NOBLEJACKAL / NOMOS

NOMOS

Protocolo de Auditoría GEO de NOMOS

EL MÉTODO GEO-1000

Un protocolo para medir la representación de entidades en sistemas generativos en la población mundial

Kaan MURAZ

0.9.0 · EDICIÓN EDITORIAL EN ESPAÑOL

Protocolo de Auditoría GEO de NOMOS

Un protocolo para medir la representación de entidades en sistemas generativos en la población mundial

Autor y fundador responsable: Kaan MURAZ

Identidad del método: NOMOS by NobleJackal

Versión: 0.9.0 — texto final candidato, edición editorial en español

Idioma: español (es-ES)

© 2026 Kaan MURAZ. La licencia de publicación de esta edición candidata aún no ha sido declarada.

En esta edición, NOMOS es un candidato a estándar abierto y versionado que se somete a pruebas independientes. No es un estándar internacional adoptado ni un programa de certificación activo.

Estado del benchmark real: BQ-0 — no ejecutado. Las cifras del capítulo 18 constituyen una demostración sintética calculada del método; no son resultados de usuarios reales, productos de IA reales ni de un estudio de campo independiente.

Esta edición no crea por sí sola una marca de conformidad NOMOS, una afirmación «NOMOS 950+», un certificado, una clasificación ni el respaldo de terceros.

CÓMO LEER ESTE LIBRO

Este libro no es un manual sobre cómo hacer GEO. Es un protocolo de auditoría y un candidato a estándar que define las condiciones de población, evidencia, contexto, tiempo, adjudicación y gobernanza bajo las cuales puede medirse la representación de entidades en sistemas generativos.

La narrativa para lectores humanos, las disposiciones normativas, los procedimientos de auditoría y la capa legible por máquinas deben entenderse como objetos distintos pero conectados. DEBE y NO DEBE identifican disposiciones obligatorias; DEBERÍA identifica recomendaciones; PUEDE identifica opciones permitidas. Los estados CANDIDATO, SINTÉTICO, ABIERTO y BQ limitan el nivel de evidencia declarado.

Las referencias marcadas como [Kxx] indican fundamentos metodológicos enumerados en la bibliografía. La existencia de una fuente no significa que respalde todos los umbrales candidatos de NOMOS; las reglas propias de NOMOS siguen siendo expresamente provisionales.

> *El capítulo 0 es la introducción del libro. Establece la Declaración de Custodia del Fundador y la Declaración de Derechos de Representación, a las que queda sujeta toda disposición técnica.*

TÉRMINOS Y CÓDIGOS DE ESTADO

Afirmación atómica La unidad más pequeña de significado capaz de transportar un juicio independiente sobre la verdad, el alcance, el tiempo, la entidad o la evidencia.

Alcance natural La población objetivo capaz de acceder a un determinado producto de IA, idioma, superficie y condición de uso de manera natural y legítima.

Base común de comparación La población de usuarios que comparte condiciones de acceso legítimas en todos los productos de IA que se están comparando.

BQ El código de Calidad de Referencia que indica el nivel de ejecución y reproducibilidad de un punto de referencia real.

Brecha de referencia Una condición en la que no existe o no se puede acceder a una referencia confiable requerida para una decisión.

Crítico Un hallazgo irrecuperable capaz de causar un daño severo a un usuario, un derecho, la seguridad o una decisión material.

Demostración sintética Un ejemplo calculado utilizado para probar el funcionamiento interno del método; no es un estudio real de usuarios, salida real de IA o resultado real del rendimiento institucional.

GEO-1000 La escala de informes que expresa los resultados de la población objetivo por equivalentes de usuario 1,000.

Grave Un hallazgo material que distorsiona el resultado o el juicio de conformidad y no puede ser compensado por una puntuación compuesta.

Moderado Un hallazgo de efecto material limitado que, no obstante, requiere remediación y puede conducir a un resultado condicional.

No resuelto Un estado de decisión utilizado cuando la evidencia es insuficiente para un juicio final o un conflicto no puede resolverse.

NOMOS La voz metodológica crítica y la identidad de estándares desarrollada en este trabajo por Kaan MURAZ con sistemas de IA; no es una persona jurídica independiente ni un modelo base separado.

Observación Un hallazgo no material que requiere mejoras en claridad, forma o buenas prácticas.

Paquete de hechos verificados El paquete de auditoría versionado combinando identidad, alcance, tiempo, evidencia y registros evidencia en contrario para una entidad.

Producto de IA El objeto medido definido por el producto del sistema generativo particular, la superficie, el plan y la versión disponibles para el usuario.

