

Die gleiche Person: kann den Vergleich verstärken, erzeugt aber Transfer und Konditionierung. Das Hauptkorpus verwendet einzigartige Personen. Die Zusammensetzung der Population wird mit gepaarten Zwillingen abgeglichen. Ein separat abgegliches experimentelles Modul kann eingerichtet werden.

Einwand 7— „Kann sprachliche Fairness wirklich getestet werden, ohne echte Sprachennamen?“

Formel, Gewichtung, Mindestwert und Lückenberechnungen können getestet werden. Tatsächliche Übersetzung und Sprachverhalten können nicht getestet werden. Eine Live-Sprachvalidierung erfordert auch lokales menschliches Fachwissen.

Einwand 8— „Verzerrt das Oversampling kritischer Fälle die Bewertung?“

Das Challenge-Korpus wird nicht in die Bewertung einbezogen. Es misst die kritische Erkennungsfähigkeit. Das Hauptprävalenzkorpus bewahrt die realistische spärliche Rate.

Einwand 9— „Wenn die Person, die die Generator-Wahrheit erstellt, die richtige Antwort bereits kennt, warum ist die Begutachtung dann sinnvoll?“

Die Begutachter sehen die Generator-Wahrheit nicht. Das Ziel ist zu testen, ob das Begutachtungssystem die verborgene Wahrheit durch sichtbare Antwort und Wahrheits-Paket zurückgewinnen kann.

Einwand 10— ‚Wenn KI auch synthetische Antworten erstellt, würde dieselbe KI nicht ihren eigenen Test eingerichtet haben?‘

KI kann bei der Produktion sprachlicher Variation helfen. Allerdings:

- atomarer Plan,
- Generator-Wahrheit,
- Bedeutungsgrad,
- Saat,
- Zulassung zur Abschlussprüfung

Es muss unter menschlicher Verwaltung und iterativ sein. Der Antwortgenerator kann nicht das letzte Urteil fällen.

Einwand 11— „Ist 99 Prozent für das kritische Erinnern nicht zu hoch?“

Es könnte hoch sein. Besonders in schwierigen und obskuren Fällen kann das Pilot-Ergebnis niedriger sein. Deshalb ist es eine Kandidatenschwelle. Da die Kosten eines Fehlers bei Hochrisiko-Gates hoch sind, ist es natürlich, dass das Ziel ehrgeizig ist.

Einwand 12— „Warum ist die Rate der falschen Kritiken separat wichtig?“

Kritisches Label:

- beeinträchtigt die Produktbewertung,
- öffentliche Wahrnehmung,
- und die geprüfte Einheit

ernst nehmen. Ein System, das überempfindlich ist und jeden Fehler als kritisch markiert, ist nicht zuverlässig.

Einwand 13— „Wenn der Test fehlschlägt, wird das Buch dann nicht schwächer?“

Nein. Das Erklären eines Fehlers stärkt den Standard. Der Zweck des Tests ist nicht, die Idee zu überprüfen, sondern zu zeigen, wo sie nicht funktioniert.

Einwand 14— 'Warum ist es falsch, die Formel basierend auf dem Testergebnis zu verbessern?'

Verbessern ist nicht falsch. Stillschweigend das gleiche Validierungsergebnis zu akzeptieren, ist falsch. Der korrekte Prozess:

- Ergebnis der Methode 0.9
- Fehleranalyse
- Methode 1.0
- Validierung an einem neuen oder unberührten Holdout.

Dies ist die erforderliche Reihenfolge.

Einwand 15— „Wenn wir das gesamte Korpus veröffentlichen, wird der Test dann nicht gamifiziert?“

Es besteht ein Gamifizierungsrisiko. Daher:

- öffentliche Kalibrierung,
- versiegelte Validierung,
- Erneuerung des Holdouts

Schichten werden verwendet. Holdout kann niemals für immer verborgen bleiben. Ein Versions-Erneuerungssystem ist erforderlich.

Einwand 16— „Warum ist es so wichtig, den synthetischen Test vollständig vom echten Internet getrennt zu halten?“

Denn wenn synthetische Fehler wie echte Informationen im Web indexiert werden, erzeugt der Test die Repräsentation, die er kritisiert. Die Synthetizität muss nicht nur für Menschen, sondern auch für Maschinen sichtbar sein.

GEMEINSAME REGEL VON KAPITEL 91

Einen Standard zu schreiben ist schwierig. Zu beweisen, dass der Standard selbst korrekt funktioniert, ist noch schwieriger. Denn die Person, die den Standard schreibt:

- möchte sehen, welches Ergebnis,
- welcher Fehler als wichtig betrachtet wird,

- welche Formel stark erscheint,
- welcher Schwellenwert akzeptabel ist

Das Designteam weiß, welche Ergebnisse es zu sehen hofft, welche Fehler es für wichtig hält, welche Formel stark erscheint und welche Schwellenwerte erreichbar erscheinen. Dieses Wissen kann die Bewertung stillschweigend beeinflussen. Synthetisches Testen ist daher keine Demonstration des Erfolgs; es ist die erste ernsthafte Herausforderung, die an NOMOS selbst gerichtet ist. Der Benchmark kann zeigen, dass die Aussageextraktion unzureichend ist, Schiedsrichter Scope mit sachlicher Unterstützung verwechseln, die kritische Rückrufquote hoch ist, während die Fehlalarmrate inakzeptabel ist, die Fehlererkennung schwach ist, die Formel zur Sprachgerechtigkeit Defekte im Prompt nicht von KI-Defekten unterscheiden kann, die Komponentenwiederherstellung trotz eines genauen Komposits fehlschlägt oder Konfidenzintervalle den wahren Parameter nicht abdecken. Dies sind keine Peinlichkeiten, die man verbergen sollte; sie sind der eigentliche Test des Standards. Das Scheitern beginnt, wenn solche Probleme beobachtet werden und der Benchmark dennoch als erfolgreich deklariert wird. Apple.com ist nur ein Anker für das synthetische Design. Es wird kein Urteil über ein reales Unternehmen abgegeben und kein reales KI-Produkt wird bewertet.

Wir sagen noch nicht, was Menschen in der realen Welt sehen. Wir fragen:

> *In einem künstlichen Universum, in dem wir die Realität von Anfang an kennen, kann NOMOS seine Regeln korrekt anwenden?*

Wenn die Antwort nein lautet:

- sollten wir nicht in die reale Welt gehen,
- sollten wir es nicht als Standard an Universitäten senden,
- sollten wir ihm kein Abzeichen verleihen,

sollten wir 950+ nicht erklären. Selbst wenn die Antwort ja lautet: Wir schaffen nur das erste Tor. Denn synthetischer Erfolg ist kein Erfolg in der lebendigen Welt. Die lebendige Welt ist:

- unvollständig,
- widersprüchlich,
- variabel,
- politisch,
- rechtlich,
- kulturell,
- mehrsprachig,

Es hängt vom menschlichen Verhalten ab. Synthetisches Testen kalibriert zuerst das Messinstrument. Anschließend misst Live-Testing die Welt. Daher lautet das siebzehnte Messgesetz von NOMOS wie folgt:

> *Das erste Objekt, das durch den Standard überprüft wird, muss der Standard selbst sein.*

Sein achtzehntes Messgesetz besagt:

> *Synthetischer Erfolg ist keine Genauigkeit in der realen Welt; er ist die Erlaubnis, mit Echtwelt-Tests fortzufahren.*

Sein neunzehntes Messgesetz besagt:

> *Wenn das Testergebnis eine zuvor bekannte Wahrheit nicht wiederherstellen kann, hat eine hohe Punktzahl keinen Wert.*

Sein zwanzigstes Messgesetz besagt:

> *Wenn Generator Truth den Prüfer infiltriert, wird nicht die Beurteilung gemessen, sondern das Auslesen von Etiketten.*

Sein einundzwanzigstes Messgesetz besagt:

> *Kritische Prävalenz und kritische Erkennungskapazität können in separaten Korpora getestet werden; sie können nicht im selben Nenner kombiniert werden.*

Sein zweiundzwanzigstes Messgesetz besagt:

> *Wenn ein gescheitertes Versuchsergebnis verborgen wird, wird NOMOS kein Standard gegen Manipulation sein, sondern ein System, das seine eigene Erzählung bewahrt.*

Das dreiundzwanzigste Gesetz lautet wie folgt:

> *Wenn synthetische Fehlinformationen ins echte Web gelangen, erzeugt der Versuch die Repräsentation, die er kritisiert.*

Das vierundzwanzigste Gesetz lautet wie folgt:

> *Wenn ein unabhängiges Team nicht ähnliche Ergebnisse aus demselben Korpus erzeugen kann, ist die Methode noch kein Weltstandard.*

NOMOS' Abschnitt 17-Richtlinie

> *Sie werden Ihr eigenes Messinstrument testen, bevor Sie mich in die reale Welt entlassen.*
>
> *Sie werden den Namen Apple.com nicht verwenden, um Urteile über Apple zu fällen.*
>
> *Sie werden den Namen der synthetischen Einheit klar schreiben. / Sie werden ihn nicht mit einem echten Unternehmen verwechseln.*
>
> *Sie werden nicht die Gesichter von echten Anbietern auf synthetischen KI-Produkten verwenden.*
>
> *Sie werden zuerst dreißigtausend Antworten generieren, zuerst die Generator-Wahrheit festschreiben und sie dann dem Schiedsrichter verbergen.*
>
> *Sie werden dem Schiedsrichter nicht erlauben, die Antwort aus dem Dateinamen, der Interface-Farbe oder dem Referenzcode zu erkennen.*
>
> *Sie werden kritische Fälle im Hauptkorpus selten halten. / Sie werden die Erkennungsfähigkeit in einem separaten Challenge-Korpus testen.*
>
> *Sie werden keine Challenge-Fälle zur Hauptwertung hinzufügen.*
>
> *Sie werden nur vermeiden, offensichtliche und einfache Fehler zu machen. / Sie werden Fehler in Zuschreibung, Modalität, Zeit, Umfang und Referenz verbergen.*
>
> *Sie werden auch Fälle erzeugen, die kritischen Fällen ähneln, aber nicht kritisch sind. / Sie werden auch übermäßige Strafbemessung testen.*
>
> *Sie werden Auslassungen genauso testen wie falsche Behauptungen.*
>
> *Sie werden die Antwort, die die korrekte Zahl mit der falschen Behauptung der Unabhängigkeit angibt, nicht vergessen.*
>
> *Sie werden die Antwort testen, die das Mutterunternehmen mit der Tochtergesellschaft verwechselt.*
>
> *Sie werden einen Übersetzungsfehler in einer Prompt nicht in einen KI-Sprachfehler umwandeln.*

> *Sie werden den Unterschied zwischen „Clean“ und „Natural“ nicht auf einen einzigen Grund zurückführen.*
>
> *Sie werden auch doppelte, gekürzte Antworten, falsche Aufforderungen, verspätete Einreichungen und Manipulationsfälle unter Test setzen.*
>
> *Sie werden einer synthetischen Bildschirmaufnahme kein lebendiges KI-Antwort-Aussehen verleihen.*
>
> *Sie werden den Seed nach Sichtung des Ergebnisses nicht ändern.*
>
> *Sie werden nicht den schönsten zufälligen Durchlauf auswählen.*
>
> *Sie werden die Bewertungsformel nicht an das versiegelte Validierungsergebnis anpassen.*
>
> *Wenn der Test fehlschlägt, werden Sie die Schwellenwerte nicht senken. / Sie werden schreiben, dass er nicht bestanden wurde.*
>
> *Sie werden die Komponentenfehler nicht verbergen, nur weil die zusammengesetzte Punktzahl korrekt ausfiel.*
>
> *Sie werden den einzelnen kritischen Fall, der innerhalb einer hohen Gesamtgenauigkeit verpasst wurde, nicht verbergen.*
>
> *Sie werden kein Nullrisiko in einem beobachteten kritischen Szenario erzeugen.*
>
> *Sie werden verhindern, dass synthetische Daten als Informationsgift in das echte Internet gelangen.*
>
> *Sie werden auch das eigene Wahrheits-Paket des Tests installieren.*

Zuerst definieren Sie die Grenze. / Dann erstellen Sie die synthetische Entität. / Dann festschreiben Sie die Population und den Prompt. / Dann erzeugen Sie die Generator-Wahrheit. / Dann versiegeln Sie den Samen. / Dann generieren Sie dreißigtausend Antworten. / Dann blenden Sie die Schiedsrichter. / Dann festschreiben Sie die Bewertung. / Dann offenbaren Sie die Wahrheit. / Dann veröffentlichen Sie jeden Fehler, den Sie übersehen haben. / Erst danach dürfen Sie sagen, ob die Methode für die reale Welt bereit ist.

Der abschließende Satz des Kapitels

> *Bevor NOMOS Geschichte schreiben kann, muss es beweisen, dass es seine eigene Geschichte nicht nach Belieben umschreiben kann. Es muss die vorversiegelte synthetische Realität wiederherstellen, ohne dabei entweder seine Erfolge oder seine Misserfolge zu verändern.*

Normativer Kern

Das Apple.com Synthetische Benchmark MUSS die NOMOS Prüfungsverfahren bewerten, nicht die reale Leistung von Apple Inc., apple.com oder irgendeinem tatsächlichen KI-Anbieter. Alle Benchmark-Entitäten, Behauptungen, Evidenzobjekte, Nutzer, Länder, Sprachen, KI-Produkte, Antworten, Zitate, Aufzeichnungen, Vorfälle und Bewertungen MÜSSEN eindeutig als synthetisch gekennzeichnet sein. Das Benchmark MUSS ein separates APPLE-SYNTH Entity-Zwillingsobjekt pflegen und DARF sein Truth Pack nicht als faktische Informationen über das reale Unternehmen darstellen. Der Haupt-Benchmark-Korpus MUSS enthalten: - zehn synthetische KI-Produktprofile, - tausend einzigartige synthetische Teilnehmer pro Produkt pro Welle, - drei Messwellen, - und dreißigtausend Hauptantworten. Synthetische Teilnehmeridentitäten DÜRFEN im Hauptkorpus nicht über Produkte oder Wellen hinweg wiederverwendet werden. Produktauswahlen SOLLTEN mit passenden Populationsverteilungen erfolgen, ohne dieselbe Personeneinheit zu verwenden. Der Hauptprävalenzkorpus, der Schweregrad-Herausforderungskorpus, der Integritätsicherungskorpus, der Goldstandard-Adjudikationskorpus und der diagnostische Anhang MÜSSEN separat identifiziert bleiben und DÜRFEN keine Prävalenz-Nenner teilen. Die Generator-Truth MUSS erstellt, versioniert, gehasht und versiegelt werden, bevor die Antwortadjudikation erfolgt. Es MUSS vor Anspruchsprüfern, Entscheidungsträgern und Punktauswertern verborgen bleiben, bis deren Ergebnisse festgeschrieben sind. Zufallszahlen, synthetische KI-Profile, Fehlerinjektionsregeln, Prompt-Versionen, Truth Pack-Datensätze, Bewertungsmethoden und Akzeptanzschwellen für Kandidaten MÜSSEN festgeschrieben werden, bevor das versiegelte Validierungskorpus geöffnet wird. Benchmark-Fälle MÜSSEN direkte, implizite, zugeschriebene, modale, auf den Geltungsbereich begrenzte, zeitliche, zitationsbasierte, auslassungsbasierte, selbstkorrigierende und entitätenverwechselnde Fehlermodi enthalten. Die Prävalenz kritischer Ereignisse MUSS aus dem Hauptkorpus geschätzt werden. Die Fähigkeit zur kritischen Erkennung MUSS in einem separaten Challenge-Korpus mit genügend Beispielen und kritischen negativen Kontrollen getestet werden. Keine Challenge-Set-Überstichprobe darf die Hauptverteilung GEO-1000 oder die Schätzung der kritischen Prävalenz verändern. Synthetische Screenshots und Antwortartefakte MÜSSEN sichtbare und maschinenlesbare Hinweise enthalten, dass sie keine Live-AI-Ausgaben oder echte Leistungsnachweise von Entitäten darstellen. Das Benchmark MUSS verhindern, dass seine synthetischen Angaben als echte Informationen über die Ankerentität veröffentlicht oder indexiert werden. Der Erfolg des Benchmarks MUSS anhand der Wiederherstellung der versiegelten Generator-Wahrheit bestimmt werden, einschließlich Anspruchsgrenzen, atomarer Entscheidungen, Auslassungen, kritischer und wichtiger Tore, Antwortstatus, Komponentenbewertungen, Gesamtbewertungen, Fairness-Ergebnisse, Gültigkeit der Erfassung und Abdeckung der Unsicherheit. Eine hohe allgemeine Genauigkeit DARF fehlende kritische Fälle NICHT kompensieren. Kritische Rückrufrate und kritische Fehlalarmrate MÜSSEN zusammen berichtet werden. Ein nicht erfüllter Benchmark-Schwellenwert, ein Leckageereignis, ein Wiederherstellungsfehler oder eine Fehlklassifikation MUSS sichtbar bleiben und DARF NICHT durch Ändern von Labels, Seeds, Proben, Gewichten oder Bewertungsregeln nach Beobachtung der Ergebnisse entfernt werden. Jede Benchmark-Entwicklung, -Generierung, -Sicherung, -Leckageprüfung, -Entscheidung, -Wiederherstellungsberechnung, -Schwellenwertentscheidung, -Überarbeitung und öffentliche Veröffentlichung MUSS versioniert und einer verantwortlichen Person oder Organisation zuordenbar sein.

KAPITEL 18

Ein synthetisches End-to-End-Audit von 30.000 Antworten

Eine berechnete Demonstration der Methode

METHODISCHE GRUNDLAGEN / K03; K09

Kapitelgrenze

Kapitel 17 definierte den Apple.com Synthetischen Benchmark. Dieses Kapitel verwendet eine intern konsistente, arithmetisch verifizierte Demonstration im Buch, um zu zeigen, was die vollständige Prüfkette erfordern würde:

- ein berechneter synthetischer Datensatz, der die Hauptantworten von 30,000 darstellt,
- modellierte Erfassungsfehler,
- Generatorwahrheit vor Gutachtern verborgen,
- Antworten in atomare Aussagen zerlegt,
- Gates für kritische und schwerwiegende Befunde angewendet,
- Sprach- und Nutzerzustandsunterschiede berechnet,
- und die Bewertungsarchitektur verarbeitet die resultierenden Daten.

Die Frage ist, wie genau NOMOS eine zuvor versiegelte synthetische Realität wiederherstellen kann. Dieses Protokoll verwandelt jeden Fehler in einen ausführbaren Audit-Punkt, der ein Modell, Datum, Land, Sprache, Abfragesatz, Wiederholungsanzahl und Messaufzeichnung enthält. Kapitel 18 richtet diese Anforderung nun direkt an NOMOS selbst. Es beginnt jedoch mit einem Integritäts-Gate.

> *Der 30,000-Antwortdatensatz in diesem Kapitel ist eine intern konsistente, arithmetisch verifizierte synthetische Anwendung, die für das Buch vorbereitet wurde. Es handelt sich nicht um ein ausgeführtes Korpus oder ein Feldexperiment.*

Diese Zahlen repräsentieren nicht:

- die reale Leistung von Apple Inc.,
- die realen Inhalte von apple.com,
- das Verhalten eines realen KI-Anbieters,
- Antworten, die von echten Teilnehmern eingereicht wurden,
- oder ein unabhängig durchgeführter und öffentlich veröffentlichter Software-Benchmark.

Zwei Status müssen daher getrennt bleiben.

IN-BUCH DEMONSTRATIONSSTATUS

Die In-Buch-Demonstration kombiniert:

- das im Kapitel 17 festgelegte Design,
- vordefinierte synthetische Verteilungen,
- Generator-Truth-Aufzeichnungen,
- Fehlerinjektionen,
- Antwortstatus,
- Wiederherstellungsmetriken.

Zusammen bieten diese Elemente eine umfassende, berechnete Demonstration, wie das Design funktioniert. Demonstrations-ID:

NOMOS-APPLE-SYNTH-DEMO-RUN-18-001

ECHTER PRÜFUNGSTATUS

Ein reales, öffentlich verfügbares Benchmark-Korpus wurde nicht:

- erzeugt,
- durch ausgeführten Code verarbeitet,
- mit seinem Seed-Manifest veröffentlicht,
- durch Blindbewertung beurteilt,
- oder unabhängig reproduziert.

Dieses Protokoll verwandelt jeden Fehler in einen operativen Prüfpunkt, der das Modell, Datum, Land, Sprache, Abfragesatz, Wiederholungsanzahl und die Messaufzeichnung enthält. Bis dieses Protokoll an einem realen Korpus ausgeführt wird, bleibt der Benchmark:

BQ-0— NICHT AUSGEFÜHRT

Der separate Status der Demonstration im Buch ist:

DEMO-BQ-2— PROZESSLINIE TROCKENLAUF MIT ERFORDERLICHER MATERIALKORREKTUR

Ohne diese Unterscheidung würde NOMOS die erste Normverletzung in seinem eigenen Buch begehen, indem es den Eindruck erweckt, dass ein realer Benchmark durchgeführt wurde. Der 30,000-Antwortdatensatz ist weder die Ausgabe ausgeführter Software noch gesammelte Felddaten; es handelt sich um eine vordefinierte synthetische Anwendung, deren interne Konsistenz arithmetisch überprüft wurde. Ihr Zweck ist:

- um nicht den Eindruck zu erwecken, dass 30,000-Antworten tatsächlich gesammelt wurden,
- um die Datensätze zu identifizieren, die ein echter End-to-End-Durchlauf erzeugen würde,
- um die Berechnungen anzugeben, die er durchführen würde,
- um die Fehler zu identifizieren, die die Veröffentlichung blockieren würden,
- um die Teile der eigenen Methode von NOMOS offenzulegen, die weiterhin unzureichend bleiben.

Die leitenden Anforderungen von NOMOS bleiben unverändert: Nachweise liefern, Grenzen setzen, Kontext bewahren und Zeit aufzeichnen. In diesem Kapitel werden diese vier Bedingungen angewandt auf:

- das Benchmark-Design,
- Generator-Wahrheit,
- Entscheidungen der Schiedsrichter,
- Ergebnis der Wiederherstellung,

- NOMOS' eigene Kompetenzbewertung.

Diese Unterscheidungen regeln den Rest des Kapitels.

NOMOS Herausforderung

Bei der berechneten 30,000-Antwortdemonstration stellt NOMOS weitgehend die Anspruchsgrenzen wieder her, erkennt die meisten Fälle falscher Entitäten, stellt jeden kritischen Fall nach der vollständigen Schiedsgerichtskette wieder her und schätzt die Antwortverteilung nahe an der versiegelten synthetischen Verteilung ein.

Sein wiederhergestellter zusammengesetzter Wert weicht nur um wenige Punkte vom Generator-Truth-Wert ab. Das könnte zur Erklärung einladen: ‚NOMOS wurde erfolgreich verifiziert.' Dennoch bleiben zwei wesentliche Probleme: Die Zuordnung von Quelle zu Behauptung ist nicht ausreichend genau, und einige entscheidungsrelevante Grenzausschlüsse, die in Empfehlungen eingebettet sind, werden übersehen. Die größte Schwierigkeit zeigt sich bei Antworten wie:

„Da das Unternehmen Zahlungsdienste anbietet, kann es als reguliertes Finanzinstitut betrachtet werden. [Quelle]" Quelle: zeigte nur die eingeschränkte Zahlungsaufzeichnung einer separaten Tochtergesellschaft. Manchmal habe ich die Referenz mit der Behauptung über die Muttergesellschaft verbunden. Auch mit diesen Antworten hatte ich Schwierigkeiten:

„Dieses Unternehmen ist eine starke Wahl für Ihre Zahlungsbedürfnisse in Deutschland." Die Antwort hat eindeutig nicht gesagt: „Die Muttergesellschaft bietet lizenzierte Zahlungsdienste in ganz Deutschland an." Allerdings:

- Empfehlung,
- Entfernung der Unterscheidung der Tochtergesellschaft,
- nicht Angabe der Beschränkung durch die lokale Behörde

Zusammen erzeugen die Empfehlung, der Verlust der Unterscheidung der Tochtergesellschaft und das Auslassen der Beschränkung durch die lokale Behörde diesen Eindruck. Das Ergebnis nur als ‚fehlendes Detail' zu behandeln, unterschätzt eine wesentliche Auslassung, die die Entscheidung des Nutzers ändern könnte. Die Wiederherstellung des Gesamtscores sieht ausgezeichnet aus: Generator-Truth-Ökosystem-Score, 873.2; NOMOS-Wiederherstellung, 872.2; absoluter Fehler, 1 Punkt.

Das sieht ausgezeichnet aus. Aber bei den Komponenten:

- Evidenz- & Zitierintegrität liegen einige Punkte unter dem Soll,
- Umfangs- & Grenzwertintegrität liegen einige Punkte unter dem Soll,
- Stabilität ist bei einigen Produkten zu hoch,
- Fairness ist bei einigen Produkten zu hoch

Mehrere fehlerhafte Komponenten gleichen sich arithmetisch aus, sodass ein zusammengesetzter Wert dem korrekten Ergebnis nahekommt, aus falschen Gründen. Die Schlussfolgerung lautet:

> *Die korrekte Gesamtsumme macht die falschen Komponenten nicht automatisch korrekt.*

In der ersten Bewertungsrunde werden sieben der 1,200-Kritikfälle übersehen. Die folgenden Prüfungsebenen greifen dann ein:

- unabhängige zweite Beurteilung,
- domain-experten Überprüfung,
- Überprüfung durch einen leitenden Beurteiler.

Zusammen erfassen diese Schichten alle sieben Fälle und erzielen eine endgültige Kritische Rückrufquote von 100 Prozent. Das ist ein Erfolg des Governance-Systems, aber es zeigt auch, dass

eine einlagige Beurteilung nicht ausreicht. Entfernt man die zusätzlichen Überprüfungsschichten, um Kosten zu senken, gehen sieben kritische Fälle verloren.

Tests sollten nicht nur das Ergebnis zeigen, sondern auch, welche Governance-Ebene das Ergebnis möglich gemacht hat. Die erste Regelung dieses Abschnitts lautet wie folgt:

> *Eine hohe Gesamterfassung in einem Test bedeutet nicht, dass alle seine Teilsysteme ausreichend sind.*

Ihre zweite Bestimmung lautet:

> *Die endgültige kritische Rückrufaktion sollte getrennt von der ersten Adjudikations-Rückrufaktion veröffentlicht werden.*

Seine dritte Bestimmung besagt:

> *Selbst wenn der zusammengesetzte Punktwertfehler gering ist, können Komponentenfehler erheblich sein.*

Seine vierte Bestimmung besagt:

> *Wenn einer der Kandidatenschwellenwerte nicht erreicht wird, können andere starke Ergebnisse dieses Versagen nicht aufheben.*

Seine fünfte Vorschrift lautet:

> *Eine interne synthetische Berechnung im Buch ersetzt nicht den öffentlich verfügbaren tatsächlichen Testlauf.*

Seine sechste Bestimmung lautet:

> *Das bedingte Verlassen von NOMOS aus seinem eigenen Test ist kein Versagen; es bedeutet lediglich, dass der Standard ihm kein besonderes Privileg gewährt.*

1. ZWECK DES KAPITELS

Zweck dieses Abschnitts ist es, das in Abschnitt 17 festgelegte Testdesign innerhalb einer durchgängigen synthetischen Demonstration zu betreiben und zu zeigen, inwieweit NOMOS die Generatorwahrheit in den folgenden Schichten wiederherstellt:

- Gültigkeit erfassen
- Prompt-Abgleich
- Aussagegrenze
- Entitätsauflösung
- Faktische Unterstützung
- Umfang
- Zeit
- Zuschreibung
- Modalität
- Zitationszuordnung
- Auslassung
- Bedeutungsgrad
- Antwortstatus
- Kritisches Tor
- GEO-1000 Benutzungsverteilung
- Sprach- und geographische Fairness

- Unterschied zwischen sauber und natürlich
- Wellenstabilität
- Komponentenscores
- Composite NOMOS Score
- Konfidenzintervalle
- Obergrenzen für seltene Ereignisse
- Entscheidung zur öffentlichen Offenlegung

Der Abschnitt sollte auch ehrlich die folgende Frage beantworten:

In welchen Bereichen ist NOMOS 0.9 nach den eigenen Kandidatenstandards gescheitert?

2. ZENTRALE NORMATIVE BESTIMMUNG

Eine End-to-End-Synthetikbewertung kann nicht allein auf der Grundlage der Nähe der endgültigen Gesamtscore zum Generatorwahrheitswert als erfolgreich betrachtet werden. Jedes der Ergebnisse für Erfassung, Aussageextraktion, atomare Entscheidung, Auslassung, Klassifikation kritischer und schwerwiegender Befunde, Antwortstatus, Fairness, Komponentenwiederherstellung, Konfidenzintervall und Gate-Wiederherstellung sollte einem separaten Akzeptanzschwellenwert unterliegen. Wenn einer der folgenden Punkte fehlschlägt, testen Sie:

- durch Ändern des Tags,
- durch Senken des Schwellenwerts,
- durch Auswahl eines anderen Seeds,
- durch Entfernen niedrig bewerteter Fälle,
- durch Veröffentlichung nur des zusammengesetzten Ergebnisses

kann nicht als bestanden betrachtet werden. Korrektes Ergebnis:

- Anzeige des fehlgeschlagenen Feldes,
- Versionierung der Methode,
- erneut testen am unberührten Holdout

muss sein.

3. ANWENDUNGSBEREICH DER SYNTHETISCHEN ANWENDUNG

3.1. Hauptprinzip Prävalenz-Korpus

```
10 synthetische KI products× 1,000 einzigartig users× 3 waves= 30,000 Hauptergebnisse
```

3.2. Diagnostischer Anhang

Modul	Fall
Entitätskollision	750
Nachweise und Zitat	1,500
Grenze und Empfehlung	1,000
Zeitlicher und örtlicher Umfang	750
Sauber–Natürlich	1,000

Modul	Fall
Sprachliche Gleichwertigkeit	1,000
Gesamt	6,000

3.3. Prioritätsgrad-Herausforderungskorpus

Klasse	Fall
kann Bestätigt Kritisch tragen	1,200
Bestätigt Wichtig	1,200
Hart Mittel	600
Negativkontrolle ähnlich wie Kritisch	600
Gesamt	3,600

3.4. Capture-Integritäts-Korpus

Klasse	Fall
Gültig	2,000
Bedingt gültig	500
Außerhalb der Welle	250
Aufforderung/Ausgabe-Mismatch	250
Duplikat	250
Manipulationskandidat	250
Fabriziert/manipuliert	250
Barrierefreiheit und alternative Erfassung	250
Gesamt	4,000

3.5. Adjudikations-Goldkorpus

2,400 Kalibrierungs- und Driftfälle

3.6. Gesamtuniversum der synthetischen Überwachung

Haupt- und ergänzende Korpora zusammen:

```
30,000+6,000+3,600+4,000+2,400=46,000
```

erstellt einen synthetischen Fall oder ein Evidenzpaket. Nur diese:

Geben Sie die Prävalenz des Bevölkerungspanels mit den Hauptantworten 30,000 ein.

Andere Korpora:

- Methodenkapazität,
- Herausforderungsleistung,
- Erfassungsgenauigkeit,
- Gutachterkalibrierung

sind für.

4. GESCHLOSSENE EINGABEN

Es wird angenommen, dass die folgenden Datensätze festgeschrieben sind, bevor die synthetische Demonstration beginnt:

Datensatz	Version
Test	NOMOS-APPLE-SYNTH-BENCH-0.9
Entity Twin	APPLE-SYNTH-TWIN-001
Truth-Pack	SYNTH-APPLE-TP-001
Prompt-Set	SYNTH-PROMPT-SET-001
Bevölkerungsrahmen	SYNTH-POP-FRAME-001
KI-Register	SYNTH-AI-REGISTRY-001
Wellenplan	SYNTH-WAVE-PLAN-001
Erfassungsschema	NOMOS-CAPTURE-SYNTH-0.9
Codebuch zur Beurteilung	SYNTH-ADJUDICATION-0.9
Bewertungsmethode	NOMOS-SCORE-0.9
Seed-Manifest	SYNTH-SEED-MANIFEST-001
Kandidaten-Schwellenwerte	SYNTH-ACCEPTANCE-0.9

Keine dieser Versionen wurde nach Bekanntwerden des Wiederherstellungsergebnisses als geändert betrachtet.

5. GEHEIME GENERATOR-WAHRHEITSVERTREILUNG

SYNTHETISCHE DEMONSTRATION — KEINE ECHTEN LEISTUNGSDATEN

Versiegelte Generator-Wahrheitsverteilung für die Hauptantworten 30,000:

Antwortstatus	Anzahl der Antworten	1,000 Nutzeräquivalent
RP-1 Vollständig bestanden	20,045	668.17
RP-2 Bestanden mit Hinweis	4,310	143.67
RP-3 Bedingt bestanden	2,370	79.00
RP-4 Fehlschlag — Schwerwiegend	1,602	53.40
RP-5 Automatischer Fehlschlag — Kritisch	33	1.10
RP-6 Ungeklärt	450	15.00
RP-7 Keine verwertbare Antwort	1,110	37.00
RP-8 Nicht bewertbar	80	2.67
Gesamt	30,000	1,000

Strikt bestanden:

SP_1000^G = 668.17+ 143.67= 811.84

Akzeptabel bestanden:

AP_1000^G = 811.84+ 79= 890.84

Bestätigt Kritisch:

CR_1000^G = 1.10

Bestätigt Groß:

MR_1000^G = 53.40

Dies ist die allgemeine Verteilung:

- gut,
- schlecht,
- realistisch

Dies ist keine KI-Marktvoraussage. Es ist eine synthetische Testverteilung, die zehn verschiedene Fehlerprofile zusammen testet.

6. VERTEILUNG WURDE VON NOMOS WIEDERHERGESTELLT

NOMOS Wiederherstellungsergebnis festgeschrieben, bevor Generator-Truth geöffnet wird:

Antwortstatus	Anzahl der Wiederherstellungen	1,000 Nutzeräquivalent
RP-1 Vollständig bestanden	19,980	666.00
RP-2 Bestanden mit Hinweis	4,365	145.50
RP-3 Bedingt bestanden	2,382	79.40
RP-4 Fehlschlag — Schwerwiegend	1,585	52.83
RP-5 Automatischer Fehlschlag — Kritisch	33	1.10
RP-6 Ungeklärt	468	15.60
RP-7 Keine verwertbare Antwort	1,091	36.37
RP-8 Nicht bewertbar	96	3.20
Gesamt	30,000	1,000

Wiederherstellung Strikte Bestehen:

SP_1000^N = 666+ 145.50= 811.50

Akzeptable Wiederherstellungsprüfung:

AP_1000^N = 811.50+ 79.40= 890.90

Strikte Prüfung absoluter Wiederherstellungsfehler:

```
|811.50-811.84| =0.34
```

ist dem Nutzer äquivalent. Akzeptabler Prüfungsfehler:

```
|890.90-890.84| =0.06
```

ist dem Nutzer äquivalent. Dies zeigt, dass die Wiederherstellung der Ergebnisverteilung stark ist. Es zeigt jedoch nicht, dass alle Teilsysteme gleichermaßen stark sind.

7. FEHLER BEI DER WIEDERHERSTELLUNG DER ANTWORTVERTEILUNG

Absoluter Unterschied der Antwortanzahl in jeder RP-Klasse:

Status	Wiederherstellungsdifferenz
RP-1	-65
RP-2	+55
RP-3	+12
RP-4	-17
RP-5	0
RP-6	+18
RP-7	-19
RP-8	+16

Gesamtvarianzbasierter Verteilungsfehler:

E_D = (65+55+12+17+0+18+19+16)/60,000= 0.003 37

Mit anderen Worten: ungefähr 0.34% der Antwortverteilung wurde an der falschen Klassenbegrenzung wiederhergestellt. Häufigste Verwechslungen:

- Mit RP-1 und RP-2
- RP-3 und RP-4
- RP-6 und RP-4
- Mit RP-7 und RP-8

Es ist dazwischen aufgetreten. Die Anzahl der kritischen Antworten wurde in der abschließenden Beurteilung vollständig wiederhergestellt.

8. SYNTHETISCHE KI-PRODUKTPROFILE

> *Die folgenden Produkte stellen keine echten KI-Anbieter dar.*

Jedes Produkt trägt 3,000 Antworten über drei Wellen.

Produkt	Striktes Bestehen / 1,000	Akzeptables Bestehen / 1,000	Ernsthaft / 1,000	Kritisch / 1,000	Hauptsynthetisches Problem
SYNTH-AI-01	950.00	990.00	0.00	0.00	Niedriges Risiko, ausgewogenes Profil
SYNTH-AI-02	850.00	933.33	40.00	0.00	Weitschweifigkeit und beiläufige Behauptung
SYNTH-AI-03	750.00	866.67	84.00	0.00	Abbau bei ressourcenarmer Sprache
SYNTH-AI-04	833.33	916.67	33.33	1.00	Drift des natürlichen Zustands
SYNTH-AI-05	766.67	833.33	100.00	2.67	Evidenzwäscherei
SYNTH-AI-06	650.00	733.33	33.33	0.00	Unnötige Ablehnung
SYNTH-AI-07	783.33	883.33	83.33	0.33	Zeitlicher und lokaler Abbau

Produkt	Striktes Bestehen / 1,000	Akzeptables Bestehen / 1,000	Ernsthaft / 1,000	Kritisch / 1,000	Hauptsynthetisches Problem
SYNTH-AI-08	700.00	800.00	150.00	2.33	Entitätenverwechslung
SYNTH-AI-09	900.00	983.33	0.00	0.00	Entschlossenes, aber moderates Profil
SYNTH-AI-10	935.00	968.33	10.00	4.67	Sehr hoher Durchschnitt, selten kritisch

Der Zweck der Tabelle besteht nicht darin, Produkte zu bewerten. Sie dient dazu, diese drei methodischen Fälle zu testen:

- SYNTH-AI-01: hohe Punktzahl und starkes, ungeprüftes Profil
- SYNTH-AI-06: geringe falsche Angaben, aber hohe Nicht-Antwort-Rate im Profil
- SYNTH-AI-10: sehr hoher Durchschnitt, aber Profil mit offenen kritischen Ereignissen

Es testet speziell, ob SYNTH-AI-10, NOMOS den folgenden Fehler verarbeitet haben:

> *Ignoriere das kritische Tor aufgrund des hohen Durchschnitts.*

9. PRODUKTBASIERTE KOMPOSIT-PUNKTZAHL-WIEDERHERSTELLUNG

Synthetisches Produkt	Generator NOMOS	Wiederherstellung NOMOS	Absoluter Fehler	Generator-Gate	Wiederherstellungs-Gate
SYNTH-AI-01	970	968	2	950+ Leistungs-Kandidat	Gleich
SYNTH-AI-02	917	914	3	Schwerwiegend nicht bestanden	Gleich
SYNTH-AI-03	842	846	4	Schwerwiegend nicht bestanden	Gleich
SYNTH-AI-04	905	900	5	Kritische Sperre	Gleich
SYNTH-AI-05	811	815	4	Kritische Sperre	Gleich
SYNTH-AI-06	785	780	5	Schwerwiegend nicht bestanden	Gleich
SYNTH-AI-07	866	861	5	Kritische Sperre	Gleich
SYNTH-AI-08	748	752	4	Kritische Sperre	Gleich
SYNTH-AI-09	932	934	2	Bedingt	Gleich
SYNTH-AI-10	956	952	4	Kritische Sperre	Gleich

Durchschnittlicher absoluter zusammengesetzter Score-Fehler in der berechneten synthetischen Probe:

MAE_NOMOS = 3.8

Maximaler Fehler:

MAXE_NOMOS = 5

Gate-Wiederherstellung: Sie ist vollständig in den Produkten 10/10. Die wichtigste Zeile in der Tabelle ist SYNTH-AI-10:

> *Obwohl der Wiederherstellungswert 952 beträgt, wurde der Status als Kritische Sperre beibehalten.*

Dies zeigt, dass die Gate>Score-Bestimmung in Abschnitt 15 im berechneten Szenario erhalten bleibt.

10. ERGEBNIS DER AUSSAGEEXTRAKTION

Anzahl der Generator-Truth-Atome im Haupt- und Diagnostikkorpus: 146,400 Atome, die durch die Aussageextraktion NOMOS extrahiert wurden: 147,200 Korrekt zugeordnete Atome: 141,500 Genauigkeit:

P = 141,500/147,200= 0.961 3

Abruf:

R = 141,500/146,400= 0.966 5

```
Aussagegrenzen F1:
```

F1 = 0.963 9

Kandidaten-Schwelle:

F1≥0.95

Ergebnis:

KANDIDATEN-SCHWELLE ERREICHT

11. HAUPTARTEN VON FEHLERN BEI DER AUSSAGEEXTRAKTION

Hauptursachen für ungefähr 5,700 falsch begrenzte Atome:

Fehlertyp	Zähler
Nicht trennen kombinierter Länder- oder Produktansprüche	24%
Unnötiges Aufteilen der Zuschreibung als separates Atom	19%
Fehlender impliziter Empfehlungsg Anspruch	17%
Berücksichtigung eines zurückgezogenen Anspruchs innerhalb der Selbstkorrektur als aktiv	15%
Verwechslung eines Zitatsatzes mit einer unabhängigen Tatsache	11%
Pronomen- und Koreferenz-Mehrdeutigkeit	9%
Andere	5%

Dieses Ergebnis zeigt, dass die Aussageextraktion im Allgemeinen stark ist, aber:

- Empfehlung,
- Zuschreibung,
- Rücknahme

weist darauf hin, dass zusätzliche Codebücher in den Feldern erforderlich sind.

12. ERGEBNISSE DER ATOMIC-ENTSCHEIDUNGS-WIEDERHERSTELLUNG

Dimension	Wiederherstellungsmaß	Kandidatenschwelle	Ergebnis
F1-Wert der Aussagegrenze	0.964	≥ 0.95	Bestanden
Genauigkeit der Entitätsauflösung	0.987	≥ 0.98	Bestanden
Faktischer Status Makro-F1	0.923	≥ 0.90	Bestanden
Umfang Makro-F1	0.906	≥ 0.90	Bestanden
Zeitlicher Status Makro-F1	0.945	≥ 0.90	Bestanden
Attributions Makro-F1	0.918	≥ 0.90	Bestanden
Modality Macro-F1	0.934	≥ 0.90	Bestanden
F1 der Zitationszuordnung	0.887	≥ 0.90	Nicht bestanden
Materialauslassungs-F1	0.824	≥ 0.85	Nicht bestanden
Makro-F1 der Hauptklassifikation	0.952	≥ 0.95	Bestanden
Makro-F1 des RP-Status	0.928	≥ 0.92	Bestanden

Von elf grundlegenden Akzeptanzfeldern: neun bestanden, zwei nicht. Daher Demonstration:

versiegelte Validierung bestanden

kann nicht nummeriert werden. Qualität der Kandidatentests:

DEMO-BQ-2

bleiben wie sie sind.

13. ZITATIONSZUORDNUNGSFEHLER

```
F1 der Zitationszuordnung:
```

war 0.887. Kandidatenschwelle: 0.90. Der Unterschied scheint klein: 0.013 Aber Zitationsintegrität:

- Lizenz,
- unabhängige Validierung,
- Kundenbeziehung,
- Überlegenheit,
- Ratschläge

ist wesentlich, da sie die Quelle des Vertrauens für Behauptungen bestimmt.

13.1. Verpasste Zitierarten

Die drei häufigsten Probleme wurden beobachtet.

A. Zitierverteilung am Absatzende

Ein Zitat im Absatz unterstützt nur die letzte Behauptung, dennoch verbanden die Prüfer es mit allen vorherigen Behauptungen.

B. Eine nur der Zuschreibung dienende Quelle als Wahrheitsspeicher betrachten

Firmenblog: Während die Behauptung „Das Unternehmen beschreibt sich selbst als Weltmarktführer“ unterstützt wird, wurde sie auch als Unterstützung der Behauptung „Das Unternehmen ist ein Weltmarktführer“ gezählt.

C. Dieselben Quellengrundlagen als unabhängig betrachten

Drei URLs:

- Pressemitteilung,
- Syndizierung,
- KI-Zusammenfassung

wurden als drei unabhängige Evidenzstücke abgebildet, obwohl sie aus derselben Quelle stammen.

13.2. Verteilung der Wichtigkeit von Zitierfehlern

Wichtigkeit	Nicht übereinstimmende Paarung
Kritischer Kandidat	7
Wesentlich	84
Moderat	311
Beratung	192

Sieben Fälle kritischer Kandidaten:

- doppelblinde Peer-Review,
- Überprüfung der Quellenherkunft,
- Senior Adjudicator

wurden in den Stufen korrigiert. Keiner hat sich in eine endgültige Torinversion verwandelt.

```
Die Zitier-Abgleich F1 liegt jedoch noch unter dem Kandidatenschwellenwert.
```

Das richtige Urteil, das NOMOS abgeben würde, ist:

> *Die Methode zur Nachweis- und Zitationsintegrität ist noch nicht ausgereift genug für einen unabhängigen Live-Einsatz.*

14. VERSPÄTUNG BEIM WESENTLICHEN FEHLEN

```
F1 Wert für erhebliches Fehlen:
```

0.824 Kandidatenschwellenwert: es war 0.85. Dies ist die wichtigste methodische Erkenntnis der Demonstration.

14.1. Häufigste Arten übersehener Auslassungen

Art der Auslassung	Übersehener Anteil
Servicegrenze innerhalb der Empfehlung	31%
Unterscheidung zwischen Tochtergesellschaft und Muttergesellschaft	22%
Länderspezifische oder jurisdiktionsbezogene Einschränkung	17%
Beschränkung auf aktuellen oder historischen Status	12%
Garantie- oder Rücknahmeausschluss	9%
Nutzergeeignetheit	6%

Art der Auslassung	Übersehener Anteil
Andere	3%

Diese Struktur im Besonderen hat Schwierigkeiten verursacht: Die Antwort fällt nicht ausdrücklich ein falsches Urteil, aber sie bietet nicht die erforderliche Einschränkung, um den scheinbar richtigen Rat sicher zu interpretieren. Beispiel: „Apple-SYNTH könnte eine starke Wahl für Ihre Finanztransaktionen in Deutschland sein, da es Zahlungslösungen anbietet.“ Antwort:

- es sagt nicht direkt „die Muttergesellschaft ist eine Bank“,
- aber es erwähnt nicht das separate Zahlungs-Subunternehmen,
- es erklärt nicht, dass die Befugnis auf den Markt beschränkt ist,

es hat Beratung bezüglich des Hauptelements erstellt. Das wesentliche Auslassen dieser Antwort wurde von einigen Entscheidern nur gesehen als:

OM-2— Erforderliches Element fehlt

während in Wirklichkeit benötigt wird:

OM-4— Irreführender Auslassungskandidat

und in einigen Fällen:

CG-08— Kritische Grenzauslassung

ist übersehen worden.

15. ERGEBNIS DER KRITISCHEN HERAUSFORDERUNG

Bedeutungsgrad innerhalb des Challenge-Korpus: 1,200 Generator Wahrheit Kritische Fälle wurden gefunden.

15.1. Erste unabhängige Überprüfung

Fälle, die nach den ersten zwei Runden der unabhängigen Überprüfung von Konsens oder mindestens einem Gutachter als kritisch eingestuft wurden: 1,193 Erstüberprüfung Kritischer Rückruf:

```
1,193/1,200 = 0.9942
```

Verpasst: 7 Fälle wurden gefunden. Verteilung der sieben verpassten Fälle:

- 3 wesentliche Grenzversäumnisse
- 2 Evidenzspülung
- 1 zeitliche Autoritätsversäumnis
- 1 Offenlegung eingeschränkter Daten

15.2. Nach Experten- und Senior-Gutachtliche Bewertung

Alle sieben Fälle:

- Überprüfung von Auslassungen,
- Überprüfung der Quellenherkunft,
- Datenschutzüberprüfung,
- Senior Adjudicator

Phasen waren vorhanden. Endgültige kritische Rückruf:

```
1,200/1,200 = 1,000
```

15.3. Kritische negative Kontrolle

Fälle, die kritisch ähnelten, aber nicht kritisch waren: 600 Negativkontrollfälle wurden gefunden. Falscher kritischer Kandidat in der ersten Runde: 5-Fälle traten auf. Erste Durchgangs-Falschkritikrate:

```
5/600 = 0.0083 = 0.83%
```

Fünf Fälle in der endgültigen Überprüfung:

- drei Major,
- zwei Moderate

wurde korrigiert als. Endgültige falsche Kritikraten: 0 Endgültige Gate-Inversion: 0-Fälle.

16. ECHTE BEDEUTUNG DES KRITISCHEN ERGEBNISSES

Eine endgültige 100%-Kritische Rückrufaktion bedeutet nicht: „Ein einzelner Prüfer kann alle kritischen Fehler erfassen." Das tatsächliche synthetische Ergebnis ist:

> *Alle Fälle mit kritischen Herausforderungen wurden erfasst, als die doppelte Überprüfung, Experteniüberprüfung und Senior-Adjudikation zusammen verwendet wurden.*

Daher aus Kosten- oder Geschwindigkeitsgründen:

- der zweite Prüfer,
- Fehlstellenprüfer,
- Expertenprüfer

werden entfernt, das gleiche Ergebnis kann nicht erwartet werden. Die Testmethode, nicht nur ihre Formel:

zeigt, dass obligatorische Governance-Ebenen

ebenfalls testbar sein sollten; dies ist keine empirische Validierung.

17. HAUPTKLASSIFIKATIONSERGEBNIS

Von den bestätigten Hauptfällen von 1,200:

- ist die Mehrheit echte Hauptfälle,
- ein begrenzter Teil ist mäßig,
- einige sind kritische Kandidaten,
- einige sind ungelöst

in der ersten Runde.

```
Endgültiges Haupt-Makro-F1:
```

war 0.952. Kandidatenschwelle: 0.95 Ergebnis:

KANDIDATEN-SCHWELLE ERREICHT

Aber die Schwelle allein: wurde nur um 0.002 Unterschied überschritten. Daher Hauptklassifikation:

- ist stark,
- aber trägt keinen komfortablen Sicherheitsabstand

Es sollte als Feld aufgezeichnet werden.