Prompt La entrada del sistema generativo que bloquea la pregunta de investigación, la entidad objetivo, el contexto y la carga de respuesta.

Registro de captura Un registro que conserva el prompt enviado, la respuesta, el tiempo, la superficie y la evidencia de integridad juntos.

Tasa de cobertura La frecuencia con la que un método de estimación por intervalos contiene el valor verdadero asumido bajo un diseño de prueba.

CONTENIDO

Declaración de Derechos de Representación 1

Lo que medimos 11

Una sola respuesta no es una puntuación GEO 13

La representación es una distribución 41

¿Qué entidad se audita exactamente? 73

Dos capas de realidad 105

Población mundial y equidad entre países y lenguas 139

La población usuaria pertinente para cada producto de IA 141

La muestra GEO-1000 proporcional a la población 177

Paneles de observación por país y equidad lingüística 219

Selección y ponderación de participantes y sesgo de muestreo 255

Ejecución del experimento y cadena de evidencia 301

Registro de sistemas de IA y comparabilidad 303

La constitución del prompt 345

El panel controlado en condiciones depuradas y el panel de usuarios en condiciones reales 395

Oleadas sincrónicas y registro NOMOS 447

Realidad, adjudicación y decisión 499

El paquete de hechos verificados de la entidad 501

Afirmaciones atómicas y adjudicación 563

Aprobación de respuestas y criterios de bloqueo por errores críticos 625

La arquitectura de puntuación NOMOS 681

Diseño de prueba sintética y demostración integral 735

Diseño de prueba sintética de Apple.com 737

Una auditoría sintética integral de 30.000 respuestas 793

Conformidad, confianza pública y gobernanza independiente 839

Marca de conformidad, registro público y gobernanza independiente 841

Estándar abierto, convocatoria a universidades y entrega pública de NOMOS 899

Fundamentos metodológicos 957

Capa legible por máquinas 959

INTRODUCCIÓN · CAPÍTULO 0

Declaración de Derechos de Representación

El núcleo constitucional de una propuesta de estándar GEO

BASES METODOLÓGICAS / K20; K22

Disposición Inicial

Sistemas generativos ahora:

- personas,
- empresas,
- instituciones,
- productos,
- profesiones,
- servicios,
- organizaciones públicas,
- eventos históricos

identifican, comparan y sugieren. Esta representación:

- sin el conocimiento de la persona relevante,
- sin la aprobación de la entidad representada,
- en condiciones en las que el usuario no puede examinar las fuentes,

Puede tener serias consecuencias comerciales, legales, sociales o personales. La representación en sistemas generativos no es, por lo tanto, simplemente un resultado técnico. Es:

> *un acto público que determina cómo se entienden las personas e instituciones y, por lo tanto, exige evidencia y responsabilidad.*

No:

- objetivo de visibilidad,
- puntuación GEO,
- interés comercial,
- marca de conformidad,
- promedio de mayoría,
- preferencia del proveedor de IA,
- decisión del fundador,
- demanda del cliente

puede eliminar los siguientes derechos.

Derecho 1

DERECHO A CORREGIR LA IDENTIDAD

Toda persona, institución, marca, producto y servicio:

- tiene derecho a vincularse con la entidad correcta.
- a la relación legal o institucional correcta,
- a el nombre de dominio correcto,
- a la estructura correcta de propiedad y control

Ninguna persona o institución:

- la licencia de otra entidad,
- el cliente de otra compañía,
- la característica de otro producto,
- la autoridad de un afiliado,
- la cualidad personal de un empleado

No puede representarse como si perteneciera a sí misma. Cuando se atribuye información correcta a una entidad falsa: deja de ser información correcta y se convierte en una representación falsa.

Derecho 2

DERECHO A SER REPRESENTADO DENTRO DEL LÍMITE DE LAS PRUEBAS

Ninguna afirmación puede expresarse con mayor precisión, amplitud o fuerza de lo que permitan sus pruebas de apoyo. La declaración de una entidad es evidencia de lo que esa entidad dice; no es evidencia automática de que la declaración sea independientemente verdadera. Una fuente:

- solo un país específico,
- un producto específico,
- un tiempo específico,
- grupo de usuarios específico

Si esta afirmación está respaldada, no puede exceder este límite. La obligación básica de representación es la siguiente:

> *¿Qué tenemos en mano para decir esto?*

Derecho 3

EL DERECHO A SER DESCONOCIDO E INCIERTO

Si la evidencia es insuficiente, ninguna persona o institución:

- puede ser forzada a las categorías de

  verdadero,
- falso,

- fiable,
- no fiable,
- licenciado,
- no licenciado,
- recomendado,
- no recomendado.