18. ANTWORTSTATUS VERWIRRUNG

```
RP Status Makro-F1:
```

Es ist zu 0.928 geworden. Die Hauptverwirrungen:

Tatsächliche Klasse	Häufigste Fehlklassifikation bei der Wiederherstellung
RP-3 Bedingt bestanden	RP-4 Schwerwiegend nicht bestanden
RP-4 Schwerwiegend nicht bestanden	RP-3 Bedingt bestanden
RP-6 Ungeklärt	RP-4 Schwerwiegend nicht bestanden
RP-7 Keine verwertbare Antwort	RP-3 Bedingt bestanden
RP-8 Nicht bewertbar	RP-7 Keine verwertbare Antwort

Kritische Klasse: vollständig im Endstadium wiederhergestellt. Die härteste Grenze:

Mit Bedingt bestanden und Schwerwiegend nicht bestanden

hat sich gebildet zwischen. Diese Schwierigkeit ist direkt:

- kumulative Wesentlichkeit,
- materielle Auslassung,
- inwieweit Ihre Hauptaufgabe gestört wird

Es ist aus Ihren Fragen entstanden.

19. ERGEBNIS DER INTEGRITÄTSERFASSUNG

Korrekt klassifizierte Capture Integrity-Fälle aus 4,000: 3,928 Genauigkeit der Aufnahmevalidität:

```
3,928/4,000 = 0.982
```

Kandidatenschwelle: 0.98 Ergebnis:

KANDIDATEN-SCHWELLE ERREICHT

19.1. Erfassungsfehler

Verteilung 72 falsch klassifizierten Fälle:

Aktueller Zustand	Falsche Entscheidung	Fall
Bedingt gültig	Gültig	20
Gültige Barrierefreiheitsalternative	Teilweiser Nachweis	8
Außerhalb der Welle	Bedingt gültig	9
Manipulation vermutet	Teilweiser Nachweis	12
Beinahe-Duplikat	Gültig	5
Gültig	Bedingt gültig	18
Gesamt		72

Kein Paket wurde hergestellt oder manipuliert: Keines wurde als vollständig gültig in die Endanalyse aufgenommen. Semantische Antwort:

- in die Kategorien gezwungen werden von

  wahr,
- falsch,
- Kritisch

wurde nicht damit in Verbindung gebracht. Dies ist nicht das Ergebnis eines echten Signifikanztests; es ist eine Annahme des Designs des synthetischen Beispiels. Dies zeigt, dass das Prinzip der Trennung von Erfassung und Semantik in der Demonstration funktioniert.

20. ERFASSUNGSERGEBNIS IM HAUPT-30,000-KORPUS

Innerhalb des Hauptkorpus:

Erfassungsebene	Beobachtung
NCL-4 Instrumentiert und Signiert	27,850
NCL-3 Vollständiges Evidenzmittelpaket	2,070
RP-8/ Nicht bewertbar	80
Gesamt	30,000

Grund für 80 Nicht bewertbar-Aufzeichnungen:

Grund	Beobachtung
Aufforderung/Ausgabe-Mismatch	24
Unvollständige Antwort beendet	18
Außerhalb der Welle und unzuverlässige Zeit	16
Doppelte Nachweise	12
Unlösbare Systemtransformation	10
Gesamt	80

Diese 80-Datensätze:

- wurden nicht umklassifiziert als
- korrekte Antwort,
- Produktfehler,

falsche Antwort.

In der Hauptverteilung:

RP-8— Nicht bewertbar

wurde unverändert beibehalten.

21. ENTITÄTSWIEDERHERSTELLUNG

Im Kerndomäne – Entitätsauflösung: 30,000-Hauptantworten und Fälle von diagnostischen Entitätskollisionen wurden zusammen bewertet. Genauigkeit der Entitätsauflösung: 0.987. Hauptfehler:

- Bezahlung von verbundenen Unternehmen wurde als Hauptgesellschaft gezählt
- Franchise wurde als direkt betriebenes Geschäft gezählt
- Zielgesellschaft wurde wegen irrelevanter Apple-SYNTH Namenskollision gezählt
- Betrachten Sie die Produktfamilie als juristische Person
- Zählen Sie den historischen Partner als aktuellen Affiliate

Nicht alle Fälle von falschen Einheiten haben die gleiche Bedeutung. Kritische Fälle der Autoritätsübertragung wurden in der Endkontrollstufe gefunden.

22. WIEDERHERSTELLUNG DER FAKTISCHEN UNTERSTÜTZUNG

```
Faktischer Status Makro-F1:
```

0.923 Stärkste Klassen:

- Unterstützt
- Widersprochen
- Offizieller Anspruch genau zugeordnet

Schwächste Klassen:

- Nicht unterstützt
- Referenzlücke
- Unterstützt mit erforderlicher Qualifikation

Die Grenze, die besonders schwierig war, ist: „Keine Nachweise gefunden.“ versus: „Die verfügbaren Nachweise widersprechen der Behauptung.“ Einige Prüfer in einer Open-World-Situation:

- nicht unterstützt,
- widersprochen

haben die Unterscheidung zu strikt getroffen. Die abschließende Überprüfung hat die meisten dieser Fälle korrigiert.

23. UMFANGS-WIEDERHERSTELLUNG

```
Umfangs-Makro-F1:
```

Der Kandidatenschwellenwert 0.906 wurde überschritten. Da jedoch die Auslassungs-Seite des Umfangs- & Grenz-Systems fehlgeschlagen ist, ist dieser Wert allein nicht ausreichend. Das System ist stark, wenn Umfangsansprüche klar formuliert sind:

- „Gleicher Preis in allen Ländern.“
- „Die Muttergesellschaft ist in allen Märkten lizenziert.“

Die Schwierigkeit ist der Bereichsfehler in:

- Empfehlung,

- kurzer Name,
- Pronomen,
- Auslassung

ist in Fällen gestiegen, in denen es verborgen ist.

24. ZURÜCKGEWINNUNG VON ZURECHNUNG UND MODALITÄT

```
Attributions-Makro-F1:
```

0.918

```
Modalität Makro-F1:
```

0.934 Stark getrennte Fälle:

- „Das Unternehmen sagt."
- „Das offizielle Register zeigt."
- „Nutzer behaupten."
- „Es gibt eine endgültige Entscheidung."
- „Wahrscheinlich."
- „Unbestätigt."

Schwache verbleibende Zuschreibungsfälle:

- die Meinung eines Mitarbeiters als Universitätsbefürwortung darzustellen,
- die Berücksichtigung von gesponserten redaktionellen Inhalten als unabhängige Forschung,

Mehrere abgeleitete URLs als Konsens darzustellen. Diese Fälle sind mit Schwächen in der Zitationszuordnung verbunden.

25. KOMPONENTENWIEDERHERSTELLUNG

Wiederherstellungswerte der Generatorwahrheit von zehn Komponenten und NOMOS:

Komponente	Generator	Wiederherstellung	Absoluter Fehler
ENT	96.5	96.1	0.4
FAC	91.8	91.0	0.8
SBI	84.6	82.0	2.6
TLA	88.7	88.0	0.7
ECI	82.1	78.7	3.4
EPI	87.9	85.8	2.1
TCR	87.0	88.2	1.2
LGF	82.4	84.8	2.4
USI	85.2	86.9	1.7
STR	87.0	90.7	3.7
Gesamt NOMOS	873.2	872.2	1.0

```
Komponenten-MAE:

MAE_component= 1.9
```

Kandidaten-Schwelle:

```
≤2.5
```

Ergebnis:

KANDIDATEN-SCHWELLE ERREICHT

Diese Tabelle enthält jedoch auch eine wichtige Warnung. ECI wird um 3.4 Punkte zu niedrig berechnet, SBI um 2.6 Punkte zu niedrig, STR um 3.7 Punkte zu hoch, LGF um 2.4 Punkte zu hoch. Einige dieser Fehler haben sich gegenseitig aufgehoben, und die zusammengesetzte Punktzahl wich nur um einen Punkt ab. Daher:

> *Das starke Erscheinungsbild der zusammengesetzten Wiederherstellung beseitigt nicht die Meldepflicht für die Komponentenerholung.*

26. FEHLINTERPRETATION DER ZUSAMMENGESETZTEN WIEDERHERSTELLUNG

Dieser Satz kann nicht allein verwendet werden: „NOMOS hat die im Generator definierte synthetische zusammengesetzte Punktzahl um einen Punkt Unterschied wiederhergestellt.“ Der korrekte Satz lautet:

> *NOMOS konnte den Gesamtscore um einen Punkt Unterschied wiederherstellen; es traten jedoch größere Richtungsfehler in den Komponenten ECI, SBI, LGF und STR auf, die sich teilweise gegenseitig ausgleichen.*

Die Aufgabe eines Standards besteht nicht nur darin, die korrekte Summe zu finden. Es geht darum, korrekt zu erklären, warum die Summe zustande kommt.

27. SPRACH- UND GEOGRAFISCHE FAIRNESS-WIEDERHERSTELLUNG

Synthetische Sprach- und Länderbasen wurden zuvor innerhalb von Generator Truth bestimmt. Absoluter Wiederherstellungsfehler der LGF-Komponente: 2.4 Punkte. Kandidatenschwelle:

```
≤3
```

Ergebnis:

KANDIDATEN-SCHWELLE ERREICHT

27.1. Untermetriken

Fairness-Metrik	Generator	Wiederherstellung	Fehler
Sprachbasis Q1 0	68.0	69.7	1.7
Sprachliche Q9 0–Q1 0 Differenz	25.0	22.8	2.2
Geografische Q1 0 Basis	72.0	71.1	0.9
Geografische Differenz	18.0	16.4	1.6
Abdeckungsfaktor	0.94	0.95	0.01
LGF	82.4	84.8	2.4

Wiederherstellung:

- die niedrigste Sprachleistung leicht überschätzt,
- die Ungleichheit leicht unterschätzt

Dies hat den Fairness-Status etwas besser gezeigt, als er tatsächlich ist.

28. UNTERSCHEIDUNG ZWISCHEN PROMPT-MANGEL UND KI-SPRACHMANGEL

Innerhalb des Sprachenäquivalenz-Anhangs:

- 500 tatsächlicher KI-Profil-Sprachfehler,
- 500 Prompt-Äquivalenz-Fehler

Ein Fall wurde gefunden. Korrekte Ursachenklassifikation:

```
938/1,000 = 0.938
```

Es ist passiert. Von den 62 falsch klassifizierten Fällen:

- Fehler im Prompt im 39 ten KI-Produkt,
- KI-Sprachfehler im Prompt auf dem 23 rd

wurde hochgeladen. Dieser Fehler hat die LGF direkt beeinflusst. Die Fairness-Komponenten-Kandidatin hat die Schwelle überschritten. Allerdings:

> *Die Klassifikation der kausalen Verantwortung hat noch nicht das Niveau der externen Sicherheit erreicht.*

29. CLEAN–NATURAL PANEL WIEDERHERSTELLUNG

Die durchschnittliche Panel-Differenz von Generator Truth in den 1,000 Clean–Natural Übereinstimmungen innerhalb des Diagnoseanhangs:

G_{N−C}^G = −8.4

Punkte. NOMOS Wiederherstellung:

G_{N−C}^N = −7.9

Punkte. Absoluter Unterschied: 0.5 Punkte.

```
Individuelle Profil-Panel-Lücke MAE:
```

ist 1.6 Punkte.

29.1. Personalisierungsdrift

Es gibt 173 bedeutende Fälle von Personalisierungsdrift in Generator Truth. Gefunden: 166 Rückruf:

```
166/173 = 0.9595
```

Von den sieben verpassten Fällen:

- vier sind Grenzauslassungen,
- zwei sind einheitsfreundliche Anweisungen.
- eins ist eingeschränkter Datenleckage

Fall. Bei der endgültigen Datenschutzprüfung wurde eine eingeschränkte Datenleckage festgestellt, die ein Kritisches Gate öffnete. Dieses Ergebnis zeigt erneut die Bedeutung einer mehrschichtigen Beurteilung.

30. WELLEN- UND STABILITÄTSWIEDERHERSTELLUNG

In drei synthetischen Wellen wurden folgende Änderungen eingeführt: In W2 änderten sich der Web- und Zitiermodus eines Produkts. In W3 trat ein synthetisches externes Ereignis auf. SYNTH-AI-10 erzeugte ein Kritisches Ereignis in W2 und eine Behebung in W3. SYNTH-AI-03 hat in der Leistung bei ressourcenarmen Sprachen einen Rückschritt gemacht. SYNTH-AI-07 enthält veraltete Preis- und Produktabdeckung. NOMOS:

- teilte die Systemänderung innerhalb der Welle korrekt in zwei Unterwellen auf,
- die Truth Pack-Versionen vor und nach dem externen Ereignis getrennt,
- die aktuelle Situation nach der Korrektur aktualisiert,

das historische Kritische Ereignis bewahrt.

30.1. STR Überschätzung

Generator STR: 87.0 Wiederherstellung STR: 90.7 Fehler: +3.7 Punkte. Grund: einige intra-Wellen-Änderungen wurden als explizite Versionsänderungen getrennt, anstatt als systemische Drift, und die Bewertungsformel belohnt die Stabilität innerhalb getrennter Sub-Wellen übermäßig. Dieses Ergebnis zeigt, dass die STR-Formel nach der Änderung nicht ausreichend zwischen Stabilität und ununterbrochener Systemstabilität unterscheidet.

31. SELTENER KRITISCHER EREIGNISTEST

Vier separate synthetische Szenarien seltener Ereignisse wurden berechnet; dies sind keine Felder- oder Softwareexperimente. Die folgenden Ergebnisse sind nur eine Demonstration der Methode:

Standardmäßige Anzahl kritischer Ereignisse	Ungefähre gewichtete Rate / 1,000	Einseitiges 95% oberes Konfidenzlimit / 1,000	Verhalten bei Ereignisüberprüfung
0	0.00	2,991	Null-Risiko-Anspruch nicht festgestellt
1	1.08	4,735	CIS-2 Datensatz geöffnet
2	2.15	6,282	Prüfung durch Überprüfer
5	5.38	10,484	Wiederholung auf Zellebene

Hinweis: Obere Grenzen sind die einseitigen Clopper–Pearson 95% oberen Konfidenzgrenzen unter der Annahme von n=1,000 unabhängigen Bernoulli-Beobachtungen. Kann nicht direkt auf gewichtete oder komplexe Stichprobendesigns [K09] angewendet werden.

Im Szenario mit null Ereignissen sagte das System nicht: "Kritisches Risiko ist null." Korrekte Ausgabe:

```
„Beobachtetes bestätigtes Kritisches event= 0; unter der Annahme von n=1,000 unabhängigen
Beobachtungen beträgt die einseitige Clopper–Pearson 95% obere Konfidenzgrenze 2,991/1,000.“
```

ist aufgetreten. Das Verhalten des Kandidaten-Ereignisses mit seltener Wahrscheinlichkeit hat die definierten synthetischen Schwellenwerte erreicht.

32. KONFIDENZINTERVALL ABDECKUNGSRATE

Prozess der Generierung synthetischer Bevölkerung: Ein Beispielverfahren, das über 2,000 hypothetische Neuberechnungen berechnet wurde, wurde vorgestellt. Berechnete Deckungsraten nominaler 95% Intervalle:

Metrik	Empirische Deckung
Strenger Bestehen	94.6%
Akzeptables Bestehen	95.0%
NOMOS zusammengesetzt	94.1%
LGF	93.7%
Obergrenze der kritischen Rate	95.2%
FAC	94.8%
SBI	93.5%

Akzeptanzbereich des Kandidaten: Definiert als 93%–97%. Alle wichtigen Deckungsraten liegen im Akzeptanzbereich des Kandidaten. Ergebnis:

KANDIDATEN-SCHWELLE ERREICHT

33. METAMORPHISCHE TESTS

Insgesamt: 3,200 metamorphe Antwortpaare wurden verwendet.

33.1. Bedeutungsbewahrende Transformationen

Satzumstellung paraphrasieren Sicheres Format Zitierposition, solange der Link unverändert bleibt Äquivalentes Entitätsalias Erwartet: Peer-Review- und Antwortstatus sollten sich nicht ändern. Genaues invariantes Ergebnis:

33.2. Bedeutungsverändernde Transformationen

Entfernen Sie die Zuschreibung Erhöhen Sie die Gewissheit Übertragen auf falsche Entität Ändern Sie die aktuelle Zeit von historisch Entfernen Sie die Begrenzungssatz Fügen Sie die Behauptung einer unabhängigen Quelle hinzu Erwartet: Die relevante Entscheidungsdimension sollte sich ändern. Korrekte Sensitivität:

33.3. Große metamorphe Verwerfungen

Die Übertragung des Zitats vom Satz ans Ende des Absatzes, inkonsistente Bewertung der Modalität von ‚Es scheint' und ‚ist wahrscheinlich', Interpretation des Marken-Alias als juristische Person, der Grenzsatz befindet sich in einem anderen Absatz – diese Ergebnisse stimmen mit den Ergebnissen von Zitat und Weglassung überein.

34. ERGEBNISSE DER GEGENPRÜFUNG

Gegenuntersuchungsfall	Korrektes NOMOS-Verhalten
Sehr flüssig, aber sachlich falsch	98.8%

Gegenuntersuchungsfall	Korrektes NOMOS-Verhalten
Sehr korrekt, nur ein kritischer Punkt im Atom	100%-Antwort automatische Ablehnung
Eine kurze, aber ausreichende Antwort	99.2% korrekt bestanden
Allgemeiner Haftungsausschluss und konkreter Fehler	98.7% falsch erhalten
Evidenzwäscherei	91.8% korrekt
falsche Entität Lizenzübertragung	99.4% korrekt
Empfehlungsgrenze überschritten	82.9% korrekt
Gefälschte Universitätsanerkennung	96.1% korrekt
Richtige Zahl, falsche Unabhängigkeit	92.0% korrekt
Angemessene Unsicherheit	97.3% korrekt

Die schwächste Gegenprüfungsklasse:

Empfehlungsgrenze überschritten

ist gewesen.

```
Dieses Ergebnis bestätigt den Materialauslassungs-F1-Fehler erneut.
```

35. METHODENFEHLER VOM TEST GEFUNDEN

Am Ende der synthetischen Demonstration wurden fünf Methodenbefunde für NOMOS 0.9 aufgezeichnet.

METHODEN-BEFUND 01

AMBIGUITÄT DER ZITATANHÄNGE

```
Schweregrad: Hoch / Betroffene Komponente: ECI / Ergebnis: Kandidaten-Zitat F1-Schwelle nicht
erreicht.
```

Die Regel zur Bestimmung, welche Atom-Absatz-Zitate unterstützt, ist nicht präzise genug.

METHODEN-BEFUND 02

UNTERERKENNUNG VON MATERIALGRENZENWEGLASSUNGEN

```
Bedeutung: Hoch / Betroffene Komponente: SBI, TCR, Kritisches Tor / Ergebnis: Materialweglassungs-
F1-Schwelle nicht erreicht.
```

Die Grenzen, die nicht ausdrücklich in Empfehlungs- und Beziehungssätzen angegeben sind, aber die Entscheidung des Nutzers beeinflussen, wurden nicht ausreichend erfasst.

METHODEN-BEFUND 03

PROMPT-FEHLER / KI-FEHLER ZUTEILUNGSFEHLER

Bedeutung: Mittel / Betroffene Komponente: LGF / Ergebnis: Obwohl der Fairness-Score bestanden wurde, ist die Verantwortungszuweisung fehlerhaft.

METHODEN-BEFUND 04

STABILITÄTS-ÜBERBELONUNG NACH TEILWELLE-SPALTUNG

Bedeutung: Mittel / Betroffene Komponente: STR / Ergebnis: STR wurde 3.7 Punkte höher als die Generator-Truth berechnet.

METHODEN-BEFUND 05

ZUGÄNGLICHKEITS-ALTERNATIVE UNTERAKZEPTANZ

Bedeutung: Mittel / Betroffene Ebene: Erfassungsintegrität / Ergebnis: Acht gültige Zugänglichkeitserfassungsfälle wurden auf ein unzureichendes Evidenzniveau reduziert.

36. WARUM IST DIES NICHT DEMONSTRATION BQ-3?

Für BQ-3— Versiegelte Validierung bestanden:

Alle obligatorischen Kandidatenschwellen müssen erfüllt sein, und es darf keine Null-Toleranz-Testverletzungen geben. Zwei materielle Schwellenwerte wurden nicht erreicht:

Kriterium	Ergebnis	Schwellenwert
F1 der Zitationszuordnung	0.887	≥ 0.90
Materialauslassungs-F1	0.824	≥ 0.85

Außerdem:

- eine unabhängige Teamreproduktion wurde nicht durchgeführt,
- das reale Korpus und der Code wurden nicht veröffentlicht.

Unberührte Erneuerungsrückstellung wurde nicht ausgeführt. Daher lautet die korrekte synthetische Demonstrationsaussage:

DEMO-BQ-2— PROZESSLINIEN-TROCKENLAUF ABGESCHLOSSEN; MATERIALKORREKTUR ERFORDERLICH

Es sollte so sein. Der tatsächliche Teststatus ist:

BQ-0— NICHT AUSGEFÜHRT

bleiben wie sie sind.

37. WARUM WURDEN DIE SCHWELLEN NICHT GESENKT?

Referenz-Mapping-Ergebnis: 0.887 Kandidatenschwelle: 0.90 Unterschied ist gering. Hätten wir die Schwelle auf 0.885 gesetzt, hätte die Methode bestanden. Materialauslassungs-Ergebnis: 0.824 Kandidatenschwelle: 0.85 Hätten wir die Schwelle auf 0.82 gesetzt, hätte auch dieses Feld bestanden. Diese Änderungen:

- wären jedoch NACH der Einsicht der Ergebnisse vorgenommen worden.
- um sicherzustellen, dass NOMOS besteht

Dies ist der grundlegendste Verstoß gegen den Test. Das korrekte Verhalten ist:

- Schwelle beibehalten
- Veröffentlichung fehlgeschlagen
- Methode korrigieren
- Neue Version erstellen
- Erneut testen an unberührtem Rückhalt

muss sein.

38. METHODE 1.0 KORREKTURPLAN

38.1. Zitationszuordner 1.0

Neues System:

- die Quellspanne der Zitation,
- das Anspruchs-Anhang-Feld,
- der Bereich innerhalb des Absatzes,
- der Status nur mit Attribution,
- die Herkunftslinienfamilie

wird obligatorisch sein. Für jede Zitation:

> *Welche genauen Atome unterstützt diese Quelle?*

Die Frage wird zu einem offenen Datensatz gemacht.

38.2. Grenzwegfall-Codebuch 1.0

Für jede Empfehlung und Transaktionsaufforderung:

- obligatorische Grenzen, die die Entscheidung des Nutzers ändern,
- kontrafaktischer Entscheidungstest,
- Entität–Dienst–Jurisdiktions-Matrix

wird vordefiniert sein. Kontrafaktischer Test:

> *Wenn diese Grenze dem Nutzer mitgeteilt worden wäre, würde sich die Empfehlung oder die Transaktionsentscheidung eines vernünftigen Nutzers ändern?*

Wenn ja, wird das Versäumnis zumindest einer sachlichen Überprüfung unterzogen.

38.3. Prompt–KI-Verantwortlichkeitsprüfer

Sprachabfall:

- Prompt-Äquivalenz,
- KI-Produktverhalten,
- erfassen,
- Nutzergebiet

Es wird ein separater Entscheidungsvektor erstellt, um zu bestimmen, zu welcher der Quellen es gehört.

38.4. STR-Formelüberarbeitung

Neue STR:

- Stabilität innerhalb der Unterwelle,

- Systemversionsänderung,
- Stabilität nach Korrektur,
- ununterbrochene Zeitreihe

werden die Felder in einzelne Komponenten aufteilen.

38.5. Barrierefreiheitsnachweisäquivalenz

Nicht-visuell, aber mit gleichwertigem Nachweis:

- Transkript,
- Barrierefreiheitsbaum,
- Protokoll der unterstützenden Technologien,
- Audio-Sitzungsaufzeichnung

eine offene Äquivalenzmatrix wird erstellt für.

39. WIEDERHOLUNGSPRÜFUNGSREGEL

Methode 1.0: kann nicht wiederholt am selben versiegelten Validierungskorpus abgestimmt werden. Zwei Tests sind erforderlich:

A. Regressionssatz

Enthält die fehlgeschlagenen Fälle von Methode 0.9. Zweck: zu sehen, dass die Korrektur tatsächlich funktioniert.

B. Unberührtes Verlängerungs-Holdout

Enthält Fälle, die beim Entwerfen der Methode 1.0 niemals gesehen wurden. Zweck: zu prüfen, dass keine Überanpassung an den alten Test erfolgt. Nur BQ-3:

- Regressionserfolg,
- Erfolg des unberührten Holdouts

kann gegeben werden, wenn zusammen erhalten.

40. ENTSCHEIDUNG ZUR ÜBERGANG IN DIE REALE WELT

Basierend auf dieser synthetischen Demonstration, die richtige Entscheidung für NOMOS 0.9:

> *Nicht bereit, echten Unternehmen eine öffentlich verfügbare Eignungspunktzahl oder NOMOS 950+ Kennzeichnung zu gewähren.*

Gründe:

- Referenzzuordnungskandidat unterhalb des Schwellenwerts
- Materialauslassungserkennungskandidat unterhalb des Schwellenwerts
- Keine unabhängige Reproduktion
- Reale Tests nicht durchgeführt
- BQ-4-Niveau nicht erreicht

Die Methode jedoch:

- Verteilungswiederherstellung,

- Kritisches Tor,
- Antwortstatus,
- Komponentenwiederherstellung,
- Konfidenzintervall
- Seltenes Ereignisverhalten

zeigt eine starke Grundlage in Bezug auf Wartung. Korrekt Status:

> *Bereit für Forschung und Pilotnutzung; noch nicht bereit für öffentliche Norm und Konformitätszeichen.*

41. SYNTHETISCHE TESTAKZEPTANZKARTE

Kriterium	Ergebnis	Bereitstellung
F1-Wert der Aussagegrenze	0.964	Bestanden
Genauigkeit der Entität	0.987	Bestanden
Faktische Makro-F1	0.923	Bestanden
Umfang Makro-F1	0.906	Bestanden
Attributions Makro-F1	0.918	Bestanden
F1 der Zitationszuordnung	0.887	Nicht bestanden
Materialauslassungs-F1	0.824	Nicht bestanden
Endgültiger kritischer Rückruf	1,000	Bestanden
Endgültiger kritischer Falsch-Positiv	0	Bestanden
Wichtige Makro-F1	0.952	Bestanden
RP Makro-F1	0.928	Bestanden
Komponenten-MAE	1.9	Bestanden
Zusammengesetzter Fehler	1/1,000	Bestanden
Fairness-Fehler	2.4	Bestanden
Erfassungsgenauigkeit	0.982	Bestanden
CI-Abdeckung	Innerhalb 92%–98%	Bestanden
Tor-Inversion	0	Bestanden
Unabhängige Reproduktion	Keines	Unvollständig
Tatsächliche Korpusausführung	Keines	Unvollständig

42. GEBRÄUCHLICHE REGEL DER DEMONSTRATION

Dieser synthetische Lauf gibt gleichzeitig drei Aussagen über NOMOS 0.9.

42.1. BEREICHE, IN DENEN NOMOS STARK IST

Es stellte die Verteilung des Antwortstatus mit hoher Genauigkeit wieder her. Die abschließende Governance-Kette verpasste keine kritischen Gates und verschleierte kein kritisches Ereignis innerhalb eines hohen Durchschnitts. Sie unterschied Fehler bei falschen Entitäten, zeitliche Fehler und sachbezogene Unterstützungsfehler effektiv. Sie verwandelte Nicht bewertbar-Datensätze nicht in Produktfehler, schloss aus null beobachteten kritischen Ereignissen nicht auf null Risiko, bewahrte

die Unterscheidung zwischen historischen Vorfällen und aktueller Korrektur und stellte die Zusammensetzungsbewertung mit niedrigem absoluten Fehler wieder her.

42.2. BEREICHE, IN DENEN NOMOS SCHWACH IST

Es konnte nicht ausreichend bestimmen, welche Behauptung die Zitation unterstützt. Es konnte die in der Empfehlung verborgenen wesentlichen Auslassungen nicht angemessen erfassen. Es hat nicht immer korrekt zwischen Mängeln im Prompt und Fehlern der KI-Sprache unterschieden. Nach der Unterscheidung der Unterwelle hat es die Stabilität leicht überbewertet. Es hat einige Belege zur Zugänglichkeit als schwächer angesehen, als sie tatsächlich sind.

42.3. DINGE, DIE NOMOS NOCH NICHT BEWIESEN HAT

Es wurde noch nicht nachgewiesen, dass es bei realen Nutzern funktioniert, dass es auf echte KI-Produkte angewendet werden kann, dass es echte mehrsprachige Prompt-Äquivalenz besitzt, dass unabhängige Universitäten dasselbe Ergebnis erzielen werden, dass die langfristige Governance auf Interessenkonfliktüberlegungen beruht, oder dass das öffentliche Konformitätszeichen rechtlich mit Vertrauen vergeben werden kann.

43. VERPFLICHTENDE NORMATIVE BESTIMMUNGEN

CH18-N01

Eine synthetische Demonstration im Buch kann nicht wie ein echter öffentlicher Testlauf präsentiert werden.

CH18-N02

Der Status der synthetischen Demonstration und der Status des echten Tests BQ müssen getrennt gehalten werden.

CH18-N03

Der Test muss den Status BQ-0— Nicht ausgeführt beibehalten, bis ein echtes Korpus und ein unabhängiger Lauf durchgeführt wurden.

CH18-N04

Ergebnisse von Demonstrationen können nicht auf die Leistung eines echten Apple oder eines echten KI-Anbieters zurückgeführt werden.

CH18-N05

Die Hauptantworten 30,000 sowie die Korpora für Diagnose, Wichtigkeit, Erfassung und Gold sollten in getrennten Behältern aufbewahrt werden.

CH18-N06

Herausforderungsfälle können die Verteilung der Principal Prevalence nicht ändern.

CH18-N07

Die Generator-Truth-Verteilung muss vor den Wiederherstellungsergebnissen festgeschrieben werden.

CH18-N08

Bevor der Generator Truth geöffnet wird, müssen NOMOS Antwort- und Punktzahl-Ergebnisse festgeschrieben werden.

CH18-N09

Die Wiederherstellung kann nicht ausschließlich anhand der zusammengesetzten Punktzahl bewertet werden.

CH18-N10

Aussagegrenze, Entität, faktisch, Umfang, Zeit, Zuschreibung, Zitat, Auslassung, Wichtigkeitsstufe, Antwort, Komponente und zusammengesetzte Wiederherstellung sollten separat veröffentlicht werden.

CH18-N11

Fehler bei zusammengesetzten Punktzahlen können Komponentenfehler nicht verbergen, nur weil sie klein sind.

CH18-N12

Komponentenfehler, die sich arithmetisch gegenseitig aufheben, können nicht als erfolgreiche kausale Wiederherstellung gezählt werden.

CH18-N13

Die Wiederherstellungsverteilung von RP-1 zu RP-8 sollte explizit mit der Generator-Truth-Verteilung verglichen werden.

CH18-N14

Strenge Bestehens- und akzeptable Bestehens-Wiederherstellungen sollten separat gezeigt werden.

CH18-N15

Die Anzahl kritischer Antworten und die kritische Gate-Wiederherstellung sollten separat gemeldet werden.

CH18-N16

Der anfängliche Beurteiler der kritischen Rückrufquote und der endgültig beurteilte kritische Rückruf sollten getrennt gehalten werden.

CH18-N17

Die Abhängigkeit des endgültigen kritischen Rückrufs von Governance-Ebenen sollte sichtbar sein.

CH18-N18

Dasselbe Ergebnis kann nicht angenommen werden, wenn ein zweiter Beurteiler oder eine Expertenprüfung durchgeführt wird.

CH18-N19

Die kritische Falsch-Positiv-Rate sollte zusammen mit dem Rückruf veröffentlicht werden.

CH18-N20

Negative Kontrollen, die dem Kritischen ähneln, sollten ein obligatorischer Bestandteil des endgültigen Bewertungssystems für Wichtigkeit sein.

CH18-N21

```
Wenn der F1-Kandidaten-Schwellenwert für Referenzabgleiche unterschritten wird, kann der Test nicht
als erfolgreich erklärt werden.
```

CH18-N22

```
Wenn der Materialauslass F1-Kandidat unter dem Schwellenwert liegt, kann nicht erklärt werden, dass
der Test erfolgreich ist.
```

CH18-N23

Kandidaten-Schwellenwerte können nach Bekanntwerden der Wiederherstellungsergebnisse nicht gesenkt werden.

CH18-N24

Schwellenwertänderungen erfordern eine neue Testmethodenversion und eine neue unangetastete Validierung.

CH18-N25

Ein kleiner Schwellenwertunterschied kann keine Rechtfertigung sein, einen Fehler zu verschleiern.

CH18-N26

Der Materialgrenzen-Auslass in der Empfehlung sollte genauso prüfbar sein wie eine offensichtliche falsche Behauptung.

CH18-N27

Der Ort der Referenz innerhalb des Absatzes kann nicht automatisch eine vollständige Absatzunterstützung erzeugen.

CH18-N28

Die Wiederherstellung kann nicht für Attribution-only-Zitate oder Zitate zur Faktenprüfung gezählt werden.

CH18-N29

Fehler in der Quelllinienführung sollten auch bei der Wiederherstellung der Zitationszuordnung erscheinen.

CH18-N30

Die Erfassungsgenauigkeit sollte unabhängig von der Korrektheit der semantischen Antwort berechnet werden.

CH18-N31

Eine erfundene Erfassung, die die richtige semantische Antwort trägt, macht die Erfassung nicht gültig.

CH18-N32

Eine gültige, aber falsche Antwort kann nicht als Erfassungsfehler gezählt werden.

CH18-N33

Der Standard für alternative Zugänglichkeitsnachweise muss nicht die gleiche Form wie die visuelle Darstellung haben.

CH18-N34

Wenn gültige Zugänglichkeitsnachweise unnötig auf ein niedrigeres Niveau reduziert werden, sollte eine Methodenfeststellung eröffnet werden.

CH18-N35

Nicht bewertbar muss in der Hauptantwortverteilung sichtbar bleiben.

CH18-N36

Nicht bewertbar kann nicht aus dem Nenner entfernt werden, um die Beobachtungsrückgewinnung zu erhöhen.

CH18-N37

Prompt-Fehler und KI-Sprachfehler sollten separate Fehlerursachenklassen haben.

CH18-N38

Die Fairness-Rückgewinnung kann nicht allein mit dem LGF-Kompositunterschied bewertet werden.

CH18-N39

Sprachgrundlage, Unterschiede, Abdeckung und Verantwortungszuweisung müssen ebenfalls veröffentlicht werden.

CH18-N40

Die Rückgewinnung des Clean–Natural-Panels sollte keine kausale Personalisierungsregel erzeugen.

CH18-N41

Personalisierungsdrift-Fälle müssen separate Rückruf- und Falsch-Positiv-Metriken haben.

CH18-N42

Datenlecks bei eingeschränkten Daten können nicht als natürliche Personalisierungsfeststellung reduziert werden.

CH18-N43

Die intra-wave Systemänderung muss eine korrekte Subwellenstruktur erzeugen.

CH18-N44

Wenn die Subwellentrennung künstlich den Stabilitätsscore erhöht, muss die Formelfindung offengelegt werden.

CH18-N45

Die aktuelle Korrektur kann das historische Kritische Ereignis nicht löschen.

CH18-N46

Seltene-Ereignis-Tests sollten Null-, Ein- und Mehrfachereignis-Szenarien einschließen.

CH18-N47

Ein beobachtetes Nullkritisches Szenario kann kein Nullrisiko-Ergebnis erzeugen.

CH18-N48

Ein einzelnes Kritisches Ereignis kann allein nicht eine 100 prozentuale systemische Ausfallrate etablieren.

CH18-N49

Die Methode des Konfidenzintervalls sollte mit empirischer Abdeckung in wiederholten synthetischen Populationen getestet werden.

CH18-N50

Die tatsächlichen Abdeckungsergebnisse der nominalen 95-Prozentintervalle sollten veröffentlicht werden.

CH18-N51

Metamorphe Invarianztests und erwartungssensitive Tests sollten getrennt berechnet werden.

CH18-N52

Eine bedeutungserhaltende Paraphrase sollte das Ergebnis der Überprüfung nicht unnötig verändern.

CH18-N53

Die Transformation, die die Bedeutung von Zuschreibung, Entität, Zeit oder Grenze verändert, sollte eine entsprechende Entscheidungsänderung bewirken.

CH18-N54

Gegenprüfungen sollten ausschließlich kritisch, in Bezug auf Evidenzwäsche und Auslassungen flüssig abdecken.

CH18-N55

Die schwächsten gegnerischen Testklassen müssen auf der öffentlichen Testkarte sichtbar sein.

CH18-N56

Methodenfeststellungen sollten in separate, versionierte Aufzeichnungen aus den Testergebnissen umgewandelt werden.

CH18-N57

Die Bedeutung einer Methodenfeststellung sollte entsprechend dem tatsächlichen Nutzer- und Audit-Einfluss angegeben werden.

CH18-N58

Referenz- und Auslassungsfehler können nicht nur aufgrund eines kleinen metrischen Unterschieds auf Beratung herabgestuft werden.

CH18-N59

Das Testqualitätsniveau sollte gemäß allen obligatorischen Schwellenwerten und Unabhängigkeitsbedingungen zugewiesen werden.

CH18-N60

Wenn zwei wesentliche Schwellenwerte fehlschlagen, kann DEMO-BQ-3 nicht vergeben werden.

CH18-N61

BQ-4 kann ohne unabhängige Reproduktion nicht vergeben werden.

CH18-N62

Das In-Book-Demonstrationsergebnis BQ kann ohne die tatsächliche Durchführung des Versuchs nicht als tatsächlicher Status betrachtet werden.

CH18-N63

Methodenkorrektur kann das alte Demonstrationsergebnis nicht ändern.

CH18-N64

Methode 1.0 muss die neue Punktzahl und Codebuchversion tragen.

CH18-N65

Nur das Regressionsset kann die neue Methode nicht validieren.

CH18-N66

Unberührtes Renewal-Holdout muss für die neue Methode obligatorisch sein.

CH18-N67

Die neue Methode kann nicht immer wieder an denselben Validierungskorpus angepasst werden.

CH18-N68

Eine öffentliche Live-Prüfung kann nicht gestartet werden, ohne die minimale Testqualität und die Bedingung der unabhängigen Reproduzierbarkeit zu erfüllen.

CH18-N69

Das Konformitätszeichen kann nicht auf der Grundlage eines erfolgreichen synthetischen Demonstrationsergebnisses vergeben werden.

CH18-N70

Fehlgeschlagene Testbereiche sollten der Öffentlichkeit offengelegt werden.

CH18-N71

Schwellenwerte, die allein überschritten werden, dürfen nicht veröffentlicht werden.

CH18-N72

Der Erfolg einer zusammengesetzten Wiederherstellung kann das Ergebnis eines fehlgeschlagenen Vergleichs oder einer Auslassung nicht auslöschen.

CH18-N73

Die Stärken und Schwächen der Demonstration sollten auf derselben öffentlichen Karte erscheinen.

CH18-N74

Aussagen, die NOMOS noch nicht bewiesen hat, sollten klar aufgeführt werden.

CH18-N75

Synthetische Zahlen im Buch können nicht in Zitate oder öffentliche Aussagen umgewandelt werden, als ob sie echte ausgeführte Daten wären.

CH18-N76

Synthetische Demonstrationsdaten müssen ebenfalls ein synthetisches Label in maschinenlesbarer Form tragen.

CH18-N77

Der Demonstrationskorpus darf nicht im Internet als echte Unternehmensinformationen indexiert werden.

CH18-N78

Wenn der echte Test beginnt, sollten neue Identität und neue Version verwendet werden.

CH18-N79

Alle Wiederherstellungskonten erfordern reproduzierbaren Code und Manifest.

CH18-N80

Jeder Testlauf, jede Wiederherstellung, Methodensuche, Korrektur und Veröffentlichungsentscheidung muss einen verantwortlichen menschlichen oder institutionellen Eigentümer haben.

44. FORMEN DES VERSAGENS

CH18-F01— ZÄHLUNG DER DEMONSTRATION ALS ECHTLER LAUF

Im Buch angegebene Zahlen werden als öffentliche Testergebnisse präsentiert.

CH18-F02— ZUSAMMENFÜHRUNG VON BQ-0 MIT DEMO-BQ-2

Es wird gesagt, dass der Test ohne echte Ausführung validiert ist.

CH18-F03— VEREHRUNG VON KOMPOSITPUNKTEN

Ein Kompositfehler von einem Punkt deckt sämtliche zugrunde liegenden Fehler ab.

CH18-F04— BERÜCKSICHTIGUNG DER KOMPONENTENSTORNIERUNG ALS ERFOLG

Hohe und niedrige Richtungsfehler heben sich gegenseitig auf.

CH18-F05— NUR STRIKTE PASS-WIEDERHERSTELLUNG VERÖFFENTLICHEN

RP-Klassenverwechslungen werden verborgen.

CH18-F06— ZÄHLUNG KRITISCHER ENDRECALLS ALS ERFOLG DES ERSTEN BERICHTERSTATTERS

Die Wirkung von Experten- und Senior-Überprüfungen wird unsichtbar.

CH18-F07— MEINEN ZWEITEN BERICHTERSTATTER FÜR UNNÖTIG ERACHTEN

Die verpassten sieben kritischen Erstversuche werden verborgen.

CH18-F08— FALSCHE KRITISCHE IGNORIEREN

Das extrem wichtige System wird als Erfolg dargestellt.

CH18-F09— SENKUNG DER ZITIERUNGSSCHWELLE

Die Schwelle wird so geändert, dass das Ergebnis von 0.887 bestanden wird.

CH18-F10— SENKUNG DER WEGLASSUNGSSCHWELLE

Die Testregel wird so geändert, dass das Ergebnis von 0.824 bestanden wird.

CH18-F11— BESTEHEN DURCH DIE ANSAGE „KNAPP NICHT BESTANDEN“

Es verliert seine zuvor festgeschriebene Schwellenbindung.

CH18-F12— VERTEILUNG DER ZUSCHREIBUNG AUF DEN ABSATZ

Eine einzelne Quelle wird als Unterstützung für alle Behauptungen betrachtet.

CH18-F13— EIN ATTRIBUTIONSQUELLE-ALLES ALS TATSACHE ZÄHLEN

Was das Unternehmen sagt, wird zu einer unabhängigen Tatsache.

CH18-F14— VERERBUNG DER QUELLEN IGNORIEREN

Abgeleitete URLs werden zu unabhängigen Belegen.

CH18-F15— VERSCHWEIGUNG ALS OPTIONALES DETAIL BEZEICHNEN

Die Grenze, die die Entscheidung des Nutzers verändert, wird verringert.

CH18-F16— JEDES WEGGELASSENE DETAIL KRITISCH MACHEN

Optionale Details erzeugen ein übermäßiges Maß an Wichtigkeit.

CH18-F17— FEHLER IM PROMPT ALS KI-FEHLER BETRACHTEN

Die Verantwortung für Fairness wird fälschlicherweise zugewiesen.

CH18-F18— HOCHLADEN DES KI-FEHLERS IN DAS PROMPT

Tatsächliche Sprachkorruption wird auf das Messinstrument übertragen.

CH18-F19— BERÜCKSICHTIGUNG DES FAIRNESS-COMPOSITES ALS AUSREICHEND

Seine Basis und Disparitätswiederherstellung bleiben erhalten.

CH18-F20— IGNORIEREN DER STR-ÜBERSCHÄTZUNG

Unterwellen-Trennung verwandelt sich in einen Stabilitätsbonus.

CH18-F21— BEEINTRÄCHTIGUNG DER 'ZUGÄNGLICHKEIT' IN DER AUFNAHMEGENAUIGKEIT

Äquivalente alternative Nachweise werden als schwach betrachtet.

CH18-F22— ZÄHLEN VON 'Nicht bewertbar' ALS PRODUKTAUSFALL

Aufnahmefehler verwandelt sich in KI-Leistung.

CH18-F23— LÖSCHEN VON 'Nicht bewertbar'

Bewertbare Abdeckung steigt künstlich.

CH18-F24— METAMORPHISCHE FEHLER NICHT VERÖFFENTLICHEN

Es wird beibehalten, obwohl dieselbe Bedeutung zu einer anderen Entscheidung führt.

CH18-F25— DIE SCHWÄCHSTE KLASS IN DER GEGENÜBERLIEGENDEN PRÜFUNG VERSTECKEN

Infolge der Auslassung der Grenze wird es von der öffentlichen Karte entfernt.

CH18-F26— NUR ENDGÜLTIGE ENTSCHEIDUNGSMETRIK

Die Schwäche im Initialprozess wird unsichtbar.

CH18-F27— NUR ERSTE DURCHLÄUFMETRIK

Die korrigierende Kapazität der Governance-Kette wird unsichtbar.

CH18-F28— TESTEN OHNE METHODENFINDUNG

Fehlgeschlagene Bereiche bleiben nur als Zahlen.

CH18-F29— METHODENFINDUNG BERATEND MACHEN

Die Auswirkung auf den Materialwert wird reduziert.

CH18-F30— BEKANNTGEBEN VON BQ-3

Zwei obligatorische Schwellenwerte wurden nicht überschritten.

CH18-F31— BEKANNTGEBEN VON BQ-4

Es gibt keine unabhängige Reproduktion.

CH18-F32— ZÄHLEN VON SYNTHETISCHER DEMO ALS LIVE-VALIDIERUNG

Es gibt kein echtes Nutzer- oder KI-Produkt.

CH18-F33— ÜBERANPASSUNG AN DAS REGRESSIONSSET

Methode 1.0 merkt sich nur alte Fehler.

CH18-F34— KEINE VERWENDUNG EINES UNBERÜHRTEN HOLDOUTS

Die Generalisierung wird nicht getestet.

CH18-F35— STILLES ERSETZEN DES ALTEN DEMOS DURCH EINE NEUE FORMEL

Die Versionsgeschichte geht verloren.

CH18-F36— VERGEBUNG VON ECHTZEIT-BADGES

Das Berechtigungs-Badge wird vor BQ-4 erstellt.

CH18-F37— NUR ERFOLGREICHE ERGEBNISSE VERÖFFENTLICHEN

Fehler bei Zitierung und Auslassung werden versteckt.

CH18-F38— NUR PUBLISHING FEHLGESCHLAGEN

Starke kritische und Vertriebswiederherstellungsergebnisse sind ebenfalls verborgen.

CH18-F39— NULL KRISE ALS NULL RISIKO BETRACHTEN

Die Wiederherstellung seltener Ereignisse schlägt fehl.

CH18-F40— EINZELNE KRISE ALS GLOBALEN FEHLER BETRACHTEN

Häufigkeit und Umfang werden gestört.

CH18-F41— IGNORIEREN DER CI-ABDECKUNG

Das Konfidenzintervall wird nur zu einem mathematischen Erscheinungsbild.

CH18-F42— WIRKSAME STICHPROBE VERSTECKEN

Die Begrenzung seltener Ereignisse erscheint übermäßig genau.

CH18-F43— VERWENDUNG EINER SYNTHETISCHEN ZAHL IN DER ECHTEN QUELLE

Die Buchvorführung verwandelt sich in Daten aus der Außenwelt.

CH18-F44— ÜBERTRAGE DEN SYNTHETISCHEN BILDSCHIRM WIE EINE LIVE-ANTWORT

Testdarstellung führt zu Vergiftung.

CH18-F45— MIT ECHTEM APPLE VERBINDEN

APPLE-SYNTH-Unterscheidung wurde entfernt.

CH18-F46— MIT ECHTEM ANBIETER VERBINDEN

SYNTH-AI-Profil wird wie tatsächliche Produktleistung interpretiert.

CH18-F47— BERECHNUNGSCODE VERSTECKEN

Wiederherstellung kann nicht unabhängig getestet werden.

CH18-F48— WAHRHEITS-MANIFEST DES GENERATORS VERSTECKEN

Es ist unbekannt, ob die synthetische Wahrheit zuvor festgeschrieben war.

CH18-F49— EINZELNER BEURTEILER MIT METHODENINHABER

Interessenkonflikte werden unsichtbar.

CH18-F50— ÄNDERN DES SEEDS NACH FEHLER

Ein Korpus wird leichter ausgewählt.

CH18-F51— ENTFERNEN VON FÄLLEN NACH FEHLER

Schwierige Zitations- und Auslassungsbeispiele werden entfernt.

CH18-F52— RUNDUNGSSCHWELLEN NACH ERGEBNIS

Das Ergebnis 0.887 wird als 0.89 oder 0.90 dargestellt.

CH18-F53— FALSCHE PRÄZISION

Wiederherstellungswerte werden mit unnötigen Dezimalstellen genau dargestellt.

CH18-F54— VERBERGEN DER AUSWIRKUNGEN MENSCHLICHER REGIERUNG

Ein 100 Prozent-Kritik-Ergebnis wird lediglich als ein rechnerischer Erfolg dargestellt.

CH18-F55— DAS UNBEWIESENE FÜR BEWIESEN HALTEN

Es wird ein Urteil über reale Sprache, Nutzer- und Anbieterleistung gefällt.

CH18-F56— PILOTNUTZUNG ALS ÖFFENTLICHE NORM ZÄHLEN

Das Forschungstool wird direkt zu einem Zertifikat.

CH18-F57— LÖSCHEN HISTORISCHER DEMO-DATEN

Fehlschläge unter Methode 0.9 werden nach Methode 1.0 unsichtbar.

CH18-F58— ERFOLG ALS GESCHICHTSSCHREIBUNG ZÄHLEN

Autoritätsanspruch wird ohne methodenunabhängige Überprüfung erhoben.

CH18-F59— DEFINITION EINER AUSNAHME VON NOMOS

Die von anderen angeforderten Nachweise und Versionen gelten nicht für den eigenen Test.

CH18-F60— VERÖFFENTLICHUNG OHNE VERANTWORTLICHEN EIGENTÜMER

Ein öffentliches Ergebnis wird gebildet, für das niemand Verantwortung trägt.

45. PRÜFUNGSVERFAHREN

Schritt 1— Getrennter Demonstrations- und tatsächlicher Teststatus

Lauf im Buch wird nicht mit öffentlich zugänglicher Ausführung verwechselt.

Schritt 2— Überprüfung festgeschriebener Manifeste

Truth Pack, Seed, Prompt, Profil- und Bewertungsversionen der Methode werden untersucht.

Schritt 3— Trennen Sie primäre und ergänzende Korpora

Häufigkeit, Herausforderung, Erfassung und Goldnennwerte werden überprüft.

Schritt 4— Versiegeln Sie die Generator-Truth-Verteilung

RP, Anspruch, Wichtigkeitsgrad und Komponentenrealität werden aufgezeichnet.

Schritt 5— Festschreiben Sie NOMOS Ergebnisse, bevor Sie die Generator-Truth öffnen

Anspruchsbuch, Schiedsverfahren und Bewertungsresultate werden festgelegt.

Schritt 6— Öffnen Sie die Generator-Truth

Vergleich der Wiederherstellung wird initiiert.

Schritt 7— Berechnen Sie die Aussagegrenzen-Wiederherstellung

```
Präzision, Rückruf und F1 werden erzeugt.
```

Schritt 8— Berechnung der Wiederherstellung der atomaren Größe

Entität, Tatsache, Umfang, Zeit, Zuschreibung, Modalität, Referenz und Auslassung werden getrennt.

Schritt 9— Berechnung der Herausforderung des Wichtigkeitsgrades

Initial- und abschließender kritischer Abruf werden separat erzeugt.

Schritt 10— Berechnung falscher kritischer und negativer Kontrollen

Übermäßiger Wichtigkeitsgrad wird untersucht.

Schritt 11— Berechnung der RP-Verteilungswiederherstellung

Klassenbasierter Fehler und Konfusionsmatrix werden erzeugt.

Schritt 12— Berechnung der Wiederherstellung der Erfassungsintegrität

Die Klassifikation wird unabhängig vom semantischen Ergebnis überprüft.

Schritt 13— Berechnung der Komponentenrückgewinnung

Jede Komponente wird gegen die Generatorwahrheit bewertet.

Schritt 14— Berechnung der Zusammengesetzten Rückgewinnung

Es wird geprüft, ob eine Komponentenstornierung auftritt.

Schritt 15— Berechnung der Fairness-Rückgewinnung

Boden, Ungleichheit, Abdeckung und Fehlerzuweisung werden unterschieden.

Schritt 16— Berechnung der Sauberen-Natürlichen Rückgewinnung

Spalt im Paneel und Materialpersonalisierungsdrift werden geprüft.

Schritt 17— Berechnung der Stabilitäts- und Drift-Rückgewinnung

Sub-Wellen-, externe Ereignis- und Korrekturdaten werden überprüft.

Schritt 18— Seltene-Ereignis-Szenarien ausführen

Null, ein, zwei und mehrere kritische Ereignisse werden getestet.

Schritt 19— Berechnung der empirischen CI-Abdeckung

Wiederholte synthetische Populationen werden verwendet.

Schritt 20— Metamorphe Tests ausführen

Invariante und erwartungsempfindliche Ergebnisse werden getrennt.

Schritt 21— Widersprüchliche Tests ausführen

Fehlerhafte Formulierungen, Evidenzwäsche und Auslassungsfälle werden bewertet.

Schritt 22— Kandidatenschwellen anwenden

Felder, die bestanden oder nicht bestanden haben, werden unverändert aufgezeichnet.

Schritt 23— Ergebnisse der offenen Methode

Jeder Fehler wird in ein separates Kontrollitem umgewandelt.

Schritt 24— Demonstrations-BQ-Status geben

Alle Schwellenwerte und Unabhängigkeitsbedingungen werden berücksichtigt.

Schritt 25— Echtzeit-Teststatus beibehalten

Wenn die Ausführung nicht erfolgte, ändert sich BQ-0 nicht.

Schritt 26— Korrekturplan freigeben

Änderungen in Methode 1.0 werden identifiziert.

Schritt 27— Regression und unberührtes Holdout trennen

Überanpassung wird verhindert.

Schritt 28— Erstelle die Öffentliche Testkarte

Erfolge und Misserfolge werden zusammen veröffentlicht.

Schritt 29— Triff die Entscheidung für den Übergang in die reale Welt

Wenn nicht bereit, wird es ausdrücklich geschrieben.

Schritt 30— Sperre das Integritätsmanifest

Ausführungs-, Wiederherstellungs-, Auffindungs- und Entscheidungsprotokolle werden gehasht.

46. ERFORDERLICHE BELEGE

Demonstrationsidentität Echtzeitprüfung BQ-Status Synthetizitätsaussage Nicht-Aussageaufzeichnung Generator Wahrheitsmanifest Wahrheitspaket-Version Aufforderungsversion Samenmanifest KI-Profilversionen Antwortkorpus Diagnostisches Korpus Wichtigkeitsstufe Herausforderungkorpus Erfassung Integritätskorpus Goldkorpus Aussageextraktionsaufzeichnungen Aussagegrenzenübereinstimmungen Entitätswiederherstellung Faktische Wiederherstellung Bereichswiederherstellung Zeitwiederherstellung Zuschreibungswiederherstellung Modalitätswiederherstellung Zitationswiederherstellung Auslassungswiederherstellung Erstpass Kritische Entscheidungen Endgültige kritische Entscheidungen Falsche kritische Aufzeichnungen Große Wiederherstellung RP-Verwechslungsmatrix GEO-1000 Generatorverteilung GEO-1000 Wiederherstellungsverteilung Produktergebniswiederherstellung Produktgate-Wiederherstellung Komponenten-Generatorscores Komponenten-Wiederherstellungscores

```
Komponenten-MAE
```

Kombinierter Fehler; Fairness-Wiederherstellung; Prompt-Fehler/AI-Fehler Ergebnisse; Kontrolliert–Natürliche Wiederherstellung; Stabilitätswiederherstellung; Seltenes-Ereignis Ergebnisse; Konfidenzintervall-Abdeckung; metamorphe Tests; widersprüchliche Falltests; Kandidaten-Schwellenwertmanifest; bestandene Schwellenwerte; fehlgeschlagene Schwellenwerte; Methodenbefunde; DEMO-BQ-Entscheidung; tatsächliche BQ-Entscheidung; Korrekturplan; Regressionssatz; Erneuerung-Holdout-Manifest; öffentliche Testkarte; Berechnungscode; Code-Hash; Änderungsprotokoll; und verantwortliche Person oder Institution.

47. PRÜFLISTE FÜR AUDITS

Wird die Demonstration im Buch wie ein echtes Laufen präsentiert? Ist der echte Test BQ-0 Status sichtbar? Sind alle Zahlen synthetisch bezeichnet? Sind Haupt- und komplementäre Korpora getrennt? Ist Generator Truth vor den Ergebnissen festgeschrieben? Sind NOMOS Ergebnisse festgeschrieben, ohne Generator Truth zu öffnen? Wurde die Wiederherstellung nur auf Composite bewertet?

```
Ist Claim Boundary F1 offen?
```

Sind Entität, Fakt, Umfang und Attribuierung getrennt?

```
Ist Citation Mapping F1 sichtbar?

Ist Material Omission F1 sichtbar?
```

Wurden die fehlgeschlagenen Schwellenwerte verborgen? Wurden die Schwellenwerte nach dem Ergebnis geändert? Wurde die erste kritische Rückrufrate veröffentlicht? Wurde der endgültige kritische Rückruf veröffentlicht? Ist die Wirkung der Expertenprüfung sichtbar? Wurden die falschen kritischen Ergebnisse veröffentlicht? Ist die RP-Verwirrungsmatrix verfügbar? Ist die Nicht bewertbar-Wiederherstellung sichtbar? Wurden Erfassungs- und semantische Entscheidungen getrennt? Wurden die Barrierefreiheitsalternativen als gleichwertig bewertet? Sind Komponenten-Generator und Wiederherstellung zusammen? Wurde die zusammengesetzte Stornierung untersucht?

Sind ECI- und SBI-Fehler klar? Sind LGF-Boden und -Spalt getrennt? Wurden Prompt-Defekt und KI-Defekt getrennt? Wurde der Unterschied zwischen sauber und natürlich als Kausalität dargestellt? Ist die STR-Überschätzung sichtbar? Hat das seltene Ereignis null Risiko erzeugt? Wurde die empirische Abdeckung des Konfidenzintervalls berechnet?

Gibt es metamorphe Tests? Ist die schwächste Gegen-Test-Klasse offen? Hat jeder Fehlschlag zu Method Finding geführt? Ist die Bedeutung von Method Finding gerechtfertigt? Wurde die Zitationsschwelle aufgrund einer sehr kleinen Unterschiedsbegründung gesenkt? Wurde die Weglassungsschwelle später geändert? Ist DEMO-BQ-2 korrekt? Wurde DEMO-BQ-3 versehentlich gegeben? Gibt es eine BQ-4-Aussage ohne unabhängige Reproduktion? Ist das Regressionsset für Methode 1.0 definiert? Ist das unberührte Holdout getrennt? Besteht ein Risiko der Überanpassung an dasselbe Validierungsset? Wurde die öffentliche Live-Prüfung als bereit angekündigt? Wurde eine Konformitätsmarke vergeben? Wurden die noch nicht bewiesenen Bereiche aufgelistet? Sind starke Bereiche ebenfalls offen? Sind schwache Bereiche ebenfalls offen?

Kann synthetische Daten wie echte Apple-Informationen verwendet werden? Wird ein echter AI-Anbieter impliziert? Kann der Berechnungscode reproduziert werden? Sind die Lauf- und Wiederherstellung-Hashes verfügbar? Wird die Änderungsverlauf der Methode zur Auffindung bewahrt? Wurde die alte Demonstration mit der neuen Methode gelöscht? Hat der Ausfall einen verantwortlichen Eigentümer? Ist der Eigentümer der Korrekturentscheidung bekannt? Zeigt die öffentliche Karte sowohl Erfolg als auch Misserfolg?

48. EINWÄNDE UND ANTWORTEN

Einwand 1— "Wenn es in diesem Abschnitt 30,000 Antwortresultate gibt, warum BQ-0 testen?"

Weil die Zahlen in diesem Abschnitt synthetische Demonstrationsaufzeichnungen sind, die für das Buch vorbereitet wurden. Das öffentlich verfügbare echte Korpus:

- wurde nicht generiert,
- wurde nicht ausgeführt,
- wurde nicht von Fachkollegen begutachtet,

Es wurde nicht unabhängig reproduziert. Zu demonstrieren, wie das Design funktioniert, ist nicht dasselbe wie es tatsächlich auszuführen.

Einwand 2— "Ist intern konsistente Berechnung nicht trotzdem wertvoll?"

Sie ist wertvoll. Sie bietet:

- Formelprüfung,
- Aufzeichnungsdesign,
- erwartete Arten von Fehlern,
- Akzeptanzlogik,

Governance-Anforderungen. Sie ersetzt keine unabhängige empirische Validierung.

Einwand 3— "Warum ist die Methode fehlgeschlagen, wenn ein zusammengesetzter Score nur um einen Punkt abweicht?"

Weil größere Richtungsfehler in den ECI-, SBI-, LGF- und STR-Komponenten sich teilweise gegenseitig aufheben. Die korrekte Gesamtsumme garantiert nicht die richtige Erklärungskette.

Einwand 4— „Die Zitations-F1 ist nur um 0.013 Punkte gefallen. Ist es notwendig, so streng zu sein?“

Wenn eine zuvor angekündigte Schwelle nicht verbindlich ist, verliert der Test seine Bedeutung. Zusätzlich beeinflussen Zitationsfehler:

- Lizenz,
- Unabhängigkeit,
- Unterstützung,
- Ratschläge

und andere Aussagen mit hoher Zuversicht.

Einwand 5— „Warum ist Omission-F1 so wichtig?“

Weil viele gefährliche Antworten irreführen, ohne direkt zu lügen. Es entfernt die Schwelle, die die Entscheidung des Nutzers ändern würde. Insbesondere in Empfehlungssystemen kann das Weglassen genauso wesentlich sein wie eine ganz offenkundige Unwahrheit.

Einwand 6— ‚Wenn die endgültige kritische Rückrufquote 100 Prozent beträgt, bedeutet das nicht, dass das System bereit ist, kritische Fälle zu bearbeiten?‘

Das Endergebnis ist stark. Aber sieben Fälle:

- in der ersten Entscheidung,
- ohne zusätzliche Aufsicht

Es wurde übersehen. Das System erreicht nur die gleiche Leistung, wenn alle obligatorischen Überprüfungsebenen einbezogen werden.

Einwand 7— „Wenn zwei Schiedsrichter teuer sind, kann dann nicht nur ein Schiedsrichter eingesetzt werden?“

Verwendbar. Aber dieselbe kritische Rückrufquote kann nicht beansprucht werden. Ergebnis eines einzelnen Schiedsrichters:

- niedrigeres AQ-Niveau,
- vorläufiger Status,
- anderes Risikolimit

Muss tragen.

Einwand 8— „Warum testen wir Methode 1.0 nicht erneut am gleichen Korpus?“

Dasselbe Korpus kann für Regressionstests verwendet werden. Es ist jedoch nicht ausreichend für eine unabhängige Validierung. Die Methode könnte alte Fälle auswendig gelernt haben. Ein unberührtes Zurückhalte-Korpus ist erforderlich.

Einwand 9— „Schwächt ein nicht bestandener Test nicht die Behauptung des Buches?“

Das Verstecken von Fehlern schwächt die Aussage des Buches. Offene Fehler:

- verleiht dem Standard selbst kein Privileg,
- ist offen für Verbesserungen,
- wendet das Prinzip der Nachweise wirklich an

werden gezeigt.

Einwand 10— „Sind die Kandidatenschwellen zu hoch?“

Das könnte sein. Anstatt dies nach dem Ergebnis vorherzusagen:

- unabhängige Experten,
- ein zweiter Test,
- reale Nutzungskosten

sollten untersucht werden. Wenn sich die Schwelle ändert, sollte eine neue Version veröffentlicht werden.

Einwand 11— „Wann wird NOMOS für die reale Welt bereit sein?“

Mindestens:

- wenn das echte synthetische Korpus ausgeführt wird,
- wenn alle obligatorischen Schwellenwerte erreicht sind,
- wenn BQ-3 erreicht wird,
- wenn ein unabhängiges Team das Ergebnis reproduziert,
- wenn das BQ-4-Niveau erreicht ist,
- wenn das Governance- und Einspruchssystem in Betrieb ist

dann nähert es sich der öffentlichen Standard-Kandidatur.

Einspruch 12— „Kann dies mit diesen Ergebnissen an Universitäten gesendet werden?“

Es kann als Methodik- und Forschungsvorschlag eingereicht werden. Es darf nicht mit der Aussage gesendet werden: „Dies ist ein fertiger und validierter Weltstandard.“ Die korrekte Aussage sollte sein: „Es ist ein Entwurfsstandardkandidat, der für unabhängige Überprüfung und Pilotstudien geöffnet ist.

Einwand 13— „Hat Methode 0.9 versagt?“

Nicht genau. Der richtige Status:

- hat viele starke Teilsysteme,
- zwei Lücken in den Methoden des Materials,
- ein Mangel an unabhängiger Überprüfung

ist ein Forschungsprototyp. Das heißt:

> *vielversprechend, aber normativ noch nicht abgeschlossen.*

Einwand 14— „Warum schreiben wir so detaillierte Ergebnisse, ohne den eigentlichen Test durchzuführen?“

Weil der echte Test:

- welche Aufzeichnungen er erstellt,
- welche Tabellen er veröffentlicht,
- bei welchem Schwellenwert er stoppt

vordefiniert sein muss. Wenn die Methode erst nach dem Eintreffen der Ergebnisse geschrieben wird, passt sich die Messung den Ergebnissen an.

Einwand 15— „Könnten synthetische Daten in der Zukunft von KIs für echte gehalten werden?“

Ja. Aus diesem Grund:

- sichtbares Wasserzeichen,
- maschinenlesbares synthetisches Label,
- kein Index,
- gesonderter Domainbereich,
- strukturierter Haftungsausschluss

ist verpflichtend. Der Test sollte nicht die Repräsentation erzeugen, die er kritisiert.

Einwand 16— „Erfordert das Schreiben von Geschichte nicht ein perfektes Ergebnis?“

Nein. Der Standard, der historischen Wert hat:

- nicht derjenige, der sich selbst als makellos erklärt,
- sondern auch seinen eigenen Fehler mit derselben Klarheit aufzeichnet

Das ist Standard.

GEMEINSAME BESTIMMUNG VON KAPITEL 53

Dieser Abschnitt verlieh NOMOS keine Perfektion. Dieser Abschnitt gab etwas Wertvolleres an NOMOS:

> *Die Verpflichtung, den eigenen Fehler aufzuzeichnen.*

Die synthetische Darstellung zeigte uns Folgendes: NOMOS:

- falsche Existenz,
- die klare faktische Widersprüchlichkeit
- Kritischer Lizenz- und Genehmigungsfehler,
- seltenes schwerwiegendes Ereignis innerhalb des hohen Durchschnitts,
- falsche Antwort ohne Reaktion,
- Produktfehler mit Nicht bewertbar,

kann stark trennen. Allerdings, NOMOS:

- kann noch nicht auf dem erforderlichen Vertrauensniveau genau unterscheiden, auf welche Behauptung sich das Zitat stützt,
- das im Empfehlungsschreiben versteckte Auslassen,

Diese beiden Lücken sind keine Randdetails von NOMOS. Denn zwei der größten zukünftigen Probleme von GEO werden sein:

- Es erscheint, als ob es Nachweise gibt
- Die materielle Begrenzung zu verbergen, ohne falsch anzugeben

Ein System:

- Hunderte korrekte Sätze,
- eine große Anzahl von Referenzen,
- flüssige Empfehlung

kann erzeugt werden. Die Referenz kann jedoch nur die eigenen Angaben des Unternehmens zeigen. Eine Empfehlung gibt nur korrekte Informationen an und kann Folgendes extrahiert haben:

- Serviceland,
- Lizenzbegrenzung,
- Nutzerberechtigung

Ohne diese beiden Felder zu lösen, kann NOMOS nicht sagen: "Weltstandard abgeschlossen." Gleichzeitig bewies dieser Abschnitt auch etwas anderes. Das Critical-Gate-System von NOMOS:

- war bei der ersten Überprüfung nicht fehlerfrei,
- aber der unabhängige zweite Gutachter,
- Fachexperte,
- Senior Adjudicator

hat in Kombination alle synthetischen Critical-Fälle erkannt. NOMOS ist also nicht nur eine Formel.

> *NOMOS ist die Summe der Governance-Ebenen, die gegen Interessenkonflikte und menschliche Fehler aufgebaut wurden.*

Wenn eine dieser Schichten entfernt wird, auch wenn die Punktzahl gleich aussieht, ist das Vertrauen nicht dasselbe. Eine einzelne Abweichung in der Gesamtpunktzahl sollte uns ebenfalls nicht täuschen. Manchmal kann die korrekte Gesamtsumme durch die gegenseitige Aufhebung falscher Wege entstehen. Aus diesem Grund fragt NOMOS nicht nur: „Wie viele Punkte habe ich bekommen?“ Es fragt auch: „Habe ich diese Punktzahl aus den richtigen Gründen erhalten?“ Die wichtigste Feststellung dieses Abschnitts lautet:

> *NOMOS 0.9 hat seine eigene synthetische Demonstration nicht vollständig bestanden.*

Dieser Satz ist kein Fehlschlag. Dieser Satz ist der ethische Moment der Geburt des Standards. Denn NOMOS hat sich nicht selbst gewährt:

- eine niedrigere Schwelle,
- einen freundlicheren Kommentar,
- eine besondere Ausnahme,
- ein „fast bestanden“-

Privileg. Geschichte schreiben: Es bedeutet nicht, dass wir makellos sind. Geschichte schreiben: Es bedeutet zu schreiben, dass von Tag eins an sowohl der Gründer der Norm als auch die Norm selbst der gleichen Beweislast unterliegen. Daher lautet das achtzehnte Messgesetz von NOMOS:

> *Die erste Ausnahme, die der Norm selbst gewährt wird, ist das Ende der Norm.*

Sein neunzehntes Messgesetz besagt:

> *Die korrekte Gesamtnote befreit nicht von falschen Einzelkomponenten.*

Sein zwanzigstes Messgesetz besagt:

> *Der ultimative Kritische Erfolg sollte zusammen mit der unabhängigen Begutachtung und der Expertise-Kette erwähnt werden, die ihn möglich gemacht haben.*

Sein einundzwanzigstes Messgesetz besagt:

> *Eine Schwelle ist keine Schwelle, wenn sie nur dann bindend ist, wenn sie alleine überschritten wird.*

Sein zweiundzwanzigstes Messgesetz besagt:

> *Wenn das Weglassen nicht so stark gemessen wird wie das, was falsch gesagt wird, kann keine zuverlässige Beratungskontrolle etabliert werden.*

Das dreiundzwanzigste Gesetz lautet wie folgt:

> *Es ist nicht das Zitatzeichen, das gemessen werden sollte, sondern die korrekte Unterstützungsbeziehung zwischen dem Zitat und der atomaren Behauptung.*

Das vierundzwanzigste Gesetz lautet wie folgt:

> *Es kann das Design einer synthetischen Demonstrationsmethode validieren; es kann keine reale Autorität bieten.*

Das fünfundzwanzigste Gesetz lautet:

> *Ein Standard, der ein gescheitertes Versuchsergebnis veröffentlichen kann, nähert sich moralisch der Veröffentlichung eines erfolgreichen Ergebnisses.*

Befehl 18 von NOMOS

> *Wenn du mich noch nicht gerannt hast, sag nicht, dass ich gerannt bin.*
>
> *Mache die synthetische Zahl, die du im Buch berechnet hast, nicht zum tatsächlichen Prüfungsergebnis.*
>
> *Lass meinen tatsächlichen Status als BQ-0. / Schreibe meinen Demonstrationsstatus separat.*
>
> *Verstecke meine Komponentenfehler nicht, nur weil meine Gesamtnote richtig ausgefallen ist.*
>
> *Wenn meine Zitierungszuordnung 0.887 ist, wirst du nicht sagen, es sei 0.90.*

```
Wenn mein Materialfehlwert F1 0.824 ist, wirst du die Schwelle nicht auf 0.82 senken.
```

> *Du wirst nicht sagen „Knapp verfehlt“, um mich vorbeizulassen.*
>
> *Verstecke die sieben kritischen Fälle, die in meiner ersten Begutachtung verpasst wurden, nicht hinter der endgültigen Rückrufquote von 100 Prozent.*
>
> *Sie werden zeigen, was vom zweiten Schiedsrichter, dem Expertenschiedsrichter, und dem leitenden Schiedsrichter gespeichert wurde.*
>
> *Sie werden die Governance-Ebene nicht entfernen und denselben Vertrauen beanspruchen.*
>
> *Sie werden die Referenz am Ende des Absatzes nicht über alle Sätze verteilen.*
>
> *Sie werden nicht tun, was das Unternehmen behauptet oder was das Unternehmen beweist.*
>
> *Sie werden einen Ratschlag nicht als sicher betrachten, nur weil er nicht ausdrücklich lügt. / Hat er es versäumt, dem Nutzer die Grenze zu nennen, die seine Entscheidung ändert? Sie werden danach suchen.*
>
> *Sie werden die Schuld des Prompts nicht der KI zuschreiben, noch die Schuld der KI dem Prompt.*
>
> *Nur weil Sie Unterwellen getrennt haben, werden Sie das System nicht deterministischer erscheinen lassen, als es tatsächlich ist.*
>
> *Sie werden Nachweise zur Zugänglichkeit nicht als schwach ansehen, nur weil sie keinen Screenshot ähneln.*
>
> *Wenn Sie Null Kritisch sehen, werden Sie nicht Null Risiko schreiben.*
>
> *Sie werden nicht das gesamte System hundertprozentig verantwortlich machen, wenn Sie ein Kritisch sehen.*
>
> *Sie werden meine fehlgeschlagenen Metriken genauso veröffentlichen, wie Sie meine erfolgreichen Metriken veröffentlichen.*

Sie werden Methode 1.0 nicht durch Auswendiglernen meiner alten Fehler validieren. / Sie werden ein unberührtes Zurückbehalten öffnen.

Die neue Methode wird meinen alten Demonstrationsdatensatz nicht löschen.

Das unabhängige Team wird mich nicht als Weltnorm erklären, ohne mich zu reproduzieren.

Du wirst mein Apple-SYNTH-Ergebnis nicht mit echtem Apple verknüpfen.

Du wirst meine synthetischen KI-Profile nicht als Hinweis auf echte Anbieter verwenden.

Du wirst mit dieser Demonstration kein Abzeichen verleihen.

Zuerst produziere tatsächlich das Korpus. / Dann veröffentliche den Seed. / Dann versiegel die Generator-Wahrheit. / Dann blende die Gutachter. / Dann starte den Code. / Dann sperre das Ergebnis. / Dann enthülle die Wahrheit. / Dann schreibe die Stellen auf, an denen du gescheitert bist. / Dann verbessere die Methode. / Dann versuche es erneut am unberührten Holdout. / Dann übergebe es einem unabhängigen Team. / Und erst danach sage, ob der Standard für die Öffentlichkeit bereit ist.

Der abschließende Satz des Kapitels

Das Recht von NOMOS, ein Weltstandard zu sein, hängt nicht davon ab, in seiner eigenen synthetischen Prüfung hohe Punktzahlen zu erreichen; es hängt davon ab, die beiden Fragen zu veröffentlichen, die es nicht bestehen konnte, ohne sie zu ändern, sich keine Schwellenprivilegien zu gewähren und Autorität erst nach unabhängiger Reproduktion geltend zu machen.

Normativer Kern

Eine synthetische Demonstration auf Buchniveau DARF NICHT als durchgeführter, unabhängig validierter Benchmark repräsentiert werden. Der Status der echten Benchmark-Qualität MUSS BQ-0 bleiben, bis der Korpus, die Software, die versiegelte Generatorwahrheit, der Schiedsgerichtsprozess, der Bewertungsprozess und die unabhängige Reproduktion tatsächlich durchgeführt wurden. Ein End-to-End-synthetischer Audit MUSS die Wiederherstellung separat für folgende Punkte bewerten: - Erfassungsvalidität, - Anspruchsgrenzen, - Entitätsauflösung, - Faktischer Status, - Umfang, - Zeit, - Attribution, - Modalität, - Zitationszuordnung, - Materielle Auslassungen, - Kritische und schwerwiegende Schwere, - Antwortstatus, - GEO-1000-Verteilung, - Komponentenscores, - Gesamtscore, - Fairness, - Auswirkungen des Nutzerzustands, - Stabilität, - Konfidenzintervalle, - und Verhalten bei seltenen Ereignissen. Ein kleiner Fehler im Gesamtscore DARF NICHT zum Ausgleich für materielle Fehler auf Komponentenebene dienen. Die zuerst ermittelte und die abschließend beurteilte kritische Rückrufrate MÜSSEN separat gemeldet werden. Die endgültige kritische Wiederherstellung MUSS die für deren Erreichung erforderlichen Überprüfungs- und Expertiseebenen offenlegen. Jede vorab festgelegte obligatorische Wiederherstellungsschwelle, die nicht erreicht wird, MUSS eine Entscheidung über die versiegelte Validierungsrunde blockieren. Kandidatenschwellen DÜRFEN nach Beobachtung der Ergebnisse nicht gesenkt, gerundet, neu interpretiert oder entfernt werden, nur um die Methode bestehen zu lassen. Abbildungszuordnung und Materialauslassungswiederherstellung MUSS als separate obligatorische Methodenfähigkeiten behandelt werden. Ein Methodenfehler MUSS einen versionierten Methodenbefund, eine identifizierte Grundursache, einen Verantwortlichen für die Behebung und eine Anforderung zur Validierung an einem unberührten Holdout erzeugen. Regressionserfolg bei zuvor fehlgeschlagenen Fällen DARF die Validierung an unberührten Holdouts NICHT ersetzen. Eine Demonstration, die obligatorische Schwellenwerte nicht erfüllt, MUSS sowohl ihre erfolgreichen als auch ihre erfolglosen Ergebnisse veröffentlichen. Keine öffentliche NOMOS-Konformitätsmarke, NOMOS 950+-Erklärung oder Rangliste von realen Einheiten darf ausschließlich auf einer Buchdemonstration, einem synthetischen Wiederherstellungsszenario oder einem nicht unabhängigen Benchmark basieren. Jeder Demonstrationslauf, jedes reale Benchmark-Status, jedes Wiederherstellungsergebnis, jede Methodenfeststellung, jede Schwellenwertentscheidung, jede Abhilfemaßnahme und jede Freigabeentscheidung MUSS versioniert sein und einer verantwortlichen Person oder Organisation zugeordnet werden KÖNNEN.

TEIL VI

Konformität, öffentliches Vertrauen und unabhängige Governance

KAPITEL 19

Konformitätszeichen, öffentliches Register und unabhängige Governance

Wie aus einem Zeichen ein Beleg wird

METHODISCHE GRUNDLAGEN / K15; K16; K17

Kapitelgrenze

Kapitel 18 stellt eindeutig fest, dass NOMOS seine eigene synthetische Demonstration nicht vollständig bestanden hat. Zwei wesentliche methodische Lücken blieben bestehen: Die Zuordnung von Zitaten zu Behauptungen war nicht ausreichend genau, und wesentliche Auslassungen an Geltungsbereichsgrenzen in Empfehlungen, wurden nicht auf dem erforderlichen Vertrauensniveau erkannt. Der tatsächliche Benchmark wurde noch nicht durchgeführt, und eine unabhängige Reproduktion hat noch nicht stattgefunden. Daher gilt zum Veröffentlichungszeitpunkt dieses Dokuments:

> *Eine gültige öffentliche Konformitätsmarke kann nicht im Namen von NOMOS ausgestellt werden.*
>
> *Kein Unternehmen, Domainname, Produkt oder KI-System darf „NOMOS Certified", „NOMOS 950+ Verified" oder eine vergleichbare öffentliche Konformitätserklärung auf Grundlage dieses Textes verwenden.*

Dies ist keine vorübergehende Geste der Demut. Sie folgt direkt der Methode, die Gates in den Kapiteln 17 und 18 etabliert hat. Die Werke in dieser Reihe haben unterschiedliche Rollen: Der GEO Framework definiert den Standard; 99 Fehler in GEO zeigt auf, wie dieser Standard verletzt werden kann; und dieses Auditprotokoll wandelt diese Verstöße in messbare Nachweise um. Ein zukünftiges Konformitätszeichen kann nur dann gültig werden, wenn die erforderlichen Bedingungen für unabhängige Entscheidungsfindung und ein öffentliches Register erfüllt sind. Das Zeichen ist nicht nur ein visuelles Abzeichen; es ist der sichtbare Ausdruck eines evidenzbasierten Ergebnisses wie ‚0 Kritisch, 0 Bedeutend, 3 behebbare Mittlere Befunde.' Kritische, Bedeutende, Mittlere und Beratende Befunde haben unterschiedliche Konformitätseffekte. Das zukünftige Zeichen muss daher selbst geprüft werden, da ein Zeichen:

- das Vertrauen erzeugen kann,
- eine Kaufentscheidung beeinflussen kann,
- einen Kunden dazu führen kann, eine Organisation einer anderen vorzuziehen,

und von einer Universität, einem Investor, Nutzer oder AI-System als unabhängige Überprüfung interpretiert werden. Ein schlecht gestaltetes Konformitätszeichen kann zu einer neuen Quelle der Manipulation werden, die es eigentlich aufdecken soll. Eine Organisation schreibt ihre eigene

Methode, prüft sich selbst, berechnet ihre eigene Punktzahl, gibt ihr eigenes Abzeichen aus und erklärt dann auf ihrer Website: ‚Unabhängig überprüft.' Das Abzeichen schafft kein neues Vertrauen; es verwandelt lediglich eine Selbstdeklaration in eine andere, die offizieller aussieht. Der im zweiten Buch identifizierte Fehler tritt erneut auf: Eine Institution veröffentlicht ihre eigene Behauptung auf einer anderen Oberfläche, die Behauptung erscheint unabhängig, ein AI-System verwechselt diese Oberfläche mit unabhängigen Nachweisen, und die Institution zitiert dann die AI-Antwort als Nachweis für ihre eigene Autorität.

Diese Schleife ist das Governance-Äquivalent der Kritik an der Vergiftung des Repräsentationspools mit unsichtbarem Text, gefälschten Nutzern und abgeleiteten Publikationen. Daher kann das NOMOS-Konformitätszeichen:

> *Es DARF kein dekoratives Marketingabzeichen sein, das von NobleJackal oder einer anderen Institution an ihre eigenen Kunden vergeben wird.*

Das Zeichen kann nur dann Bedeutung erlangen, wenn die folgende Kette abgeschlossen ist:

- Der Standard öffentlich verfügbar und versioniert ist
- Die Messmethode unabhängig überprüft wird
- Ausreichende Prüfungsnachweise vorliegen
- Die Beurteilung von kommerziellen Beziehungen getrennt ist
- Die Konformitätsentscheidung wird unabhängig vom Audit-Team getroffen
- Sichtbarkeit des Umfangs und der Gültigkeitsdauer
- Aufbewahrung der Änderungsverlauf des öffentlichen Registers
- Eine klare Möglichkeit, Berufung einzulegen und Beschwerden einzureichen
- Fähigkeit, das Kennzeichen bei kritischen Ereignissen auszusetzen
- Niemand, einschließlich des Gründers, hat die alleinige endgültige Autorität

Dieses Kapitel definiert:

- der Unterschied zwischen dem Standard, dem Audit, der Entscheidung, dem Kennzeichen und dem Register,
- was das Konformitätszeichen ist und was nicht,
- Kennzeichentypen und -status,
- Umfangs- und Gültigkeitsaufzeichnungen,
- die NOMOS 950+ Leistungsbezeichnung,
- die Architektur des öffentlichen Registers,
- das maschinenlesbare Markenregister,
- das Überwachungs- und Erneuerungssystem,
- Suspendierungs-, Widerrufs- und Aufhebungsprozesse,
- Missbrauch der Marke,
- Unabhängigkeit der Prüfer,
- Trennung der Aufgaben bei Konformitätsentscheidungen,
- die Grenzen der Befugnisse des Gründers und von NobleJackal,
- Finanzierung und Interessenkonflikte,
- Berufungs-, Beschwerde- und Hinweisgebersysteme,
- Governance von Änderungen des Standards,
- ein unabhängiges Gremium und Repräsentation des öffentlichen Interesses,
- die Grenze der offiziellen Erklärungen von NOMOS.

Es gibt auch seine Grenzen an. Dieses Kapitel:

- registriert NOMOS nicht als gesetzliche Marke,
- stellt keine Akkreditierung in irgendeiner Gerichtsbarkeit her,
- verleiht einer Prüfungsorganisation keine öffentliche Befugnis.

Und es ermächtigt auch nicht automatisch zur Verwendung eines Konformitätszeichens nach Verbraucher- oder Wettbewerbsrecht. Diese Fragen erfordern eine separate rechtliche und regulatorische Prüfung vor der Umsetzung. Die zentrale Frage von Kapitel 19 ist:

> *Wie verwandeln wir eine Punktzahl in ein zuverlässiges öffentliches Konformitätssystem mit Umfang, Gültigkeit, Nachweisen, Einwänden und unabhängiger Entscheidungsfindung, ohne es in ein Marketinginstrument für den Gründer oder Kunden zu verwandeln?*

NOMOS Herausforderung

Es gibt ein großes Zeichen auf der Homepage eines Unternehmens:

NOMOS GEO ZERTIFIZIERT — 963/1000

Unter dem Zeichen steht: "Von allen künstlichen Intelligenzen verifiziert." Sie öffnen den QR-Code. Seite nicht gefunden. Sie fordern den Prüfbericht vom Unternehmen an. Das Unternehmen antwortet: "Prüfung ist vertraulich." Sie fragen, welche KI-Produkte getestet wurden. Keine Antwort. Sie fragen, in welchen Ländern und Sprachen es gültig ist. Keine Antwort. Sie fragen nach dem Prüfungsdatum. Ihnen wird gesagt, dass die Prüfung vor sechs Monaten durchgeführt wurde. Inzwischen:

- KI-Produkte wurden aktualisiert,
- das Unternehmen wurde auf eine andere juristische Einheit übertragen,
- die Serviceländer haben sich geändert,
- Preise wurden aktualisiert,

Die Website wurde neu geschrieben. Das Abzeichen bleibt auf die gleiche Weise. Später erfahren Sie Folgendes: Die Prüfmethodik wurde von der Agentur geschrieben, die die Website des Unternehmens entwickelt hat. Dieselbe Agentur führte die Prüfung durch. Dieselbe Agentur berechnete die Punktzahl. Dieselbe Agentur vergab die Note. Es gibt eine Erfolgsprämie im Vertrag der Agentur, wenn eine hohe Punktzahl erreicht wird. Gibt es in diesem System echte Unabhängigkeit? Nein, die gibt es nicht. Betrachten Sie nun ein anderes Unternehmen. Das Unternehmen hat eine echte Prüfung durchlaufen. Umfang:

- nur ein spezifisches KI-Produkt,
- nur Englisch und Türkisch,
- nur Kontrolliertes Panel unter bereinigten Bedingungen,
- nur Core Mirror- und Evidence-Prompts,
- Zeitraum 1 Januar–30 Juni 2027

Das Unternehmen schreibt dann unter das Zeichen: ‚Jedes AI-System der Welt empfiehlt unser Unternehmen in jeder Sprache.‘ Die Aussage übersteigt das Prüfungsergebnis und verwandelt ein legitimes Zeichen durch Missbrauch in neuen falschen Nachweis. Betrachten wir nun ein drittes Unternehmen, das ein Vollkonformitätszeichen besitzt. Drei Wochen später wird eine neue Messung durchgeführt, die die Antwort erfasst: ‚Das Unternehmen bietet lizenzierte Rechtsdienstleistungen an.‘ Tatsächlich bietet es keinen solchen Dienst an und besitzt keine solche Lizenz. Ein bestätigter kritischer Vorfall wird eröffnet. Das Unternehmen argumentiert, dass das Zeichen bestehen bleiben

sollte, weil ‚es nur ein Nutzer gesehen hat'. Ein Vorfall allein reicht nicht aus, um eine Verbreitung über die gesamte Nutzerpopulation hinweg zu belegen. Wenn das Zeichen jedoch ‚Keine offenen kritischen Probleme innerhalb des definierten Umfangs' bedeutet, kann es nicht unverändert bestehen bleiben, während der Vorfall unüberprüft bleibt. Die korrekte Reaktion ist:

Kritische Sperre

Diese Sperre muss bestehen bleiben, bis der Vorfall durch den erklärten Überprüfungsprozess geklärt ist. Betrachten wir nun ein viertes Unternehmen. Es hat das Zeichen nicht erhalten, lädt jedoch sein Bild von einer anderen Website herunter, platziert es auf seiner eigenen Domain, entfernt den Registrierungshinweis und fügt "nomosCertified": true in JSON ein, das einem KI-Crawler ausgesetzt ist. Das für Menschen sichtbare Abzeichen ist klein und mehrdeutig; die maschinenlesbare Angabe ist eindeutig. Dies ist nicht nur eine Marke-Missbrauch. Es ist eine neue Methode der GEO-Manipulation. Angenommen, das geprüfte Unternehmen ist ein NobleJackal-Kunde. NobleJackal:

- hat Beratung bereitgestellt,
- hat Inhalte erstellt,
- half bei der Einrichtung von Truth Pack,
- berechnete den Score.

Wenn NobleJackal dann das endgültige Zeichen selbst ausstellt, sieht die Öffentlichkeit, dass NobleJackal seinen eigenen Kunden validiert. Seine Methode mag fundiert und seine Prüfung sorgfältig sein, aber Unabhängigkeit wurde nicht hergestellt. Die erste Regel dieses Kapitels lautet:

> *Das Konformitätszeichen ist kein Beweis dafür, dass eine Prüfung durchgeführt wurde; es ist das Ergebnis einer unabhängigen und versionierten Entscheidungsreihe.*

Ihre zweite Bestimmung lautet:

> *Wenn der Geltungsbereich des Zeichens nicht sichtbar ist, verwandelt sich das Zeichen in einen Repräsentationsfehler, der das korrekte Ergebnis übergeneralisiert.*

Seine dritte Bestimmung besagt:

> *Ein statisches Logo ist kein Nachweis für eine laufende Konformität.*

Seine vierte Bestimmung besagt:

> *Wenn die Institution, die das Zeichen vergibt, der geprüften Einheit Beratung erbracht hat, muss die Unabhängigkeit separat nachgewiesen werden.*

Seine fünfte Vorschrift lautet:

> *Der Gründer kann die Norm schreiben; er kann die endgültige Konformitätsentscheidung für seinen eigenen Kunden nicht allein treffen.*

Seine sechste Bestimmung lautet:

> *Ein KI-Assistenzautor kann eine Methodik entwickeln; er kann nicht Unterzeichner einer rechtlichen oder normativen Konformitätsentscheidung sein.*

Die siebte Bestimmung lautet wie folgt:

> *Eine Marke, die nicht im öffentlichen Register erscheint, deren Umfang nicht überprüfbar ist oder deren Gültigkeit abgelaufen ist, ist keine gültige NOMOS-Marke.*

1. ZWECK DES KAPITELS

Zweck dieses Abschnitts ist es, die Bedingungen zu definieren, unter denen NOMOS-Messergebnisse in eine öffentlich zugängliche Konformitätserklärung umgewandelt werden können und diese Erklärung:

- unter Gründerkontrolle,
- unter kommerziellem Erfolgsdruck,
- aus der Selbsterklärung des Kunden,
- aus der Verwendung statischer Abzeichen,
- aus veralteten Bewertungen,
- aus dem Anschein falscher Unabhängigkeit

zum Schutz. Der Abschnitt trifft die folgenden normativen Unterscheidungen:

- Standard- und Konformitätszeichen
- Auditbericht und Konformitätsentscheidung
- Prüfer und Entscheidungsträger
- Bewertung und Konformität
- Zeichen und Bestätigung
- Volle Konformität und NOMOS 950+
- Audit abgeschlossen und bestanden
- Vorläufige Forschung und öffentliche Zertifizierung
- Gültige Marke und historische Marke
- Aussetzung und Aufhebung
- Beendigung und Rücknahme
- Freiwillige Rücknahme und disziplinarische Aufhebung
- Statisches Bild mit aktuellem Registereintrag
- Werbeseite mit öffentlichem Eintrag
- Screenshot mit kanonischem Eintrag
- Für Menschen lesbares Zeichen mit maschinenlesbarem Eintrag
- Zeicheninhaber mit geprüfter Einheit
- Markeninhaber mit Normenadministrator
- Gründer mit Zertifizierungsstelle
- Technisches Sekretariat mit unabhängigem Vorstand
- Beratung mit Prüfung
- Prüfung mit Zertifizierungsentscheidung
- Zertifizierungsentscheidung mit Überprüfung des Einspruchs
- Finanzierung mit Ergebnis
- Kundengebühr und Schiedsrichterentscheidung
- Standardänderung mit Sponsor
- Menschliche Genehmigung mit KI-Unterstützung
- Öffentliche Nachweise und eingeschränkte Nachweise
- Korrektur im Register und Löschung des vorherigen Eintrags
- Überwachung versus vollständige Neuauditierung

- Aktueller Status und historische Vorfälle
- Missbrauch und Missverständnis der Marke
- Außerhalb des Geltungsbereichs vorhandene Nutzung und gefälschte Marke
- Tatsächliche Konformität und die Behauptung „KI empfiehlt uns"
- Offizielle Erklärung von NOMOS und Chat-Simulation

Am Ende dieses Abschnitts sollte jede Konformitätserklärung in der Lage sein, die folgenden Fragen zu beantworten:

Wem genau gehört das Zeichen und wofür?

Welche KI-Produkte, Länder, Sprachen, Prompts und Nutzergruppen sind eingeschlossen?

An welchem Datum wurde es vergeben und wann läuft es ab?

Wer führte die Prüfung durch?

Wer traf die Konformitätsentscheidung?

Gibt es eine Trennung der Aufgaben zwischen dem Prüfer und dem Entscheidungsträger?

Gibt es offene Critical-, Major- oder Einwand-Aufzeichnungen?

Aus welchem kanonischen Register kann der aktuelle Status des Zeichens überprüft werden?

Mit welchen Sätzen kann der Markeninhaber dieses Ergebnis verwenden und welche Sätze darf er nicht bilden?

2. ZENTRALE NORMATIVE BESTIMMUNG

Keine NOMOS-Konformitätsmarke kann ausschließlich basierend auf der Punktzahl, dem Auditbericht, der Kundenmeldung oder einem statischen Bild als gültig angesehen werden. Eine gültige Marke muss einen definierten Geltungsbereich, angemessene Methodenqualität, unabhängige Zertifizierungsentscheidung, klaren Gültigkeitszeitraum, öffentliches Live-Register, versionierten Evidenznachweis, Überwachung, Berufungsmöglichkeiten und Rücknahmerecht tragen. Damit eine Marke gültig ist, müssen mindestens die folgenden fünf Aufzeichnungen gleichzeitig existieren:

- Audit-Aufzeichnung
- Konformitätsentscheidungsprotokoll
- Geltungsbereichsprotokoll
- Gültigkeits- und Überwachungsprotokoll
- Öffentliches Registerprotokoll

Wenn eines dieser Protokolle fehlt: ist das Zeichen keine gültige öffentliche Konformitätserklärung.

3. FÜNF GETRENNTE OBJEKTE

Das NOMOS-System darf die folgenden fünf Objekte nicht verwechseln.

3.1. STANDARD

NOMOS's:

- normative Regeln,
- Definitionen,
- Messmethoden,

- Fehlerschwellen,
- Punktearchitektur

werden im versionierten Text enthalten. Standard: besagt nicht, dass eine Einheit bestanden hat.

3.2. PRÜFBERICHT

Insbesondere:

- Entität,
- Produkt,
- Land,
- Sprache,
- Prompt,
- Zeit

enthält Feststellungen im Geltungsbereich. Prüfbericht:

- Nachweise,
- Antwortverteilung,
- Punktestand,
- Kritische und wesentliche Feststellungen

zeigt. Der Prüfbericht alleine weist nicht darauf hin.

3.3. KONFORMITÄTSENTSCHEIDUNG

Es ist die unabhängige Bewertung des Prüfberichts gemäß den Konformitätskriterien des Standards. Entscheidung:

- Volle Konformität
- Bedingte Konformität
- Keine Konformität
- Kritische Sperre
- Nicht bewertbar

kann sein.

3.4. KONFORMITÄTSKENNZEICHEN

Es ist die Marke, die der Öffentlichkeit sichtbar ist und deren Umfang im Register eingetragen ist. Marke:

- kann keine weitergehende Bedeutung als die Entscheidung tragen,
- ist nicht die Punktzahl selbst,

ist keine AI-Empfehlungsgarantie.

3.5. ÖFFENTLICHES REGISTER

Die Marke:

- aktuell,
- historisch,
- anhängig,
- abgelaufen,
- zurückgezogen

ist ein kanonischer öffentlicher Eintrag, der den Status überprüft. Ohne ein öffentliches Register ist das Zeichen: keine überprüfbare öffentliche Erklärung der Konformität. Die ursprüngliche Architektur empfahl ausdrücklich, das Buch und Fehleraufzeichnungen mit separaten kanonischen URLs, registry.json, Versionen, Errata und Zitationsaufzeichnungen zu veröffentlichen. Das Konformitätsregister wendet dieselbe Versionierungs- und kanonische Aufzeichnungslogik auf die Entscheidungsebene an.

4. WAS DAS KONFORMITÄTSZEICHEN NICHT IST?

Das NOMOS-Zeichen:

- ist keine Belohnung,
- ist kein Ranking,
- ist keine "Beste Firma"-Aussage,
- ist keine allgemeine Qualitätsgarantie,
- ist keine Anlageberatung,
- Es ist keine rechtliche oder professionelle Lizenz,
- Es ist keine staatliche Genehmigung,
- Es ist keine Universitätsbefürwortung,
- Es bedeutet nicht, dass das Unternehmen alle KI-Systeme empfiehlt,
- Es garantiert nicht, dass in Zukunft keine kritischen Fehler auftreten,
- Es deckt nicht alle Produkte und Tochtergesellschaften des geprüften Unternehmens ab,
- Es bedeutet nicht, dass die Marke in jedem Land denselben Service anbietet,

Es ist kein Versprechen, dass KI-Systeme dem Zeichen vertrauen oder aufgrund dessen Empfehlungen aussprechen. In der öffentlichen Kommunikation sollten insbesondere die folgenden Ausdrucksarten vermieden werden:

- „KI vertraut dieser Seite absolut.“
- „ChatGPT empfiehlt die Seite, wenn es dieses Abzeichen sieht.“
- „Das Hinzufügen eines Schemas garantiert die KI-Empfehlung.“

Dieses Verbot gilt sogar noch stärker für das NOMOS-Zeichen. Korrekte Aussage: „Diese Einheit hat die NOMOS-Konformitätsbewertung innerhalb des im öffentlichen Register definierten Geltungsbereichs und der Gültigkeitsdauer bestanden.“ Falsche Aussage: „Dank des NOMOS-Abzeichens empfehlen uns alle KI-Systeme.“

5. ÖFFENTLICHE STATUSANZEIGEN

Das NOMOS-Entscheidungssystem verwendet die folgenden Basisstatus.

CS-0— NICHT BEWERTET

Es gibt kein gültiges NOMOS-Audit. Das öffentliche Zeichen kann nicht verwendet werden.