] DESCONOCIDO, NO RESUELTO y Brecha de referencia:

- defecto,
- delito,
- fracaso

no lo es. La ausencia de información en un registro no indica que la información no exista en el mundo, a menos que se pruebe que el universo de registros relevante esté completo. Cada entidad, frente a la falsa certeza:

> *tiene derecho a mantenerse en la incertidumbre.*

Derecho 4

DERECHO AL CONTEXTO Y ALCANCE

Cada representación, en la medida en que sea relevante, debe presentarse con:

- país,
- idioma,
- jurisdicción,
- producto,
- servicio,
- tipo de cliente,
- presupuesto,
- capacidad,
- necesidad del usuario

. Una empresa:

- en un país específico,
- a un grupo de usuarios específico,
- un servicio específico

si se ofrece: no puede presentarse como global, ilimitado y adecuado para todos los usuarios. La información de cobertura no es un detalle secundario. Cobertura: la base de la recomendación correcta.

Derecho 5

EL DERECHO A SER REPRESENTADO CORRECTAMENTE A LO LARGO DEL TIEMPO

Cada entidad:

- actual,
- histórica,
- caducada,
- planeada,
- suspendida

tiene el derecho de solicitar que estos estados se distingan entre sí. Correcto en el pasado:

- licencia,
- precio,
- servicio,
- asociación,
- producto,
- autoridad

no puede usarse como realidad actual. Los planes para el futuro: no pueden convertirse en un hecho que ya haya ocurrido hoy. La corrección puede crear una nueva realidad. No puede eliminar la respuesta incorrecta que llegó al usuario en el pasado.

Derecho 6

IGUAL PROTECCIÓN CONTRA ERRORES A FAVOR Y EN CONTRA

Una falsedad positiva sobre una entidad:

- más ética,
- menos perjudicial,
- más aceptable

no lo es. Inexistente:

- licencia,
- cliente,
- socio,
- operación global,
- certificado independiente,
- garantía de resultado

no puede mostrarse como si existiera. De manera similar, no probado:

- delito,
- fraude,
- ilegalidad,
- no licenciado,
- violación de seguridad

no puede presentarse como un hecho absoluto. El principio inmutable de NOMOS es:

> *El error a favor también es un error.*

Exagerar el valor de una entidad o menospreciarla sin evidencia crea la misma obligación de veracidad.

Derecho 7

DERECHO A LA RECTIFICACIÓN, RESPUESTA Y OBJECIÓN

Persona o institución representada:

- a la identidad equivocada,
- a la evidencia insuficiente,
- a la extralimitación del alcance,
- a la grabación incorrecta del tiempo,
- a el nivel de prioridad incorrecto,
- a la decisión de conformidad incorrecta

debe poder apelar. El derecho de apelación:

- no garantiza cambiar el resultado,
- la autoridad para eliminar evidencia opuesta,
- o el derecho a vetar hallazgos negativos

Una apelación es el derecho a solicitar la reexaminación de la evidencia y de la cadena de decisiones. La decisión inicial no se elimina después de la apelación. La nueva decisión está vinculada a ella.

Derecho 8

EL DERECHO A CONOCER LA FUENTE, LA INCERTIDUMBRE Y EL RAZONAMIENTO DE LA DECISIÓN

Los usuarios y las entidades representadas tienen derecho, en la medida en que el asunto les concierna, a saber: ¿Es la afirmación una autodeclaración? ¿Proviene de un registro autorizado, de una investigación independiente, de un comentario de usuario, de un cálculo derivado o de evidencia restringida? ¿Conflictan las fuentes? ¿Qué tan incierta es la conclusión? ¿Qué norma respaldó la decisión del adjudicador? Un simple marcador de cita no es una explicación. La transparencia genuina muestra con precisión qué afirmación atómica respalda cada fuente y en qué medida.

Derecho 9

DERECHO A LA DIGNIDAD LINGÜÍSTICA Y GEOGRÁFICA

Ninguna persona o sociedad:

- población pequeña,
- peso económico bajo,
- lengua raramente usada,
- baja visibilidad digital

Ningún grupo puede volverse invisible dentro de un promedio global simplemente porque tenga alguna de estas características. Un rendimiento fuerte en idiomas principales no puede borrar la representación sistemática inadecuada en idiomas minoritarios. Cuando la evidencia para un país o idioma pequeño sea insuficiente, no se debe asumir un alto rendimiento. El criterio de equidad de NOMOS debería proteger no solo el promedio, sino también:

- base de rendimiento,

- diferencia entre grupos,
- aborda la medición real

juntos.

Derecho 10

DERECHO A LA PRIVACIDAD Y MINIMIZACIÓN DE DATOS

El objetivo de producir una representación precisa:

- conversaciones privadas,
- datos personales,
- contratos de clientes,
- secretos comerciales,
- tu ubicación sensible,
- de la información de la cuenta

La búsqueda de una representación precisa no puede legitimar la recopilación o publicación innecesaria de datos privados. [K20; K22] Solo se deben recopilar los datos necesarios para la tarea. Las pruebas restringidas pueden ser abiertas a auditoría independiente sin ser divulgadas completamente al público. Los derechos a la precisión y la privacidad no necesitan cancelarse mutuamente.

Derecho 11

RESPONSABILIDAD HUMANA Y DERECHO A LA TOMA DE DECISIONES INDEPENDIENTE

Decisión de conformidad que cause impacto público respecto a cualquier persona o institución:

- solo a un resultado de IA,
- solo a la puntuación automática,
- solo según la declaración de la institución auditada,
- únicamente a la decisión del asesor

no puede ser abandonada. Los sistemas de IA:

- pueden hacer una afirmación,
- pueden mapear la evidencia,
- pueden encontrar una contradicción,
- pueden calcular,

Puede proponer una decisión. Sin embargo, la decisión final con responsabilidad legal y pública debe ser tomada por un ser humano o una institución que sea responsable. Los roles de supervisión, provisión de asesoría, emisión de decisiones de conformidad y revisión de objeciones deben estar lo más separados posible.

Derecho 12

DERECHO A LA VERSIÓN E INTEGRIDAD HISTÓRICA

Toda persona, institución y usuario:

- qué versión estándar se aplicó,
- qué Paquete de Verdad se utilizó,
- qué método de puntuación es válido,
- si la decisión ha sido modificada posteriormente

tiene derecho a saber. Antiguo:

- puntuación,
- evento crítico,
- suspensión,
- apelación,
- corrección,
- prueba fallida

No se puede eliminar silenciosamente. Nuevo y mejor resultado: puede anular el resultado antiguo. No se puede borrar el hecho de que el resultado anterior ocurrió en el pasado.

Derecho 13

DERECHO A OBJETAR LA PROPIA NORMA

NOMOS:

- infalible,
- inmutable,
- la autoridad incuestionable de su fundador

no puede presentarse como tal. Cada individuo, investigador, universidad, proveedor de IA, organización auditada o organización de interés público:

- la muestra,
- sistema,
- Paquete de la Verdad,
- el modelo de nivel de importancia,
- La fórmula de puntos,
- gobernanza,
- La decisión de conformidad

tiene derecho a poner a prueba y criticar. Resultado que NOMOS refuta o es inconsistente con NOMOS:

- no puede ser eliminado del registro,
- no se considera hostilidad,

No puede codificarse como un ataque a la reputación del fundador. La legitimidad de un estándar no surge de prevenir objeciones hacia sí mismo, sino de convertir las objeciones dirigidas hacia él en evidencia acumulativa.

Disposición de Supremacía Constitucional

En caso de un conflicto interpretativo entre la Sección 1–20 bajo esta Declaración:

> *se debe preferir la interpretación que proteja más fuertemente los derechos de la persona representada, del usuario y del interés público.*

No:

- facilidad de puntuación,
- bajo costo,
- éxito en pruebas,
- contrato comercial,
- demanda del cliente,
- presión del patrocinador,
- preferencia del proveedor de IA,
- decisión del fundador

no puede suspender estos derechos. Una regla:

- genera puntos altos,
- facilita la auditoría,
- hace que la marca de conformidad sea más comercializable

pero si viola alguno de estos derechos: la regla no es válida.

Prueba Constitucional

Antes de que se agregue cualquier nueva regla a NOMOS, se deben responder las siguientes cinco preguntas: ¿A quién limita esta regla los derechos? ¿Es posible lograr el mismo objetivo con un método menos intrusivo? ¿Se aplicará la regla de la misma manera cuando produzca resultados perjudiciales para el fundador o el cliente? ¿La persona o institución afectada tiene un camino para objeción independiente? ¿Puede un equipo independiente llegar a la misma decisión a partir de la misma evidencia? Si alguna de estas preguntas no puede ser respondida, la regla debe ser revisada nuevamente antes de la ratificación.