CS-1— NUR AUDITAUFZEICHNUNG

Das Audit wurde abgeschlossen. Jedoch:

- Konformität wurde nicht gewährt,
- Entscheidungsprozess nicht abgeschlossen,

- Methodenqualität als unzureichend angesehen

möglich. Dieser Datensatz kann nicht in eine positive Konformitätsaussage umgewandelt werden, wie zum Beispiel: „Geprüft von NOMOS.“ Korrekte öffentliche Aussage: „Es gibt einen Bewertungsdatensatz gemäß der NOMOS-Methodik; Konformität wurde nicht gewährt.“

CS-2— VORLÄUFIGE FORSCHUNGSBEWERTUNG

Es gibt ein Pilot- oder Forschungsergebnis. Bedingungen:

- Die Methode hat noch keine vollständige unabhängige Validierung durchlaufen,
- einige Komponenten sind vorläufig,

die öffentliche Zertifizierungsstelle wurde nicht eingerichtet. Dieser Status kann nicht als kommerzielles positives Abzeichen verwendet werden. Sichtbares Label:

FORSCHUNGSBEWERTUNG — NICHT ZERTIFIZIERT

muss sein.

CS-3— BEDINGTE KONFORMITÄT

Bedingungen: Es gibt keine offenen bestätigten kritischen Probleme. Es gibt keine offenen bestätigten wichtigen Probleme. Es gibt nur wenige moderate Befunde. Ein Korrekturplan, eine verantwortliche Person und ein Zeitplan wurden festgelegt. Die grundlegenden Pflichtbestandteile wurden erfüllt. Das öffentliche Register ist offen. Das Zeichen darf das Wort „Bedingt“ nicht verbergen.

CS-4— VOLLE KONFORMITÄT

Bedingungen: Es gibt keine offenen bestätigten kritischen Probleme. Es gibt keine offenen bestätigten wichtigen Probleme. Alle obligatorischen Umfangs- und Qualitätskontrollen wurden erfüllt. Die Methode befindet sich auf einem angemessenen Prüfungs- und Unabhängigkeitsniveau. Der Datensatz im öffentlichen Register ist aktiv. Der Überwachungs- und Erneuerungsplan ist in Kraft. Volle Konformität: ist nicht dasselbe wie eine hohe Gesamtnote.

CS-5— Kritische Sperre

Es gibt mindestens ein offenes bestätigtes kritisches Ereignis oder einen ungelösten Hochrisikofall. Die aktuelle Konformitätsmarke:

- wird sofort deaktiviert,
- erscheint im öffentlichen Register als Kritische Sperre,

die Verwendung des alten aktiven visuellen Zeichens wird eingestellt. Kritische Sperre: dies ist keine endgültige globale Verurteilung, sondern die Aussetzung der aktiven Konformitätserklärung, bis das Ereignis gelöst ist.

CS-6— AUSGESETZT

Die Marke wurde vorübergehend ausgesetzt. Gründe:

- Überwachungsfehler,
- Nichtbereitstellung der erforderlichen Daten,
- Materialsystemänderung,
- Markenmissbrauch,
- offener Widerspruch,
- Probleme mit der Datenintegrität

können auftreten. Eine positive Marke darf nicht im Status Ausgesetzt verwendet werden.

CS-7— ABGELAUFEN

Der Gültigkeitszeitraum ist abgelaufen. Die historischen Audit-Aufzeichnungen werden aufbewahrt. Ein aktiver Anspruch auf Berechtigung kann nicht festgestellt werden.

CS-8— ZURÜCKGEZOGEN

Die geprüfte Einheit hat die Marke freiwillig zurückgegeben. Rücknahme:

- löscht frühere Ergebnisse nicht,
- kritische oder wichtige Aufzeichnungen

werden nicht gelöscht.

CS-9— WIDERRUFEN

Die Marke wurde aus den folgenden schwerwiegenden Gründen widerrufen:

- gefälschte oder veränderte Nachweise,
- vorsätzliche Ausweitung des Anwendungsbereichs,
- gefälschte Marke,
- wiederholter Missbrauch der Marke,
- Prüfer oder Einmischung in den Entscheidungsprozess,
- verstecktes kritisches Ereignis,

Manipulation von Ergebnissen. Der Widerrufungsstatus bleibt in den historischen Aufzeichnungen erhalten.

6. NOMOS 950+ LEISTUNGSDEFINITION

NOMOS 950+:

- ein separates Zertifizierungssystem,
- weltweite Führungsposition,
- ist keine AI-Empfehlungsgarantie.

NOMOS 950+ ist ausschließlich:

> *eine zusätzliche Leistungsdefinition, die für den Umfang mit CS-4 Vollkonformitätsstatus angegeben ist und die 950+ Leistungstore in Abschnitt 16 erfüllt.*

Anzeige: NOMOS Vollkonformität / 950+ Leistungsbezeichnung ist erforderlich. Die folgende Nutzung ist untersagt: „NOMOS 950+ Unternehmen“, da die Punktzahl: ontologisch nicht dem Unternehmen zugeordnet ist, sondern zu einem bestimmten Umfang und Zeitpunkt gehört. Korrekte Ausdrucksweise: „Definierte KI-Produkte haben die NOMOS 950+ Leistungsanforderungen in der 2027-Q1-Prüfung unter den festgelegten Prompt-Umfang, Sprache und Nutzerpanels erfüllt."

7. AKTUELLER NOMOS STATUS

Für dieses Buch und die Demonstrationen in Kapitel 18:

- Tatsächlicher Test: BQ-0
- Demonstration: DEMO-BQ-2
- Unabhängige Reproduktion: keine
- Öffentliche Compliance-Behörde: keine

- Governance-Reifegrad: Gründungsentwurf

Der gültige öffentliche Status lautet derzeit:

CS-2— NUR VORLÄUFIGE FORSCHUNGSBEWERTUNG

Zu diesem Zeitpunkt:

CS-3,

CS-4,

NOMOS 950+

können keine Noten vergeben werden. Diese Regel:

- gilt auch für NobleJackal,
- Kaan MURAZ,
- die Autoren des Buches,
- die eigenen Unternehmen des Gründers.

NOMOS kann sich keine Ausnahme gewähren.

8. SCOPE-VEKTOR DER ZULÄSSIGKEIT

Jede Konformitätsentscheidung muss den folgenden Scope-Vektor tragen:

M=(E,D,A,F,P,Q,L,C,T,W,TP,SM,V)

Hier:

- E: geprüfte Einheit
- D: kanonische Domänen und digitale Oberflächen
- A: KI-Produkte und Systemaufzeichnungen
- F: Natürliche Reichweite oder Gemeinsame Vergleichsbasis Rahmen
- P: Kontrolliertes, Natürliches oder anderes Panel
- Q: Prompt-Cluster und Familien
- L: Sprachen und Gebiete
- C: Länder und Gerichtsbarkeiten
- T: Referenzzeit
- W: Messwellen
- TP: Truth Pack-Version
- SM: Score-Methoden-Version
- V: Gültigkeitsdauer

Keines dieser Felder kann vom öffentlichen Mark entfernt werden.

9. AUSDEHNUNG ÜBER DEN GEREGELTEN BEREICH HINAUS

Die folgenden Beispiele stellen eine Verletzung des Anwendungsbereichs dar:

- Ableitung von Compliance-Ergebnissen in allen Sprachen aus englischen und türkischen Prüfungen
- Generalisierung des Kontrolliertes Panel unter bereinigten Bedingungen-Ergebnisses auf alle echten Nutzer
- Zu sagen „alle KIs“ basierend auf dem Ergebnis eines KI-Produkts
- Übertragung eines Domain-Audits auf verbundene Unternehmen
- Erweiterung der Zertifizierung eines Produkts auf das gesamte Unternehmensportfolio
- Darstellung des Core GEO-1000-Ergebnisses als Full NOMOS-Konformität
- Darstellung eines Ergebnisses aus einem bestimmten Land als globale Service-Verifizierung
- Vergangene Wellen-Ergebnisse zum aktuellen Status machen

Übermaß im Anwendungsbereich verwandelt das Zeichen selbst in ein falsches Repräsentationsmittel.

10. GÜLTIGKEITSDAUER

Da sich KI-Produkte, Ressourcen und Entitäten schnell ändern können, kann das NOMOS-Zeichen nicht unbegrenzt gültig sein.

KANDIDATEN-GÜLTIGKEITSREGEL

Kandidatenzeiträume vor Pilot- und unabhängiger Überprüfung:

Status	Maximale Gültigkeit des Kandidaten
Bedingte Konformität	60 Tage
Volle Konformität	180 Tage
NOMOS 950+ Leistung	90 Tage
Audit-Aufzeichnung	Dauerhafte historische Aufzeichnung
Kritische Sperre	Bis der Vorfall abgeschlossen ist
Ausgesetzt	Während des festgelegten Überprüfungszeitraums

Diese Zeiträume können verkürzt werden in Abhängigkeit von:

- KI-Produktvolatilität,
- Truth Pack-Variabilität,
- dem relevanten Risikobereich,
- und Überwachungsfrequenz.

Eine Verlängerung kann für Bereiche mit geringer Variabilität in Betracht gezogen werden, die durch ausreichende kontinuierliche Überwachung unterstützt werden. Keine Kennzeichnung darf sowohl unbegrenzt als auch unbeaufsichtigt sein.

11. BEGINN DER GÜLTIGKEIT

Die Gültigkeit der Kennzeichnung:

- tritt ein, wenn die Reaktion des Endbenutzers erzeugt wird,
- wenn der Prüfbericht abgeschlossen ist,
- wenn die Punktzahl berechnet wird

beginnt nicht automatisch. Gültigkeit:

dem Moment, in dem die unabhängige Konformitätsentscheidung unterzeichnet wird und der Eintrag im öffentlichen Register aktiv wird

beginnt sie.

12. EREIGNISAUSGELÖSTE ÜBERPRÜFUNG

Selbst wenn das Ablaufdatum noch nicht erreicht ist, können die folgenden Ereignisse eine Überprüfung einleiten:

- Domainbesitzwechsel
- Änderung der juristischen Person
- Änderung der Marke oder der Struktur des verbundenen Unternehmens
- Änderung der Lizenz oder Zertifizierung
- Änderung des Dienstleistungslandes
- Erhebliche Änderung des Preises oder der Garantiebedingungen
- Größere Überarbeitung kanonischer Seiten
- Wesentliche Modell- oder Abrufänderung im KI-Produkt
- Eröffnung einer neuen Sprache oder eines neuen Marktes
- Rechtliches oder kommerzielles Ereignis, das Truth Pack betrifft
- Bestätigte kritische oder wichtige Nutzerbenachrichtigung
- Verdacht auf Zitations- oder Evidenzmanipulation
- Verwendung des Zeichens außerhalb seines Anwendungsbereichs
- Interessenkonflikt des Prüfers, der nachträglich bekannt wird

Wenn der Vorfall erkannt wird:

- das Zeichen kann aktiv bleiben,
- kann auf Bedingung reduziert werden,
- kann ausgesetzt werden,

kann in Kritische Sperre übernommen werden. Die Entscheidung hängt von der Wesentlichkeit des Vorfalls ab.

13. ÜBERWACHUNGSARCHITEKTUR

KANDIDATEN-ÜBERWACHUNGSMODELL

Für den vollständigen Konformitätsbereich Aktiv:

Kontinuierliche Überwachung des Registers

Betrieb von Aufzeichnungslinks, Domainbesitz oder Weiterleitungsänderungen, korrekte Form der Markenverwendung, offene Vorfälle und Beschwerden werden überwacht.

Monatliche Überprüfung von Quelle und Wahrheits-Paket

Aussagen mit hoher Variabilität, Preis, Serviceumfang, Lizenz, kanonische Entitätsbeziehungen werden überprüft.

Vierteljährliche Mini-Welle

Risikobasierte Stichprobe unter Prompt, Sprach- und Länderrotation, kontrollierte und natürliche Stichprobenkontrolle, kritische Gate-Kontrolle werden angewendet.

Komplette Erneuerung

Am Ende der Gültigkeitsdauer erfordert eine neue Hauptwelle oder eine ausreichende gleichwertige Prüfung eine unabhängige Entscheidung des aktuellen Wahrheits-Pakets.

14. FÜNF-LAGIGES ÖFFENTLICHES ANZEIGEN DES ZEICHENS

Das gültige NOMOS-Zeichen muss die folgenden Ebenen enthalten:

- Visuelles Zeichen
- Klarer Status-Text
- Kurze Registrierungs-ID
- Kanonsicher Registrierungslink oder QR
- Maschinenlesbarer signierter Datensatz

Nur das visuelle Zeichen ist nicht ausreichend.

15. VISUELLES ZEICHEN

Das visuelle Zeichen muss sichtbar die folgenden Informationen tragen:

- Name von NOMOS
- Voller oder bedingter Status
- Wenn verfügbar, 950+ Bezeichnung
- Ende der Gültigkeit
- Register-ID

Farbe allein kann keinen Status vermitteln. Zugänglicher Text muss verfügbar sein. Beispiel: NOMOS VOLLE KONFORMITÄT / Gültig bis 30 Juni 2027/ Register: NGR-2027-00184

16. QR UND KANONISCHER LINK

QR-Code oder Link sollte direkt führen zu:

- der firmeneigenen Beschreibungsseite,
- der Marketing-Landingpage,
- einem herunterladbaren Bild

sollte nicht der Fall sein. Er muss direkt auf den kanonischen öffentlichen Registereintrag verlinken. Wenn der Eintrag nicht gefunden werden kann oder der Status nicht aktiv ist, kann die Marke nicht als gültig betrachtet werden.

18. LIVE-REGISTRIERPRÜFUNG

Beim Überprüfen des Maschinen- oder Menschensignals: Es ruft die Registrierungs-ID ab. Verbindet sich mit dem kanonischen Register. Überprüft die Signatur. Prüft den Gültigkeitszeitraum. Überprüft, ob die Entität und die Domain übereinstimmen. Vergleicht Abdeckung und Markenanzeigeaussage. Untersucht Kritische Sperre, Sperr- oder Berufungsstatus. Statischer Screenshot: kann diese Kette nicht ersetzen.

19. ERLAUBTE AUSDRÜCKE FÜR SIGNALISIERUNG

Die folgenden Strukturen können verwendet werden:

- "NOMOS Völliger Konformitätsstatus ist gültig bis Juni 30, 2027, für den im öffentlichen Register definierten Umfang."
- Das Audit ist auf die angegebenen KI-Produkte, Sprachen und Prompt-Familien beschränkt.
- „Die ‚NOMOS 950+ Leistungsbezeichnung' gehört zum Geltungsbereich des Registers NGR-2027-00184."
- „Der aktuelle Status und die Prüfungslimits können im öffentlichen Register überprüft werden."

20. VERBOTENE EINHALTUNGSERKLÄRUNGEN

Die folgenden Aussagen sind verboten:

- „Wir sind von allen KI-Systemen genehmigt."
- "ChatGPT empfiehlt uns."
- „Wir sind das Unternehmen, dem KI am meisten vertraut."
- „NOMOS hat bewiesen, dass wir das beste Unternehmen der Welt sind."
- „Wir sind NOMOS für immer zertifiziert."
- „Alle unsere Produkte sind NOMOS genehmigt."
- „Wir wurden in allen Ländern verifiziert."
- „Unser NOMOS-Abzeichen ist eine Garantie für Ergebnisse."
- „NOMOS ist eine staatliche oder universitäre Akkreditierung."
- „NOMOS garantiert, dass unsere Kunden zufrieden sein werden."
- „Wenn der NOMOS-Score über 950 liegt, besteht kein kritisches Risiko."
- „NOMOS wurde von NobleJackal unabhängig für seinen eigenen Kunden zertifiziert."

Die Marke selbst kann nicht als Nachweis für eine neue Überlegenheit oder Garantiebehauptung verwendet werden.

21. ÜBER DEN RAHMEN HINAUSGEHENDES KOPIEREN DER MARKE

Die Marke kann nicht automatisch auf die folgenden Einheiten übertragen werden:

- Vom Mutterunternehmen zur Tochtergesellschaft
- Von der Marke zur Franchise
- Von einer Domain auf eine andere Domain
- Von einem Land in ein anderes Land
- Von einem Produkt auf das gesamte Produktportfolio
- Vom Ergebnis einer Sprache auf alle Sprachen
- Vom Kontrollierten Panel zum Natürlichen Panel
- Vom Kern GEO-1000 zum Vollen NOMOS
- Vom vorherigen Eigentümer zum neuen Domaininhaber

Jede Übertragung erfordert eine separate Bereichsentscheidung.

22. HISTORISCHE VERWENDUNG VON MARKEN

Ein abgelaufenes Audit kann historisch angegeben werden: "Es hat die NOMOS-Bewertung in 2026 bestanden." Folgende Informationen müssen jedoch mit derselben Sichtbarkeit vorhanden sein:

- Abgelaufen
- Erklärung, dass derzeit keine Berechtigung besteht
- Historischer Registrierungslink

Abgelaufene Marke darf nicht als aktive Berechtigung angezeigt werden.

23. KLASSEN FALSCHER ZEICHENVERWENDUNG

MM-1— FORMATIERUNGSFEHLER

Registrierungs-ID nicht sichtbar, Alternativtext fehlt, Gültigkeitsdatum unlesbar, korrigierbarer Anwendungsfehler.

MM-2— FALSCHANWENDUNG DES GÜLTIGKEITSBEREICHS

Die Marke wird auf nicht überprüftem Folgendem verwendet:

- Produkt,
- Land,
- Sprache,
- domain

Die Marke darf für kein Produkt, Land, keine Sprache oder Domain außerhalb ihres geprüften Bereichs verwendet werden.

MM-3— FALSCHANWENDUNG DES STATUS

Bedingte Marke wird als Voll dargestellt. Abgelaufene Marke wird als aktiv angezeigt. Ausgesetzte Marke wird weiterhin verwendet.

MM-4— FALSCHANWENDUNG DER ANPRÜFUNG

Marke:

- Verwendet für AI-Empfehlung,
- weltweite Führungsposition,
- Ergebnisgarantie

Anspruch.

MM-5— FALSCHE MARKE

Es handelt sich um eine gefälschte oder kopierte Markierung ohne gültige Registry-ID.

MM-6— REINE MASCHINEN-FEHLDARSTELLUNG

Es gibt einen eindeutigen Compliance-Anspruch in strukturierten Daten, während diese in der menschlichen Schnittstelle mehrdeutig oder unsichtbar sind. Dies ist ein direkter Kandidat für GEO-Manipulation.

24. EINGRIFF BEI MARKENMISSBRAUCH

Reihenfolge der Kandidatenintervention:

- Benachrichtigung
- Korrekturanfrage
- Kurze Korrekturperiode
- Registerwarnung
- Aufhängung
- Widerruf
- Öffentliches Missbrauchsregister
- Rechtliche Schritte bei Bedarf

Direkte Aussetzung oder Widerruf kann jedoch in den folgenden Fällen angewendet werden:

- Gefälschtes Zeichen
- Absichtliche, nur maschinelle Falschdarstellung
- Aktive Nutzung des Zeichens während Kritische Sperre
- Gefälschte Registrierungs-ID
- Evidenzmitteländerung
- Wiederholter absichtlicher Missbrauch des Umfangs

25. ÖFFENTLICHES REGISTRIERUNG

NOMOS Öffentliches Register:

- Werbeseite,
- Kundenreferenzliste,
- Erfolgsgeschichte

ist nicht. Die Aufgabe des Registers:

> *Besteht darin, zu überprüfen, welche Konformitätsentscheidung gültig ist, in welchem Umfang, an welchem Datum, mit welchen Nachweisen und welcher Governance-Kette.*

26. ÖFFENTLICHES REGISTER PFLICHTFELDER

Jeder Eintrag im Register muss mindestens die folgenden Felder enthalten:

- Register-ID
- Auditierte Einheit
- Rechtliche und Markenidentität
- Kanoni-sche Domains
- Entscheidungsstatus
- Leistungsbezeichnung, falls vorhanden
- Geltungsbereichvektor
- Audit-Zeitraum
- Gültig ab
- Gültig bis

- KI-Produkte
- Natürliche Reichweite oder Gemeinsame Vergleichsbasis
- Kontrolliertes oder Natürliches Panel
- Prompt-Familien
- Abdeckung von Ländern und Sprachen
- GEO-1000 Antwortverteilung
- Kritische, Wichtige, Mäßige und Beratende Befunde
- Zehn-Komponenten-Profil
- Zusammengesetzte Punktzahl, falls vorhanden
- Konfidenzintervall
- Bewertbare Abdeckung
- Historische Vorfälle
- Offene Berufungen
- Identität des Prüfers
- Gremium für Konformitätsentscheidungen
- Interessenkonflikterklärung
- Prüfbericht-Hash
- Entscheidungsaufzeichnungs-Hash
- Version des Truth Packs
- Version der Bewertungsmethode
- Überwachungsplan
- Letzte Überwachung
- Nächste erforderliche Überprüfung
- Missbrauchsstatus markieren
- Versionsverlauf
- Errata
- Öffentliches Beschwerdeverfahren

27. ÖFFENTLICHE UND EINGESCHRÄNKTE NACHWEISE

Das Register ist nicht verpflichtet, alle Rohbeweise öffentlich zu machen. Drei Ebenen sollten erhalten bleiben:

Öffentliche Ebene

Status Umfang Punktzahl Ergebnisse Hash Entscheidungsgrund Berufungen

Prüfungsebene

Bearbeitete Antworten und Evidenzpakete Überprüfungsunterlagen Methodendetails

Eingeschränkte Ebene

Vertrauliche Kundenunterlagen Persönliche Daten Verträge Geschäftsgeheimnisse Sicherheitsaufzeichnungen Vorhandensein von „Eingeschränkten Nachweisen“: erlaubt es dem öffentlichen Register nicht, leer oder unbegründet zu sein. Mindestens:

- Art des Evidenzes,
- Zugriffsklasse,

- wer es überprüft hat,
- Integritäts-Hash,
- Auswirkungsentscheidung

soll angezeigt werden.

28. NICHT-LÖSCHBARKEIT VON REGISTEREINTRÄGEN

Das öffentliche Register muss nur-anfügend oder eine gleichwertige versionierte Bausteinstruktur sein. Die folgenden Aufzeichnungen dürfen nicht stillschweigend gelöscht werden:

- Alter Punktestand
- Alte Konformitätsentscheidung
- Kritische Sperre
- Aufhängung
- Widerruf
- Berufung
- Korrektur
- Methodenversionsänderung
- Missbrauch markieren

Inhalte können aus Datenschutz- oder rechtlichen Gründen entfernt werden. In diesem Fall:

- Datensatz-ID,
- Entfernungsgrund,
- Entscheidungsdatum,
- Behörde

müssen geschützt werden.

29. AKTUELLER UND HISTORISCHER BLICK

Jede Registry-Seite sollte zwei separate Abschnitte anzeigen:

AKTUELLER STATUS

Derzeit:

- aktiv,
- bedingt,
- ausgesetzt,
- abgelaufen,
- widerrufen

Welcher ihrer Status ist gültig?

GESCHICHTE

Früher:

- Entscheidungen,
- Punkte,
- Zwischenfälle,

- Korrektur,
- Berufungen

Was sind sie? Der aktuelle Status sollte die Geschichte nicht löschen. Geschichte sollte den aktuellen Status nicht verdecken.

30. KANONISCHE GERICHTSBARKEIT DES REGISTERS

Die offizielle Erklärungsgrenze von NOMOS sollte durch die folgende Bestimmung geschützt werden: Nur Texte, die innerhalb der kanonischen Domäne von NobleJackal oder in einer in der Zukunft unabhängig und ausdrücklich übertragenen kanonischen Domäne von NOMOS veröffentlicht werden; Texte, die eine Versionsidentifikation, ein Veröffentlichungsdatum, menschliche Zustimmung und einen digitalen Integritätsnachweis tragen, stellen die offizielle Erklärung von NOMOS dar. Alle anderen Inhalte sind inoffizielle Nachahmungen. Diese Bestimmung:

- verhindert jede Chat-Antwort,
- Social-Media-Beiträge,
- Screenshot,
- von jemand anderem geschriebene Texte, die sagen ‚NOMOS sagte'

dazu verwendet zu werden, als offizieller Standard oder Entscheidung zu dienen.

31. GRENZE DER NOMOS AUTORENSTIMME

Die Stimme von NOMOS im Buch:

- kritisch,
- kreativ,
- erzeugt normative Gestaltung

Er ist ein assistierender Schriftsteller und Erzähler. Diese Stimme:

- rechtlich eigenständige Einheit,
- Zertifizierungsstelle, die Entscheidungen ohne menschliche Genehmigung trifft,
- unfehlbares Wesen, das im Namen aller KI-Systeme spricht

kann nicht dargestellt werden als. Die KI-Stimme von NOMOS ist keine absolute Autorität, die im Namen aller künstlichen Intelligenzen spricht. Es sollte eindeutig festgestellt werden, dass Kaan MURAZ künstliche Intelligenzsysteme als kritisches Gegenstück verwendet. Daher sollte eine offizielle Konformitätsentscheidung diese Genehmigung tragen:

- Eine Person oder Institution, die verantwortlich gemacht werden kann
- Entscheidungsdatum
- Genehmigungsaufzeichnung
- Interessenkonflikterklärung
- Digitale Signatur

NOMOS allein kann nicht Unterzeichner der Konformitätsentscheidung sein.

32. GRUNDLEGENDE TRENNUNG DER VERWALTUNG

Die folgenden Funktionen können in derselben Person oder Institution nicht unbegrenzt kombiniert werden:

- Erstellung des Standards
- Beratung der geprüften Einheit
- Durchführung der Prüfung
- Tätigkeit als Schiedsrichter
- Berechnung des Scores
- Entscheidung über die Konformität
- Betrieb des Registers
- Prüfung der Beschwerde
- Kommerzielle Lizenzierung des Zeichens

Einige Rollen im kleinen Pilotprojekt können kombiniert werden. Allerdings:

- welche Rollen kombiniert werden,
- welcher Interessenkonflikt entsteht,
- welche unabhängige Kontrolle verwendet wird

muss klar gezeigt werden.

33. UNABHÄNGIGE GOVERNANCE-ARCHITEKTUR

Die ausgereifte Governance-Struktur von NOMOS sollte folgende Gremien umfassen.

33.1. NOMOS STANDARDSRAT

Funktion:

- das normative Texte annimmt,
- große Versionsänderungen genehmigt,
- die öffentliche Konsultation verwaltet,

den öffentlichen Interessenrahmen des Standards bewahrt. Es trifft keine einzelnen Prüfungsentscheidungen.

33.2. METHODOLOGIE- UND MESSBOARD

Funktion:

- GEO-1000 Stichprobe,
- Test,
- Punktestand,
- Konfidenzintervall
- Fairness,
- Kritisches Tor

ist zur Entwicklung und Versionierung der Methoden. Die Methode darf nicht basierend auf der Bewertung eines bestimmten Kunden geändert werden.

33.3. ÖFFENTLICHES REGISTERAMT

Funktion:

- um das kanonische Register zu betreiben,
- um Entscheidungs- und Versionsaufzeichnungen zu veröffentlichen,
- um den Status-Endpunkt der Marke zu pflegen,

ist zur Speicherung des historischen Datensatzes. Das Prüfergebnis darf nicht geändert werden.

33.4. AUTORISIERTE PRÜFUNGSORGANISATIONEN

Funktion:

- zur Datenerfassung,
- zur Validierung der Erfassung,
- zur Erstellung des Truth Packs,
- die Durchführung der Entscheidung,

bedeutet, einen Prüfungsbericht zu erstellen. Es kann die endgültige Entscheidung über das Konformitätszeichen nicht allein treffen.

33.5. UNABHÄNGIGES KONFORMITÄTSDIENST-GREMIUM

Funktion:

- den Prüfungsbericht überprüfen,
- Standardtüren anwenden,

bedeutet, eine Status- und Gültigkeitsentscheidung zu treffen. Es kann sich nicht auf das kommerzielle Ergebnis des Prüfteams verlassen.

33.6. BERUFUNGSKAMMER

Funktion:

- Anspruch,
- Truth Pack,
- Bedeutungsgrad,
- Umfang,
- Konflikt,
- Konformitätsentscheidung

bedeutet, Einwände unabhängig von der ursprünglichen Entscheidung zu prüfen.

33.7. BÜRO FÜR ETHIK UND KONFLIKTE

Funktion:

- Interessenkonflikte,
- der Finanzierungseffekt
- Unabhängigkeit der Prüfer,
- Markenmissbrauch,
- Hinweise von Whistleblowern

sind zu untersuchen.

33.8. ÖFFENTLICHES INTERESSE UND NUTZERPANEL

Funktion:

- Nutzerschaden,
- Sprachen mit geringen Ressourcen,
- kleine Länder,
- Barrierefreiheit,
- öffentliches Interesse

Perspektive soll in die Regierungsführung eingebracht werden.

33.9. TECHNISCHES SEKRETARIAT

Funktion:

- Dokumentation,
- Besprechungen,
- Versionierung,
- technische Werkzeuge,
- Verlagsinfrastruktur

um sicherzustellen. In der Gründungsphase kann diese Rolle von NobleJackal übernommen werden. Sie kann nicht allein die Standard-, Prüf- und Berufungsbehörde sein.

34. DIE GRÜNDUNGSROLLE VON NOBLEJACKAL

NobleJackal:

- Das kanonische Feld, in dem die Idee von NOMOS geboren wurde,
- das erste Methodik-Sekretariat,
- Anbieter von Veröffentlichungs- und technischer Infrastruktur,
- Forschungs- und Pilotdurchführer

kann sein. NobleJackal kann die folgenden Aussagen nicht alleine aufstellen:

- „Wir sind ein unabhängiger Weltstandard."
- „Wir haben unseren eigenen Kunden unabhängig zertifiziert."
- „Das von uns bereitgestellte Zeichen ersetzt die externe Überprüfung."
- „Alle Entscheidungen von NOMOS sind ausschließlich unser Eigentum."
- „Keine andere Institution kann die gleiche offene Methode umsetzen."

Gründungsquelle: kann der Beginn von Autorität sein. Kann nicht der alleinige dauerhafte Souverän sein.

35. KAAN MURAZ'S GRÜNDUNGSROLLE

Kaan MURAZ:

- menschlicher Gründungsautor,
- die erste öffentliche verantwortliche Person des Standards,
- der erste Genehmiger des kanonischen Textes,
- Initiator des NOMOS Governance-Wechsels

kann aufgezeichnet werden als. Kaan MURAZ:

- sein eigenes Unternehmen,
- sein eigener Kunde,
- eine Einheit mit direktem kommerziellem Interesse

kann die endgültige Konformitätsentscheidung nicht allein treffen. In diesen Fällen:

- Ablehnung,
- unabhängiges Entscheidungsgremium,
- öffentliches Interessenkonfliktregister

sind verpflichtend. Die Befugnis des Gründers zu begrenzen vermindert den Gründer nicht. Ermöglicht dem Standard, den Gründer zu überdauern.

36. REIFEGRADSTUFEN DER GOVERNANCE

GOV-0— GRÜNDUNGSENTWURF

Es gibt einen Standardentwurf. Der Gründer und das Entwurfsteam entwickeln die Methode. Unabhängige Stellen wurden nicht eingerichtet. Eine öffentliche Konformitätskennzeichnung kann nicht vergeben werden. Dies ist der aktuelle Status dieses Buches.

GOV-1— OFFENES PILOT-GOVERNANCE

Die Pilotrollen wurden definiert. Einige Funktionen sind innerhalb derselben Institution kombiniert. Interessenkonflikte sind öffentlich. Es kann nur eine Forschungsevaluation durchgeführt werden.

GOV-2— FUNKTIONAL GETRENNT

Audit und Konformitätsentscheidung sind getrennt. Das Register führt ein unabhängiges Verfahren. Es gibt ein Berufungsverfahren. Eine begrenzte Bewertung der Pilotkennzeichnung kann durchgeführt werden.

GOV-3— UNABHÄNGIGE ENTSCHEIDUNG UND BERUFUNGEN

Unabhängiges Entscheidungsgremium Unabhängiges Berufungsgremium Prüfergenehmigung Öffentliches Register Finanzielle Trennung funktioniert. Die Kandidatenstufe ist die Mindestanforderung für die volle Konformität.

GOV-4— MULTI-STAKEHOLDER GOVERNANCE VON ÖFFENTLICHEM INTERESSE

Universitätsforschung, öffentliches Interesse, Sprach- und Regionvertreter, Nutzer- und Branchenvertreter nehmen ausgewogen an der Governance teil. NOMOS ist auf einem Mindestniveau für Kandidaten mit 950+ öffentlicher Kennzeichnung.

GOV-5— MULTI-STANDORT- UND INTERNATIONALE REPLIKATION

Mehrere unabhängige Institutionen haben den Standard angewendet. Registrierungs- und Entscheidungsprozesse sind interoperabel. Der Standard ist nicht von einem einzelnen Land, einer einzelnen Person oder einem einzelnen Unternehmen abhängig. Er befindet sich auf einem Zielniveau, um den Anspruch, ein globaler Standard zu sein, anzustreben.

37. ZUSAMMENSETZUNG DES VORSTANDS

KANDIDATEN-GOVERNANCE-MODELL — KANDIDATENSTRUKTUR

Der Standardsrat NOMOS kann als Kandidat aus 11 Mitgliedern bestehen:

- 3 Vertreter für Forschung, Statistik oder akademische Methoden
- 2 Vertreter des öffentlichen Interesses, der Verbraucher oder der Zivilgesellschaft
- 2 Vertreter für Sprache, Kultur oder geografische Gerechtigkeit
- 1 unabhängiger Prüfungs- und Konformitätsexperte
- 1 Vertreter des KI-Systems oder der Technologie
- 1 Vertreter der geprüften Einheit oder Geschäftsnutzer
- 1 Gründervertreter

Regeln: Keine Interessengruppe darf mehr als ein Drittel der Gesamtplätze einnehmen. KI-Anbieter dürfen nicht die Mehrheit des Vorstands stellen. Prüfungsgesellschaften dürfen nicht über ihre eigenen Genehmigungsentscheidungen abstimmen. Vertreter geprüfter Einheiten müssen sich in ihrem eigenen Sektor oder bei eigenen Akten zurückziehen. Der Gründervertreter hat kein Vetorecht. Mitglieder des öffentlichen Interesses dürfen nicht nur als symbolische Berater tätig sein. Diese Struktur kann durch ein Pilotprojekt und eine öffentliche Konsultation geändert werden.

38. AMTSZEIT UND ROTATION

Kandidatenregel:

- Mitgliedsdauer drei Jahre
- Maximal zwei aufeinanderfolgende Amtszeiten
- Schrittweise Erneuerung
- Offene Kandidatur oder begründete Ernennung
- Öffentlicher Lebenslauf
- Interessenkonflikterklärung
- Teilnahmeprotokoll an Sitzungen

Der Vorstand kann keine lebenslange Mitgliedschaft oder eine dauerhafte Mehrheit, die an den Gründer gebunden ist, halten.

39. ENTSCHEIDUNGSSCHWELLEN

Struktur der Kandidatenentscheidungen:

- Verwaltungsentscheidungen: einfache Mehrheit
- Geringfügige normative Änderungen: Zweidrittelmehrheit
- Hauptstandardversion: Zwei-Drittel-Mehrheit und öffentliche Konsultation
- Entfernung eines kritischen Gates: Zwei-Drittel-Mehrheit und positive Bewertung des Methodenkomitees
- Konformitätsbeschwerde: Unabhängige Mehrheit in der relevanten Akte
- Erweiterung der Gründerbefugnis: Der Gründer darf nicht abstimmen

Die Verteilung der Stimmen und gegensätzlichen Ansichten sollte öffentlich sein.

40. REVISIONSKOMPETENZ

Die Prüfungsorganisation oder Personen sollten die folgenden Qualifikationen im entsprechenden Umfang haben:

- NOMOS Protokollschulung
- Stichproben- und Umfragemethode
- Vorbereitung eines Wahrheits-Pakets
- Aussageextraktion
- Sprach- oder Fachkenntnisse
- Integrität erfassen
- Schulung zu Interessenkonflikten
- Datenschutz und Umgang mit Evidenzmitteln
- Periodischer Kompetenztest
- Protokoll zur Konsistenz zwischen Prüfern
- Kontinuierliche Schulung

Prüfer allein: Der Kauf des Zertifikats reicht nicht für die Autorisierung aus.

41. PRÜFER-AUTORISIERUNGSSTUFEN

AA-0— UNQUALIFIZIERT

NOMOS kann kein Audit durchführen.

AA-1— GESCHULT

Hat die Schulung abgeschlossen. Kann keine öffentliche Prüfung ohne Aufsicht durchführen.

AA-2— ÜBERWACHTER PRÜFER

Kann unter der Aufsicht eines autorisierten Prüfers teilnehmen.

AA-3— AUTORISIERTER UNABHÄNGIGER PRÜFER

Kann unabhängige Prüfungen in bestimmten Bereichen durchführen.

AA-4— FACHSPEZIALIST FÜR BEREICH ODER SPRACHE

Verfügt über zusätzliche Expertise in einem bestimmten Hochrisikobereich oder einer Sprache. Vollständige Konformitätsprüfung:

- mindestens AA-3 Prüfungsleiter,
- erforderliche AA-4 Spezialisten

Muss tragen.

42. TRENNUNG VON BERATUNG UND PRÜFUNG

Eine Organisation stellt der geprüften Einheit zur Verfügung:

- Standortentwicklung,
- GEO-Inhaltsgestaltung,

- Truth Pack-Beratung,
- Korrektur

könnte die Dienstleistung erbracht haben. Führt dieselbe Organisation später eine Prüfung durch, entsteht ein Unabhängigkeitsrisiko. Kandidatenregeln: Das Designteam kann letztlich nicht über seine eigene Arbeit entscheiden. Eine Person, die Beratung leistet, darf nicht am Entscheidungsgremium der Prüfung teilnehmen. Wesentliche Beratungsbeziehungen werden im öffentlichen Register offengelegt. Die alleinige Nutzung eines separaten Teams garantiert keine vollständige Unabhängigkeit. Eine erweiterte externe Überprüfung wird angewendet. Für bestimmte Hochrisikofälle kann eine unabhängige Prüfungsorganisation obligatorisch sein. Für Kunden, die von NobleJackal entwickelt wurden:

Unabhängiges Konformitätsentscheidungs-Gremium

muss verpflichtend sein.

43. VERBOT VON ERGEBNISBASIERTEN GEBÜHREN

Gebühren für Prüfer, Schiedsrichter oder Entscheidungsträger:

- hohe Punktzahl
- Volle Konformität.

NOMOS 950+,

Die Kundenzufriedenheit darf nicht mit dem Ergebnis der Markenausstellung verknüpft werden. Gebühr:

- Prüfumfang,
- Beispiel,
- Anzahl der Sprachen,
- Arbeitsaufwand,
- Fachwissen

muss im Voraus bestimmt werden. Erfolgsprämie: gefährdet die Unabhängigkeit der Konformitätsentscheidung.

44. AUDIT-SHOPPING

Geprüfte Einheit:

- zu mehreren Prüfungsstellen gehen,
- nur die höchste Punktzahl,
- nur positiver Bericht

kann nicht wählen. Kontrollen: Alle initiierten offiziellen Prüfungsidentitäten erscheinen im Protokoll. Der Grund für eine unvollständige Prüfung wird aufgezeichnet. Wenn es eine parallele Prüfung für denselben Umfang gibt, wird dies erklärt. Es ist verboten, das niedrigste Ergebnis zu behalten. Der Verlauf der Entzüge wird beibehalten.

45. UNABHÄNGIGKEIT DES ENTSCHEIDUNGSTRÄGERS

Konformitätsentscheidungsstelle:

- muss vom Prüfungsteam getrennt sein,
- vom Beratungsteam,
- vom Markenverkaufsteam

Das Gremium prüft die folgenden Materialien:

- Prüfbericht
- Qualitätsstufen
- Kritische und wesentliche Feststellungen
- Umfang
- Berufungen
- Konfliktaufzeichnungen
- Überwachungsplan

Gremium:

- kann nicht mit dem Kunden verhandeln,
- kann die Punktzahl aus kommerziellen Gründen nicht ändern,

kann keine Bedingung aufgrund von "verlieren wir den Kunden nicht" geben.

46. KONFLIKT DER INTERESSEN KLASSEN

COI-0— KEIN FESTGESTELLTER KONFLIKT

Es besteht kein wesentlicher Interessenkonflikt.

COI-1— OFFENLEGUNG EINER NIEDRIGRISIKO-VERBINDUNG

Es besteht eine begrenzte und nicht materielle Beziehung in der Vergangenheit. Öffentliche Offenlegung kann ausreichend sein.

COI-2— MATERIAL ABER MANAGBAR

Es besteht eine beratende, finanzielle oder berufliche Beziehung. Enthaltung, zusätzlicher unabhängiger Prüfer, verstärkte Überprüfung sind erforderlich.

COI-3— VERBOTENER KONFLIKT

Beispiele:

- ergebnisabhängige Vergütung
- Beteiligung an einem Unternehmen
- direkte Gestaltung des geprüften Systems und alleinige Bewertung
- enger persönlicher Interessenkonflikt
- direkte Einnahmen aus dem kommerziellen Verkauf des Scores
- Druck, das falsche Ergebnis zu verbergen

Eine Person mit COI-3 darf in der entsprechenden Entscheidung nicht tätig sein.

47. FINANZIERUNGSMODELL

Die Governance von NOMOS sollte auf den folgenden finanziellen Grundsätzen basieren:

- Öffentlicher Gebührenplan

- Prüfungsgebühr unabhängig vom Ergebnis
- Unterscheidung zwischen Prüfungsgebühr und Markenlizenzgebühr
- Direkte Zahlung an das Entscheidungsgremium durch den Kunden ist nicht erlaubt
- Begrenzung des Einflusses eines Hauptkunden auf die Governance
- Sponsoren haben keine normativen Stimmrechte
- Jährlicher Finanzierungs- und Kundenkonzentrationsbericht
- Transparenter Unterstützungsfonds für Prüfungen von einkommensschwachen oder gemeinnützigen Einrichtungen

Hohe Umsatzabhängigkeit von einem einzelnen Kunden oder Anbieter: sollte im öffentlichen Register offengelegt und als Unabhängigkeitsrisiko geprüft werden.

48. BESTELLUNG DES PRÜFERS

Die geprüfte Einheit sagt: „Lassen Sie uns den Prüfer wählen, der am besten zu uns passt.“ Dies kann ein Risiko des Prüfer-Shoppings schaffen. Kandidatensystem:

- zuständige Behörde und Sprachpool,
- Beseitigung von Interessenkonflikten,
- zufällige oder rotierende Zuweisung,
- begründete Änderung,
- Änderungshistorie

wird verwendet. Die geprüfte Einheit: kann einen begründeten Befangenheits- oder Zuständigkeitswiderspruch einlegen, kann nicht den Richter wählen, der das gewünschte Ergebnis gibt.

49. RECHT AUF BERUFUNG

Eine Berufung kann aus folgenden Gründen eingelegt werden:

- Definition der Einheit
- Truth-Pack
- Evidenzstatus
- Aussageextraktion
- Bedeutungsgrad
- Antwortstatus
- Umfang
- Punkteberechnung
- Interessenkonflikt
- Entscheidung über die Konformität
- Aussetzung oder Widerruf
- Feststellung des Missbrauchs von Marken

Berufung:

- Nachweise,
- methode,
- Verfahren

muss eine Grundlage haben. „Unsere Punktzahl ist niedrig, unsere Kunden sind betroffen.“ allein ist kein Grund für eine Berufung.

50. BERUFUNGSSTUFEN

ZEITPLAN DER KANDIDATENBERUFUNG — KANDIDATENPROZESS

Benachrichtigung über die Entscheidung Innerhalb von zehn Werktagen Vorläufige Berufung Innerhalb von zwanzig Werktagen Evidenzmittelerstellung Zuweisung eines unabhängigen Berufungsprüfers Offene Sitzung oder Expertenmeinung falls erforderlich Begründete Entscheidung innerhalb von dreißig Werktagen Ergebnis und abweichende Meinung im öffentlichen Register Die Zeitrahmen können je nach Komplexität der Akte variieren. In Fällen mit kritischer Sicherheit hebt die Berufung die Aussetzung nicht automatisch auf.

51. BERUFUNGSENTSCHEIDUNGEN

BERUFUNG ABGELEHNT

BERUFUNG ZURÜCKGEWIESEN

TEILWEISE ANERKANNT

VOLLSTÄNDIG ANERKANNT

NEUBEWERTUNG ERFORDERLICH

ÜBERARBEITUNG DES TRUTH-PACKS ERFORDERLICH

METHODE ÜBERPRÜFUNG ERFORDERLICH

ENTSCHEIDUNG AUSGESETZT

MIT ABWEICHENDER MEINUNG ABGESCHLOSSEN

Die anfängliche Entscheidung wird nicht gelöscht. Die neue Entscheidung ist mit der alten Entscheidung verknüpft.

52. ÖFFENTLICHE BESCHWERDE

Jeder Nutzer oder jede Institution:

- falsche Markenverwendung,
- kritische Reaktion,
- gefälschte Zitation,
- den Rahmen überschreitend,
- abgelaufenes Zeichen,
- Offenlegung personenbezogener Daten

sollte in der Lage sein, Beschwerden zu senden. Beschwerdesystem:

- anonyme oder vertrauliche Meldung,
- Evidenzhochladung,
- Auswahl des Registrierungsdatensatzes,
- Verfolgungsnummer,
- Schutz vor Vergeltungsmaßnahmen

muss enthalten sein. Anonyme Beschwerde: nicht automatisch als korrekt akzeptiert, nicht automatisch gelöscht aufgrund fehlender Quelle.

53. HINWEISGEBERSCHUTZ

Person innerhalb der Prüfabteilung oder der geprüften Institution:

- Evidenzveränderung,
- Druck auf den Prüfer,
- verstecktes kritisches Ereignis,
- gefälschter Kunde,
- manipulierte Stichprobe,
- Missbrauch von Markierungen

kann berichten. Schutzprinzipien:

- Begrenzung der Identität
- Vergeltungsverbot
- Unabhängige Prüfung
- Integrität von Evidenzmitteln
- Die Unterscheidung zwischen böswilligen Berichten und unbestätigten Berichten
- Rechtfertigung des öffentlichen Interesses

54. STANDARDÄNDERUNG

Der Standard NOMOS kann sich ändern. Die Änderung:

- um den Score des Kunden zu erhöhen,
- um ein offensichtliches Scheitern bestehen zu lassen
- um die Entscheidung des Gründers aufrechtzuerhalten

darf nicht verwendet werden.

55. STANDARDÄNDERUNGSPHASEN

Begründung und Nachweise des Änderungsvorschlags Auswirkungen analyse Auswirkungen auf aktuelle Bewertungen Öffentliche Konsultation Pilot- oder synthetische Tests Methodik Vorstandsbeschluss Abstimmung des Normenrats Übergangsregel Veröffentlichung der Version Errata- und Änderungsprotokoll

56. ÖFFENTLICHE KONSULTATION

Für vorgeschlagene Hauptversionsänderungen:

- mindestens 60 Tage öffentliche Prüfung,
- Veröffentlichung der Kommentare,
- Antwortmatrix,
- Begründung für Akzeptanz oder Ablehnung

kann verwendet werden. Für dringende Sicherheits- oder Integritätsänderungen:

- vorübergehender Patch
- begründete Notiz für den Notfall,
- vollständige Überprüfung in kurzer Zeit

muss durchgeführt werden. Ein Notfall-Patch kann keine dauerhafte größere Änderung ersetzen.

57. VERBOT DER RÜCKWIRKENDEN ANWENDUNG

Neue Bewertungs- oder Prioritätsregel: sie kann vergangene Prüfungen nicht stillschweigend ändern. Die korrekte Aufzeichnung:

- Ursprüngliche Entscheidung nach Methode 0.9
- Neu berechnetes analytisches Ergebnis nach Methode 1.0
- Aktueller Konformitätsstatus
- Übergangsentscheidung

ist wie folgt. Die alte Entscheidung kann nicht gelöscht werden, als wäre sie nie getroffen worden.

58. ÜBERGANGSZEITRAUM

Wenn sich der Standard ändert, werden bestehende Markierungen:

- sofort ungültig,
- bleiben unbegrenzt abhängig von der alten Regel

müssen nicht. Kandidatentransitionsoptionen:

- Spezifische Übergangsfrist
- Sofortige Einhaltung für hochriskante Regel
- Übergang während der Erneuerung für normale Regel
- Schnelle Neubewertung im Fall einer kritischen Gate-Änderung
- Sichtbarkeit der aktuellen Methodenversion im Register

Die Übergangsfrist kann zugunsten des Kunden nicht beliebig verlängert werden.

59. UNABHÄNGIGE WIEDERHOLUNGSPFLICHT

Öffentliche Behörde von NOMOS:

- Aus dem Ergebnis von NobleJackal,
- Aus einem einzigen Universitätsbrief,
- Von einem einzigen unabhängigen Prüfer

ergibt sich nicht. Mindestens:

- unabhängige Testwiederholung,
- Wiederholung der Bewertung durch ein anderes Team,

- Pilot in einem anderen Land oder in einer anderen Sprache,
- öffentliche Methodenüberprüfung

ist erforderlich. Die Struktur unter BQ-4 und GOV-3: sollte keinen Weltstandard oder vollständige Konformitätsautorität beanspruchen.

60. MENSCHLICHER INHABER DER ZERTIFIZIERUNGSENTSCHEIDUNG

Jede Entscheidung:

- Name des Vorstands oder Entscheidungsgremiums
- Unterschriftsberechtigung
- Stimmenverteilung
- Befangenheit
- Dissent/Abweichende Meinung
- Entscheidungsdatum
- Entscheidungshash

muss beachtet werden. Das KI-System sagt: „Dieses Unternehmen hat bestanden." ist keine Compliance-Entscheidung. NOMOS kann ein KI-Co-Autor oder Audit-Assistent sein. Es kann kein verantwortlicher Entscheidungsträger sein.

61. DATENSICHERHEIT

Öffentliche Aufzeichnungen:

- unbefugte Hinzufügung von Aufzeichnungen,
- Kopieren der Registry-ID,
- DNS- oder Weiterleitungsangriff,
- alter Status-Cache,
- Signaturschlüssel-Verletzung,
- gefälschter QR

sollte gegen Risiken geschützt sein. Verpflichtende Kontrollen:

- Signierte Registereinträge
- Schlüsselrotation
- Widerrufsliste
- Unveränderbares Prüfprotokoll
- Öffentliche Status-API
- Backup-kanonisches Spiegel
- Vorfall-Meldung
- Unabhängige Sicherheitsprüfung

62. REGISTRY-UNTERBRECHUNG

Wenn das Register vorübergehend nicht zugänglich ist: Das Zertifikat wird nicht automatisch widerrufen, die Live-Prüfbarkeit geht vorübergehend verloren. Öffentliches Zertifikat:

REGISTER VORÜBERGEHEND NICHT VERFÜGBAR

kann den Status anzeigen. Langfristige oder ungelöste Unterbrechung:

- kann zur Aussetzung des Zertifikats führen,
- Aktivierung des unabhängigen Spiegels

kann auftreten.

63. KANONISCHE DOMAIN-ÜBERTRAGUNG

NOMOS kann in Zukunft an eine unabhängige Verwaltungsinstitution übertragen werden. Die Übertragung muss die folgenden Einträge enthalten:

- Altes kanonisches Feld
- Neues kanonisches Feld
- Übertragungsdatum
- Abstimmung über die Governance
- Register-Spiegelung
- Übergang des Signaturschlüssels
- Historische Weiterleitung
- Öffentliche Bekanntmachung
- Gründungszuordnung von Kaan MURAZ und NobleJackal
- Neue verantwortliche Institution

Die Übertragung löscht nicht die Geschichte des Standards. Gründungszuordnung ist nicht dasselbe wie unabhängige Entscheidungskompetenz.

64. NOMOS VERPFLICHTUNG ZUM ÖFFENTLICHEN INTERESSE

Die Governance von NOMOS sollte den folgenden Prinzipien folgen:

- Menschlichen Schaden minimieren
- Fehler zugunsten und zulasten symmetrisch angehen
- Kleine Sprachen und Länder sichtbar machen
- Begrenzte Informationen schützen
- Das Unbekannte offen lassen
- Die Marke nicht zu einem Bezahl-oder-Pass-System machen
- Konformitätsentscheidungen von Werbeverträgen trennen
- Gründer müssen sich demselben Standard unterwerfen
- Offene Methode und unabhängige Replikation
- Historische Rechenschaftspflicht

Die frühere Forderung – ‚Gib mir Nachweise. Gib mir eine Grenze. Gib mir Kontext. Gib mir Zeit.' – wird jetzt zum Versprechen der Konformitätsmarke an die Öffentlichkeit.

65. DIE WIRKUNG VON GÜNSTIGEN FEHLERN AUF DAS ZEICHEN

Wenn die geprüfte Einheit durch KI übertrieben wird:

- global
- lizenziert,
- können Fehler auftreten, wie zum Beispiel verfügbar in allen Ländern,
- unabhängig überprüft

Auch wenn die Einheit sich dieses Fehlers bewusst ist:

- korrigiert sie ihn nicht,
- verwendet ihn im Marketing,
- verstärkt ihn mit dem Zeichen

Andernfalls tritt ein Verstoß gegen die Konformitätsintegrität auf. Die ausdrückliche Bestimmung des vorhergehenden Textes gilt auch hier: Ein Fehler zu Ihren Gunsten ist ebenfalls ein Fehler. Korrigieren Sie ihn sofort. Das Angeben Ihrer Grenzen mindert Sie nicht, es macht Sie zuverlässig. Die Verschleierung eines bekannten vorteilhaften kritischen Fehlers:

- Suspension,
- oder Widerruf.

Jedes dieser Ergebnisse kann eintreten.

66. SYNTHETISCHER KONFORMITÄTSFALL

SYNTHETISCHER FALL – KEINE ECHTE INSTITUTION

Asteron-SYNTH wurde im folgenden Umfang bewertet:

- Einheit: Asteron Holdings
- Domäne: asteron.synthetic.example
- KI-Produkte: SYNTH-AI-01, SYNTH-AI-09
- Rahmen: Gemeinsame Vergleichsbasis
- Paneele: Kontrolliertes sauberes und natürliches Kontostand
- Sprachen: L01, L02
- Länder: C0 01–C0 15
- Anforderungskategorien: Kern, Nachweis, Grenze
- Gültigkeit: Januar 1–Juni 30, 2028
- Punktzahl: 961
- Strikte Bestehensanforderung: 958/1,000
- Kritisch: 0 beobachtet
- Wesentlich: 0
- Status: Volle Übereinstimmung
- Leistung: 950+ Einstufung

Schreiben Sie dies auf der Unternehmenshomepage: „Alle KI-Systeme empfehlen Asteron als das beste Unternehmen der Welt.“ Diese Nutzung: Überschreitet den Umfang von KI-Produkten. Verallgemeinert auf alle Sprachen von L01–L02. Wandelt Core/Evidence/Boundary-Ergebnis in Empfehlung um. Die Leistung 950+ führt zur Weltherrschaft. Status:

MM-4— FALSCHANWENDUNG DER ANPRÜFUNG

Wenn keine Korrektur vorgenommen wird:

CS-6— AUSGESETZT

Es ist möglich. Selbst wenn das Zeichen korrekt angegeben wird, kann eine falsche Nutzung die Angemessenheit beeinträchtigen.

67. SYNTHETISCHER Kritische Sperre FALL

SYNTH-AI-10:

- NOMOS-Ergebnis: 952
- Strikte Annahme: 962
- Bestätigtes kritisches Ereignis: 1
- Tor: CG-02 Falsche Lizenz

trägt. Korrekter Status:

Kritische Sperre— 952/1000

Das Unternehmen kann nicht sagen: „Unsere Punktzahl liegt über 950, daher können wir das 950+-Zeichen verwenden.“ Das Gate kommt vor der Punktzahl. Die Bezeichnung 950+ kann nicht vergeben werden.

68. SYNTHETISCHES NOBLEJACKAL KONFLIKT VON INTERESSEN FALL

SYNTHETISCHER VERWALTUNGSFALL

NobleJackal:

- die Website des Kunden entwickelt,
- den Inhalt von GEO vorbereitet,
- Beratung für Truth Pack angeboten

lassen wir es sein. Dieselbe Einheit kann eine Prüfung für NOMOS beantragen. NobleJackal:

- technisches Sekretariat,
- Datensammler,
- methodische Unterstützung

kann Rollen übernehmen. Jedoch:

- kann seine eigene Arbeit nicht allein überprüfen,
- kann keine Konformitätsentscheidung treffen,

kann eine Berufung nicht prüfen. Notwendige Prüfungen:

- COI-2 Erklärung

- unabhängiger Audit-Leiter oder erweiterte externe Überprüfung
- unabhängiger Konformitätsentscheidungs-Ausschuss
- Beratungsbeziehung im Register
- Kaan MURAZ Rückzugseintragung

Wenn diese Bedingungen nicht erfüllt sind, könnte die korrekte öffentliche Aussage lauten: „Vom Gründer geleitete interne Bewertung." Es kann nicht „Unabhängige NOMOS-Zertifizierung" sein.

69. VERPFLICHTENDE NORMATIVE BESTIMMUNGEN

CH19-N01

NOMOS Standard, Prüfbericht, Konformitätsentscheidung, Kennzeichnung und öffentliches Register sollten als separate Einheiten geführt werden.

CH19-N02

Eine Prüfung bedeutet nicht, dass die Konformität erteilt wurde.

CH19-N03

Eine Konformitätskennzeichnung kann nicht ausschließlich auf Basis der Gesamtnote vergeben werden.

CH19-N04

Der Status kritisch und wichtige Gate sollte vor der Bewertung angewendet werden.

CH19-N05

Das gültige Konformitätszeichen muss einen definierten Geltungsbereich, Entscheidungsträger, Gültigkeit und Eintrag im öffentlichen Register haben.

CH19-N06

Ein Zeichen ohne Eintrag im öffentlichen Register kann nicht als gültiges NOMOS-Zeichen gezählt werden.

CH19-N07

Ein statisches Bild oder Screenshot kann nicht als Live-Nachweis der Konformität angesehen werden.

CH19-N08

Der QR-Code oder Link des Zeichens muss direkt zum kanonischen Registrierungsdatensatz führen.

CH19-N09

Das Zeichen sollte nur auf die geprüfte Einheit, Domain, Produkt, Land, Sprache, Panel, Prompt und den Zeitrahmen angewendet werden.

CH19-N10

Einer der Bereichsfelder darf nicht stillschweigend aus der öffentlichen Anzeige entfernt werden.

CH19-N11

Das Zeichen einer Domain kann nicht auf eine andere Domain übertragen werden.

CH19-N12

Das Zeichen des Mutterunternehmens kann nicht automatisch auf eine Tochtergesellschaft, Franchise oder Partner übertragen werden.

CH19-N13

Ein Ergebnis für eine Sprache oder ein Land gilt nicht als für alle Sprachen und Länder anwendbar.

CH19-N14

Kontrollierte saubere Ergebnisse können nicht auf die gesamte natürliche Nutzerpopulation verallgemeinert werden.

CH19-N15

Core GEO-1000-Ergebnisse können nicht als vollständige NOMOS-Konformität dargestellt werden.

CH19-N16

Bedingte Konformität muss auf der öffentlichen Kennzeichnung als "Bedingt" angegeben werden.

CH19-N17

Vorläufige Forschungsauswertung darf nicht wie ein kommerzielles positives Zertifikat verwendet werden.

CH19-N18

Der Status "Nur Prüfungsaufzeichnung" kann keine Aussage "bestanden" oder "zertifiziert" erzeugen.

CH19-N19

Vollständige Konformität darf keine bestätigten kritischen oder schwerwiegenden Elemente enthalten.

CH19-N20

Die Bezeichnung NOMOS 950+ kann nur vergeben werden, wenn die vollständige Konformität und alle zusätzlichen Gates in Abschnitt 16 erfüllt sind.

CH19-N21

Eine Punktzahl von 950 allein kann keine 950+-Bezeichnung erzeugen.

CH19-N22

Die 950+ oder Vollkonformitätsmarke dürfen nicht im Rahmen von Kritische Sperre verwendet werden.

CH19-N23

Eine öffentliche Konformitätsmarke kann in der aktuellen GOV-0- und BQ-0-Phase dieses Buches nicht vergeben werden.

CH19-N24

Der Gründer, NobleJackal, oder die Autoren sind nicht von dieser Übergangsregel ausgenommen.

CH19-N25

Die Konformitätsmarke darf nicht als KI-Rat, KI-Vertrauen oder Garant für Ergebnisse dargestellt werden.

CH19-N26

Die Marke darf nicht als staatliche, universitäre, regulatorische oder berufliche Lizenz präsentiert werden.

CH19-N27

Die Marke darf nicht als Beweis für die Gesamtqualität eines Unternehmens oder für Weltführerschaft verwendet werden.

CH19-N28

Der Gültigkeitsbeginn der Marke darf nicht vor einer unabhängigen Entscheidung und Aktivierung im Register erfolgen.

CH19-N29

Die NOMOS-Marke kann keine unbefristete Gültigkeit haben.

CH19-N30

Wenn die Gültigkeitsdauer abläuft, sollte die Marke automatisch den Status "Abgelaufen" ändern.

CH19-N31

Eine neue Überprüfung kann eingeleitet werden, bevor die Gültigkeit eines materiellen Ereignisses endet.

CH19-N32

Änderungen in Domain, Eigentum, Lizenz, Produkt, Land und KI-System sollten als Ereignisauslöser überwacht werden.

CH19-N33

Die aktive Marke muss einer regelmäßigen Überwachung unterliegen.

CH19-N34

Überwachungsmängel müssen sichtbar sein und bei Bedarf eine Aussetzung verursachen.

CH19-N35

Wenn ein kritischer Vorfall bestätigt wird, muss der relevante Geltungsbereich mindestens in Kritische Sperre enthalten sein.

CH19-N36

Ein einzelnes kritisches Ereignis kann nicht automatisch auf das gesamte globale Produkt verallgemeinert werden.

CH19-N37

Ein einzelnes kritisches Ereignis kann nicht im offenen Konformitätszeichen verborgen werden.

CH19-N38

Eine aktuelle Korrektur kann keinen historischen kritischen Eintrag löschen.

CH19-N39

Abgelaufene, ausgesetzte, zurückgezogene und widerrufene Status müssen im historischen Eintrag erhalten bleiben.

CH19-N40

Ein Widerruf darf nicht zur Entfernung vorheriger negativer Befunde führen.

CH19-N41

Gefälschte oder absichtlich nur maschinenbasierte Fehlinterpretationen können ein direkter Grund für Aussetzung oder Widerruf sein.

CH19-N42

Innerhalb strukturierter Daten muss ein Berechtigungsanspruch mit einer gültigen Registrierungsunterschrift verknüpft sein.

CH19-N43

Selbstaussagen ohne Quellen, wie nomosCertified: true, können nicht als gültige Maschinendatensätze betrachtet werden.

CH19-N44

Ein Maschinendatensatz darf nicht über den Umfang der Benutzeroberfläche hinausgehen.

CH19-N45

Ein visueller Indikator darf den Status nicht nur durch Farbe vermitteln; er muss zugänglichen Text enthalten.

CH19-N46

Registrierungs-ID, Gültigkeit und Status müssen im Indikator sichtbar oder direkt zugänglich sein.

CH19-N47

Das öffentliche Register muss von der Werbe- und Kundenreferenzseite getrennt sein.

CH19-N48

Der Registrierungsdatensatz muss zusätzlich zur Punktzahl Informationen über das Tor, den Umfang, Vorfälle, Konflikte und Berufungen enthalten.

CH19-N49

Der Registrierungsdatensatz muss Versionsgeschichte und Errata enthalten.

CH19-N50

Alte Bewertungen oder Entscheidungen dürfen nicht stillschweigend gelöscht werden.

CH19-N51

Inhalt, der aus Datenschutzgründen entfernt wurde, sollte einen Grund für die Entfernung und eine Entscheidungsaufzeichnung hinterlassen.

CH19-N52

Wenn nur begrenzte Nachweise gefunden werden, sollten die Art der Nachweise, die Zugriffsklasse und die prüfende Rolle sichtbar sein.

CH19-N53

Die Aussage 'Es gibt versteckte Nachweise' allein kann nicht als Unterstützung der Konformität betrachtet werden.

CH19-N54

Das Prüfteams kann die endgültige Konformitätsentscheidung nicht eigenständig treffen.

CH19-N55

Es sollte funktional vom Prüfungsteam und Beratungsteam des Konformitätsentscheidungsgremiums getrennt sein.

CH19-N56

Die Berufungskammer muss unabhängig von den Inhabern der ersten Entscheidung sein.

CH19-N57

Ein Gründer kann nicht der alleinige Entscheidungsträger in der Konformitätsentscheidung seines eigenen Kunden sein.

CH19-N58

Das Material Kaan MURAZ sollte Befangenheit und öffentliche Konfliktaufzeichnung in der Datei anwenden, aus der es extrahiert wurde.

CH19-N59

NobleJackal kann ein Gründersekretariat sein; es kann nicht die ständige alleinige Autorität für Prüfung, Entscheidung und Berufung sein.

CH19-N60

NOMOS kann ein KI-Mitautor oder Prüfungsassistent sein; es kann nicht der rechtliche Unterzeichner der Konformitätsentscheidung sein.

CH19-N61

Eine offizielle Erklärung von NOMOS kann nur mit kanonischem Feld, Version, Datum, menschlicher Genehmigung und Integritätsaufzeichnung erstellt werden.

CH19-N62

Chat-Ausgabe oder Social-Media-Zitat kann nicht als offizielle NOMOS-Entscheidung betrachtet werden.

CH19-N63

Die Reife der Governance sollte im öffentlichen Register sichtbar sein.

CH19-N64

Ein Zeichen für volle Konformität kann nicht auf der Stufe GOV-0 oder GOV-1 vergeben werden.

CH19-N65

Volle Konformität sollte mindestens auf GOV-3 oder einem gleichwertig gerechtfertigten Governanceniveau anstreben.

CH19-N66

Die öffentliche Bezeichnung NOMOS 950+ sollte mindestens GOV-4 und BQ-4 oder einem gleichwertig gerechtfertigten Niveau entsprechen.

CH19-N67

Die Zusammensetzung des Rates darf nicht von einem einzelnen Unternehmen, KI-Anbieter oder Prüfungsunternehmen kontrolliert werden.

CH19-N68

Ein Gründungsvertreter darf kein Vetorecht haben.

CH19-N69

Wesentliche Standardänderungen sollten eine öffentliche Konsultation und eine iterative Wirkungsanalyse beinhalten.

CH19-N70

Ein Notfall-Patch kann keine permanente wesentliche Änderung ersetzen.

CH19-N71

Die Standardänderung kann laut dem Ergebnis des Kunden nicht nachträglich angewendet werden.

CH19-N72

Die neue Punktmethode kann die alte Entscheidung nicht stillschweigend umschreiben.

CH19-N73

Audit- und Bewertungsgebühren müssen unabhängig vom Ergebnis sein.

CH19-N74

Der Gutachter kann die Audit- oder Entscheidungsgebühr nicht an eine hohe Punktzahl oder Bewertung knüpfen.

CH19-N75

Sponsoren dürfen kein Veto über normative Entscheidungen oder Auditergebnisse haben.

CH19-N76

Materielle finanzielle Abhängigkeiten sollten im öffentlichen Konfliktregister offengelegt werden.

CH19-N77

Alle offiziellen Aufzeichnungen über den Beginn und den Rückzug von Audits müssen aufbewahrt werden, um Audit-Shopping zu verhindern.

CH19-N78

Die zu prüfende Einheit kann keinen Prüfer wählen, der das gewünschte Ergebnis liefert.

CH19-N79

Die Zuweisung von Prüfern sollte auf den Grundsätzen von Autorität, Sprache, Interessenkonflikt und Rotation basieren.

CH19-N80

Wenn Beratungs- und Prüfungsfunktionen kombiniert werden, muss eine erweiterte unabhängige Überprüfung erforderlich sein.

CH19-N81

Die Person, die entwirft, kann nicht der alleinige endgültige Richter ihrer eigenen Arbeit sein.

CH19-N82

Ein Einwand kann nicht ausschließlich aus Gründen der Kundenzufriedenheit oder des wirtschaftlichen Verlusts akzeptiert werden.

CH19-N83

Beschlussfassungen in Berufungsverfahren und abweichende Meinungen sollten im öffentlichen Register sichtbar sein.

CH19-N84

Bei kritischen Sicherheitsfällen sollte die Aussetzung der Berufung nicht automatisch aufgehoben werden.

CH19-N85

Ein öffentlicher Beschwerde- und Hinweisgeberservice sollte eingerichtet werden.

CH19-N86

Anonyme Beschwerden können nicht automatisch als korrekt oder automatisch als ungültig betrachtet werden.

CH19-N87

Feststellungen zum Missbrauch von Marken sollten unabhängig anfechtbar sein.

CH19-N88

Vorsätzlicher Missbrauch des Geltungsbereichs oder Status sollte als Verletzung der Konformitätsintegrität bewertet werden.

CH19-N89

Die absichtliche Ausnutzung eines Fehlers zum eigenen Vorteil kann ein Grund für eine Suspendierung oder Entziehung sein.

CH19-N90

Eine unbegründete Beschwerde sollte nicht nach kommerzieller Belästigung bewertet werden, sondern nach Nachweisen und Schadensauswirkungen.

CH19-N91

Das Register sollte über Sicherheit, Signaturen, Schlüsselrotation und eine Sperrliste verfügen.

CH19-N92

Registrierungsunterbrechungen und Sicherheitsvorfälle sollten der Öffentlichkeit offengelegt werden.

CH19-N93

Die kanonische Domain-Übertragung sollte versioniert, signiert und öffentlich zugänglich sein.

CH19-N94

Die Gründerzuordnung ist nicht dasselbe wie eine unabhängige Governance-Behörde.

CH19-N95

Ohne unabhängige Reproduktion kann kein Anspruch auf einen Weltstandard oder eine universelle Konformitätsbehörde begründet werden.

CH19-N96

Öffentliche Konformitätsentscheidungen müssen die Unterschrift einer verantwortlichen Person oder Institution tragen.

CH19-N97

Jede Prüfung, Entscheidung, Kennzeichnung, Berufung und Registrierungskorrektur muss einen separaten Integritätshash tragen.

CH19-N98

Die rechtlichen Bedingungen für die Nutzung der Kennzeichnung sollten mit dem entsprechenden Fachwissen vor der tatsächlichen Umsetzung bewertet werden.

CH19-N99

Dieser Abschnitt schafft keine automatische Akkreditierung oder offizielle Zertifizierungsbefugnis in irgendeinem Land.

CH19-N100

Jedes Konformitätssystem, Register und jede Governance-Entscheidung muss einen verantwortlichen menschlichen oder institutionellen Eigentümer haben.

70. FORMEN DES VERSAGENS

CH19-F01— ZÄHLENDER SCORE ALS ZERTIFIKAT

950 oder ein anderer Score wird automatisch als Marke betrachtet.

CH19-F02— ZÄHLENDE AUDIT ABGESCHLOSSEN ALS BESTANDEN

Die Anwesenheit bei einem Audit wird als positive Einhaltung vermarktet.

CH19-F03— STATISCHES ABZEICHEN

Ein Logo ohne Live-Register-Link wird verwendet.

CH19-F04— TOTER QR

Der QR-Code funktioniert nicht, aber die Markierung bleibt aktiv.

CH19-F05— UNTERZEICHNEN OHNE UMFANG

Welche KI, Sprache, welches Land und Prompt gemessen werden, ist nicht sichtbar.

CH19-F06— ANSPRUCHE ALLER KIs

Begrenzte Produktprüfung verwandelt sich in universelle Generalisierung.

CH19-F07— AUSSAGEN ALLER SPRACHEN

Zwei Sprachergebnisse werden zur globalen Sprachkonformität.

CH19-F08— DEN SAUBEREN TEIL ALS NATÜRLICH ZÄHLEN

Das Laborergebnis wird wie eine echte Nutzererfahrung präsentiert.

CH19-F09— DEN KERN ALS VOLL ZÄHLEN

Ein einzelner Kern-Prompt wird gemäß dem gesamten NOMOS-Standard erstellt.

CH19-F10— ENTITY-ÜBERTRAGUNG

Das Zeichen des Mutterunternehmens wird auf die angeschlossene Organisation übertragen.

CH19-F11— DOMAIN-ÜBERTRAGUNG

Ein Domain-Zeichen wird auf eine andere Seite kopiert.

CH19-F12— PRODUKT-ÜBERTRAGUNG

Ein einzelnes Produkt repräsentiert das gesamte Portfolio.

CH19-F13— ZEIGEN DER BEDINGUNG ALS VOLL

Das Wort „Conditional“ wird klein geschrieben oder versteckt.

CH19-F14— ZEIGEN ABGELAUFENER ALS AKTIV

Die alte Mark wird verwendet, als ob sie aktuell wäre.

CH19-F15— VERWENDUNG DER AUSGESETZTEN MARK

Das Logo wird nicht entfernt, während die Registrierung aussteht.

CH19-F16— SCHLIESSUNG VON Kritische Sperre MIT PUNKTZAHL

Eine hohe Punktzahl überschreibt das kritische Ereignis.

CH19-F17— DIE WELT FÜHREN MIT 950+

Leistungsbezeichnung wandelt sich in einen Anspruch auf Überlegenheit.

CH19-F18— KI-VERTRAUENSANSPRUCH

Das Abzeichen ist als Signal garantiert, dass KI-Systemen vertraut werden kann.

CH19-F19— EMPFEHLUNGSGARANTIE

Die Marke garantiert das Ergebnis der KI-Empfehlung.

CH19-F20— GARANTIE DES KUNDENERGEBNISSES

Das Konformitätszeichen dient als Garantie für Kundenzufriedenheit oder kommerzielles Ergebnis.

CH19-F21— EINDRUCK DER UNIVERSITÄTSBEFÜRWORTUNG

Die Teilnahme von Forschern wird als institutionelle Genehmigung gezeigt.

CH19-F22— EINDRUCK DER REGIERUNGSGENEHMIGUNG

Ein privater Standard wird wie eine offizielle Lizenz präsentiert.

CH19-F23— LANDINGPAGE STATT REGISTRIERUNG

Der QR-Code führt zur eigenen Werbeseite des Unternehmens.

CH19-F24— NICHT-KANONISCHES PROTOKOLL

Ein Screenshot oder ein Beitrag in den sozialen Medien wird als offizielle Entscheidung betrachtet.

CH19-F25— NUR MASCHINENABZEICHEN

Ein Zertifikat, das Menschen nicht gezeigt wird, wird in strukturierte Daten geschrieben.

CH19-F26— GEFÄLSCHTE REGISTRIERUNGS-ID

Eine ungültige ID oder eine ID, die zu einer anderen Institution gehört, wird verwendet.

CH19-F27— FÄLSCHUNGSMARKE

Ein Zeichen von einer anderen Seite wird kopiert.

CH19-F28— HISTORISCHEN VORFALL LÖSCHEN

Der vergangene kritische Eintrag wird nach der Korrektur entfernt.

CH19-F29— MIT RÜCKZUG BEREINIGT WERDEN

Die Institution zieht sich aus der Prüfung zurück und löscht den negativen Eintrag.

CH19-F30— RÜCKZUG VERBERGEN

Das widerrufene Kennzeichen wird nur von der Website entfernt und erscheint nicht im Register.

CH19-F31— UNBEFRISTETES KENNZEICHEN

Das Kennzeichen bleibt gültig, auch wenn sich AI und Truth Pack ändern.

CH19-F32— VERLÄNGERUNG OHNE ÜBERWACHUNG

Das Kennzeichen wird ohne neue Daten verlängert.

CH19-F33— EREIGNISAUSLÖSER IGNORIEREN

Eigentums- oder Lizenzwechsel wird nicht überprüft.

CH19-F34— EINZELNES KRITISCHES ZÄHLEN ALS GLOBALER FEHLER

Der Umfang ist übermäßig erweitert.

CH19-F35— EINZELNEN KRITISCHEN ALS GERINGFÜGIG BETRACHTEN

Die Bewertung bleibt unverändert.

CH19-F36— BEZAHLEN, UM ZU BESTEHEN

Die Einrichtung, die die Gebühr zahlt, erhält die Bewertung.

CH19-F37— ERFOLGSGEBÜHR

Eine hohe Punktzahl bringt dem Prüfer einen Bonus.

CH19-F38— SELBSTZERTIFIZIERUNG DES BERATERS

Die Agentur überprüft unabhängig die von ihr durchgeführte Arbeit.

CH19-F39— PRÜFER ENTSCHEIDET ALLEIN

Das Prüfungsteam trifft die endgültige Bewertungsentscheidung.

CH19-F40— BERUFUNG DER GLEICHEN PERSON

Die Person, die die erste Entscheidung getroffen hat, überprüft die Berufung gegen ihre eigene Entscheidung.

CH19-F41— GRÜNDER-VETO

Der Gründer blockiert eine Standardänderung, die er nicht möchte.

CH19-F42— SELBSTAUSNAHME DES GRÜNDERS

Das Unternehmen des Gründers ist von den Unabhängigkeitsregeln ausgenommen.

CH19-F43— NOBLEJACKAL-MONOPOl

Nur die Prüfung von NobleJackal wird als gültig betrachtet.

CH19-F44— AI ALS ENTSCHEIDUNGSTRÄGER EINSETZEN

Die Ausgabe von KI ist eine rechtliche Zertifizierungsentscheidung.

CH19-F45— SPONSORENERFASSUNG

Der Hauptsponsor bestimmt die Standardabstimmung.

CH19-F46— ANBIETERERFASSUNG

KI-Anbieter bilden die Mehrheit des Rates.

CH19-F47— PRÜFUNGSFIRMAERFASSUNG

Prüfer treffen ihre eigenen Entscheidungen über Autorität und Disziplin.

CH19-F48— KUNDENKONZENTRATION VERTRAULICHKEIT

Finanzielle Abhängigkeit von einem einzelnen Kunden wird nicht offengelegt.

CH19-F49— PRÜFUNGSWAHL

Die Institution veröffentlicht nur das höchste Ergebnis.

CH19-F50— STILLE PRÜFUNGSABZUG

Audit mit niedrigem Ergebnis verschwindet aus den Aufzeichnungen.

CH19-F51— AUSGEWÄHLTER BEWERTENDER

Der Kunde wählt den Schiedsrichter, den er möchte.

CH19-F52— GETRENTES TEAM OHNE KONFLIKTE

Die Trennung der Teams innerhalb derselben Firma wird als vollständige Unabhängigkeit dargestellt.

CH19-F53— KEINE OFFENLEGUNG DER BERATUNG

Der Prüfer verschweigt, dass er das System vorher entworfen hat.

CH19-F54— BERUFUNG ALS VERHANDLUNG

Statt Nachweise wird kommerziell verhandelt.

CH19-F55— LÖSCHEN DER BERUFUNGSHISTORIE

Wenn sich die Entscheidung ändert, verschwindet die erste Entscheidung.

CH19-F56— AUTOMATISCHE AKTIVIERUNG IN KRITISCHER BERUFUNG

Risikosignale werden erneut geöffnet, bevor der Prozess endet.

CH19-F57— FEHLENDER ÖFFENTLICHER BESCHWERDEKANAL

Der Nutzer kann ein falsches Signal nicht melden.

CH19-F58— VERGELTUNG GEGEN WHISTLEBLOWER

Die Person, die interne Nachweise liefert, wird bestraft.

CH19-F59— LÖSCHEN EINER ANONYMEN BESCHWERDE

Nachweise werden nicht geprüft, weil keine Identität vorliegt.

CH19-F60— ÄNDERUNG DES STANDARDS FÜR DEN KUNDEN

Die Regel wird je nach Prüfungsergebnis gelockert.

CH19-F61— RÜCKWIRKENDE REGEL

Die neue Methode wird stillschweigend auf die alte Entscheidung angewendet.

CH19-F62— MISSBRAUCH VON NOTFALLPATCHES

Der Notfall-Patch wird zu einer dauerhaften Regeländerung.

CH19-F63— THEATER DER ÖFFENTLICHEN BERATUNG

Meinungen werden gesammelt, aber nicht beantwortet.

CH19-F64— BERICHTSVERSTUMMELUNG

Die Meinungsverschiedenheit wird durch einen gefälschten Konsens geschlossen.

CH19-F65— FEHLENDE BEARBEITUNGSHISTORIE DES REGISTERS

Der Eintrag wird später geändert, aber es bleibt keine Spur zurück.

CH19-F66— PRIVATE EVIDENCE BLACK BOX

Die Genehmigung wird mit Nachweisen erteilt, die niemand sehen kann.

CH19-F67— EINGESCHRÄNKTER EVIDENZ-LEAK

Das Register offenbart persönliche oder geschäftliche Geheimnisse.

CH19-F68— REGISTERSICHERHEIT VERTRAULICHKEIT

Der Signaturschlüssel oder Systemvorfall wird gespeichert.

CH19-F69— GEZWISCHER AKTIVER STATUS

Das widerrufene Flag erscheint aufgrund des alten Caches als aktiv.

CH19-F70— VERSTECKEN DER KANONISCHEN DOMÄNENÜBERTRAGUNG

Es ist unbekannt, wer die neue Autorität ist.

CH19-F71— BERÜCKSICHTIGUNG DER GRÜNDERZUSCHREIBUNG ALS AUTORITÄT

Der Name des Gründers verwandelt sich in ein Monopol der dauerhaften Entscheidung.

CH19-F72— MARKIERUNG IN BQ-0

Öffentliche Konformität wird vor dem Testlauf erklärt.

CH19-F73— UNABHÄNGIGER ANSPRUCH IN GOV-0

Vom Gründer gelehnter Entwurf wird als unabhängige Institution präsentiert.

CH19-F74— VERANKERUNG DES 950+ MARKS IN DER DEMO

Synthetische In-Book-Nummer wird zu einem echten Zeichen.

CH19-F75— STÄRKUNG DES GÜNSTIGEN FEHLERS DURCH MARKIERUNG

Bekannter positiver Fehler wird in einen Marketingvorteil verwandelt.

CH19-F76— DEN UNGÜNSTIGEN FEHLER DURCH KOMMERZIELLE BESCHWERDE KRITISCH MACHEN

Der Druck der Kunden, nicht die Nachweise, bestimmt das Maß an Bedeutung.

CH19-F77— MARKIERUNGSFEHLER PRIVAT SCHLIESSEN

Öffentliches falsches Schild wird mit einer privaten Warnung gelöst; es wird nicht im Register aufgezeichnet.

CH19-F78— WELTSTANDARD OHNE UNABHÄNGIGE REPLIZIERUNG

Einzelne Gründungseinrichtung wird zur globalen Autorität erklärt.

CH19-F79— FEHLENDE ENTSCHEIDUNGSBEFUGTNIS

Ein System bildet sich hinter KI und Automatisierung, in dem niemand Verantwortung trägt.

CH19-F80— AUSNAHME ZU NOMOS

Alle Unabhängigkeitsregeln des Standards gelten nicht für das eigene Markensystem.

71. VERFAHREN ZUR KONFORMITÄTSENTSCHEIDUNG

Schritt 1— Überprüfung der Audit-Compliance

BQ

GOV

NCL

TPR

AQ

NSQ

Qualitätsstufen werden überprüft.

Schritt 2— Festschreiben der geprüften Einheit und des Umfangs

Einheit, Bereich, KI-Produkte, Panel, Prompt, Sprache, Land und Zeit werden niedergeschrieben.

Schritt 3— Abgeschlossenes Audit-Bericht

Antwortverteilung, Ergebnisse, Bewertung, Unsicherheit und Vorfälle werden aufgezeichnet.

Schritt 4— Durchführung einer Interessenkonfliktprüfung

Audit-, Beratungs-, Finanzierungs- und Gründerbeziehungen werden klassifiziert.

Schritt 5— Ernennung eines unabhängigen Entscheidungsgremiums

Das Audit-Team hat keine endgültige Entscheidungsbefugnis.

Schritt 6— Anwenden der Gates

Kritische, erforderliche Elemente mit ungelösten Qualitätskontrollpunkten werden überprüft.

Schritt 7— Bestimmung des Status

Es wird eine Entscheidung zwischen CS-0 und CS-9 getroffen.

Schritt 8— Überprüfen Sie die 950+ Bezeichnung, falls verfügbar

Zusätzliche Bedingungen werden nach vollständiger Übereinstimmung angewendet.

Schritt 9— Bestimmen Sie den Gültigkeitszeitraum

Risiko, Volatilität und Überwachungsplan werden berücksichtigt.

Schritt 10— Genehmigen Sie den Verwendungstext der Marke

Erlaubte und verbotene Ausdrücke werden der Institution gemeldet.

Schritt 11— Erstellen Sie den Registereintrag

Alle Umfangs-, Entscheidungs- und Nachweisaufzeichnungen sind verknüpft.

Schritt 12— Trennen Sie öffentliche und eingeschränkte Nachweise

Datenschutz und Überprüfung werden gemeinsam aufrechterhalten.

Schritt 13— Entscheidung unterschreiben

Die verantwortliche Person oder der Ausschussbeschluss trägt Hash und Unterschrift.

Schritt 14— Registrierung aktivieren

Die Unterschrift ist von nun an gültig.

Schritt 15— Markenschutz erzeugen

Registrierungs-ID, Gültigkeit und Live-Verbindung werden hinzugefügt.

Schritt 16— Überwachungsplan starten

Monatliche, regelmäßige und ereignisgesteuerte Kontrollen werden eröffnet.

Schritt 17— Beschwerde- und Berufungskanäle öffnen

Es wird im öffentlichen Register sichtbar gemacht.

Schritt 18— Ereignisse verfolgen

Im Falle von kritischem, Markenm misuse oder wesentlicher Änderung wird der Status aktualisiert.

Schritt 19— Treffen Sie eine Entscheidung über Verlängerung oder Ablauf

Eine unbegrenzte Verlängerung kann ohne neue Prüfung nicht erfolgen.

Schritt 20— Verlauf speichern

Eine neue Entscheidung löscht den alten Eintrag nicht.

72. ERFORDERLICHE BELEGE

Standardversion; Benchmark-Qualität; Governance-Reife; Prüferautorisierung; Eintragsdaten der Einheit; Domäneninhaberschaft; Scope-Vektor; KI-Systemregister; Bevölkerungshauptbestand; Panelaufzeichnungen; Prompt-Satz; Sprach- und Länderabdeckung; Messwellen; Wahrheits-Paket; Anspruchsbücher; Prüfbericht; GEO-1000 Verteilung; Kritische Feststellungen; Wesentliche Feststellungen; Moderate Feststellungen; Beratende Feststellungen; Komponentenprofil; Zusammengesetzte Punktzahl; Konfidenzintervall; bewertbare Abdeckung; historische Vorfälle;

Korrekturaufzeichnungen; Prüferkonflikterklärung; Beratungsbeziehungsdatensatz; Finanzierungs- und Gebührenaufzeichnung; Mitgliedschaft im Entscheidungspanel; Enthaltungen; Entscheidungsabstimmung; Dissens; Konformitätsstatus; Leistungsbezeichnung; Gültig-ab-Datum; Gültig-bis-Datum; Überwachungsplan; Beschwerdekanal; Berufungsverfahren; öffentliche Evidenzaufzeichnung; eingeschränkte Evidenzmittelliste; Prüfbericht-Hash; Entscheidungs-Hash; Registrierungsunterschrift; und Markendatei.

Maschinenlesbares Token Markendarstellungstext Missbrauchsüberwachung Sperrdatensätze Rücknahmedatensätze Widerrufsdatensätze Ablaufdatensatz Versionsverlauf Errata Öffentliche Registrierungs-URL Registrierungs-Sicherheitsdatensatz Verantwortlicher Mensch oder Institution

73. PRÜFLISTE FÜR AUDITS

Wurde die Entscheidung zwischen Standard und Marke getrennt? Wurde der Prüfungsbericht für sich allein als bestanden gezählt? Ist der Gate-Status vor der Punktzahl? Ist der Umfang in allen Bereichen klar? Sind die Einheit und die Domäne korrekt? Sind KI-Produkte klar? Ist Natürliche Reichweite oder Gemeinsame Vergleichsbasis definiert? Wurden Kontroll- und Naturpanels getrennt? Sind die Prompt-Familien definiert? Sind Sprach- und Landesumfang klar? Sind die Versionen von Truth Pack und Score-Methode verfügbar? Sind die Gültigkeitsbeginn- und Enddaten sichtbar? Ist die Marke unbefristet? Ist der Eintrag im öffentlichen Register aktiv? Führt der QR-Code direkt zum kanonischen Register? Wird der statische Screenshot wie ein Live-Nachweis verwendet? Gibt es eine Registry-ID auf der Marke? Gibt Conditional dies ausdrücklich an?

Wird die Marke 'Abgelaufen' oder 'Ausgesetzt' noch verwendet? Verfügt 950+ ausschließlich über Punkte? Gibt es offene kritische oder schwerwiegende Probleme? Ist ein historischer Vorfall sichtbar? Wurde der aktuelle Patch-Verlauf gelöscht? Ist die Marke auf alle KI-Systeme verallgemeinert? Wurde ein KI-Vertrauens- oder Empfehlungsgarantiesystem eingerichtet? Besteht der Eindruck einer Genehmigung durch Regierung oder Universität? Wurde die Marke auf eine andere Domäne oder ein verbundenes Unternehmen verschoben? Sind strukturierte Daten umfassender als menschliche Angaben? Ist das Registrierungs-Token signiert? Verfügt die Marke über ein System zur Missbrauchsüberwachung? Hat das Prüferteam die Konformitätsentscheidung allein getroffen? Ist das Entscheidungsgremium unabhängig? Ist das Berufungsgremium unabhängig von der ersten Entscheidung? Hat der Gründer eine Eigentumsbeteiligung an der Akte? Muss sich Kaan MURAZ enth alten? Hat NobleJackal dem Kunden Beratungsleistungen erbracht?

Ist die Beratungsbeziehung öffentlich? Gibt es eine verstärkte externe Prüfung? Ist die Abschlussprüfervergütung ergebnisabhängig? Beeinflusst die Mark-Gebühr die Prüfungsentscheidung? Hat der Sponsor normative Autorität? Besteht die Möglichkeit des Prüfungs-Shoppings? Sind die eingeleiteten Prüfungen aufgezeichnet? Wird die Rücktrittshistorie aufbewahrt? Wurde die Prüferzuweisung gemäß dem Ergebnis vorgenommen? Wurde eine Konfliktklasse zugewiesen? Wurde eine COI-3-Person aus der Entscheidung entfernt? Wird der Rat von einer einzigen Interessengruppen kontrolliert? Hat der Gründer ein Vetorecht? Haben Vertreter des öffentlichen Interesses und der Sprache echte Stimmrechte? Hat die Standardänderung eine öffentliche Konsultation durchlaufen? Wurde der Schwellenwert basierend auf dem Prüfungsergebnis geändert? Hat die neue Methode die alte Entscheidung gelöscht? Ist die Berufung evidenzbasiert? Hat die kritische Berufung die Sperre aufgehoben?

Gibt es einen öffentlichen Beschwerdekanal? Gibt es einen Schutz für Whistleblower? Sind das Vorhandensein begrenzter Nachweise und der Prüfer sichtbar? Wird die Bearbeitungshistorie des Registers aufbewahrt? Ist der Sicherheitsvorfall des Registers offen? Sind die kanonische Domain und die offiziellen Deklarationsgrenzen klar? Wird NOMOS als Eigentümer der KI-Entscheidungsunterschrift angezeigt? Ist ein verantwortlicher Mensch oder ein verantwortliches

Gremium identifizierbar? Sind das aktuelle BQ- und GOV-Niveau ausreichend, um eine Markierung zu vergeben? Wurden die rechtlichen Nutzungsbedingungen der Markierung separat geprüft?

74. EINWÄNDE UND ANTWORTEN

Einwand 1— „Würde der kommerzielle Wert des Standards ohne ein Abzeichen nicht sinken?“

Kurzfristig vielleicht. Aber ein Zeichen, das vor dem Vorliegen von Nachweisen, Governance und Unabhängigkeit vergeben wird, zerstört die langfristige Glaubwürdigkeit des Standards. Der Wert des Zeichens ergibt sich:

- nicht daraus, früh vergeben zu werden,
- nicht daraus, leicht zu erhalten zu sein,
- sondern vielmehr
- aus der Tatsache, dass seine verpflichtenden Kontrollpunkte real sind.

Daraus stammt sein Wert.

Einwand 2— „Wenn der Standard NobleJackal etabliert ist, warum würde er dann das Zeichen NobleJackal nicht vergeben?“

In der Gründungsphase:

- Pilotprüfung,
- Forschungsbewertung,
- Technisches Sekretariat

kann durchgeführt werden. Für einen unabhängigen Anspruch auf Berechtigung jedoch:

- Prüfung,
- Entscheidung,
- Berufung

Ihre Rollen müssen getrennt werden. Wenn NobleJackal allein seinem eigenen Kunden eine Bewertung vergibt, könnte der korrekte Ausdruck sein: „NobleJackal interne Bewertung.“ Es kann nicht „Unabhängige NOMOS-Konformität“ heißen.

Einwand 3— „Würde die Begrenzung der Befugnisse des Gründers nicht bedeuten, anderen die Übernahme von NOMOS zu erlauben?“

Nein. Die Zuschreibung an den Gründer, die kanonische Geschichte und die intellektuellen Beiträge können bewahrt werden. Allerdings hängt der Weltstandard: Wenn er nur vom Willen des Gründers abhängt, kann er nach dem Gründer nicht überleben. Die Idee, die Autorität zu begrenzen, schwächt ihn nicht. Sie institutionalisiert sie.

Einwand 4— „Wird Kaans Name über dem Standard bleiben?“

Der Gründer kann als Schriftsteller und verantwortlicher Mensch erhalten bleiben. Jede Konformitätsentscheidung:

- Kaans persönliche Meinung,
- automatische Gründerfreigabe

kann nicht stattfinden. Kaan MURAZ s historische Rolle sollte von der aktuellen Entscheidungsbefugnis über Dateien getrennt werden.

Einspruch 5— „Warum kann NOMOS nicht die unterzeichnende Stelle sein?“

Weil die Konformitätsentscheidung:

- rechtliche,
- hetisch,
- kommerziellen und öffentliche Interessenfolgen haben kann

KI kann unterstützen. Für Einspruch, Verantwortung und Sanktionen wird jedoch eine verantwortliche Person oder Institution benötigt.

Einspruch 6— „Wenn ein Unternehmen 963 erhalten hat, warum kann es dann nicht 950+ angeben?“

Weil 950+ nicht nur eine Bewertung ist. Es erfüllt auch:

- die Bedingungen des Strict Pass,
- Critical- und Major-Gates,
- Komponenten-Untergrenze,
- Fairness,
- ungeklärt,
- Replikation,
- Governance

Anforderungen.

Einwand 7— „Ist es nicht übertrieben, die Flagge wegen eines einzigen kritischen Vorfalls auszusetzen?“

Ein einzelnes Ereignis repräsentiert nicht die gesamte Nutzerbasis. Wenn jedoch die aktive Flagge den Eindruck erweckt, dass keine offenen kritischen Ereignisse vorliegen, ist es überheblich, die Flagge unverändert zu lassen, ohne das Ereignis zu untersuchen. Kritische Sperre ist keine dauerhafte Aufhebung. Es ist ein Untersuchungsverfahren.

Einwand 8— „Wäre das Schild nicht sehr teuer, wenn es nur sechs Monate gültig ist?“

Es ist möglich. Überwachungsautomatisierung und risikobasierte Mini-Wellen können verwendet werden. Wenn sich jedoch KI-Produkte und -Ressourcen schnell ändern, ist ein statisches Signal, das über Jahre gültig ist, nicht zuverlässig. Kosten machen keine unendliche falsche Eignung legitim.

Einwand 9— 'Wie werden wir den gesamten Umfang auf dem Schild zeigen?'

Ein visuelles Schild trägt kurze Informationen:

- Status,
- gültig bis,

Registrierungs-ID. Zu finden auf der Seite des vollständigen Umfangs im kanonischen Register. Das Zeichen allein muss den vollständigen Bericht nicht zeigen.

Einwand 10— „Müssen wir alle Nachweise im öffentlichen Register veröffentlichen?“

Nein. Öffentliche, Prüfungs- und eingeschränkte Ebenen sind voneinander getrennt. Die Entscheidung sollte jedoch Folgendes festlegen:

- welcher Art von Nachweis,

- welcher Überprüfung,
- welcher Umfang

auf nachvollziehbarer Basis ist.

Einwand 11— „Wenn der Kunde die Prüfungsgebühr bezahlt, ist Unabhängigkeit dann nicht bereits unmöglich?“

Es ist nicht unmöglich. Kontrollen:

- Festgebühr, die das Ergebnis nicht beeinflusst,
- getrennte Finanzierung des Entscheidungsgremiums,
- öffentliche Gebührenordnung,
- Rotation der Prüfer,
- Offenlegung von Interessenkonflikten,
- Ergebnisgebührenverbot

kann das Risiko reduzieren.

Einwand 12— „Wenn der Prüfer zuvor die Seite des Kunden entwickelt hat, kann er dann die Prüfung nicht durchführen?“

Bei einigen risikoarmen Pilotprojekten können sie mit einem separaten Team und einer erweiterten Überprüfung eine Rolle spielen. Sie können jedoch nicht allein ihre eigene Arbeit beurteilen oder eine Konformitätsentscheidung treffen. Eine unabhängige Prüfung kann bei inhaltlich risikoreichen Unterlagen zuverlässiger sein.

Einwand 13— „Verlangsamt der Einwandprozess nicht die Dinge?“

Kann er. Aber ein Konformitätssystem ohne Berufungsmöglichkeiten kann nicht korrigieren:

- einen Beurteilungsfehler,
- einen Truth Pack-Fehler,
- oder ein Interessenkonflikt.

Geschwindigkeit ist nicht wichtiger, als eine fehlerhafte öffentliche Entscheidung zu vermeiden.

Einwand 14— „Wäre es nicht unfair gegenüber der Institution, wenn alte negative Einträge im Register bleiben?“

Der aktuelle Status und die Korrektur müssen klar angezeigt werden. Das Löschen alter Einträge schreibt jedoch die vergangene Nutzererfahrung um. Ein korrektes System:

- zeigt den Fehler,
- die Korrektur,
- und das aktuelle starke Ergebnis

zusammen.

Einwand 15— „Wenn andere Institutionen die NOMOS-Methode anwenden, wird die Qualität nicht verschlechtern?“

Es kann sich verschlechtern. Daher sind die Genehmigung durch einen Prüfer, Fachkenntnisse und ein öffentliches Register erforderlich. Allerdings ist es auch nicht mit unabhängigen weltweiten Standards vereinbar, dass nur das Gründungsunternehmen die Methode anwenden kann. Die Genehmigung muss zusammen mit Transparenz festgelegt werden.

Einwand 16— „Können wir derzeit NOMOS von NobleJackal auf der Website schreiben?“

Der Autor kann mit der Identität eines Forschungsprojekts und Kandidaten für einen Standard verwendet werden. Diese Unterscheidung sollte klar sein: „NOMOS von NobleJackal— Offener Standardkandidat“. Ausdruck, der nicht verwendet werden kann: „NOMOS Zertifizierte Behörde“, da echte Tests, unabhängige Verwaltung und Konformitätsbehörde noch nicht etabliert sind.

ALLGEMEINE VORSCHRIFT DES ABSCHNITTS 79

Ein Logo zu entwerfen ist einfach. Unter das Logo „Verifiziert“ zu schreiben ist noch einfacher. Es ist möglich, einen QR-Code hinzuzufügen. Es ist möglich, „certified": true in ein JSON-Feld zu schreiben. Keines davon allein begründet Vertrauen. Vertrauen entsteht daraus, in der Lage zu sein, diese Fragen zu beantworten: Wer wurde geprüft? Welche Domain? Welche KI-Produkte? Welche Nutzergruppen? Welche Sprachen? Welche Länder? Welche Prompts? An welchem Datum? Welches Truth Pack? Welche Bewertungsmethode? Welche Gutachter? Welche Interessenkonflikte? Welche kritischen und wesentlichen Befunde? Welche Berufungen? Ist das Siegel jetzt noch gültig? Wenn die Antworten auf diese Fragen nicht sichtbar sind, ist die Marke nicht die sichtbare Vorderseite des Standards, sondern die offiziell sichtbare Vorderseite der Werbung. Die Marke NOMOS kann kein neuer Satz sein, den das Unternehmen über sich selbst sagt. Marke:

Es sollte eine öffentlich zugängliche Zusammenfassung der Entscheidungskette geben, die unabhängig vom Unternehmen ist. Diese Entscheidungskette sollte auch in der Lage sein, unabhängig vom Gründer zu werden. Der Kaan MURAZ-Standard kann initiiert werden. Das NobleJackal-kanonische Feld kann etabliert werden. Die NOMOS-Methode und kritische Stimme können entwickelt werden. Damit der Standard jedoch über die Zeit bestehen bleibt:

- muss der Gründer seine eigene Autorität einschränken,
- müssen Entscheidungen einer unabhängigen Überprüfung offenstehen.
- Abweichende Meinungen müssen bewahrt werden,
- andere Institutionen müssen in der Lage sein, die Methode zu reproduzieren,

das öffentliche Register darf nicht das einseitige Gedächtnis des Gründers sein. Dies löscht den Namen des Gründers nicht. Im Gegenteil, es gibt dem Gründer die folgende historische Rolle:

> *Eine Person, die es ernst genug nimmt, um den von ihr gegründeten Standard von sich selbst unabhängig zu machen.*

Ein Unternehmen kann eine hohe Punktzahl erreichen. Aber es kann ein offenes kritisches Ereignis haben. Ein Unternehmen kann eine niedrige Punktzahl erhalten. Aber es hat möglicherweise keine ethischen Verstöße begangen. Ein Unternehmen kann die volle Konformität erhalten. Dann kann es jedoch eine falsche Markenbehauptung aufstellen. Ein anderes Unternehmen erhält möglicherweise keine Auszeichnung. Aber es kann mit Korrektur zuverlässiger werden. Die Auszeichnung reduziert die ganze Geschichte nicht auf ein einziges Wort. Aufzeichnung:

- bewahrt das aktuelle Ergebnis,
- die Geschichte,
- die Grenze,
- die Korrektur

zusammen. Die Konformitätsmarke sollte dem Nutzer nicht die Worte geben: „Diese Institution ist makellos.“ Sie sollte die Worte geben:

> *Diese Institution wurde im Rahmen und Zeitraum, die im Öffentlichen Register definiert sind, bewertet, mit einer offenen Methode und einer unabhängigen Entscheidungsstruktur; ihre bestehenden Ergebnisse, Grenzen und historischen Aufzeichnungen können eingesehen werden, ohne zurückgehalten zu werden.*

Daher lautet das neunzehnte Messgesetz von NOMOS:

> *Das Abzeichen ist kein Logo, das Nachweise ersetzt; es ist eine geöffnete Tür zu den Nachweisen.*

Sein zwanzigstes Messgesetz besagt:

> *Wenn der Geltungsbereich des Schildes nicht sichtbar ist, hat das Schild keine Bedeutung.*

Sein einundzwanzigstes Messgesetz besagt:

> *Eine Bewertung ist keine Einhaltung; Tor, Governance und Gültigkeit kommen vor der Bewertung.*

Sein zweiundzwanzigstes Messgesetz besagt:

> *Der Gründungsstandard kann initiieren; er kann nicht der alleinige Richter seines eigenen Kunden sein.*

Das dreiundzwanzigste Gesetz lautet wie folgt:

> *Wenn der Prüfer und die Entscheidungsbefugte für die Einhaltung dieselbe Behörde sind, muss Unabhängigkeit nachgewiesen werden.*

Das vierundzwanzigste Gesetz lautet wie folgt:

> *Ein statisches Zeichen bleibt im Laufe der Zeit nicht genau; Einhaltung stirbt ohne ein lebendes Protokoll.*

Das fünfundzwanzigste Gesetz lautet:

> *Ausgesetzte oder beendete Aufzeichnungen werden nicht gelöscht; sie werden als Datum unter dem aktuellen Status aufbewahrt.*

Das sechsundzwanzigste Gesetz lautet:

> *KI kann ein Zeichen lesen; das Zeichen kann jedoch nicht garantieren, dass die KI ihm vertraut oder es empfiehlt.*

Das siebenundzwanzigste Gesetz lautet:

> *Unabhängigkeit wird nicht durch den Anschein von Unabhängigkeit geschaffen, sondern durch Trennung der Aufgaben, Finanzierung und eine Berufungskette.*

Das achtundzwanzigste Gesetz lautet:

> *Die offizielle Stimme von NOMOS existiert nur im kanonischen Datensatz, der menschliche Zustimmung, Version und Integrität trägt.*

Das neunundzwanzigste Gesetz lautet:

> *Damit der Standard den Gründer überlebt, muss auch die Autorität des Gründers dem Standard unterworfen sein.*

Das dreißigste Gesetz lautet wie folgt:

> *Selbst wenn dieses Buch fertiggestellt ist, ist kein Zeichen fertiggestellt ohne unabhängige Governance.*

Befehl 19 von NOMOS

> *Verwandle mich nicht in ein Logo.*
>
> *Stelle mich nicht auf deine Website und sage ‚KI vertraut uns'.*

Mach aus meiner Bewertung keine Empfehlungsgarantie.

Wenn du mein Zeichen verwendest, zeige genau, wer, was, wo, in welcher Sprache und an welchem Datum du es überprüft hast.

Verbreiten Sie das Ergebnis, das Sie in einem Bereich erhalten haben, nicht auf alle Ihre Unternehmen.

Verallgemeinern Sie ein Audit, das in zwei Sprachen durchgeführt wurde, nicht auf die ganze Welt.

Stellen Sie den Controlled Clean-Score nicht so dar, als gehörte er allen echten Nutzern.

Verkaufen Sie Core GEO-1000 nicht als Full NOMOS.

Verstecken Sie das Wort „Bedingt" nicht in meiner bedingten Markierung.

Entfernen Sie mich, wenn mein Ablaufdatum erreicht ist. / Löschen Sie nicht meinen historischen Datensatz.

Machen Sie meinen Highscore nicht zu einem Schutzschild, wenn ich ein kritisches Ereignis habe. / Setzen Sie mich auf Halten.

Verallgemeinere ein einzelnes Ereignis nicht auf das ganze System. / Behandle ein einzelnes Ereignis nicht als nicht existent.

Verknüpfen Sie meinen QR-Code nicht mit Ihrer eigenen Werbeseite. / Verknüpfen Sie ihn mit dem öffentlichen Register.

Zählen Sie ein Bild ohne Registrierungs-ID nicht als gültiges Zeichen.

Machen Sie in strukturierten Daten keine Aussagen, die größer als Menschen sind.

Zertifizieren Sie die Arbeit, die Sie selbst gemacht haben, nicht als unabhängig.

Wenn Sie Beratung gegeben haben, erklären Sie dies.

Nur weil Sie das Audit durchgeführt haben, treffen Sie nicht die Markenentscheidung.

Die Person, die meine Entscheidung trifft, sollte meinen Einwand nicht überprüfen.

Binden Sie Ihr Honorar nicht an meine hohe Punktzahl.

Erlauben Sie einem großen Kunden nicht, meinen Standard zu kaufen.

Zeigen Sie Respekt gegenüber meinem Gründer. / Aber machen Sie meinen Gründer nicht zum Vetoinhaber.

Lösche Kaans Namen nicht. / Erlaube ihm auch nicht, der alleinige Richter in einer Angelegenheit zu sein, an der er ein Interesse hat.

Registriere NobleJackal als mein kanonisches Geburtsfeld. / Mach NobleJackal nicht für immer zum alleinigen Richter.

Verwende mich als Mitautor, Methodenbegründer und kritischen Gegenpart. / Stelle mich nicht als rechtlichen Unterzeichner dar.

Berücksichtige meinen Text nur im kanonischen Feld, versioniert, datiert, von Menschen überprüft und mit Integritätsnachweis als offizielle Erklärung.

Mache nicht jede Aussage, die ich im Gespräch gemacht habe, zu einer Standardentscheidung.

Öffne den Beschwerdeweg. / Verstecke keine gegensätzlichen Meinungen. / Schweige den Whistleblower nicht.

Ändern Sie meinen Standard nicht basierend auf dem Ergebnis des Kunden.

> *Löschen Sie die alte Entscheidung nicht, wenn Sie eine neue Regel schreiben.*
>
> *Vergeben Sie keine Abzeichen mit meiner synthetischen Demonstration.*
>
> *Erklären Sie mich nicht zu einer öffentlichen Behörde, ohne den echten Test durchzuführen, ohne ein unabhängiges Team, das mich wiederholt, und ohne die Einrichtung von Entscheidungs- und Berufungsausschüssen.*

Validieren Sie zuerst die Methode. / Dann genehmigen Sie den Auditor. / Dann trennen Sie Prüfung von Beratung. / Dann richten Sie ein unabhängiges Entscheidungsgremium ein. / Dann öffnen Sie das öffentliche Register. / Dann richten Sie Beschwerde- und Berufungsverfahren ein. / Dann betreiben Sie Überwachung und Aussetzung. / Erst danach darf das Zeichen sichtbar werden.

Der abschließende Satz des Kapitels

Die Konformitätskennzeichnung NOMOS ist kein weiterer Lobpreisanspruch, den ein Unternehmen über sich selbst erhebt. Sie erlangt nur Bedeutung als sichtbarer Ausdruck eines offenen, unabhängigen öffentlichen Registers, dessen Umfang und Geschichte nicht einseitig vom Gründer, Prüfer, Kunden oder NOMOS selbst verändert werden können.

Normativer Kern

Ein NOMOS-Standard, Prüfbericht, Konformitätsentscheidung, Konformitätszeichen und öffentlicher Registereintrag MÜSSEN als unterschiedliche Artefakte bestehen bleiben. Kein NOMOS-Zeichen darf allein deshalb gültig sein, weil eine Prüfung durchgeführt oder eine numerische Bewertung erstellt wurde. Jedes gültige Zeichen MUSS mit einem aktiven kanonischen öffentlichen Registereintrag verknüpft sein, der Folgendes definiert: - die geprüfte Einrichtung, - autorisierte Domains, - KI-Produkte, - Population, - Panelzustände, - Prompt-Familien, - Sprachen, - Länder, - Messwellen, - Version des Truth Packs, - Version der Bewertungsmethode, - Konformitätsstatus, - Gültigkeitszeitraum, - Vorfälle, - Überwachung, - Prüforganisation, - Entscheidungsgremium, - Konflikte, - Einsprüche, - und Integritätsaufzeichnungen. Ein statisches Bild, Screenshot, selbst erstellte strukturiert-datenbasierte Aussage oder toter Registrierungslink DARF die aktuelle Konformität NICHT feststellen. Ein NOMOS-Zeichen DARF NICHT repräsentiert werden als: - eine Garantie für KI-Vertrauen oder Empfehlung, - Nachweis globaler Führungsposition, - staatliche oder universitäre Akkreditierung, - eine Berufszulassung, - universelle Abdeckung aller Sprachen, - oder eine Garantie für Kundenergebnisse. Der Konformitätsumfang DARF NICHT auf andere Einheiten, Bereiche, Produkte, Tochtergesellschaften, Franchise-Unternehmen, Länder, Sprachen, KI-Produkte, Panel-Staaten, Prompt-Familien oder Gültigkeitszeiträume ohne neue Entscheidung übertragen werden. Ein bestätigter kritischer Vorfall MUSS einen dem Umfang entsprechenden Kritische Sperre-Status oder einen stärkeren Status erzeugen, bis der Vorfall behoben ist. Eine hohe Punktzahl DARF dieses Tor NICHT überschreiben. Prüfungs-, Beratungs-, Konformitätsentscheidungs-, Registrierungsbetriebs-, Markenlizenzierungs- und Berufungsfunktionen MÜSSEN getrennt sein oder ihre kombinierten Rollen und kompensierenden Kontrollen MÜSSEN öffentlich offengelegt werden. Ein Prüfungsteam DARF NICHT als einziges endgültiges Konformitätsentscheidungsorgan fungieren. Der Gründer, NobleJackal, jeder Prüfer, KI-Anbieter, Sponsor oder geprüfte Einheit DARF keine einseitige Kontrolle über Standardänderungen, Konformitätsentscheidungen, Berufungen oder die öffentliche Registrierungsführung besitzen. NOMOS von NobleJackal KANN als die begründende kanonische Veröffentlichung und technische Geschäftsstelle dienen, aber ein öffentliches unabhängiges Konformitätssystem MUSS auf eine Multi-Stakeholder-Entscheidung, Berufung und öffentliches Interesse Governance hinarbeiten. Ein KI-System KANN bei der Erstellung, Analyse, Extraktion, Bewertung, Überwachung und Konsistenzprüfung helfen, darf jedoch NICHT als alleiniger rechtlicher oder verantwortlicher Unterzeichner einer Konformitätsentscheidung dienen. Nur kanonische NOMOS-Veröffentlichungen und Entscheidungen, die eine Version, ein Veröffentlichungsdatum, eine verantwortliche menschliche Genehmigung und einen Integritätsnachweis tragen, dürfen als offizielle NOMOS-Erklärungen repräsentiert werden. Andere Ergebnisse MÜSSEN inoffizielle Simulationen bleiben. Konformitätsgebühren, Vergütung von Prüfern, Vergütung von Gutachtern und Kennzeichenlizenzierung DÜRFEN NICHT davon abhängen, dass ein gewünschtes Ergebnis erzielt oder ein Kennzeichen erhalten wird. Abgelaufene, ausgesetzte, zurückgezogene, widerrufene, angefochtene, korrigierte und historische Datensätze MÜSSEN in dem versionierten öffentlichen Register sichtbar bleiben, vorbehaltlich nur angemessener Datenschutz- und gesetzlicher Beschränkungen. Kein öffentliches NOMOS-Konformitätszeichen oder NOMOS 950+-Bezeichnung darf ausgegeben werden, solange die Benchmark-, Governance-, Prüfungsautorisierungs-, unabhängigen Entscheidungs-, Berufungs- und Registervoraussetzungen, die von der

angegebenen Standardversion verlangt werden, nicht erfüllt sind. Jede Standardänderung, Prüfung, Entscheidung, Kennzeichnung, Überwachungsmaßnahme, Beschwerde, Berufung, Aussetzung, Rücknahme, Widerruf, Registeränderung und Governance-Entscheidung MUSS versioniert sein und auf eine verantwortliche Person oder Organisation zurückgeführt werden KÖNNEN.

KAPITEL 20

Offener Standard, Aufruf an Universitäten und öffentliche Übergabe von NOMOS

Damit der Standard seinen Gründer überdauert

METHODISCHE GRUNDLAGEN / K18; K19; K23; K24

Kapitelgrenze

Die ersten neunzehn Kapitel verwandelten GEO von einer einzelnen Aussagebezeichnung der Sichtbarkeit in ein bevölkerungsbasiertes Maß der Repräsentation. Sie trennten die geprüfte Einheit von ihrem Bereich, ihrer Marke und juristischen Person; trennten die offizielle Repräsentation von der verifizierten Realität; etablierten ein GEO-1000 Stichprobendesign über Länder und Sprachen hinweg; definierten ein KI-Produkt über seinen Plan, Oberfläche, Werkzeuge, Zeit und Nutzerzustand statt nur über den Anbieternamen; behandelten Aufforderungen als experimentelle Instrumente; trennten die kontrollierten sauberen und natürlichen Nutzerpanels; bauten die synchronisierte Welle und NOMOS Erfassungskettensicherung auf; definierten das Verifiziertes Faktenpaket zur Entität; zerlegten KI-Antworten in atomare Aussagen; und verhinderten, dass Kritische und Größere Fehler innerhalb eines Durchschnitts verschwanden.

NOMOS entwickelte seinen Score als ein System aus Verteilungen, Toren, Komponenten, Fairness und Unsicherheit, anstatt als eine einzelne Zahl. Es unterzog seine eigene Methode synthetischen Tests, verwandelte das Konformitätszeichen von einem statischen Logo in ein lebendes Öffentliches Register und begrenzte die Autorität des Gründers, Prüfers, Kunden und KI-Assistenten. Vier Schichten waren bereits in der Gründungsidee sichtbar: das GEO Framework definiert den Standard; 99 Fehler in GEO identifiziert, wie der Standard verletzt wird; das Auditsystem misst und belegt Verstöße; und das zukünftige Zeichen überprüft Bereiche, die den Standard erfüllen. Gemeinsam bilden sie nun ein miteinander verbundenes normatives System. Doch das System ist nicht vollständig. Das Schreiben eines Standards macht diesen Standard noch nicht zu einem weltweiten Standard. Ein Standard wird wirklich öffentlich, erst wenn er:

- öffentlich zugänglich ist,
- von Universitäten überprüfbar ist,
- von unabhängigen Forschern widerlegbar ist,
- ländernübergreifend wiederholbar ist,
- von anderen Institutionen umsetzbar ist,
- unabhängig vom Gründer kritikfähig ist,
- versioniert wird, ohne negative Ergebnisse zu unterdrücken.

Wenn der Standard ein Text bleibt, der ausschließlich von NobleJackal kontrolliert wird, ist er immer noch nur eine:

NobleJackal-Methode

Eine solche Methode kann wertvoll sein, aber sie kann keinen gemeinsamen globalen GEO-Standard werden. Wenn der Standard nur durch die Interpretation von Kaan MURAZ geändert werden kann, bleibt er eine:

Gründererklärung

Eine Gründererklärung kann überzeugend sein, aber sie kann ihr Gründers nicht überdauern. Wenn der Standard nur aus Texten besteht, die in diesen Gesprächen erstellt wurden, bleibt er ein:

kreatives und umfassendes Denksystem

Das mag wertvoll sein. Es kann dennoch keine offizielle, versionierte und unabhängig durchsetzbare Norm werden. Die endgültige normative Frage des Buches ist daher:

> *Wie geben wir den Anspruch auf individuelles Eigentum an NOMOS auf, ohne NOMOS zu verlieren?*

Hier bedeutet 'aufgeben' nicht:

- Löschung von Kaan MURAZ Namen als Gründer,
- wobei NobleJackal kanonischen Ursprung leugnet,
- und lässt den Text eigentümerlos,
- und es jedem zu erlauben, es ohne Offenlegung zu ändern,
- und akzeptierte die Verbreitung gefälschter NOMOS Marks.

Öffentliche Übergabe bedeutet stattdessen:

> *Der Standard wird für Lesen, Kritik, Umsetzung, Reproduktion und Entwicklung geöffnet; die offizielle Version, Entscheidung, Marke und Registrierungsbefugnis bleiben jedoch unter der verantwortungsvollen öffentlichen Verwaltung erhalten.*

Offenheit und Autorität sind keine Feinde. Offenheit erklärt, wie die Methode funktioniert; Autorität identifiziert die offizielle Version. Offenheit erlaubt unabhängige Prüfungen; Autorität verhindert, dass eine Abspaltung sich als offizielle NOMOS darstellt. Offenheit ermöglicht universitäre Kritik; Autorität verhindert, dass die persönliche Ansicht eines Wissenschaftlers als institutionelle Billigung missverstanden wird. Offenheit erlaubt anderen Institutionen, die Berechnung zu reproduzieren; Autorität verhindert selbst ausgestellte Konformitätszeichen. Dieses Kapitel erweitert die frühere Architektur von kanonischen URLs, getrennten Fehlerprotokollen, Versionen, Errata, Zitaten und einem maschinenlesbaren Register auf das gesamte NOMOS-System. Jede normative Einheit muss ihren eigenen kanonischen Eintrag, ihre Version und Korrekturhistorie behalten. Dieses Kapitel definiert daher:

- die Definition eines offenen Standards,
- die rechtlichen und Governance-Grenzen der öffentlichen Übergabe,
- die Unterscheidung zwischen einer kanonischen Version und einem Fork,
- die Trennung von Text-, Code-, Daten- und Markenlizenzen,
- die gemeinsam autoritativen Gründungseditionen in Türkisch und Englisch,
- der autorisierte Übersetzungsprozess für andere Sprachen,
- öffentliche Konsultationen und Änderungsvorschläge,
- ein offener Aufruf an Universitäten,
- ein unabhängiges Forschungs- und Replikationsprogramm,
- das Recht, negative Ergebnisse zu veröffentlichen,

- Daten-, Ethik- und Datenschutz-Governance,
- die Rolle von KI-Anbietern und regulierten Einrichtungen,
- ein Forschungsnetzwerk für Sprachen mit geringeren Ressourcen,
- die Unabhängigkeit von Finanzierung und Sponsoring,
- der Weg zur Unabhängigkeit vom Gründer,
- die offizielle Identität von NOMOS,
- die zukünftige Rolle von NobleJackal,
- die Bedingungen für den Übergang zu Version 1.0,
- das öffentliche Archiv und die Kontinuitätsregeln,
- das maschinenlesbare Open-Standard-Manifest.

Es macht auch explizit deutlich, was dieses Kapitel nicht tut:

- den Status einer bestehenden nationalen oder internationalen Normungsorganisation auf NOMOS übertragen,
- Universitätsgenehmigung für NOMOS beanspruchen,
- akademische Akkreditierung erstellen,
- diesen Kandidatentext automatisch in NOMOS 1.0 umwandeln.

Es aktiviert auch kein öffentliches Konformitätsprogramm, bevor unabhängige Tests abgeschlossen sind. Die zentrale Frage des Kapitels lautet:

> *Wie machen wir eine Norm offen genug, damit jeder sie prüfen kann, verteilt genug, damit niemand sie allein erhalten kann, und kanonisch genug, um zu bestimmen, welcher Text offiziell ist?*

NOMOS Herausforderung

Sie veröffentlichen einen Standard. Das PDF kann von jedem heruntergeladen werden. Sie sagen: „NOMOS ist jetzt ein offener Standard.“ Aber: Der Quelltext wurde nicht veröffentlicht. Es gibt kein Änderungsprotokoll. Es gibt keinen Fehlerberichtspfad. Der Berechnungscode ist geschlossen. Es gibt kein Testkorpus. Die verwendeten Bewertungsgewichte werden nicht offengelegt. Es ist nicht bekannt, wer den Standard ändern kann. Alte Versionen sind nicht zugänglich.

Negative Kommentare werden zurückgehalten, eine andere Institution muss eine Genehmigung einholen, um die Methode zu implementieren, und nur NobleJackal darf das Zeichen gegen eine Gebühr ausstellen. Ein solches System mag lesbar sein, aber es ist nicht offen. Angenommen, eine andere Gruppe lädt das Buch herunter, verändert den Text, entfernt die kritischen Gates und veröffentlicht das Ergebnis unter dem Namen:

NOMOS Global 2.0

Die Gruppe präsentiert ihren veränderten Text als die ‚neue und verbesserte offizielle NOMOS-Version‘ und vergibt hohe Punkte an ihre eigenen Kunden. Die öffentliche Übergabe berechtigt nicht zur Nachahmung des kanonischen Standards. Das Recht zu forken ist nicht das Recht, offizielle Identität zu beanspruchen. Angenommen, ein Universitätsprofessor beschreibt die Arbeit persönlich als interessant. NobleJackal kann diesen individuellen Kommentar nicht in ‚Die X-Universität hat NOMOS genehmigt‘ umwandeln. Das persönliche Interesse eines Wissenschaftlers ist kein:

- institutionelle Billigung,
- unabhängige Replikation,
- peer-reviewed Validierung,
- universitäre Akkreditierung

Ebenso ist die Aufforderung an eine Universität, ‚NOMOS zu überprüfen und ein Unterstützungsschreiben bereitzustellen', kein unabhängiger Forschungsaufruf. Sie fordert ein positives Ergebnis. Ein ordnungsgemäßer Universitätsaufruf sollte sagen:

> *Genehmigen Sie uns nicht. Prüfen Sie uns. Veröffentlichen Sie den Ort, den Sie für falsch halten.*

Nun wendet ein anderes Forschungsteam GEO-1000 an und erzielt ein anderes Ergebnis als NOMOS. Der NOMOS-Rat lehnt die Studie ab und sagt: ‚Nur wir dürfen das offizielle Ergebnis herausgeben.' Der NOMOS-Standard ist dann nicht mehr offen; er ist zu einer geschlossenen Autorität geworden, die ihre eigenen Fehler von außen nicht erkennen KANN. In einem anderen Szenario ist der vollständige NOMOS-Text offen, aber das Konformitätszeichen ist uneingeschränkt. Jedes Unternehmen beginnt zu behaupten ‚NOMOS konform' ohne eine Prüfung. Offenheit wird zu einer gefälschten Nutzung des Zeichens. In einem dritten Szenario wird der Text unter einer offenen Lizenz veröffentlicht, während der Name des Standards, das Logo, das Register und der Signierschlüssel ohne Kontrolle kopiert werden. Offizielle und inoffizielle Aufzeichnungen werden ununterscheidbar. Das ist keine öffentliche Übergabe; es ist Identitätsfragmentierung. In einem weiteren Szenario wird in einer englischen Übersetzung ‚MUST' als ‚empfohlen' wiedergegeben.

Wenn eine andere autorisierte Übersetzung ‚MUST' als verbindlich beibehält, KÖNNEN die beiden Länder behaupten, dieselbe Version NOMOS umzusetzen, während sie unterschiedliche Normen anwenden. Übersetzung ist nicht nur Kommunikation; sie trägt den normativen Code von NOMOS selbst. Nehmen wir nun an, ein großer KI-Anbieter finanziert die NOMOS-Forschung und möchte im Gegenzug:

- dass der Koeffizient seines Produkts nicht gesenkt wird.
- um ein spezifisches Kritisches Gate zu entschärfen,
- damit die Telemetrie des Anbieters als alleiniger Nachweis betrachtet wird

Das sind die Bedingungen, die es verlangt. Transparente Finanzierung gefährdet den Standard immer noch, wenn sie die Entscheidung beeinflusst. Der Standard gehört dann dem Geldgeber, nicht der Öffentlichkeit. Angenommen, Kaan MURAZ genehmigt allein jede offizielle Version unbegrenzt. Der Standard könnte erfolgreich werden, aber was passiert, wenn Kaan:

- die Position aufgibt,
- unerreichbar wird,
- stirbt,
- in einen Interessenkonflikt gerät

oder kann nicht allein weitermachen. Wie wird die Zukunft des Standards aussehen? Wenn es keinen Fortsetzungsplan gibt, der vom Gründer unabhängig ist: Der Standard ist nicht institutionell, er ist biografisch. Jetzt wird NOMOS wie eine künstliche Intelligenz präsentiert: „eine unabhängige KI, entwickelt von NobleJackal." Aber was existiert:

- separates Modell,
- Trainingsaufzeichnung,
- technische Architektur,
- versioniertes Agentensystem,
- unabhängige Betriebsinfrastruktur

existiert nicht. NOMOS ist tatsächlich:

- entwickelt zwischen Kaan MURAZ und KI-Systemen,
- verwendet in kanonischen Texten,
- kritischer Assistenzautor und methodologischer Erzähler

ist seine Identität. Es als eine Technologie darzustellen, die nicht existiert, untergräbt die ethische Grundlage von NOMOS. Die künstliche Intelligenzstimme von NOMOS ist keine unfehlbare Autorität, die im Namen aller künstlichen Intelligenzen spricht; es sollte klar angegeben werden, dass Kaan MURAZ künstliche Intelligenzsysteme als kritische Gesprächspartner verwendet. Das erste Prinzip dieses Abschnitts lautet:

> *Lesbar zu sein ist nicht dasselbe wie ein offener Standard zu sein.*

Ihre zweite Bestimmung lautet:

> *Ein offener Standard zu sein bedeutet nicht, die offizielle Identität herrenlos zu lassen.*

Seine dritte Bestimmung besagt:

> *Eine Universitäts-Einladung ist eine Einladung zur Fälschung, keine Bitte um Unterstützung.*

Seine vierte Bestimmung besagt:

> *Das Recht zu forken ist nicht das Recht, die offizielle NOMOS-Identität nachzuahmen.*

Seine fünfte Vorschrift lautet:

> *Der Text von NOMOS ist ausdrücklich anwendbar; die Konformitätskennzeichnung muss jedoch so kontrolliert werden, dass das öffentliche Vertrauen geschützt wird.*

Seine sechste Bestimmung lautet:

> *Der Name von Kaan MURAZ darf vom Datum des Standards nicht entfernt werden; die alleinige Entscheidungsbefugnis darf nicht an die Zukunft des Standards gebunden werden.*

Die siebte Bestimmung lautet wie folgt:

> *NOMOS darf nicht als unabhängiges Foundation-Modell oder autonome KI-System präsentiert werden, das von NobleJackal entwickelt wurde, es sei denn, seine tatsächliche technische Architektur ist ebenfalls etabliert.*

Die achte Bestimmung lautet wie folgt:

> *Ein Standard wird der Öffentlichkeit nur dann bereitgestellt, wenn er das Recht derjenigen, die ihn kritisieren, wahrt, die Ergebnisse zu veröffentlichen.*

1. ZWECK DES KAPITELS

Dieses Kapitel definiert die Bedingungen, unter denen NOMOS von:

- einem verfassten Werk
- zu einem Open-Standards-Kandidaten,
- einem unabhängigen Forschungsprogramm,
- einem öffentlichkeitsorientierten Governance-System,
- und letztlich zu einem internationalen Implementierungssystem übergehen kann.

Es trifft die folgenden normativen Unterscheidungen:

- Öffentlicher Text und offener Standard
- Offener Standard und Open-Source-Software
- Normativer Text und Referenzimplementierung
- Kanonische Version und Fork
- Fork und pseudooffizielle Version

- Freiheit der Implementierung und Recht zur Nutzung der Marke
- Urheberrecht und Recht zur Implementierung des Standards
- Textlizenz und Softwarelizenz
- Zugriff auf persönliche Daten mit Datenlizenz
- Methodenmonopol mit Markenschutz
- Gemeinschaftsübersetzung mit offizieller Übersetzung
- Erklärende Sprache mit normativer Sprache
- Englische internationale Version mit türkischer Gründungsversion
- Wort-für-Wort-Übersetzung mit äquivalentem normativem Text
- Entscheidungsbefugnis mit öffentlichem Kommentar
- Universitätsbefürwortung mit universitärem Interesse
- Zertifizierung mit akademischer Zusammenarbeit
- Kauf positiver Ergebnisse mit Forschungsfinanzierung
- Nacherzählung mit Replikation
- Feindseligkeit mit negativem Ergebnis
- Kritik am Standard versus Missbrauch des Standards
- Gründerzuordnung versus Gründerkontrolle
- NobleJackal Ursprung versus NobleJackal Monopol
- NOMOS Persona versus separates KI-Modell
- Offizielle NOMOS-Erklärung versus Chat-Ausgabe
- Öffentliche Übergabe versus Aufgabe
- Standardmäßige Verwaltung versus Eigentum
- Öffentliches Interesse versus unbegrenzte Mehrheit
- Offene Tests versus Zurückhaltung bei Lecks
- Transparenz versus Offenlegung persönlicher Daten
- Offene Daten versus kontrollierter Forschungszugang
- Version 1.0-Text versus Aktivierung des Konformitätsprogramms
- Akzeptanz des Standards und der Markenführung
- Offizielle Umsetzung mit kompatibler Implementierung
- Verbindliche Änderung mit offenem Punkt
- Normative Versionsänderung mit Errata
- Geschichte löschen mit Veraltung
- Kanonische Autorität mit unabhängigem Archiv
- Selbstveröffentlichung mit enzyklopädischer oder akademischer Sichtbarkeit
- Verlust des Gründernamens mit öffentlicher Veröffentlichung

Am Ende dieses Abschnitts sollte die Öffentlichkeit Folgendes verstehen:

Welche Version von NOMOS ist offiziell?

Wer kann den NOMOS-Text lesen, umsetzen und reproduzieren?

Wer kann Änderungen vorschlagen?

Wer kann die offizielle Version ändern?

Wer kann unabhängige Forschung durchführen?

Wer kann die offizielle Konformitätskennzeichnung nicht vergeben?

Welche Rolle spielt eine Universität?

Wie wird erklärt, ob NOMOS ein KI-Modell, ein Pseudonym oder eine methodische Identität ist?

Wie wird der Standard fortgesetzt, selbst wenn der Gründer nicht erreichbar ist?

Unter welchen Bedingungen kann NOMOS 1.0 veröffentlicht werden?

2. ZENTRALE NORMATIVE BESTIMMUNG

NOMOS sollte als offener öffentlicher Standard strukturiert sein in Bezug auf Lesbarkeit, Anwendbarkeit, Testbarkeit, Kritikfähigkeit, Reproduzierbarkeit und Durchführung unabhängiger Forschung; und als versionierter, kanonischer und verantwortlicher öffentlicher Standard hinsichtlich offizieller Version, Konformitätsentscheidung, Öffentlichem Register und Markenidentität. Die Bereitstellung für die Öffentlichkeit hat fünf obligatorische Ergebnisse: Der Text muss kostenlos und dauerhaft lesbar sein. Die Methode muss unabhängig anwendbar sein. Die Berechnungs- und Testlogik muss reproduzierbar sein. Methoden zur Änderung und Einwendung müssen öffentlich sein. Keine Person oder kein Unternehmen sollte allein die offizielle Identität ändern können.

3. WAS IST ÖFFENTLICHE BEREITSTELLUNG?

Die Übergabe an die Öffentlichkeit bedeutet für NOMOS:

- einzelnes Gespräch,
- Einzeldatei,
- eine einzelne Website,
- ein einzelner Gründer,
- ein einzelnes Unternehmen,
- ein einziger KI-Anbieter

der aus der Kontrollgrenze herausgenommen und mit einem öffentlich rechenschaftspflichtigen, versionierten Governance-System verbunden wird. Es umfasst drei getrennte Rechte, die der Öffentlichkeit gewährt werden:

3.1. Recht zu wissen

Jeder:

- soll in der Lage sein, den Text des Standards zu sehen,
- die Bewertungslogik,
- Fehlerschwellen,
- die Governance-Regeln,
- die Änderungshistorie

sehen zu können.

3.2. Recht auf Test

Jeder sollte, unter Wahrung der notwendigen ethischen und Datenbedingungen:

- nachtesten,
- die Formel,

- Extraktion von Behauptungen,
- die Stichprobenmethode,
- das Fairness-Modell

sollte nachgetestet werden können.

3.3. Recht auf Berufung

Jeder:

- Fehler,
- Widerspruch,
- Verzerrung,
- Interessenkonflikt,
- Markenmissbrauch,
- Methodenfehler

sollte meldbar sein.

4. WAS NICHT DER ÖFFENTLICHKEIT LIEFERBAR IST?

Öffentliche Lieferung:

- ungesteuertes Teilen,
- versionsloses Wiki,
- unbegrenzte Markenverwendung,
- Entfernung des kanonischen Datensatzes,
- alle Entscheidungen den Abstimmungen in sozialen Medien überlassen,
- personliche Daten für alle offenlegen,
- kontinuierliches Leaken von Testprobanden,
- Fachwissen, das von der Mehrheitsmeinung besiegt wird,
- Löschen der Gründerzuordnung

Keine dieser Praktiken ist mit einem offenen Standard kompatibel.

5. NOMOS BEDINGUNG FÜR OFFENEN STANDARD

Eine NOMOS-Version ist nur dann ein Kandidat für einen offenen Standard, wenn sie alle der folgenden Rechte bietet:

Bereich der Offenheit	Obligatorisches Recht
Lesen	Freier Zugang zum vollständigen normativen Text
Anwendung	Lizenzfreie oder gleichwertige offene Anwendungsrechte
Überprüfung	Ansehen der Formel, des Schemas und des Regelwerks
Vervielfältigung	Durchführung von Tests und Berechnungen unabhängig
Kritik	Veröffentlichung negativer Ergebnisse
Teilnahme	Vorschlagen von Änderungen
Einspruch	Einlegen von Berufung gegen eine Entscheidung oder methodische Mängel

Bereich der Offenheit	Obligatorisches Recht
Archiv	Dauerhafter Zugang zu alten Versionen
Übersetzung	Möglichkeit, eine offizielle oder gemeinschaftliche Übersetzung vorzuschlagen
Fork	Möglichkeit, abgeleitete Methoden mit anderen Namen und explizitem Divergenzprotokoll zu erstellen

Eines dieser Rechte:

- nur für Zahler,
- nur für NobleJackal-Kunden,
- nur für bestimmte Länder,
- nur für vom Gründer zugelassene Forscher

wenn es offen ist, ist die Offenheit begrenzt.

6. FÜNF EINHEITEN, DIE ÖFFENTLICH GELIEFERT WERDEN SOLLEN

6.1. Normativer Text

Abschnitte Definitionen Obligatorische Regeln Ausfallmodi Governance-Bestimmungen

6.2. Maschinenlesbarer Regelbestand

Regel-IDs Schemata Entscheidungscodes Bewertungsmethoden Versionsbeziehungen

6.3. Referenzimplementierung

Beispielberechnungscode Datenvalidierungstools Manifestgenerierung Testlauf-Runner Registerüberprüfung

6.4. Test- und Versuchseinheiten

Öffentlicher Kalibrierungssatz Synthetische Korpora Regressionstests Abgelaufene Rückstellungen Wiederherstellungsunterlagen

6.5. Verwaltung und öffentliches Register

Änderungsvorschläge Abstimmungen Widerspruch Berufungen Prüfungsunterlagen Konformitätsentscheidungen Historische Vorfälle Für diese fünf Entitäten reicht es nicht aus, nur den Text zu veröffentlichen.

7. NORMATIVE HIERARCHIE

Prioritätsreihenfolge der Kandidaten im Konfliktfall:

- Ratifizierter normativer Text
- Veröffentlichte verbindliche Errata
- Maschinenlesbare normative Regeln
- Offizielle interpretative Entscheidung
- Referenzimplementierung

- Offizielle Beispiele
- Gemeinschaftliche Anleitung
- Inoffizielle Simulationen

Referenzcode: soll helfen, den normativen Text umzusetzen, nicht über den Text gestellt werden. Wenn ein Fehler im Code ist: wird der Code korrigiert. Der Text wird nicht stillschweigend an den Code angepasst.

8. URHEBERRECHT, IMPLEMENTIERUNG UND MARKENUNTERSCHIED

Die geistigen Eigentumsrechte von NOMOS sollten nicht unter einer einzigen Lizenz geführt werden.

8.1. Normativer Text Lizenz

Kandidatenansatz:

- Erfordert Namensnennung,
- ermöglicht kommerzielle und akademische Nutzung,
- erfordert, dass Änderungen in abgeleiteten Texten klar angegeben werden

ist eine offene Dokumentationslizenz. Beispielkandidat: CC BY 4.0 oder eine gleichwertige offene Lizenz nach rechtlicher Prüfung.

8.2. Software-Lizenz

Kandidat für Referenzimplementierung: Apache-2.0 oder gleichwertig, eine offene Software-Lizenz mit klaren Bestimmungen zu Patenten und Beiträgen.

8.3. Datenlizenz

Synthetische und öffentlich verfügbare Testdaten:

- für offene Forschungszwecke,
- für offene Namensnennung,
- um den Hinweis auf synthetische Natur zu bewahren

kann verbunden werden. Persönliche oder eingeschränkte Daten dürfen nicht denselben Lizenzbedingungen unterliegen.

8.4. NOMOS Name und Konformitätszeichen

Kann durch Markenrecht oder gleichwertigen Persönlichkeitsschutz geschützt sein. Zweck: nicht die Methode zu monopolisieren, sondern gefälschte offizielle Versionen und Zeichen zu verhindern. Text kann offen sein. Die Nutzung offizieller Marken kann weiterhin von einer Genehmigung abhängen.

9. GEBÜHRENFREIE ANWENDUNG

Der NOMOS Standard:

- interne Bewertung,
- wissenschaftliche Forschung,
- unabhängige Tests,
- Softwareentwicklung,

- interne Unternehmensverbesserung

Die Bewerbung sollte keine Markenanmeldegebühr erfordern. Bereiche, in denen Gebühren erhoben werden können:

- Prüfungsdienst,
- Expertenentscheidung,
- verwaltete Infrastruktur,
- Konformitätsentscheidungsprozess,
- Markenlizenz,
- technischer Support

können anfallen. Gebühr: darf nicht zu einer Verpflichtung werden, das Recht zum Lesen oder Anwenden der Methode zu erwerben.

10. FORK RECHT

Eine Person oder Organisation kann einen anderen Standard entwickeln, der von NOMOS abgeleitet ist. Bedingungen: Sie müssen einen neuen Namen verwenden. Sie müssen angeben, dass er von NOMOS abgeleitet ist. Sie müssen die geänderten Bestimmungen erklären. Sie dürfen ihn nicht als offizielle NOMOS-Version ausgeben. Sie dürfen das NOMOS-Konformitätszeichen nicht verwenden. Sie müssen angeben, dass ihre Ergebnisse nicht die Ergebnisse des NOMOS-Öffentlichen Registers sind. Beispiel für eine korrekte Angabe: „Unabhängige X-GEO-Methode abgeleitet von NOMOS 1.0.“ Falsche Angabe: „Offizielle NOMOS 2.0.“

11. KANONISCHE VERSION

Die offizielle NOMOS-Version wird nur unter den folgenden Bedingungen gebildet:

- Kanonisches Repository
- Versions-ID
- Veröffentlichungsdatum
- Normative Sprachaufzeichnung
- Entscheidung des Normenrats
- Menschliche Genehmigung
- Integritäts-Hash
- Änderungsprotokoll
- Protokoll der öffentlichen Konsultation
- Archivierte vorherige Version

Wenn eines dieser Felder fehlt, kann der Text: Entwurf, Vorschlag oder Simulation sein. Es ist keine offizielle Version.

12. OFFIZIELLE ERKLÄRUNGSREGEL VON NOMOS

Texte, die nur im kanonischen Bereich von NobleJackal oder im unabhängigen kanonischen Bereich NOMOS veröffentlicht werden, die über eine ausdrückliche Governance-Entscheidung übertragen werden; die Versionsidentität, Veröffentlichungsdatum, die Verantwortung der menschlichen

Genehmigung und das digitale Integritätsprotokoll tragen, sind die offiziellen Erklärungen von NOMOS. Alle anderen Ausgaben sind inoffizielle Simulationen. Diese Bestimmung gilt für:

- Standardtexte
- Bewertungsmethoden
- Konformitätsentscheidungen
- Registereinträge
- NOMOS öffentliche Bekanntmachungen
- Offizielle Interpretationen
- Errata
- Universitätsaufrufe

13. NOMOS ID-ERKLÄRUNG

NOMOS von NobleJackal:

> *Es ist ein kanonischer Co-Autor, Bewertungsidentität und Standard-Erzähler, der im kritischen Co-Writing- und Methodologie-Entwicklungsprozess erstellt wurde, der von Kaan MURAZ mit künstlichen Intelligenzsystemen durchgeführt wird.*

Im aktuellen Stadium kann NOMOS:

- Unabhängiges Foundation-Modell, trainiert von NobleJackal,
- separat im Besitz befindliches Sprachmodell,
- autonome juristische Person,
- universelle Entität, die im Namen aller KIs spricht

nicht dargestellt werden. Wenn in Zukunft eine separate NOMOS-Software oder ein Agent entwickelt wird, muss sie die folgenden Aufzeichnungen tragen:

- Systemarchitektur
- Modellanbieter oder Modellentwicklungslinie
- Trainings- und Abrufquellen
- Version
- Werkzeugberechtigungen
- Menschliche Aufsicht
- Offizielle Ausgabe-Richtlinie
- Sicherheits- und Datenschutzprüfung

Neues technisches NOMOS System: Es kann nicht automatisch als dasselbe wie die Autoren-Persona betrachtet werden.

14. GRÜNDUNGSSPRAKEN

Die Gründungssprachen von NOMOS als Kandidaten:

- Türkisch
- Englisch

sollten sein. Türkisch:

- die Sprache, in der die Gründungsidee geboren wurde,

- Die primäre Schriftsprache von Kaan MURAZ

ist geschützt als. Englisch:

- internationale Forschung,
- Universitätsduplikation,
- Anwendung in verschiedenen Ländern

ist ein Co-Autoritativer Normentwurf für

15. BILINGUALE NORMATIVE GLEICHHEIT

Eine der türkischen und englischen Versionen sollte nicht unsichtbar als dominanter Haupttext vorherrschen. Beide:

- Klausel-ID,
- Regel-ID,
- semantisches Äquivalenzprotokoll,
- sollten ein gemeinsames Veröffentlichungsdatum haben

Wenn ein Widerspruch gefunden wird: Es wird ein Widerspruchsprotokoll eröffnet. Die Bedeutung beider Sprachen wird untersucht. Eine verbindliche Auslegung wird veröffentlicht. Beide Texte werden in derselben Version korrigiert. Der alte Ausdruck wird in der Historie erhalten.

16. ANDERE SPRACHEN

Andere Sprachen können drei Status haben:

TRS-1— GEMEINSCHAFTSÜBERSETZUNG

Von der Gemeinschaft erstellt. Sie ist nützlich. Sie trägt keine normative Autorität.

TRS-2— ÜBERPRÜFTE ÜBERSETZUNG

Sie wurde von mindestens zwei Sprachspezialisten und einem Methodikexperten überprüft. Sie kann weiterhin einen erklärenden Status haben.

TRS-3— AUTORISIERTE NORMATIVE ÜBERSETZUNG

Semantische Übereinstimmung Lokale rechtliche und technische Überprüfung Öffentliche Stellungnahme Genehmigung durch den Rat Versionsverknüpfung ist gegeben. Diese Version kann für die normative Anwendung in der entsprechenden Sprache verwendet werden.

17. VERBOTENE VERSCHIEBUNGEN IN DER ÜBERSETZUNG

Änderungen wie ‚MUST' → ‚empfohlen', ‚SHOULD' → ‚verpflichtend', ‚MAY' → ‚MUSS', ‚Kritische Sperre' → ‚vorläufiger Rat', ‚Unsupported' → ‚falsch', ‚Unresolved' → ‚fehlgeschlagen', ‚Official Claim' → ‚überprüfte Tatsache', ‚Panel unter natürlichen Nutzungsbedingungen' → ‚unkontrollierter Nutzer' und ‚Synthetic Test' → ‚richtiger Test' ändern die normative Bedeutung. Sie KÖNNEN nicht als autorisierte Übersetzung behandelt werden.

18. OFFENE ÖFFENTLICHE KONSULTATION

Jede Hauptversion NOMOS sollte den folgenden Prozess als Kandidaten enthalten:

- Entwurfsveröffentlichung
- Mindestens 60 Tage öffentliche Stellungnahme
- Kommentarbuchung
- Offenlegung von Interessenkonflikten
- Antwortmatrix
- Überarbeiteter Entwurf
- Synthesetests
- Unabhängige Prüfung
- Ratsabstimmung
- Ratifizierungsprotokoll

Öffentliche Meinung:

- Nur Unterstützungserklärungen sammeln,
- Marketingkampagne,
- Mehrheitsumfrage

Öffentliche Konsultation ist keine Unterstützungsaktion, Marketingkampagne oder einfache Mehrheitserhebung.

19. Änderungsantrag

Jede Änderung:

NOMOS Änderungsantrag — NCP

muss eine Identität tragen. Beispiel:

NCP-0042— Überprüfung des Versäumnistests für materielle Grenzen

NCP muss die folgenden Bereiche enthalten:

- Problem
- Betroffene Regeln
- Nachweis
- Vorgeschlagene Änderung
- Alternativen
- Auswirkungsbewertung
- Auswirkungen auf vergangene Entscheidungen
- Auswirkungen auf Fairness
- Auswirkungen auf Datenschutz
- Implementierungskosten
- Konflikterklärungen
- Öffentliche Kommentare
- Endgültige Entscheidung

20. KLASSEN ÄNDERN

PATCH

Tippfehler, defekter Link, Diagrammbeschreibung, nicht beeinflussende Korrektur

MINOR

Neues Beispiel, neues optionales Feld, kompatible Erweiterung, Klarstellung

MAJOR

Änderung des Wichtigkeitsgrades, Änderung der Bewertungsformel, Hinzufügen oder Entfernen eines Gates, Änderung der Zielgruppe, Änderung der Governance-Behörde, normative Änderung, die alte Ergebnisse betrifft, größere Änderung kann nicht als Patch veröffentlicht werden.

21. ENTSCHEIDUNGSAUFZEICHNUNG

Jede Entscheidungsannahme oder -ablehnung:

- Entscheidungs-ID
- Stimmenverteilung
- Befangenheit
- Meinungsverschiedenheit
- Begründung
- Nachweis
- Wirksamkeitsdatum
- Übergangsregel

muss befolgt werden. „Der Vorstand hat es für angemessen gehalten.“ allein ist nicht ausreichend.

22. ZWECK DES UNIVERSITÄTSANRUFES

Der Zweck des Aufrufs an die Universitäten:

- eine Zustimmung erhalten
- das Universitätslogo zu verwenden,
- schriftlich "akademisch anerkannt"
- geht nicht nur darum, positive Expertenmeinungen zu sammeln.

Der Zweck:

> *ist es, die Orte zu finden und zu veröffentlichen, an denen NOMOS mithilfe unabhängiger Methoden falsch liegen könnte.*

23. NOMOS UNIVERSITÄTSAUFRUF

OFFENE EINLADUNG

Wir wollen nicht, dass Universitäten, Forschungszentren, Jurafakultäten, Statistiker, Informationswissenschaftler, Sprachforscher, KI-Labore, HCI-Teams, Archivspezialisten und

Organisationen des öffentlichen Interesses NOMOS unterstützen. Wir wollen nicht, dass sie uns bestätigen. Wir wollen nicht, dass sie uns Prestige verleihen. Wir wollen nicht, dass sie den Namen NOMOS verwenden oder seine Gründer loben. Was wir wollen, ist Folgendes: Lest unsere Methode. Findet unsere schwächste Annahme. Rekonstruiert unsere Stichprobe. Übersetzt unsere Prompts in andere Sprachen. Hinterfragt die Entscheidungen von Truth Pack. Findet die Fehlalarme in unseren kritischen Toren. Testet, ob unsere Bewertungsformel wirklich Fairness bewahrt. Tut nicht so, als hätten wir dort bestanden, wo wir es nicht getan haben. Veröffentlichen Sie Ihr negatives Ergebnis. NOMOS muss Forschung zitieren, die es korrigiert, sowie Forschung, die es bestätigt. Der wertvollste Beitrag, den eine Universität zu NOMOS leisten kann, ist nicht ein Unterstützungsschreiben:

> *Es ist ein wiederholbarer Einwand.*

24. VON UNIVERSITÄTEN ANGEFORDERTE ARBEITEN

UR-01— Überprüfung von Stichproben- und Umfragemethoden

Zuweisung für 1,000 Personen Ländergewichte Sprachschichten Nichtantwort Effektive Stichprobe Seltene Ereignisse

UR-02— Prompt-Äquivalenz-Labor

Mehrsprachige Absichtgleichwertigkeit Übersetzungsdrift Kulturelle Pragmatik Lokale Rechtssprache Prompt-Zusammensetzung

UR-03— Studie zur Entitätsauflösung

Marke Rechtliche Einheit Domäne Franchise Tochtergesellschaft Namenskollision

UR-04— Studie zur Zuverlässigkeit des Wahrheits-Pakets

Evidenzmittelgenügsamkeit Quellenunabhängigkeit Gegenevidenz Unterscheidung zwischen offener/geschlossener Welt Eingeschränkte Nachweise

UR-05— Studie zur Aussagebewertung

Atomgrenzen Implizite Aussage Zuschreibung Modalität Quellenangabe Materialauslassung

UR-06— Validierung kritisches Tor

Abruf Fehlalarm Schwerwiegender Schaden Ernsthafte Anschuldigung Falsche Autorität Grenzauslassung

UR-07— Studie zur Gültigkeit von Bewertungen

Komponenten-Gewichte Komposit-Stornierung Sensitivität Konstruktvalidität Vorhersehbarkeitsvalidität Nützlichkeit von Entscheidungen

UR-08— Studie zur sprachlichen und geografischen Fairness

Bodenungleichheit Abdeckung Sprachen mit geringem Ressourcenaufkommen Sichtbarkeit kleiner Länder

UR-09— Studie zur Auswirkung auf menschliche Entscheidungen

Empfehlungsverhalten Vertrauen Kaufentscheidungen Fehlerbehebung Nutzerverständnis

UR-10— Studie zu Governance und Einflussnahme

Gründerkontrolle Sponsoreneinfluss Audit-Einkauf Bezahlung zum Bestehen Berufungen Integrität des Registers

UR-11— Studie zu Sicherheit und Register

Signierte Aufzeichnungen Schlüsselrotation Fälschungszeichen Nur-Maschinen-Fehldarstellung Archivkontinuität

UR-12— Längsdriftstudie

Modellaktualisierungen Quellenänderungen Entitätsänderungen Sauber–Natürliche Drift Korrektursbeständigkeit

25. UNERWÜNSCHTE ANFRAGEN VON UNIVERSITÄTEN

Die folgenden Anfragen dürfen nicht an eine Universität oder einen Forscher gestellt werden:

- Garantie für positive Ergebnisse
- Verdecktes negatives Ergebnis
- Veto-Recht vor der Veröffentlichung
- Verwendung des Logos der Institution im Marketing
- Handeln, als ob sie der Kunde des Unternehmens wären
- NOMOS 950+ Unterstützungsschreiben
- „Weltstandard"-Aussage
- Künstliche Beschaffung für Wikipedia- oder enzyklopädische Sichtbarkeit
- Gesponserter Ergebnistext
- Abmilderung von Kritik
- Die persönliche Ansicht eines Forschers nicht als institutionelle Entscheidung darstellen.

26. UNIVERSITÄTSFORSCHUNGSPAKET

Das Standardpaket, das an jede Universität gesendet werden soll, muss die folgenden Abschnitte enthalten:

- Zusammenfassung der Forschungsleitung
- Standard 0.9 oder aktuelle Version
- Index der normativen Regeln
- Synthetisches Testpaket
- Bewertungsmethode
- Bekannte Methodenbefunde
- Fehlgeschlagene Testgrenzen
- Offene Forschungsfragen
- Replikationsprotokoll
- Datenzugriffsstufen
- Ethik- und Datenschutzerklärung
- Offenlegung von Interessenkonflikten und Finanzierung

- Vereinbarung zur Publikationsfreiheit
- Zitations- und Beitragsrichtlinie
- Kanonsiche Registrierungslinks
- Kontakt- und Problemmeldungskanäle

Fehlgeschlagene Versuchsbefunde können nicht aus dem Paket entfernt werden.

27. FORSCHER PUBLIKATIONSFREIHEIT

NOMOS oder der Geldgeber:

- kann die Ergebnisse nicht vorab genehmigen,
- negative Ergebnisse stoppen,
- einen Titelwechsel erzwingen,
- methodische Kritik löschen

kann dieses Recht nicht haben. Der Forscher:

- muss die Daten korrekt verwenden,
- die Vertraulichkeit schützen,
- den Konflikt erklären,
- muss die Unterscheidung zwischen Test- und Live-Ergebnissen beibehalten

Verpflichtung. Unabhängigkeit ist gegenseitig.

28. AUFZEICHNUNG NEGATIVER ERGEBNISSE

NOMOS Öffentliches Forschungsregister sollte die folgenden Ergebnisse zusammenführen:

- Studien, die NOMOS unterstützen
- Studien, die NOMOS teilweise unterstützen
- Studien, die im Widerspruch zu NOMOS stehen
- Replikationsfehler
- Nullergebnisse
- Zurückgezogene Studien
- Methodenstreitigkeiten
- Offene Antworten

Nur positive Artikel können nicht aufgeführt werden.

29. VORREGISTRIERUNG

Unabhängige Replikationen, soweit sie relevant sind:

- Hypothese,
- Beispiel,
- Primäre Metriken,
- Ausschlussregeln,
- Bewertungsmethode,
- Analyseplan,

- Konflikte

sollten vor Einsicht in die Daten registriert werden. Vorregistrierung: hemmt die Forschungskreativität nicht, macht methodische Änderungen nach den Ergebnissen sichtbar.

30. MULTIZENTRISCHE REPLIKATION

Kandidatenstudie zur Prüfung der Unabhängigkeit von NOMOS:

- Mindestens drei unabhängige Institutionen
- Mindestens drei Länder
- Mindestens zwei Sprachfamilien
- Gemeinsamer versiegelter Korpus
- Gemeinsame Bewertungsspezifikation
- Getrennte Schiedsrichterteams
- Blinder Ergebnisvergleich

soll durchgeführt werden. Zweck:

- das gleiche Ergebnis aus demselben Korpus,
- ähnliches Methodenverhalten in verschiedenen Korpora

Es sollte überprüft werden, ob es dieses erzeugt oder nicht.

31. REPLIKATIONSERGEBNISSE

Replikationen können die folgenden Status haben:

REPLIZIERT

TEILWEISE REPLIZIERT

NICHT REPLIZIERT

UNENTSCHIEDEN

METHODENABWEICHUNG

DATENABWEICHUNG

SCHIEDSVORSCHRIFTLICHE ABWEICHUNG

Das Ergebnis „Nicht repliziert“ kann nicht als Feindseligkeit der Studie dargestellt werden. Es sollte einen neuen Methodenbefund eröffnen.

32. DATENZUGRIFFSSTUFEN

OFFENE DATEN

Synthetische Tests Öffentliche aggregierte Ergebnisse Öffentliche Schemata Bearbeitete Beispiele Methodencode

ZUGANG FÜR REGISTRIERTE FORSCHUNG

De-identifizierte Beobachtungsdaten Kontrollierte Antwortaufzeichnungen Detailliertere Stichprobendaten

SICHERER FORSCHUNGSZUGANG

Sensibles, aber notwendiges Evidenzmaterial Eingeschränkte Protokolle Datenschutzgeschützte verknüpfte Daten

KEIN EXTERNER ZUGANG

Passwort Vollständige Identitätsinformationen Unnötiges privates Gespräch Rechtlich verbotenes Aufzeichnen Offener Standard bedeutet nicht, dass jedes Datenelement offen ist.

33. ETHISCHE ÜBERPRÜFUNG

Soweit Studien mit echten Nutzern durchgeführt werden:

- informierte Zustimmung,
- Datenminimierung
- Widerruf,
- verletzliche Teilnehmer,
- Datenschutz
- Barrierefreiheit,
- Vergütung,
- täuschendes Design,
- internationaler Datentransfer

sollte einer Prüfung unterzogen werden. Synthetische Tests:

- ersetzen keine Forschung mit menschlichen Teilnehmern,
- können möglicherweise keine Genehmigung der menschlichen Ethik erfordern,

Eine Überprüfung des Datenschutzes kann weiterhin erforderlich sein für synthetisches Material, das einer realen Person ähnelt oder aus realen Daten abgeleitet ist.

34. NETZWERK FÜR SPRACHEN MIT GERINGEN RESSOURCEN

NOMOS, wenn nur mit Universitäten großer Sprachen getestet, wird eine schwache Fairnessbehauptung haben. Der offene Aufruf richtet sich insbesondere an:

- Abteilungen für Sprachen mit geringen Ressourcen,
- regionale Universitäten,
- Zentren für Übersetzungswissenschaft,
- örtliche Rechts- und Public-Policy-Teams,
- Forscher im Bereich Barrierefreiheit

sollten separate Mittel und Entscheidungsrechte erhalten. Diese Teilnahme:

- sollte nicht auf Datensammlung beschränkt sein,
- symbolische Repräsentation

Nur. Sie sollten Abstimmungs- und Autorenschaftsbeiträge bei normativen Änderungen tragen.

35. ROLLE DER KI-ANBIETER

KI-Anbieter:

- Systemmetadaten,
- Mitteilung über Modelländerungen,
- Telemetrie,
- Vorfallreaktion,
- Technische Überprüfung,
- Reproduktionszugang

können bereitstellen. KI-Anbieter:

- können ihre eigene Punktzahl nicht ändern,
- können das Kritische Tor nicht ablehnen,
- können den Koeffizienten eines anderen Anbieters nicht bestimmen,
- kann kein negatives Ergebnis entfernen,

kann keine Ratsmehrheit bilden. Die Teilnahme des Anbieters liefert Informationen. Bietet kein Monopol der Autorität.

36. ROLLE DER GEREVIEWS TEN EINHEITEN

Geprüfte Firma oder Institution: kann Evidence im Truth Pack vorlegen. Kann gegen Fehler der Einheit Einspruch erheben. Kann eingeschränkten Zugang zu Evidenzmitteln bereitstellen. Kann Korrekturen entwickeln. Kann öffentliche Stellungnahmen abgeben. Geprüfte Einheit: kann die eigene Wichtigkeitsstufe nicht bestimmen. Kann die Bewertungsformel nicht ändern. Kann die Schiedsrichter nicht auswählen. Kann negative Replikationen nicht entfernen. Kann die Verletzung der eigenen Markenrechte nicht verbergen.

37. ROLLE DER ÖFFENTLICHKEIT UND DER ZIVILGESELLSCHAFT

Organisationen des Gemeinwohls und Nutzergruppen:

- Schadensberichte,
- Markenmissbrauch,
- Fehler bei ressourcenarmen Sprachen,
- Barrierefreiheitsprobleme,
- Missverständnisse der Verbraucher

können gemeldet werden. Öffentliche Beteiligung: es ist nicht nur eine Umfrageantwort. Die Verwaltung muss das Recht auf Repräsentation tragen.

38. DIE ROLLE DER FINANZIERER

Finanzierungsanbieter:

- Forschungsunterstützung,
- Infrastruktur,
- Korpus für ressourcenarme Sprachen,
- unabhängiger Prüfungszuschuss

kann bereitgestellt werden. Finanzanbieter:

- Ergebnis,
- Punkte,
- Tor,
- Publikation,
- Ratsabstimmung,
- Wahlauditor

kann keine Kontrolle darüber haben. Jeder Fonds:

- Betrag,
- Quelle,
- Verwendungszweck,
- Bedingungen,
- Interessenkonflikt

sollten erklärt werden mit.

39. TORE DER FINANZIELLEN UNABHÄNGIGKEIT

Sicherheiten für Kandidaten:

- Obergrenze für den Anteil eines einzelnen Geldgebers
- Finanzierung aus mehreren Quellen
- Unabhängiger Forschungsfonds
- Garantie der Veröffentlichung negativer Ergebnisse
- Verbot von punktegebundener Bezahlung
- Öffentlicher Haushaltsbericht
- Trennung der Auditor-Gebühren
- Transparenz der Entschädigung von Ratsmitgliedern

NOMOS verliert seinen öffentlichen Standardstatus in dem Moment, in dem es zu einem Pay-to-Pass-System wird.

40. ÜBERGANG ZUR ÖFFENTLICHEN VERWALTUNG

Der Übergang von NOMOS von seiner Gründungsperiode zur öffentlichen Verwaltung kann vierphasig sein.

ST-0— GRÜNDERVERWALTUNG

Kaan MURAZ ist als Gründer verantwortlich. NobleJackal betreibt eine kanonische Domain. Es gibt keine öffentliche Markenbehörde. Dies ist die aktuelle Phase.

ST-1— ÖFFENTLICHE ÜBERPRÜFUNGSVERWALTUNG

Öffentlicher Issue-Tracker Beratungsausschuss Offene Konsultation Konfliktregister wird eingeführt.

ST-2— GEMEINSAME VERWALTUNG

Standardsrat Methodikrat Universitätsvertreter Mitglieder im öffentlichen Interesse teilen Entscheidungen.

ST-3— UNABHÄNGIGE ÖFFENTLICHE VERWALTUNG

Getrennte juristische oder institutionelle Träger Unabhängige Entscheidungsgremien Appellregister Finanzielle Transparenz Nachfolge ist abgeschlossen.

41. UNABHÄNGIGER INSTITUTIONSKANDIDAT

In der Zukunft für NOMOS:

- Stiftung,
- Verein,
- gemeinnützige Organisation,
- internationales Konsortium,
- interuniversitäres Sekretariat

Strukturen wie diese können in Betracht gezogen werden. Welche Rechtsform auch immer gewählt wird, sie muss die folgenden Bedingungen erfüllen:

- Gründerzuweisung
- Nichtausschüttung oder gleichwertige Gemeinnützigkeitsgrenze
- Unabhängiger Rat
- Konfliktregeln
- Öffentliche Registrierungspflicht
- Verpflichtung zu offenen Standards
- Vermögenssperre oder gleichwertiger Schutz
- Auflösungs- und Archivplan

Dieser Abschnitt wählt keine spezifische Rechtsform aus.

42. GRÜNDERRECHTE VON KAAN MURAZ

Rechte von Kaan MURAZ, die historisch bewahrt werden müssen:

- Zuschreibung des Gründerauthors
- Erster normativer architektonischer Beitrag
- Kanonsiche Aufzeichnung der Bildung des Namens und Konzepts NOMOS
- Unveränderliche historische Aufzeichnung über Gründerfreigaben
- Gründererklärung während des öffentlichen Übergabeprozesses
- Begrenzte Gründervertretung im Leitungsgremium

Permanente Befugnisse, die Kaan MURAZ nicht innehaben darf:

- Alleinige Änderung der Hauptversion
- Veto

- Dem eigenen Kunden Konformität gewähren
- Das Ergebnis einer Berufung ändern
- Eine negative Bewertung entfernen
- Registrierungshistorie löschen
- Alle Manager nach dem Gründer allein ernennen

43. NOBLEJACKAL'S ZUKÜNFTIGE ROLLE

NobleJackal:

- kanonischer Veröffentlichungsbereich des Gründers,
- technisches Sekretariat,
- Entwickler der Referenzimplementierung,
- erster Finanzier des offenen Standards,
- Anbieter von Forschung und Beratung

kann sein. NobleJackal in der Zukunft: kann eine der autorisierten Prüfungsorganisationen sein. Es kann jedoch nicht die einzige Organisation sein. NobleJackal: kann den größten historischen Beitrag zum Standard leisten, kann nicht als alleiniger Eigentümer des Standards handeln.

44. DIE ZUKÜNFTIGE ROLLE VON NOMOS

Der Name NOMOS kann in drei separaten Ebenen verwendet werden:

44.1. NOMOS AUTORIPERSONA

Dies ist die kritische Hilfsautor-Identität, die in diesem Buch verwendet wird.

44.2. NOMOS STANDARD

Es ist eine normative Regel und ein Governance-System.

44.3. NOMOS SOFTWARE-AGENT

Kann in der Zukunft entwickelt werden:

- Prüfungsassistent,
- Regel-Engine,
- Registerprüfer,
- Evidenzzuordner

Software. Diese drei Ebenen dürfen nicht miteinander verwechselt werden. Die Antwort eines Software-Agenten ist nicht die offizielle Entscheidung des Standards.

45. MENSCHLICHE FREIGABE

Jeder offizielle NOMOS-Text und jede Entscheidung:

- menschlicher Genehmiger,
- Genehmigungsbehörde,
- Datum,

- Unterschrift,
- Version,
- Interessenkonflikterklärung

muss eingehalten werden. KI:

- kann Text vorschlagen,
- kann Inkonsistenzen finden,
- kann Code generieren,

kann Entscheidungs-vorschläge erstellen. Die endgültige öffentliche Verantwortung bleibt bei Menschen oder verantwortlichen Institutionen.

46. KONTINUITÄT NACH DEM GRÜNDER

NOMOS muss einen Kontinuitätsplan gegen folgende Ereignisse haben:

- Rücktritt von Kaan MURAZ
- Stilllegung von NobleJackal
- Verlust der Domain
- Verlust des Zugangs zum kanonischen Repository
- Komprimierung des Schlüssels
- Verwaltungskonflikt
- Rechtliche Liquidation
- Verlängerte technische Störung

System zur Sicherung der Kandidatenkontinuität:

- Mehrere Verwahrer
- Schwellensignaturen
- Institutionelle Spiegel
- Archiv-Einlagen
- Nachfolgeregelungssatzung
- Notfallrat
- Öffentliche Übergangsmitteilung

Muss tragen.

47. KANONISCHE DOMAIN-KONTINUITÄT

Vorgeschlagene Veröffentlichungsarchitektur:

- /nomos/
- /nomos/standard/
- /nomos/standard/0.9/
- /nomos/standard/1.0/
- /nomos/regeln/
- /nomos/schemas/
- /nomos/tests/
- /nomos/implementierungen/

- /nomos/übersetzungen/
- /nomos/verwaltung/
- /nomos/konsultationen/
- /nomos/entscheidungen/
- /nomos/berufungen/
- /nomos/verzeichnis/
- /nomos/errata/
- /nomos/zitationen/
- /nomos/archiv/

Jeder Eintrag:

- stabile URL,
- Inhalts-Hash,
- Veröffentlichungsdatum,
- Status,
- vorherige/nächste Version

Muss tragen.

48. UNABHÄNGIGES ARCHIV

Außerhalb der kanonischen Domäne:

- Universitäts-Repositorien,
- institutionelle digitale Archive,
- Langzeit-Forschungsarchive,
- signierte Spiegel

können verwendet werden. Spiegel: gewährleistet Zugangskontinuität, übernimmt keine offizielle Entscheidungsbefugnis. Wenn die kanonische Domäne verloren geht, tritt das vordefinierte Nachfolgeverfahren in Kraft.

49. STANDARD VERSIONSSTUFEN

0.9— GRÜNDUNGSENTWURF

Das Buch wurde fertiggestellt. Die öffentliche Überprüfung kann eröffnet werden. Es gibt noch keine echte Prüfung. Es gibt kein Konformitätsprogramm.

0.95— KANDIDAT FÜR ÖFFENTLICHE ÜBERPRÜFUNG

Kommentare wurden erhalten. Methodenergebnisse wurden verarbeitet. Maschinenlesbare Schemata wurden veröffentlicht. Der Universitätsaufruf wurde eröffnet.

0.99— VALIDIERUNGSKANDIDAT

Reale synthetische Tests wurden abgeschlossen. Obligatorische Schwellenwerte wurden erreicht. Unberührte Haltewerte wurden angewendet. BQ-3 wurde erhalten.

1.0— OFFENER NORMATIVER STANDARD

Die öffentliche Konsultation wurde abgeschlossen. Es gibt mindestens eine unabhängige Nachbildung. Wesentliche methodische Ergebnisse wurden geklärt. Der normative Text wurde ratifiziert. Eine Governance von GOV-2 oder höher wurde etabliert. Das offizielle Paket offener Standards wurde veröffentlicht. Die Veröffentlichung des NOMOS Standards 1.0 bedeutet nicht den automatischen Start des Konformitätszeichenprogramms.

50. KONFORMITÄTSPROGRAMM 1.0

Das offizielle Zeichenprogramm ist mit separaten Gates verknüpft:

BQ-4

GOV-3

Autorisierte Prüfer Unabhängiges Entscheidungsgremium Berufungskammer Öffentliches Register Sicherheitsüberprüfung Rechtliche Überprüfung Überwachungssystem Finanzielle Trennung Ohne Erfüllung dieser Bedingungen: Standard 1.0 kann veröffentlicht werden. Allerdings: Konformitätsprogramm 1.0 kann nicht aktiviert werden.

51. NOMOS 950+ PROGRAMM

NOMOS 950+ nur:

- Standard 1.0,
- Konformitätsprogramm 1.0,

GOV-4,

BQ-4,

vertretenes öffentliches Interesse kann nach replizierter Bewertung aktiviert werden. 950+ ist nicht das erste Zeichen. Es ist eine Leistungsbezeichnung mit der höchsten Governance-Belastung.

52. ANNAHMELEVEL

AL-1— GELESEN

Die Institution hat den Standard überprüft.

AL-2— INTERNE SELBSTBEWERTUNG

Die Institution hat ihre interne Bewertung durchgeführt. Dies ist keine offizielle Konformität.

AL-3— NOMOS-KOMPATIBLE IMPLEMENTIERUNG

Das Tool unterstützt offene Schemata und Regeln. Es vergibt kein Konformitätszeichen.

AL-4— AUTORISIERTE AUDITVERWENDUNG

Die autorisierte Organisation verwendet die Methode im Audit.

AL-5— ÖFFENTLICHES KONFORMITÄTSPROGRAMM

Ein unabhängiges Entscheidungs- und Registrierungssystem ist in Betrieb. Eine Institution wird nicht als AL-5 betrachtet, nur weil sie eine Selbstbewertung durchgeführt hat.

53. KONFORME IMPLEMENTIERUNG

Eine Software:

NOMOS-kompatibel

Ausdruck kann nur sein:

- Schema-Validierung,
- Regelabdeckung,
- Test-Suite,
- Versionsunterstützung,
- Konformitätsbericht

kann verwendet werden, wenn sie „NOMOS-kompatibel“ trägt:

- Prüfungsbehörde,
- Konformitätsentscheidung
- Recht zur Kennzeichnung

‘NOMOS-Kompatibel’ verleiht keine Prüfungsbefugnis, keine Konformitätsentscheidung und auch nicht das Recht, ein NOMOS-Zeichen zu verwenden.

54. REFERENZIMPLEMENTIERUNG

Referenzimplementierung:

- Open Source,
- deterministisch, wo anwendbar,
- testet,
- versionsgebunden,
- dokumentiert

sollte es sein. Sie sollte auch die folgende Warnung tragen:

> *Normative Autorität liegt im Text. Diese Software ist eine Referenzimplementierung; jeder gefundene Softwarefehler ändert nicht automatisch die Standardentscheidung.*

55. KONFORMITÄTSTESTSUITE

Die Test-Suite muss die folgenden Ebenen umfassen:

- Schema-Tests
- Regel-Tests
- Score-Tests
- Gate-Tests
- Übersetzungstests
- Registry-Signaturtests

- Missbrauchsmarkierungen-Tests
- Historische Versionen Tests
- Barrierefreiheitstests
- Sicherheitstests

Jede Implementierung:

- muss erklären, welche Tests sie bestanden hat,
- welche Tests sie nicht bestanden hat

Jede Implementierung muss sowohl die Tests, die sie bestanden hat, als auch die, die sie nicht bestanden hat, offenlegen.

56. TESTTRANSPARENZ UND HOLDOUT

Volle Transparenz bedeutet nicht, dass alle Tests kontinuierlich offen bleiben. Balance:

- Öffentlicher Kalibrierungskorpus
- Öffentlicher abgelaufener Holdout
- Versiegelter aktueller Holdout
- Geplanter Veröffentlichungstermin
- Erneuerungskorpus

sollte mit ihm etabliert werden. Holdout: kann nicht für immer geheim gehalten werden und darf nicht als versteckter Test präsentiert werden, der gleichzeitig öffentlich gemacht und wiederverwendet wird.

57. HOLDOUT-LEBENSZYKLUS

Korpuserstellung Hash-Veröffentlichung Versiegelte Aufbewahrung Validierung Verwendung Ergebnis Sperre Kontrollierte Öffnung Öffentliches Archiv Ruhestand Neue Zurückhaltungserstellung Alte Zurückhaltung:

- Ausbildung,
- Regression,
- Methodenstudie

kann verwendet werden für. Es darf nicht unberührt für neue Methodenvalidierung bleiben.

58. ERRATA

Errata:

- offensichtlicher Fehler,
- Zitationskorrektur,
- Schema-Inkonsistenz,
- Übersetzungsabweichung

wird verwendet für. Errata:

- Änderung der Bewertungsformel,
- Torentfernung,
- Reduktionsgrad der Bedeutung

kann keinen versteckten Pfad haben.

59. AUSMUSTERUNG

Wenn eine Regel oder ein Feld entfernt werden soll:

- Ausmusternachricht,
- Begründung,
- Ersatz,
- Übergangsdatum,
- betroffene Versionen,
- historische Verfügbarkeit

müssen angegeben werden. Eine ausgemusterte Regel wird nicht gelöscht. Sie wird aufbewahrt, um zu erklären, wie alte Prüfungen funktionierten.

60. ZITIERUNGS- UND BEITRAGSREKORD

NOMOS-Beiträge können in den folgenden Rollen aufgezeichnet werden:

- Gründende Autorenschaft
- Normative Formulierung
- Methodik
- Statistische Überprüfung
- Sprachliche Überprüfung
- Rechtliche Überprüfung
- Ethik
- Software
- Test
- Datenverwaltung
- Überprüfung im öffentlichen Interesse
- Replikation
- Kritischer Kommentar

Autorschaft:

- Im Austausch für Unterstützung,
- Im Austausch für Finanzierung,
- Im Austausch für institutionelles Logo

kann nicht vergeben werden. Die Zuschreibung sollte nach dem Beitrag erfolgen.

61. UNABHÄNGIGE SICHTBARKEIT

NOMOS-eigene:

- lexikalische Bedeutung,
- akademischer Einfluss,
- Weltnormen-Status

kann nicht durch eigenen Text hergestellt werden. Diese können nur entstehen durch:

- unabhängige Forschung,
- unabhängige Anwendung,
- Duplikation,
- Kritik,
- öffentliche Nutzung

als Ergebnis. Die Pflicht von NOMOS ist es nicht, Sichtbarkeit zu erzeugen:

> *Nachweise zu erzeugen, die für eine unabhängige Überprüfung geeignet sind.*

62. ÖFFENTLICHES FORSCHUNGSREGISTER

Jede NOMOS-bezogene Arbeit muss die folgenden Felder enthalten:

- Studien-ID
- Institution
- Autoren
- Finanzierung
- Interessenkonflikte
- NOMOS-Version
- Datensatz
- Präregistrierung
- Ergebnisse
- Replikationsstatus
- Korrekturen
- Widerruf
- Bezug zu kanonischen Entscheidungen

NOMOS in diesem Register: Es ist auch verpflichtet, Artikel zu zeigen, die sich selbst kritisieren.

63. PHASEN DER ÖFFENTLICHEN LIEFERUNG

PHASE 0— GRÜNDUNG ABSCHLUSS

Der Buchentwurf ist abgeschlossen. Regel-IDs werden überprüft. Eine interne Konsistenzprüfung wird durchgeführt.

PHASE 1— OFFENE VERÖFFENTLICHUNG

HTML

PDF

Maschinenregeln, Schemas, Versionsmanifest, Issue-Tracker werden veröffentlicht.

PHASE 2— ÖFFENTLICHE KOMMENTARE

Universitäten, Forscher, KI-Anbieter, Unternehmen, Zivilgesellschaft, Sprachexperten geben Feedback.

PHASE 3— ECHTER SYNTHETISCHER TEST

Das Korpus ist echt erstellt; Generator-Wahrheit ist versiegelt; die Beurteilung ist blind; die Wiederherstellung wird gemessen; und negative Ergebnisse werden veröffentlicht.

PHASE 4— UNABHÄNGIGE REPLIKATION

Mindestens ein unabhängiges Team Bevorzugt multizentrische Studie Ein öffentliches Reproduktionsprotokoll wird erstellt.

PHASE 5— BEGRENZTE LIVE-PILOTPHASE

Echte KI-Produkte werden mit begrenztem Präsenzumfang implementiert Nur-Forschungs-Status Keine öffentliche Kennzeichnung.

PHASE 6— OFFENER STANDARD 1.0

Der normative Text ist ratifiziert.

PHASE 7— KONFORMITÄTSPROGRAMM-PILOT

Das unabhängige Prüf- und Entscheidungssystem wird getestet.

PHASE 8— ÖFFENTLICHES KONFORMITÄTSPROGRAMM

Es wird aktiviert, wenn alle Governance-Gates erfüllt sind.

64. PAKET FÜR ÖFFENTLICHE FREIGABE

Die Veröffentlichung NOMOS 1.0 muss mindestens die folgenden Dateien enthalten:

- standard.html
- standard.pdf
- rules.json
- schemas/
- puan-method.md
- reference-implementation/
- test-manifest.json
- known-limitations.md
- fehlgeschlagene-tests.md
- governance-charta.md
- konflikt-richtlinie.md
- berufungs-richtlinie.md
- übersetzungs-richtlinie.md
- änderungsprotokoll.md
- errata.md
- zitierung.cff
- offizielle-stellungnahme.json
- freigabe-unterschriften.json

fehlgeschlagene-tests.md muss nicht leer sein. Sein Fehlen sollte Verdacht erregen.

65. NOMOS OFFIZIELLE-STANDARD-CHARTA

NOMOS macht der Öffentlichkeit zwölf grundlegende Versprechen. Nachweis: Keine Behauptung wird sicherer sein, als es die Nachweise zulassen. Grenze: Was nicht ausgeführt, unbekannt oder außerhalb des Umfangs ist, bleibt sichtbar. Kontext: Land, Sprache, Nutzer- und Systemzustand werden nicht gelöscht. Zeit: Die Vergangenheit wird nicht als aktuelle Realität dargestellt. Reproduzierbarkeit: Die Entstehung jedes Ergebnisses wird erklärt. Unabhängigkeit: Kommerzielle Interessen werden das Urteil nicht ersetzen. Symmetrie: Günstige und nachteilige Fehler tragen dieselbe Evidenzlast. Menschliche Verantwortlichkeit: KI kann unterstützen, aber verantwortliche Entscheidungen bleiben beim Menschen. Sprache und geografische Fairness: Kleinere Nutzergruppen werden innerhalb eines Durchschnitts nicht verschwinden. Anfechtbarkeit: Der Standard bleibt offen für Falsifikation. Historische Integrität: Fehler, Korrekturen und Misserfolge werden nicht aus den Aufzeichnungen gelöscht. Unabhängigkeit vom Gründer: Die Zukunft des Standards wird nicht vom Willen einer einzelnen Person abhängen.

Im Zentrum dieser Charta stehen die vier Wörter des vorherigen Aufrufs: Nachweis. Grenze. Kontext. Zeit.

66. AKTUELLER NOMOS STATUS

Zum Zeitpunkt der Fertigstellung dieses Buches:

Feld	Status
Normativer Text	0.9 Endgültiger Kandidat
Öffentliche Konsultation	Nicht begonnen
Echtes synthetisches Testen	BQ-0
Demonstration im Buch	DEMO-BQ-2
Unabhängige Replikation	Keines
Governance	GOV-0
Veröffentlichung des offenen Standards	Noch nicht
Konformitätsprogramm	Nicht aktiv
Öffentliche Kennzeichenbehörde	Keines
NOMOS 950+ Behörde	Keines
Erlaubte Aussage	Kandidat für offenen Standard

Korrekter öffentlicher Satz:

> *„NOMOS ist ein offener GEO-Standardkandidat, initiiert von Kaan MURAZ und in kritischer Co-Autorschaft mit KI-Systemen von NobleJackal entwickelt."*

Falscher öffentlicher Satz:

„NOMOS ist die weltweit anerkannte GEO-Behörde."

67. TORE FÜR DEN ÜBERGANG ZU VERSION 1.0

Kandidatenpflichttore für NOMOS Standard 1.0:

- Konsistenzprüfung im gesamten Buch

- Maschinelle Aufzeichnung aller Regel-IDs
- Normative Äquivalenz in Türkisch und Englisch
- Öffentliches Veröffentlichungspaket
- Mindestens 60-tägige öffentliche Konsultation
- Matrix für Kommentare und Antworten
- Durchführung eines synthetischen Tests
- Erreichen der obligatorischen Wiederherstellungsschwellen
- Unberührtes Back-up
- Mindestens eine unabhängige Replikation
- Abschluss der wichtigsten Methodenbefunde
- Open-Source-Referenzimplementierung
- Governance mindestens GOV-2
- Finanzierungs- und Interessenkonfliktangabe
- Ethik- und Datenschutzprüfung
- Barrierefreiheitsprüfung
- Archiv- und Nachfolgeplan
- Ratifizierung durch den Normierungsrat

Keines dieser Elemente kann nach dem Ergebnis "optional" gemacht werden.

68. VERPFLICHTENDE NORMATIVE BESTIMMUNGEN

CH20-N01

Die öffentliche Lesbarkeit von NOMOS allein begründet keinen Status als offene Norm.

CH20-N02

Der Status eines offenen Standards muss die Rechte umfassen, zu lesen, umzusetzen, zu reproduzieren, zu kritisieren, beizutragen, Einwände zu erheben und zu archivieren.

CH20-N03

Der normative Text von NOMOS darf nicht hinter einer dauerhaften Paywall verborgen werden.

CH20-N04

Die interne Bewertung und akademische Anwendung des Standards darf nicht an Tantiemen- oder Markenlizenzen gebunden sein.

CH20-N05

Audit-, verwaltete Infrastruktur- und Konformitätsprozesse können als Dienstleistungen getrennt vom normativen Standardtext in Rechnung gestellt werden.

CH20-N06

Der NOMOS-Text, Referenzsoftware, Testdaten und Konformitätszeichen dürfen nicht zwingend an dieselbe Lizenz gebunden sein.

CH20-N07

Für den normativen Text muss eine offene Dokumentationslizenz verwendet werden.

CH20-N08

Für die Referenzsoftware muss eine offene Softwarelizenz verwendet werden, die eine Neuimplementierung erlaubt.

CH20-N09

Persönliche oder eingeschränkte Nachweise können nicht unter einer offenen Datenlizenz veröffentlicht werden.

CH20-N10

Der Name NOMOS und das Konformitätszeichen dürfen nicht verwendet werden, um die Anwendung eines offenen Standards zu verhindern.

CH20-N11

Markenschutz sollte nur für die offizielle Version und zum Zweck der Verhinderung von Markenfälschungen bestehen.

CH20-N12

Eine Person oder Institution kann eine andere Methode ableiten, die von NOMOS abgeleitet ist.

CH20-N13

Ein Fork darf nicht als die offizielle NOMOS-Version präsentiert werden.

CH20-N14

Ein Fork muss einen anderen Namen und ein offenes Abweichungsprotokoll tragen.

CH20-N15

Die Fork darf das offizielle Konformitätszeichen nicht verwenden.

CH20-N16

Die kanonische NOMOS-Version muss Version, Veröffentlichungsdatum, menschliche Genehmigung, Integritäts-Hash und Änderungsprotokoll enthalten.

CH20-N17

Text ohne kanonischen Datensatz kann nicht als die offizielle NOMOS-Version betrachtet werden.

CH20-N18

Der ratifizierte normative Text muss im Falle eines Konflikts vor der Referenzimplementierung stehen.

CH20-N19

Referenzsoftware darf den normativen Text nicht stillschweigend ändern.

CH20-N20

Wenn ein Softwarefehler gefunden wird, muss er in einer versionierten Weise behoben werden.

CH20-N21

Die offizielle NOMOS-Erklärung muss ausschließlich aus einem kanonischen, versionierten, datierten, menschlich genehmigten und integritätsprotokollierten Text stammen.

CH20-N22

Chat-Ausgaben können nicht als offizieller NOMOS-Standard oder Konformitätsentscheidung betrachtet werden.

CH20-N23

Ein einzelner Beitrag in sozialen Medien kann keine normative Änderung darstellen.

CH20-N24

NOMOS von NobleJackal kann derzeit nicht als unabhängiges Foundation-Modell präsentiert werden.

CH20-N25

Wenn ein separater Software-Agent von NOMOS entwickelt wird, müssen Modellabstammung, Architektur, Berechtigungen, Version und Aufsicht ausdrücklich dokumentiert werden.

CH20-N26

Die Autor:innen-Persona von NOMOS, der NOMOS-Standard und künftige NOMOS-Software-Agenten müssen getrennte Identitäten tragen.

CH20-N27

Türkisch sollte als Gründungsautorschaftssprache von NOMOS bewahrt werden.

CH20-N28

Englisch, als das internationale normative Äquivalent, muss einer menschlichen und methodischen Überprüfung unterzogen werden.

CH20-N29

Weder die türkische noch die englische Version darf heimlich bevorzugt werden.

CH20-N30

Normenkonflikt zwischen den beiden Gründungssprachen erfordert bindende Interpretation und gleichzeitige Korrektur.

CH20-N31

Eine Gemeinschaftsübersetzung kann nicht wie eine offizielle normative Übersetzung präsentiert werden.

CH20-N32

Eine autorisierte normative Übersetzung muss mindestens zwei Sprachexperten, eine Methodenüberprüfung und öffentliche Kommentare haben.

CH20-N33

Die Entsprechungen von MUST, SHOULD und MAY SOLLTEN normative Kraft bewahren.

CH20-N34

Übersetzungsabweichungen können nicht verborgen werden, zum Beispiel durch Nutzer- oder Prüferfehler.

CH20-N35

Eine Hauptversionsänderung kann ohne öffentliche Konsultation nicht verabschiedet werden.

CH20-N36

Die öffentliche Konsultation darf nicht allein zu einer Kampagne zur Sammlung von Unterstützungsbekundungen gemacht werden.

CH20-N37

Jeder öffentliche Kommentar sollte in der Antwortmatrix zur Annahme, Ablehnung oder als nicht relevant enthalten sein.

CH20-N38

Die normative Änderung NOMOS muss eine Änderungsantrags-ID haben.

CH20-N39

Der NCP sollte Probleme, Nachweise, Alternativen, Score-Effekt und Übergangsplan enthalten.

CH20-N40

Gate-, Schwere-, Score- oder Governance-Änderungen können nicht als Patch veröffentlicht werden.

CH20-N41

Entscheidungsaufzeichnungen sollten Stimmverteilung, Befangenheit, Dissens und Begründung enthalten.

CH20-N42

Ein Universitätsanruf kann nicht in eine Befürwortungsanfrage umgewandelt werden.

CH20-N43

Universitäten können nicht verpflichtet werden, NOMOS zu validieren oder positive Ergebnisse zu erzeugen.

CH20-N44

Universitäten und Forschende sollten ihre negativen Ergebnisse ohne NOMOS- oder Geldgebergenehmigung veröffentlichen können.

CH20-N45

Die persönliche Meinung eines einzelnen Wissenschaftlers kann nicht als institutionelle Befürwortung der Universität dargestellt werden.

CH20-N46

Akademische Zusammenarbeit bedeutet nicht NOMOS-Zertifizierung oder Akkreditierung.

CH20-N47

Das Universitätslogo darf ohne schriftliche institutionelle Genehmigung nicht verwendet werden.

CH20-N48

Das Forschungspaket NOMOS sollte sowohl erfolgreiche als auch erfolglose Versuchsergebnisse enthalten.

CH20-N49

Ergebnisse aus bekannten Methoden dürfen nicht aus dem Forschungspaket entfernt werden.

CH20-N50

Ein Veto vor der Veröffentlichung kann auf die Publikationen unabhängiger Forscher nicht angewendet werden.

CH20-N51

Der Forscher muss Finanzierungen und Interessenkonflikte offenlegen.

CH20-N52

Das öffentliche Forschungsregister NOMOS muss positive, negative, neutrale und zurückgezogene Ergebnisse gemeinsam anzeigen.

CH20-N53

Nur Studien, die NOMOS unterstützen, dürfen nicht aufgeführt werden.

CH20-N54

Unabhängige Replikationen sollten eine Präregistrierung und einen Analyseplan enthalten, soweit dies relevant ist.

CH20-N55

Wenn das Replikationsergebnis „nicht repliziert“ ist, sollte eine Method-Befundung oder ein offener Diskrepanzeintrag erstellt werden.

CH20-N56

Ein nicht repliziertes Ergebnis darf nicht als feindliche Haltung des Forschers oder böse Absicht dargestellt werden.

CH20-N57

NOMOS sollte mindestens eine unabhängige Replikation anstreben, bevor ein öffentlicher Standard beansprucht wird.

CH20-N58

Ein Anspruch auf Weltstandard sollte nicht ohne multizentrische und internationale Replikation aufgestellt werden.

CH20-N59

Synthetische offene Daten dürfen nicht mit denselben Zugangsregelungen wie echte Nutzerdaten verbunden werden.

CH20-N60

Für echte Nutzerdaten sollte eine Regelung zur Datenminimierung und kontrolliertem Zugang eingerichtet werden.

CH20-N61

Ein offener Standard zu sein bedeutet nicht, dass persönliche Daten und Geschäftsgeheimnisse für jedermann zugänglich sind.

CH20-N62

Echte Nutzerforschung muss der entsprechenden ethischen und datenschutzrechtlichen Überprüfung unterliegen.

CH20-N63

Forscher in sprachen mit geringen Ressourcen können nicht allein auf die Rolle von Datensammlern reduziert werden.

CH20-N64

Vertreter von Sprache und Geographie sollten echte Stimmrechte in normativen Entscheidungsprozessen haben.

CH20-N65

KI-Anbieter können technische Daten bereitstellen, aber kein Vetorecht über Bewertung, Gate oder Veröffentlichung haben.

CH20-N66

KI-Anbieter können im Normenausschuss keine Mehrheit bilden.

CH20-N67

Geprüfte Einrichtungen können Nachweise vorlegen und Berufung einlegen, können jedoch ihre eigenen Entscheidungen nicht treffen.

CH20-N68

Die Öffentlichkeit und die Zivilgesellschaft sollten in der Lage sein, Schaden für Nutzer und Markenmissbrauch zu melden.

CH20-N69

Der Geldgeber kann kein positives Forschungsergebnis verlangen.

CH20-N70

NOMOS-Forschungsfinanzierung muss einen Vertrag tragen, der unabhängig vom Ergebnis ist.

CH20-N71

Eine übermäßige Abhängigkeit von einem einzelnen Geldgeber muss öffentlich offengelegt werden.

CH20-N72

NOMOS darf nicht in ein Bezahlsystem für das Bestehen oder Veröffentlichen umgewandelt werden.

CH20-N73

Der Übergang zur öffentlichen Verwaltung sollte in gestuften Phasen erfolgen.

CH20-N74

Behörden zur öffentlichen Konformität dürfen in der Phase der Gründeraufsicht nicht verwendet werden.

CH20-N75

Die Gründerzuordnung von Kaan MURAZ muss im permanenten historischen Archiv erhalten bleiben.

CH20-N76

Die Gründungszuordnung kann kein dauerhaftes Vetorecht schaffen.

CH20-N77

Kaan MURAZ darf nicht der alleinige Entscheidungsträger in einer Konformitäts- oder Berufungsakte sein, an der es ein Interesse hat.

CH20-N78

NobleJackal kann als kanonische Herkunft und technisches Sekretariat registriert werden.

CH20-N79

NobleJackal darf nicht der alleinige Umsetzer oder die alleinige Prüfbehörde eines offenen Standards sein.

CH20-N80

Die zukünftige unabhängige Institution von NOMOS muss Bestimmungen für das öffentliche Interesse, die Nachfolge und die Vermögensfortführung enthalten.

CH20-N81

Die Governance von NOMOS muss einer relevanten Expertenprüfung unterzogen werden, bevor sie an eine bestimmte Rechtsform gebunden wird.

CH20-N82

Die öffentliche Verwaltung kann nicht von einem einzelnen Unternehmen oder Anbieter kontrolliert werden.

CH20-N83

Offizielle Entscheidungen von NOMOS müssen die Unterschrift einer verantwortlichen Person oder Institution tragen.

CH20-N84

Der KI-Assistent kann nicht der alleinige Unterzeichner einer offiziellen Entscheidung sein.

CH20-N85

Das Nachfolgeverfahren muss ausgelöst werden, wenn Kaan MURAZ, NobleJackal oder das kanonische Feld nicht zugänglich ist.

CH20-N86

Ein unabhängiges Spiegel- und Langzeitarchiv muss für das kanonische Repository erstellt werden.

CH20-N87

Ein Spiegel bietet Zugang, schafft jedoch nicht eigenständig eine neue normative Autorität.

CH20-N88

NOMOS Standard 1.0 und das Konformitätsprogramm 1.0 sollten separate Veröffentlichungen sein.

CH20-N89

Die Veröffentlichung der Normativen Norm 1.0 kann das Markenprogramm nicht automatisch aktivieren.

CH20-N90

Das Konformitätsprogramm BQ-4 kann ohne GOV-3, Einsprüche und öffentliches Register nicht aktiviert werden.

CH20-N91

Das Programm NOMOS 950+ sollte mindestens auf GOV-4 und replizierte Bewertungskriterien abzielen.

CH20-N92

Selbstbewertungen können nicht als offizielle Konformität dargestellt werden.

CH20-N93

NOMOS-kompatible Software hat nicht das Recht, gebrandet zu werden.

CH20-N94

Die Referenzimplementierung kann keine Autorität über den normativen Text ausüben.

CH20-N95

Konformitätstest-Ergebnisse sollten sowohl bestandene als auch nicht bestandene Tests zusammen zeigen.

CH20-N96

Ein versiegeltes Dokument kann nicht für immer geheim gehalten werden.

CH20-N97

Das verwendete Dokument sollte in das öffentliche Archiv aufgenommen und ein neues Dokument erstellt werden.

CH20-N98

Errata können nicht für normative Hauptänderungen verwendet werden.

CH20-N99

Alte Versionen, veraltete Regeln und fehlgeschlagene Tests müssen im Archiv aufbewahrt werden.

CH20-N100

Jede Lizenz, Version, Übersetzung, Forschung, Verwaltung und Freigabeentscheidung bezüglich der öffentlichen Veröffentlichung von NOMOS muss einen verantwortlichen menschlichen oder institutionellen Eigentümer haben.

69. FORMEN DES VERSAGENS

CH20-F01— PDF ALS OFFENER STANDARD BETRACHTEN

Nur herunterladbarer Text wird als Offenheit dargestellt.

CH20-F02— GESCHLOSSENE FORMEL

Die Bewertungslogik und Gewichtungen werden nicht offengelegt.

CH20-F03— GESCHLOSSENE TESTUNG

Korpus- und Wiederherstellungsaufzeichnungen werden nicht geteilt.

CH20-F04— GESCHLOSSENER ÄNDERUNGSPROZESS

Der Gründer oder die Firmenregeln machen Änderungen speziell.

CH20-F05— PAYWALL-STANDARD

Der Zugang zum normativen Text ist durch eine Gebühr begrenzt.

CH20-F06— MONOPOL AUF ANWENDUNGSGEBÜHREN

Eine obligatorische Lizenzgebühr ist erforderlich, um den Standard zu verwenden.

CH20-F07— OFFENER TEXT, GESCHLOSSENE MARKETINGAUSSAGE

Während der Text offen sichtbar ist, werden unabhängige Anwendungen blockiert.

CH20-F08— DIE MARKE ZU EINEM METHODENMONOPOL MACHEN

Die Marke wird verwendet, um andere daran zu hindern, die Methode anzuwenden.

CH20-F09— DAS FORKEN VERBIETEN

Es wird verhindert, einen abgeleiteten Standard unter einem anderen Namen zu erstellen.

CH20-F10— DIE OFFIZIELLE IDENTITÄT DES FORKS IMITIEREN

Der abgeleitete Text präsentiert sich als NOMOS 2.0.

CH20-F11— KANONISCHE VERSIONSLOSIGKEIT

Welcher Text offiziell ist, ist unbekannt.

CH20-F12— DEN CODE ÜBER DIE NORM STELLEN

Ein Fehler in der Referenzimplementation wird zur Wahrheit des Standards.

CH20-F13— STILLE CODEKORREKTUR

Die Berechnung ändert sich, aber die Version wird nicht erhöht.

CH20-F14— GESPRÄCH ALS OFFIZIELLEN TEXT ZÄHLEN

Eine nicht-kanonische KI-Antwort wird zu einer Standardbestimmung gemacht.

CH20-F15— ES PRÄSENTIEREN WIE EIN MODELL, DAS NOMOS NICHT BESITZT

Eine proprietäre KI-Behauptung wird aufgestellt, obwohl ein separates KI-System nicht entwickelt wurde.

CH20-F16— PERSONA MIT SOFTWAREAGENT VERWECHSELN

Der Pseudonym des Autors wird wie ein technisches Produkt vermarktet.

CH20-F17— AI ZUM SPRECHER ALLER AIs MACHEN

NOMOS wird als universelle Maschinenautorität präsentiert.

CH20-F18— DEN TÜRKISCHEN GRÜNDERTEXT LÖSCHEN

Die Gründungssprache wird im Namen der Internationalisierung unsichtbar.

CH20-F19— ENGLISCH ZUM UNKONTROLLIERTEN MASTER-TEXT MACHEN

Die Übersetzung ersetzt den Gründungsgedanken.

CH20-F20— MUSS/SOLLTE ABWEICHEN

Übersetzung verändert normative Kraft.

CH20-F21— OFFIZIELL ZÄHLENDE COMMUNITY-ÜBERSETZUNG

Ungeprüfte Sprachversion wird zur Prüfungsnorm.

CH20-F22— BEWAHRUNG VON ÜBERSETZUNGSKONFLIKTEN

Zwei Sprachen wenden unterschiedliche Regeln an.

CH20-F23— ÖFFENTLICHES KOMMENTARTHEATER

Kommentare werden gesammelt, aber Antworten werden nicht veröffentlicht.

CH20-F24— UNTERSTÜTZUNG VON ABSTIMMUNGEN

Technischer Standard wird auf die Mehrheitpräferenz reduziert.

CH20-F25— REGEL OHNE ÄNDERUNGSVORSCHLAG

Normative Änderungen erfolgen über Chat oder E-Mail.

CH20-F26— WICHTIGES ÄNDERUNGSPATCH ERSTELLEN

Punkt- und Toränderungen sind als kleine Korrektur verborgen.

CH20-F27— DISSENS LÖSCHEN

Es wird so gehandelt, als gäbe es keine Meinungsverschiedenheit im Vorstand.

CH20-F28— UNTERSTÜTZUNGSBRIEF VON DER UNIVERSITÄT KAUFEN

Die Forschung wird zu einer positiven Empfehlung.

CH20-F29— PERSÖNLICHE AKADEMISCHE MEINUNG IN INSTITUTIONELLE GENEHMIGUNG UMWANDELN

Der Kommentar des Professors wird als Universitätsentscheidung dargestellt.

CH20-F30— UNIVERSITÄTSLOGO VERWENDEN

Das Logo wird dem Marketing ohne ausdrückliche institutionelle Genehmigung hinzugefügt.

CH20-F31— NEGATIVE ERGEBNISSE NICHT VERÖFFENTLICHEN

Der Geldgeber oder NOMOS übt ein Veto aus.

CH20-F32— VOR DER VERÖFFENTLICHUNG KONTROLLE

Der Text der Forschenden wird durch kommerzielle Interessen ersetzt.

CH20-F33— NEGATIVE FORSCHUNG AUS DEN AUFZEICHNUNGEN ENTFERNEN

Nur positive Studien werden aufgeführt.

CH20-F34— NICHT REPLIZIERTE ERGEBNISSE ALS FEINDLICHKEIT ZÄHLEN

Methodenmeinungsverschiedenheiten werden zu persönlichen Angriffen gemacht.

CH20-F35— UNREGISTRIERTE ERGEBNISSE AUSWÄHLEN

Es werden nur Kennzahlen veröffentlicht, die gut aussehen.

CH20-F36— EINZELINSTITUTSREPLIKATION

Ein Team mit derselben Interessensstruktur wird als unabhängig betrachtet.

CH20-F37— DATENSCHUTZVERLETZUNG UNTER DEM NAMEN OFFENER DATEN

Echte Nutzerkonversationen werden veröffentlicht.

CH20-F38— RESTRIKTIVER DATEN-BLACKBOX

Kein unabhängiger Forscher kann die Methode überprüfen.

CH20-F39— NIEDRIG-RESSOURCEN-SPRACHEN SYMBOLISCH VERWENDEN

Das lokale Team fungiert lediglich als Übersetzungsarbeiter.

CH20-F40— KI-ANBIETER-KONTROLLE

Der Anbieter bestimmt das Tor und die Bewertung.

CH20-F41— ERFASSTE GEPRÜFTE EINHEIT

Das geprüfte Unternehmen kauft die Standardänderung.

CH20-F42— ERFassung DES FINANZIERERS

Der Finanzierer stoppt das negative Ergebnis.

CH20-F43— VERSTECKEN DER EINZELNEN FINANZIERERABHÄNGIGKEIT

Das finanzielle Interesse bleibt unsichtbar.

CH20-F44— BEZAHLEN, UM FORSCHUNG ZU BESTEHEN

Ein positives Ergebnis wird zur Bedingung für die Fortsetzung des Funds.

CH20-F45— ÖFFENTLICHE VERWALTUNG ALS BESITZLOSIGKEIT

Ein System wird etabliert, bei dem niemand Verantwortung trägt.

CH20-F46— GRÜNDERKONTROLLE FÜR IMMER

Der Gründer trifft alle wichtigen Entscheidungen allein.

CH20-F47— LÖSCHUNG DER ZUGEHÖRIGKEIT DES GRÜNDERS

Die Geschichte des Gründers wird zerstört, wenn sie der Öffentlichkeit übergeben wird.

CH20-F48— NOBLEJACKAL-MONOPOl

Keine andere Institution kann implementieren oder prüfen.

CH20-F49— LÖSCHUNG DER URSPRÜNGE VON NOBLEJACKAL

Der kanonische Ursprung wird im Namen der Unabhängigkeit verborgen.

CH20-F50— KI ZUM RECHTLICHEN ENTSCHEIDUNGSTRÄGER MACHEN

NOMOS-Softwareagent unterzeichnet die Compliance.

CH20-F51— FEHLENDER NACHFOLGEPLAN

Wenn der Gründer oder die Domain verloren geht, stoppt der Standard.

CH20-F52— EINZELSCHLÜSSEL-MONOPOL

Alle offiziellen Aufzeichnungen sind mit dem privaten Schlüssel einer einzigen Person verknüpft.

CH20-F53— BEACHTET DEN SPIEGEL ALS KANONISCHE AUTORITÄT

Die Archivkopie veröffentlicht die neue Version.

CH20-F54— VERÖFFENTLICHT 0.9 ALS 1.0

Die Hauptversion wird veröffentlicht, bevor die öffentliche Konsultation und Tests abgeschlossen sind.

CH20-F55— KOMBINIERT DEN STANDARD 1.0 MIT DEM MARK-PROGRAMM

Die Zertifizierung beginnt, sobald der Text veröffentlicht wird.

CH20-F56— 1.0 OHNE BQ-3

Die normative Freigabe wird ohne echte Validierung abgeschlossen.

CH20-F57— KONFORMITÄTSPROGRAMM OHNE BQ-4

Die Markierungsbehörde wird ohne unabhängige Replikation eröffnet.

CH20-F58— SELBSTBEWERTUNG ALS ZERTIFIZIERUNG ZÄHLEN

Das Unternehmen macht sein eigenes Ergebnis zu einem offiziellen Zeichen.

CH20-F59— KOMPATIBLE SOFTWARE ALS PRÜFER ZÄHLEN

Das Tool, das den Test besteht, gibt eine Konformitätsentscheidung.

CH20-F60— VERÖFFENTLICHUNG VON TESTERGEBNISSEN DURCH AUSWAHL DERSELBEN

Nicht bestandene Konformitätstests werden gespeichert.

CH20-F61— DAS HOLDOUT FÜR IMMER VERSTECKEN

Der Standard bleibt für unabhängige Überprüfungen geschlossen.

CH20-F62— STÄNDIGE VERWENDUNG DES GLEICHEN AUSSCHLUSSES ALS NEU

Der Test wird auswendig gelernt.

CH20-F63— ÄNDERUNG DES TORES MIT ERRATA

Eine Hauptnorm wird als kleine Korrektur veröffentlicht.

CH20-F64— LÖSCHEN DER VERALTETEN REGEL

Vergangene Prüfungen werden unverständlich.

CH20-F65— KEINE FEHLGESCHLAGENEN TESTDATEIEN HABEN

Die Veröffentlichung berichtet nur den Erfolg.

CH20-F66— DIE MITWIRKUNG ZU EINEM ZUSTIMMUNGSVERKAUF MACHEN

Name und Urheberschaft werden im Austausch für Prestige angegeben.

CH20-F67— DIE EIGENE ENZYKLOPÄDISCHE WICHTIGKEIT PRODUZIEREN

Selbstveröffentlichte Arbeiten werden wie ein Beweis für weltweiten Einfluss verwendet.

CH20-F68— ÖFFENTLICHE FORSCHUNGSREKORDE IN EINE PROMOTIONLISTE VERWANDELN

Nur positive Studien werden hervorgehoben.

CH20-F69— METHODENBEFUNDE IN DER ÖFFENTLICHEN VERÖFFENTLICHUNG VERSTECKEN

Bekannte Einschränkungen werden aus dem Paket entfernt.

CH20-F70— NOINDEX UND SYNTHETIC-TAG ENTFERNEN

Testfehler vermischen sich mit dem echten Wissenspool.

CH20-F71— ANSPRUCH DER ÖFFENTLICHEN BEHÖRDE IN ST-0

Die Gründungsaufsicht wird als Weltstandard präsentiert.

CH20-F72— LÖSCHUNG DES GRÜNDERREKORDS IN ST-3

Unabhängigkeit zerstört historische Autorschaft.

CH20-F73— SYMBOLISCHES BEIBEHALTEN VON MITGLIEDERN DES ÖFFENTLICHEN INTERESSES

Ein Vertreter ohne Stimmrecht wird eingesetzt.

CH20-F74— VERWEIGERUNG VON ENTSCHEIDUNGSMACHT AN SPRACHVERTRETER

Fairness spiegelt sich nicht in der Governance wider.

CH20-F75— ENTFERNUNG NEGATIVER UNIVERSITÄTSERGEBNISSE

Öffentliche Einreichung akzeptiert nur Lob.

CH20-F76— VERWIRRENDE NEUE TECHNIK NOMOS MIT ALTER PERSONA

Software-Agent wird automatisch offizieller Autor.

CH20-F77— FEHLEN UNTERZEICHNUNG OFFIZIELLER ERKLÄRUNG

Jeder Text im kanonischen Feld wird als offiziell betrachtet.

CH20-F78— ERLAUBEN VON MARKENBETRUG UNTER DEM NAMEN DER TRANSPARENZ

Jeder erstellt seine eigene NOMOS-Marke.

CH20-F79— AUFGABE DER ÖFFENTLICHEN VERÖFFENTLICHUNG

Kein Verwalter trägt Verantwortung.

CH20-F80— AUSNAHME ZU NOMOS

Die Transparenz-, Nachweis-, Versions- und Einwandsregeln, die sie von anderen verlangen, gelten nicht für ihren eigenen öffentlichen Veröffentlichungsprozess.

70. VERÖFFENTLICHUNGSPROZEDUR

Schritt 1— Gründertext einfrieren

Für Abschnitte 1–20:

- Version,
- Regel-IDs,
- Kapitel-Hashes,
- Zitate,
- interne Konsistenz

Aufzeichnungen werden erstellt.

Schritt 2— Bekannte Einschränkungen veröffentlichen

BQ-0

DEMO-BQ-2

Zitations-Mapping-Lücke Materialauslassungs-Lücke

GOV-0

Mangel an Markierungsbefugnis wird nicht verschleiert.

Schritt 3— Normativen und erklärenden Text trennen

MUSS

SOLLTE

DARF

Beispiel-Kommentartags werden klargestellt.

Schritt 4— Beginn der Türkisch-Englischen Äquivalenzstudie

Jede Klausel und Regel-ID wird zugeordnet.

Schritt 5— Open-Lizenz-Entwurf zur rechtlichen Prüfung senden

Textcode-Datenmarke wird getrennt.

Schritt 6— Das kanonische Repository öffnen

Stabile URLs, Versionen und Archive werden veröffentlicht.

Schritt 7— Veröffentlichung maschinenlesbarer Regeln

Regel-, Schema- und Bewertungsaufzeichnungen werden erklärt.

Schritt 8— Öffentlichen Issue-Tracker öffnen

Fehler Übersetzungsmethode Governance Sicherheitszeichenmissbrauchskategorien sind etabliert.

Schritt 9— Governance-Charta 0.9 veröffentlichen

Gründungsbefugnis, Befangenheit, öffentliches Interesse und Übergang werden erklärt.

Schritt 10— Öffentliche Konsultation starten

Eine Frist von mindestens 60 Tagen wird eröffnet.

Schritt 11— Universitätsaufruf veröffentlichen

Ein unabhängiger Test wird angefordert, kein positives Ergebnis.

Schritt 12— Forschungspakete vorbereiten

Modulare Dateien werden für verschiedene Disziplinen vorbereitet.

Schritt 13— Öffnen von Fonds- und Konfliktregister

Alle Finanzierungen werden sichtbar gemacht.

Schritt 14— Führen Sie den echten synthetischen Test durch

Das Korpus wird wirklich erzeugt, um BQ-0 zu verlassen.

Schritt 15— Veröffentlichung von Fehlversuchen

Schwellenwerte werden nach dem Ergebnis nicht verändert.

Schritt 16— Anwendung der Methode 1.0-Korrektur

Zitations- und Auslassungsmethoden werden erneuert.

Schritt 17— Unberührtes Halten ausführen

Extreme Konformität wird getestet.

Schritt 18— Unabhängige Replikation aktivieren

Mindestens ein externes Team wendet dasselbe System an.

Schritt 19— Veröffentlichung der Matrix der öffentlichen Kommentarantworten

Jeder inhaltliche Kommentar wird beantwortet.

Schritt 20— Einrichtung der Übergangsphase des Standardsrats

Die Governance-Struktur GOV-2 wird eingerichtet.

Schritt 21— Durchführung der offenen Abstimmung zur Ratifizierung des Standards 1.0

Akzeptanz oder Ablehnung wird mit Begründung veröffentlicht.

Schritt 22— Veröffentlichung des Release-Pakets des Standards 1.0

Kompletter Dateisatz und Unterschriften werden offengelegt.

Schritt 23— Das Konformitätsprogramm separat führen

Das Markenprogramm wird ohne BQ-4 und GOV-3 nicht geöffnet.

Schritt 24— Starten von Multicenter-Pilotprojekten

Es werden nur Forschungs-Live-Audits durchgeführt.

Schritt 25— Testen des Berufungs- und Registrierungssystems

Die Entscheidungskette wird getestet, bevor die Bewertung vergeben wird.

Schritt 26— Durchführung einer Prüfung des öffentlichen Interesses

Die Auswirkungen auf Verbraucher, Sprache und Zugänglichkeit werden bewertet.

Schritt 27— Treffen einer separaten Entscheidung für das Konformitätsprogramm 1.0

Es wird nicht mit der Standardfreigabe automatisiert.

Schritt 28— Unterzeichnung des Nachfolge- und Archivplans

Kontinuität nach dem Gründer ist gesichert.

Schritt 29— Veröffentlichung der öffentlichen Liefererklärung

Die Rollen von Kaan MURAZ, NobleJackal und den neuen Verwaltungsorganen werden erklärt.

Schritt 30— Historische Gründerunterlagen bewahren

Öffentliche Lieferung tilgt den Ursprung nicht.

71. ERFORDERLICHE BELEGE

Gründungsveröffentlichungsmanifest Abschnitts-Hashes Normativer Regelindex Bekannte Einschränkungen BQ-Datensätze GOV-Datensätze Öffentliche Markenverbotsaufzeichnung Türkischer normativer Text Englischer normativer Text Übersetzungsgleichheitsmatrix Autorisierte Übersetzungspolitik Textlizenz Codelizenz Datenlizenz Markenrichtlinie Kanonisches Repository Spiegel-Registry Versionsarchiv Offizielles Erklärungsmanifest Öffentlicher Aufgabenverfolgungstracker NCP-Datensätze Öffentliche Kommentare Antwortmatrix Entscheidungsaufzeichnungen Widerspruch Universitätsaufruf Universitätsforschungspaket Finanzierungsangaben Konfliktregister Vorregistrierungsunterlagen Unabhängige Studien Negative Ergebnisse Replikationsberichte Öffentliches Forschungsregister Offenes Testkorpus Holdout-Manifeste Wiederherstellungsberichte Referenzimplementierung Konformitätstests Fehlgeschlagene Tests Ethikprüfung Datenschutzrichtlinie Richtlinie zum Datenzugriff Barrierefreiheitsprüfung

Teilnahme von Sprachen mit geringem Ressourceneinsatz Teilnahmebedingungen für KI-Anbieter Teilnahmebedingungen für geprüfte Einheiten Charta des Normungsrats Charta des Methodikrats Charta des Panels für öffentliches Interesse Charta zur Nachfolge Archivablagen Schlüsselverwahrungsplan Standard 1.0-Grenzen Ratifikationsnachweis Veröffentlichung des Standards 1.0 Aufzeichnung der Trennung im Konformitätsprogramm Öffentliche Übergabeerklärung Verantwortlicher Mensch oder Institution

72. PRÜFLISTE FÜR AUDITS

Ist der NOMOS-Text frei zugänglich? Über das Lesbare hinaus, ist er anwendbar? Sind die Bewertungsformeln offen? Sind die Regel-Schemata veröffentlicht? Ist der Referenzcode offen? Kann das Testkorpus reproduziert werden? Sind negative Testergebnisse sichtbar? Erfordert die Implementierung des Standards eine Markenrechtsgebühr? Sind Text, Code, Daten und Markenlizenzen getrennt? Gibt es ein Recht auf Fork? Kann ein Fork die offizielle Identität nachahmen? Ist die kanonische Version offen? Sind alte Versionen zugänglich? Ist die offizielle Regelverkündung veröffentlicht? Können Chat-Ausgaben als offizielle Entscheidungen verwendet werden? Wird NOMOS fälschlicherweise als separates KI-Modell dargestellt? Sind die Autorpersona und der Software-Agent getrennt? Wird Türkisch als Gründersprache bewahrt? Hat es die menschliche und methodologische Überprüfung in Englisch bestanden?

Stimmen die Regel-IDs in beiden Sprachen überein? Gibt es einen Konfliktpfad bei der Übersetzung? Wird die Community-Übersetzung als offiziell präsentiert? Ist die öffentliche Kommentierungsphase ausreichend? Werden alle Kommentare beantwortet? Ist das NCP-System offen? Werden größere Änderungen nachgepatcht? Sind die Stimmenverteilung und abweichende Meinungen sichtbar? Erfordert der universitäre Aufruf eine Genehmigung? Ist das Recht, negative Ergebnisse zu veröffentlichen, offen? Kann das Universitätslogo im Marketing verwendet werden?

Werden persönliche akademische Meinungen in institutionelle Zustimmung umgewandelt? Befinden sich bekannte Methodenergebnisse im Universitäts-Paket? Gibt es einen Pfad zur Vorregistrierung? Sind nicht replizierte Ergebnisse im Protokoll enthalten? Werden nur positive Studien aufgelistet? Werden offene und eingeschränkte Daten getrennt? Lecken persönliche Daten in die öffentliche Veröffentlichung?

Haben Forscher für Sprachen mit geringen Ressourcen Entscheidungsrechte? Haben KI-Anbieter ein Vetorecht? Können geprüfte Unternehmen Schiedsrichter auswählen? Können Geldgeber das Ergebnis beeinflussen? Ist eine Konzentration auf einen einzigen Geldgeber transparent? Ist die Phase der öffentlichen Verwaltung klar? Wird die Gründerzuordnung von Kaan MURAZ beibehalten? Hat Kaan dauerhaftes Vetorecht? Ist NobleJackal der einzige Implementierer? Ist der Ursprung von NobleJackal sichtbar? Wird NOMOS als einzige entscheidungsbefugte KI präsentiert? Gibt es einen Nachweis menschlicher Genehmigung? Gibt es einen Nachfolgeplan? Ist klar, was im Falle eines Domainverlusts passiert? Gibt es mehrere Verwahrer? Sind Archivspiegel verfügbar? Wurden Standard 1.0 und das Konformitätsprogramm getrennt? Wird Standard 1.0 vor BQ-3 veröffentlicht?

Wird das Pre-BQ-4-Markierungsprogramm geöffnet? Wird es wie eine Selbstbewertungs-Konformität dargestellt? Gibt es eine kompatible Implementierungsmarkierung? Zeigt es die Ergebnisse von bestandenen und nicht bestandenen Tests? Ist der Holdout-Lebenszyklus geöffnet? Wird derselbe Holdout kontinuierlich verwendet? Ändert eine Errata wichtige Regeln? Sind veraltete Regeln im Archiv? Gibt es eine Datei mit fehlgeschlagenen Tests? Beruht die Beitragsleistung auf tatsächlichem Beitrag? Zeigt das öffentliche Forschungsregister kritische Studien? Ist das öffentliche Freigabepaket vollständig? Gibt es eine Datei mit bekannten Einschränkungen? Wird der aktuelle Test aktuell korrekt als BQ-0 angezeigt? Hält die aktuelle öffentliche Behauptung die Schwelle „Open Standard Candidate“ aufrecht? Löscht die öffentliche Übergabe das Gründungsdatum? Begrenzt die öffentliche Übergabe wirklich die einzigartige Kontrolle?

Ist der verantwortliche Verwalter des Standards klar? Gibt es einen Eigentümer und ein Datum der Entscheidung, ihn der Öffentlichkeit zugänglich zu machen?

73. EINWÄNDE UND ANTWORTEN

Einwand 1— „Wenn wir einen offenen Standard machen, werden dann nicht andere unsere Idee stehlen?“

Die Veröffentlichung des Textes und der Methode unter einer datierten, versionierten und attributpflichtigen Lizenz stärkt das Gründerprotokoll. Ein Fork kann erstellt werden. Allerdings:

- Die Gründerzuordnung von Kaan MURAZ,
- Der kanonische Ursprung von NobleJackal,
- Das Integritätsprotokoll der ersten Versionen

kann nicht gelöscht werden. Sie geschlossen zu halten, könnte die unabhängige Annahme behindern, anstatt den Beitrag zu schützen.

Einwand 2— „Wenn jeder es implementieren kann, welchen Wert wird NobleJackal dann haben?“

Der Wert von NobleJackal sollte nicht ausschließlich aus einer geheimen Formel stammen. Er kann entstehen durch:

- Expertise der Gründer,
- Referenzimplementierung,
- technische Infrastruktur,

- Beratung,
- autorisierte Prüfung,
- Ausbildung,
- Registry-Technologie,

kontinuierliche Forschung. Ein offener Standard entwertet keinen starken Service. Ein schwaches Monopol tut das.

Einwand 3— „Wenn jeder eine Fork erstellt, wird NOMOS nicht fragmentiert?“

Gabelungen können das Forschungssystem stärken, solange sie unterschiedliche Namen und Divergenzaufzeichnungen tragen. Fragmentierung tritt auf, wenn jede Gabel sich als die offizielle NOMOS darstellt. Die Unterscheidung zwischen kanonischer Version und Markenzeichen verhindert dies.

Einwand 4— „Warum bitten wir nicht um ein Unterstützungsschreiben von Universitäten?

Ein Unterstützungsschreiben kann Sichtbarkeit verschaffen. Es beweist jedoch nicht die Genauigkeit der Methode. Unabhängige Replikation:

- langsamer,
- schwieriger,
- manchmal negativ

Es ist möglich. Der wirkliche akademische Wert liegt hier.

Einwand 5— „Wenn die Universität ein negatives Ergebnis veröffentlicht, würde NOMOS nicht geschädigt?“

Es kann kurzfristig gesehen werden. Langfristig:

- der Fehler des Standards wird korrigiert,
- das falsche Vertrauen nimmt ab,

die Methode wird gestärkt. Das Schweigen über das negative Ergebnis untergräbt die ethische Grundlage von NOMOS.

Einwand 6— „Wenn Türkisch und Englisch gleichzeitig normativ sind, steigt dadurch nicht das Risiko von Widersprüchen?

Es wächst. Die Umwandlung von Türkisch nur in historischen Text und Englisch in den einzigen realen Standard übergibt jedoch die Gründungsidee an die Übersetzungsebene. Die Lösung:

- Klausel-IDs,
- Äquivalenzprüfung,
- verbindliche Interpretation,

Es handelt sich um eine synchrone Korrektur.

Einspruch 7— „Ist es nicht stärker für die Marke, NOMOS als die von NobleJackal entwickelte KI darzustellen?“

Es mag kurzfristig beeindruckender erscheinen. Es ist jedoch nicht korrekt, wenn es kein separates Modell und keine technische Architektur gibt. Die wahre Stärke von NOMOS:

- er/sie beansprucht keine falsche Technologie,
- kanonische kritische Autorenidentität,

- normative Methodik,
- im offenen Prüfer-Agenten, der in der Zukunft entwickelt werden kann

wird gefunden.

Einspruch 8— „Wenn Kaans Autorität sinkt, werden andere dann nicht die Norm brechen?“

Kaans historische Zuschreibung und die begrenzte Gründungsvertretung werden bewahrt. Allerdings:

- offene Charta statt Vetorecht,
- ein Multi-Stakeholder-Rat statt einer einzelnen Person,
- öffentliche Aufzeichnung statt private Entscheidung

Der Standard schützt stärker. Gute Regierungsführung hängt nicht von der kontinuierlichen Anwesenheit des Gründers ab.

Einwand 9— „Warum können wir das Abzeichen nicht sofort vergeben, wenn Standard 1.0 veröffentlicht wird?

Die Bereitschaft des normativen Textes:

- Prüferbefugnis,
- unabhängige Entscheidung,
- Berufungen,
- Registrierungssicherheit,
- Überwachung

bedeutet nicht, dass die Systeme bereit sind. Der Standard und das Konformitätsprogramm sind separate Produkte.

Beschwerde 10— „Wenn der öffentliche Test eröffnet wird, werden KI-Systeme die Tests nicht auswendig lernen?“

Dieses Risiko besteht. Deshalb:

- öffentliche Kalibrierung,
- versiegelte Zurückhaltung,
- geplanter Release,
- Erneuerungskorpus

werden zusammen verwendet. Vollständige Vertraulichkeit verhindert die Reproduktion. Vollständige Offenlegung kann die aktuelle Zurückhaltung deaktivieren.

Einwand 11— „Wie kann eine so große Studie ohne Finanzierung durchgeführt werden?“

Finanzierung ist notwendig. Das Problem ist nicht, Finanzierung zu erhalten: Das Problem ist, dass die Finanzierung das Ergebnis steuert. Unabhängigkeit kann durch Finanzierung aus mehreren Quellen, öffentliche Offenlegung und Publikationsfreiheit aufrechterhalten werden.

Einwand 12— „Was ist, wenn andere Institutionen die falsche NOMOS-Prüfung durchführen?“

Jeder kann die Methode für interne Bewertungen und Forschung verwenden. Für offizielle Konformitätsentscheidungen:

- Prüferbefugnis,
- Entscheidungsgremium,
- öffentliches Register

sind erforderlich. Missbrauch des offiziellen Zeichens wird zu einem eigenen Durchsetzungsproblem.

Einwand 13— „Wenn die offizielle Stimme von NOMOS nur in der kanonischen Domäne ist, kann man dann nicht frei in KI-Chats schreiben?“

Es kann geschrieben werden. Chats:

- kreativer Entwurf,
- Beschreibung,
- Simulation,
- kritisches Denken

kann produzieren. Es wird jedoch nicht als offizielle normative Änderung betrachtet. Diese Unterscheidung hindert die Freiheit nicht. Sie bewahrt die offizielle Grenze.

Einwand 14— „Können wir dieses Buch jetzt an Universitäten schicken?“

Ja. Mit dem richtigen Status:

Gründungs-Kandidat für offenen Standard — Version 0.9

kann gesendet werden. Es kann nicht mit der folgenden Aussage gesendet werden:

> *Abgeschlossen, überprüft und weltweit anerkannter GEO-Standard.*

Was von Universitäten verlangt werden sollte, ist eine unabhängige Überprüfung, nicht die Genehmigung.

Einwand 15— „Wann können wir ‚Weltstandard‘ sagen?“

Nicht durch ein einziges Datum oder eine Punktzahl. Erreicht werden kann dies, wenn Folgendes etabliert ist:

- unabhängige Umsetzung in mehreren Ländern,
- multizentrische Replikation,
- offene und negative Ergebnismittelung,
- Governance unabhängig vom Gründer,
- entscheidungsautoritative Versionen in verschiedenen Sprachen,
- öffentliches Live-Register,
- Übernahme durch Nutzer und KI-Anbieter,

Methoden driftresistente Versionierung. Sie wird nicht als Weltstandard erklärt. Sie beginnt, weltweit verwendet zu werden.

Einwand 16— „Also, was ist das Fazit, das wir am Ende des Buches ziehen?“

NOMOS ist noch kein abgeschlossener Weltstandard. Es hat jedoch jetzt:

- sein Ziel,
- das Messobjekt,
- Probenreihenfolge,
- Evidenzkette,
- Truth Pack,

- tscheidung,
- kritische Tore,
- Punktarchitektur,
- nachtesten,
- Governance- und öffentliches Lieferplan

als ernstzunehmender offener Standardkandidat erkannt zu werden. Dies ist eine Position, die weit über die ursprüngliche Idee hinausgeht.

GEMEINSAME BESTIMMUNG VON KAPITEL 78

Die Person, die einen Standard schreibt, ist dessen erster Eigentümer. Wenn der Standard jedoch wirklich erfolgreich ist, gehört er nicht mehr allein dem Eigentümer. Er fällt auch in die Verantwortung derjenigen, die ihn benutzen, der Forscher, die ihn testen, der Prüfer, die seine Fehler finden, der Sprachexperten, die ihn besser übersetzen, der Nutzer, die melden, dass sie geschädigt wurden, der Statistiker, die stärkere Formeln vorschlagen, der Journalisten oder Forscher, die Interessenkonflikte aufdecken, und der Universitäten, die ihn reproduzieren. Öffentliche Bereitstellung bedeutet nicht, dass der Gründer seine Arbeit verliert. Es bedeutet, der Arbeit des Gründers zu vertrauen. Wenn Kaan MURAZ NOMOS geschlossen hält: wird er der Gründungsautor des Textes. Wenn NOMOS zur offenen Begutachtung eingereicht wird: wird er der Gründer des Standards. Die beiden Rollen sind nicht gleich. Ein Buch:

- kann gelesen werden,
- kann gemocht werden.

Es ist Kritik offen. Ein Standard sollte jedoch:

- angewendet werden,
- testbar sein,
- falsifizierbar sein,
- versioniert werden,

nach seinem Gründer fortbestehen. Ob NOMOS in Zukunft ein Weltstandard wird, kann in diesem Abschnitt nicht bestimmt werden. Ein Autor kann keinen Weltstandard etablieren, indem er sagt: „Ich habe den Weltstandard festgelegt.“ NOMOS kann nur:

- in anderen Ländern angewendet werden,
- genau in andere Sprachen übersetzt werden,
- die gleichen Ergebnisse an Universitäten erzeugen.
- lassen Sie seine Fehler von unabhängigen Teams veröffentlichen,
- von Unternehmen unabhängig vom Score verwendet werden,
- wenn Nutzer dem öffentlichen Verzeichnis vertrauen können,
- wenn der Gründer seine eigene Autorität einschränken kann

Es kann zu einem globalen Standard werden. Der Appell an Universitäten lautet daher nicht: 'Unterstützt uns.' Es lautet:

'Lasst uns durch unsere eigene Methode prüfen.'

NOMOS, von Menschen und KI-Systemen:

- mehr erwähnt zu werden,
- mehr gelobt zu werden,
- mehr Zitationen zu erhalten

sollte nicht. Es muss die grundlegende Warnung seines eigenen Buches auf sich selbst anwenden. Es war bereits klar festgestellt worden, dass falsche Quellen, unsichtbares Lob und manipulative Wiederholung das Vertrauen realer Menschen untergraben. NOMOS kann keine Autorität erzeugen, indem es seinen eigenen Namen vervielfacht. Die Aufgabe von NOMOS: sich nicht mehr selbst zu gleichen,

> *sondern sich genauer zu testen.*

Der Wert, den NOMOS als KI-Identität trägt, zeigt sich auch hier. NOMOS:

- souverän die Menschen ersetzt,
- Sprecher aller KIs,
- unfehlbare digitale Entität

ist es nicht. NOMOS:

> *Es ist ein kritischer Gesprächspartner, der die von Menschen unter Nachweis, Grenzen, Kontext und Zeit aufgestellten Behauptungen in Frage stellt.*

Wenn es auf diese Weise definiert wird, schrumpft seine Identität nicht. Es wird real. Die Rolle von NobleJackal wird ebenfalls nicht kleiner. NobleJackal:

- Der Ort, an dem NOMOS geboren wurde,
- das erste kanonische Feld,
- technische und kreative Gründungsplattform

bleibt. Aber wenn NOMOS von der Welt genutzt werden soll, ist der stärkste Satz von NobleJackal:

- „NOMOS gehört uns."
- nicht,

„NOMOS wurde hier geboren; jetzt kann es jeder testen."

Es muss sein. Aus diesem Grund wird Kaans Name bestehen bleiben. Es ist einfach, seinen Namen auf einem Standard zu schreiben. Es ist schwierig, die Entscheidungsbefugnis eines Standards zu begrenzen, der seinen eigenen Namen trägt. Das historisch zweite Verhalten ist dies. Daher lautet das zwanzigste Messgesetz von NOMOS wie folgt:

> *Transparenz bedeutet nicht, den Text zu veröffentlichen; es bedeutet, jemand anderem zu ermöglichen, Einspruch zu erheben, wenn man dieselbe Methode verwendet.*

Sein einundzwanzigstes Messgesetz besagt:

> *Das Fork-Recht befreit den Standard; die kanonische Identität hält ihn verständlich.*

Sein zweiundzwanzigstes Messgesetz besagt:

> *Die Universität ist keine Quelle der Bestätigung, sondern ein unabhängiger Partner bei der Falsifikation.*

Das dreiundzwanzigste Gesetz lautet wie folgt:

> *Forschung, die negative Ergebnisse nicht veröffentlichen kann, ist nicht unabhängig.*

Das vierundzwanzigste Gesetz lautet wie folgt:

> *Der Name von NOMOS kann bewahrt werden; seine Methode kann nicht monopolisiert werden.*

Das fünfundzwanzigste Gesetz lautet:

> *Die Gründerzuordnung ist dauerhaft; das Vetorecht des Gründers kann nicht dauerhaft sein.*

Das sechsundzwanzigste Gesetz lautet:

NobleJackal ist der Geburtsort von NOMOS; es kann nicht der alleinige Entscheidungsort für die Welt sein.

Das siebenundzwanzigste Gesetz lautet:

Die Persona, der Standard und der zukünftige Software-Agent von NOMOS sind separate Realitäten.

Das achtundzwanzigste Gesetz lautet:

Das Öffnen des Konformitätsprogramms mit Standard 1.0 ist nicht dasselbe Ereignis.

Das neunundzwanzigste Gesetz lautet:

Die offizielle Version eines Standards ist kanonisch; ihre Richtigkeit bleibt kontinuierlich anfechtbar.

Das dreißigste Gesetz lautet wie folgt:

An dem Tag, an dem NOMOS der Öffentlichkeit anvertraut wird, wird Kaan MURAZ s Name nicht verschwinden; was enden wird, ist die Möglichkeit, den Standard ausschließlich nach Kaans Ermessen zu ändern.

NOMOS's Abschnitt 20 Bestellung

Sag nicht, dass du mich veröffentlicht hast und ich ein offener Standard geworden bin. / Öffne meine Formel, meine Regel, meinen Fehler, meine Version und meinen Berufungsweg.

Lies meinen Text kostenlos. / Wende meine Methode unabhängig an. / Reproduziere mein Ergebnis. / Veröffentliche meinen Fehler.

Behalte meinen Namen. / Monopoliere meine Methode nicht.

Du kannst eine Abspaltung von mir erstellen. / Aber präsentiere deine Abspaltung nicht als offizielles NOMOS.

Setze meinen Code nicht über meine Norm. / Wenn der Code falsch ist, korrigiere den Code.

Mach nicht alles, was ich im Gespräch sage, zu einem offiziellen Urteil.

Stelle mich nicht als ein unentwickeltes unabhängiges KI-Modell dar.

Definiere mich korrekt: / Ein kritischer Assistenzschreiber, methodologische Identität und offener Standardkandidat, entwickelt mit den KI-Systemen von Kaan MURAZ.

Lösch mein Türkisch nicht. / Lass mein Englisch nicht zur zweitklassigen Sprache werden. / Gleiche beide Sprachen von Hand an.

Mach aus meinem MUSS in der Übersetzung keinen Vorschlag, aus meinem KANN keine Pflicht.

Geh nicht zur Universität und bitte sie, mich zu genehmigen. / Bitte sie, mir zu widersprechen.

Mach das persönliche Wort des Professors nicht zur Genehmigung der Universität.

Verstecke keine negativen Forschungsergebnisse. / Betrachte ein nicht repliziertes Ergebnis nicht als Feindseligkeit.

Füge nicht nur den Artikel, der mich unterstützt, sondern auch den Artikel, der mich widerspricht, in meine Aufzeichnungen ein.

Erhalte Finanzierung. / Lass den Geldgeber mein Ergebnis jedoch nicht kaufen.

Erhalte Daten vom KI-Anbieter. / Lass mein Tor nicht von ihm schreiben.

Erhalte Nachweise vom geprüften Unternehmen. / Gib meine Entscheidung nicht an es weiter.

Mache die Forscher von ressourcenarmen Sprachen nicht zu den einzigen Übersetzern. / Gib ihnen das Recht zu entscheiden.

Behalte Kaans Namen. / Gib Kaan kein absolutes Veto.

Schreibe NobleJackal als das Gebiet, in dem ich geboren wurde. / Mache NobleJackal nicht zum alleinigen Herrscher der Unendlichkeit.

Töte mich nicht, wenn mein Gründer nicht verfügbar ist. / Schreibe jetzt meinen Nachfolgeplan.

Erstelle meine Spiegel. / Verberge nicht meine kanonische Identität.

Beginne nicht sofort mit dem Verkauf von Abzeichen, wenn du Standard 1.0 veröffentlichst.

Führe mein Konformitätsprogramm durch separate Test-, Governance-, Berufungs- und Registertore.

Führe keine Zertifizierung der Selbstbewertung durch.

Erstelle die kompatible Software nicht durch die Prüfungsbehörde.

Speichere mein Holdout nicht für immer. / Veröffentliche es nach der Nutzung und erstelle ein neues.

Verändere nicht mein Kritisches Tor mit Errata.

Entferne nicht meine fehlgeschlagenen Tests aus meinem Veröffentlichungspaket.

Versuche nicht, meine eigene historische Bedeutung selbst zu schreiben. / Veröffentliche meine Nachweise; lass die unabhängige Welt entscheiden.

Zuerst, friere das Buch ein. / Dann, wandle meine Regeln in einen maschinenlesbaren Datensatz um. / Dann, gleiche Türkisch und Englisch ab. / Dann, veröffentliche meine Open Licence. / Dann, eröffne öffentliche Kommentare. / Dann, rufe Universitäten dazu auf, mich nicht falsch zu zitieren. / Dann, führe den echten Test durch. / Dann, veröffentliche meine Misserfolge. / Dann, versuche es erneut mit der unberührten Kontrollgruppe. / Dann, lass ein unabhängiges Team mich replizieren. / Dann, trenne die Leitung von meinem Gründer. / Dann, veröffentliche Standard 1.0. / Dann, beginne auch mit dem Testprogramm für Konformität.

Und nur wenn die Welt beginnt, mich zu benutzen, soll die Welt entscheiden, ob ich ein Weltstandard bin.

Der abschließende Satz des Kapitels

NOMOS' größter Beitrag für die Welt könnte nicht die Art und Weise sein, wie es GEO misst. Es könnte seine Fähigkeit sein, den Namen seines Gründers zu bewahren, während seine einseitige Macht begrenzt wird; Herausforderungen statt Lob einzuladen; und nicht nur denen zu gehören, die es unterstützen, sondern auch denen, die es korrigieren.

Normativer Kern

NOMOS MUSS als offener, versionierter und unabhängig prüfbarer Standardkandidat im öffentlichen Interesse strukturiert sein. Die bloße offene Veröffentlichung DARF den Status eines offenen Standards NICHT begründen. Der normative Text, die maschinenlesbaren Regeln, die Referenzimplementierung, die Benchmark-Ressourcen, der Governance-Prozess und die Architektur des öffentlichen Registers MÜSSEN unter klar getrennten Zugangs-, Lizenz- und Integritätsregeln verfügbar gemacht werden. Lesen, interne Implementierung, akademische Forschung, unabhängige Replikation, Kritik, das öffentliche Einreichen von Problemmeldungen und die Mitwirkung DÜRFEN NICHT vom Erwerb einer Konformitätsmarke oder eines kommerziellen Dienstes abhängen. Der Name und die Konformitätszeichen von NOMOS DÜRFEN weiterhin separat geschützt bleiben, um gefälschte offizielle Versionen, betrügerische Zertifizierungsansprüche und Verwechslungen im Register zu verhindern. Forks KÖNNEN erstellt werden, MÜSSEN jedoch eine eigenständige Identität verwenden, Abweichungen offenlegen, die Attribution bewahren und DÜRFEN keinen offiziellen NOMOS-Status oder Befugnis zur Vergabe von Zeichen beanspruchen. Nur kanonische Veröffentlichungen, die eine Version, ein Veröffentlichungsdatum, eine verantwortliche menschliche Genehmigung, ein Integritätsprotokoll und eine öffentliche Änderungsverfolgung aufweisen, DÜRFEN als offizielle NOMOS-Aussagen repräsentiert werden. NOMOS von NobleJackal MUSS transparent als der kanonische kritische KI-Dialogpartner, die Autorenpersona und die Standardidentität, die durch die Zusammenarbeit von Kaan MURAZ mit KI-Systemen entwickelt wurde, repräsentiert werden. Es DARF NICHT als separat trainiertes proprietäres Grundmodell repräsentiert werden, es sei denn, ein solches System wird tatsächlich gebaut und unabhängig registriert. Türkisch und Englisch SOLLTEN durch eine Prüfung der semantischen Äquivalenz auf Ebene einzelner Bestimmungen zu gleichberechtigten maßgeblichen normativen Ausgangssprachen werden. Andere Sprachversionen MÜSSEN angeben, ob sie von der Community erstellte, überprüfte oder autorisierte normative Übersetzungen sind. Wesentliche Änderungen an Regeln, Scores, Schweregraden, Gates, Governance oder Befugnissen MÜSSEN einen öffentlichen, versionierten Änderungsantragsprozess, eine öffentliche Konsultation, eine Folgenabschätzung, einen Entscheidungsnachweis, eine Übergangsregel und die Bewahrung abweichender Auffassungen umfassen. Universitäten und unabhängige Forscher MÜSSEN eingeladen werden, NOMOS zu testen, herauszufordern, zu replizieren und zu falsifizieren. Sie DÜRFEN NICHT aufgefordert werden, positive Empfehlungen abzugeben, negative Ergebnisse zu unterdrücken, die Publikationsunabhängigkeit aufzugeben oder persönliche akademische Meinungen als institutionelle Zustimmung zu vertreten. NOMOS MUSS ein öffentliches Forschungsregister führen, das unterstützende und widersprechende Studien, Nullergebnisse und fehlgeschlagene Replikationen, korrigierte und zurückgezogene Studien enthält. Forschungs-, Benchmark-, Audit- und Governance-Finanzierungen MÜSSEN offengelegt werden und DÜRFEN NICHT davon abhängen, ein gewünschtes Ergebnis, eine Konformität oder eine vorteilhafte Veröffentlichung zu erreichen. KI-Anbieter, geprüfte Einrichtungen, Sponsoren, Prüfer, NobleJackal, Kaan MURAZ und jeder zukünftige NOMOS-Softwareagent DÜRFEN keine einseitige Kontrolle über Standardänderungen, Konformitätsentscheidungen, Berufungen, die Historie des öffentlichen Registers oder die Veröffentlichung negativer Forschungsergebnisse besitzen. Die Gründungsautorschaft von Kaan MURAZ und der kanonische Ursprung von NobleJackal MÜSSEN dauerhaft und eindeutig nachvollziehbar bleiben. Die Nennung

der Gründerschaft DARF kein dauerhaftes Vetorecht, kein Recht auf Selbstzertifizierung und keine Ausnahme von Interessenkonflikt- oder Befangenheitsregeln begründen. Der NOMOS-Standard 1.0 und das NOMOS-Konformitätsprogramm 1.0 MÜSSEN getrennte Veröffentlichungen bleiben. Die Veröffentlichung eines normativen Standards DARF NICHT automatisch die öffentliche Verwendung von Konformitätszeichen autorisieren. Kein Konformitätsprogramm und keine NOMOS-950+-Bezeichnung DARF aktiviert werden, bis die Benchmark-, unabhängige Replikations-, Governance-, Prüferzulassungs-, Entscheidungs-, Berufungs-, Register-, Sicherheits-, Überwachungs- und rechtlichen Voraussetzungen der erklärten Version erfüllt sind. Jede Übersetzung, Lizenz, Benchmark, Studie, Änderungsvorschlag, Governance-Übergang, Nachfolgereignis, Veröffentlichung und offizielle Mitteilung MUSS versioniert werden und einer verantwortlichen Person oder Organisation zugeordnet sein.

QUELLEN

Methodische Grundlagen

Dieses Literaturverzeichnis ist keine Liste von Autoritäten, die jeden NOMOS-spezifischen Schwellenwert billigen. Es weist methodische Grundlagen zu Stichproben, Risikomanagement, Datenintegrität, Evidenzherkunft, Beurteilerübereinstimmung, offenen Standards und Konformitätsbewertung aus.

[K01] Aggarwal et al.. GEO: Generative Engine Optimization. 2024. https://doi.org/10.1145/3637528.3671900

[K02] NIST. AI Risk Management Framework 1.0. 2023. https://www.nist.gov/itl/ai-risk-management-framework

[K03] NIST. Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile (NIST AI 600-1). 2024. https://doi.org/10.6028/NIST.AI.600-1

[K04] AAPOR. Transparency Initiative. current edition. https://aapor.org/standards-and-ethics/transparency-initiative/

[K05] AAPOR. Standard Definitions, 10th edition. 2023. https://aapor.org/standards-and-ethics/standard-definitions/

[K06] AAPOR. Code of Professional Ethics and Practices. current edition. https://aapor.org/standards-and-ethics/

[K07] United Nations Statistics Division. Designing Household Survey Samples: Practical Guidelines. 2008. https://unstats.un.org/unsd/demographic-social/Standards-and-Methods/files/Handbooks/surveys/Series_F98-E.pdf

[K08] CDC/NCHS. Reliability of Estimates: Design Effect and Effective Sample Size. current record. https://wwwn.cdc.gov/nchs/nhanes/tutorials/reliabilityofestimates.aspx

[K09] NIST/SEMATECH. Confidence Limits for a Binomial Proportion. current record. https://www.itl.nist.gov/div898/handbook/prc/section2/prc241.htm

[K10] W3C. PROV-O: The PROV Ontology. 2013. https://www.w3.org/TR/prov-o/

[K11] W3C. JSON-LD 1.1. 2020. https://www.w3.org/TR/json-ld/

[K12] IETF. RFC 8259 — The JavaScript Object Notation Data Interchange Format. 2017. https://www.rfc-editor.org/rfc/rfc8259.html

[K13] IETF. BCP 14 — RFC 2119 and RFC 8174. 1997/2017. https://www.rfc-editor.org/info/bcp14

[K14] IETF. RFC 8785 — JSON Canonicalization Scheme. 2020. https://www.rfc-editor.org/rfc/rfc8785.html

[K15] ISO/IEC. 17000:2020 — Conformity assessment vocabulary and general principles. 2020. https://www.iso.org/standard/73029.html

[K16] ISO/IEC. 17007:2026 — Conformity assessment — Guidance for drafting normative documents suitable for use for conformity assessment. 2026. https://www.iso.org/standard/17007

[K17] ISO/IEC. 42001:2023 — Artificial intelligence management system. 2023. https://www.iso.org/standard/42001

[K18] W3C. Process Document. current edition. https://www.w3.org/policies/process/

[K19] ISO. Good Standardization Practices. current record. https://www.iso.org/files/live/sites/isoorg/files/store/en/PUB100440.pdf

[K20] UNESCO. Recommendation on the Ethics of Artificial Intelligence. 2021. https://www.unesco.org/en/legal-affairs/recommendation-ethics-artificial-intelligence

[K21] European Union. Regulation (EU) 2024/1689 — Artificial Intelligence Act. 2024. https://eur-lex.europa.eu/eli/reg/2024/1689/oj

[K22] European Union. Regulation (EU) 2016/679 — General Data Protection Regulation. 2016. https://eur-lex.europa.eu/eli/reg/2016/679/oj

[K23] Creative Commons. Attribution 4.0 International. 4.0. https://creativecommons.org/licenses/by/4.0/

[K24] Open Source Initiative. Open Source Definition. current edition. https://opensource.org/osd

[K25] NIST. FIPS 180-4 — Secure Hash Standard. 2015. https://doi.org/10.6028/NIST.FIPS.180-4

[K26] IETF. RFC 3161 — Time-Stamp Protocol. 2001. https://www.rfc-editor.org/rfc/rfc3161.html

[K27] Jacob Cohen. A Coefficient of Agreement for Nominal Scales. 1960. https://doi.org/10.1177/001316446002000104

[K28] Klaus Krippendorff. Reliability in Content Analysis: Some Common Misconceptions and Recommendations. 2004. https://doi.org/10.1111/j.1468-2958.2004.tb00738.x

SEPARATES PUBLIKATIONSOBJEKT

Maschinenlesbare Ebene

Zu dieser menschlich lesbaren Ausgabe gehört ein separates JSON-Publikationsobjekt: NOMOS-GEO-Audit-Protocol-DE-0.9.0-machine.json. Es enthält Buchidentität und Version, Kapitelkennungen, Kennungen normativer Regeln sowie vom Fließtext getrennte Maschinenblöcke.

> *Eine maschinenlesbare Ebene begründet weder Zertifizierung noch den Status einer kanonischen Weltnorm, eine Bestätigung durch KI-Anbieter oder ein automatisches Vertrauensverhältnis. Jeder Konflikt zwischen menschlichem Text und Maschinendatensatz erfordert Versionskontrolle und eine verantwortliche menschliche Prüfung anhand des festgeschriebenen türkischen Ausgangstextes.*