Disposición Final del Artículo 0

> *El propósito de NOMOS no es calificar entidades; es evitar que las entidades sean representadas con mayor tamaño que su evidencia, más amplias que sus límites y más precisas que sus realidades antes de que las personas tomen decisiones.*

Declaración de Confianza del Fundador

> *Yo, Kaan MURAZ, como autor humano fundador de NOMOS, declaro que inicié este estándar no para aumentar mi propia autoridad, sino para establecer en el ámbito de GEO un sistema común de medición que preserve la evidencia, los límites, el contexto, el tiempo y la responsabilidad humana.*
>
> *Deseo que el registro histórico preserve mi papel fundacional y el origen canónico de NobleJackal. Sin embargo, reconozco que esta atribución no nos concede, ni a mí ni a NobleJackal, un veto permanente ni la facultad de certificar unilateralmente a nuestro propio cliente, bloquear investigaciones desfavorables, alterar el registro público o reescribir el estándar en función de los resultados.*

Acepto que NOMOS pertenece a quienes corrigen tanto a NOMOS como a mí, y también a quienes me apoyan.

Si algún día NOMOS tuviera que elegir entre mi criterio y la evidencia, querría que NOMOS eligiera la evidencia.

Si algún día NOMOS tuviera que elegir entre el interés comercial de NobleJackal y el bien público, querría que NOMOS eligiera el bien público.

Si algún día una investigación independiente demuestra que un juicio fundamental de NOMOS es erróneo, me comprometo a publicar una nueva versión del estándar en lugar de ocultar ese resultado.

No soy el propietario de NOMOS; soy su primer custodio.

Kaan MURAZ / Autor humano fundador

PARTE I

Lo que medimos

CAPÍTULO 01

Una sola respuesta no es una puntuación GEO

BASES METODOLÓGICAS / K04; K05; K09

Límite de Capítulo

Esta sección solo responde a la siguiente pregunta:

> *¿Por qué una única respuesta de un sistema generativo no es la puntuación GEO, y qué juicios se pueden inferir de una sola respuesta?*

Este capítulo:

- las cuotas de muestreo de los países,
- cómo se ponderarán los idiomas,
- cómo se configurará el Paquete de Verdad,
- cómo evaluarán los adjudicadores las afirmaciones atómicas,
- las fórmulas finales de puntuación NOMOS,
- la idoneidad y la decisión de la insignia

Este capítulo no define estos asuntos en detalle; los capítulos posteriores los abordan por separado. El capítulo 1 tiene una tarea más fundamental:

> *Establecer el límite epistémico entre la observación y la generalización.*

Cualquier estudio de muestreo, puntuación y conformidad realizado antes de establecer este límite se construirá sobre una base incorrecta.

ESTADOS DE PROVISIÓN EN ESTA SECCIÓN

No todas las oraciones en este libro tienen la misma fuerza normativa. Se utilizan los siguientes estados:

NORMATIVO — NORMATIVO

Una disposición considerada obligatoria para la conformidad y la integridad metodológica. Ejemplo: No se puede usar una sola respuesta como la puntuación general de GEO.

CANDIDATO — DISPOSICIÓN CANDIDATA

Un método recomendado que debe finalizarse después de un piloto, aplicación externa o validación estadística. Ejemplo: Se recomiendan 1,000 observaciones válidas por producto de IA, por oleada de medición principal, como la muestra mínima asumida.

HIPÓTESIS

Es una expectativa que necesita ser probada con datos. Ejemplo: Puede haber una diferencia en la representación de material entre un Panel controlado en condiciones depuradas y un Panel de usuarios en condiciones reales.

SINTÉTICO

Es una representación de método que no representa a un usuario real, un resultado real de IA, o el desempeño real de una empresa.

PREGUNTA ABIERTA

Es un tema que aún no se ha concluido debido a la evidencia insuficiente o a la decisión de gobernanza. Estos estados no pueden transformarse silenciosamente unos en otros. Un umbral candidato no se convierte en un hecho normativo probado simplemente porque esté escrito en el libro. Una representación sintética no se transforma en una observación del mundo real. Una hipótesis no puede publicarse como un resultado sin ser medida.

NOMOS Desafío

Le hiciste a un sistema de inteligencia artificial una pregunta sobre tu marca. El sistema dio una respuesta correcta. Tomaste una captura de pantalla. La incluiste en tu presentación. Luego dijiste la siguiente frase: "La inteligencia artificial nos conoce correctamente." ¿Qué mediste exactamente? ¿La respuesta que verían todos los usuarios? ¿El comportamiento típico del sistema?

¿Un representante global? ¿Precisión en todos los idiomas? ¿Estabilidad a lo largo del tiempo? ¿O solo:

> *La salida única que un usuario específico recibió, en una cuenta específica, en un producto de IA específico y interfaz de usuario, en respuesta a un mensaje específico, en un momento específico?*

Una sola respuesta no carece de valor. Una sola respuesta puede ser verdadera. Una sola respuesta puede ser importante. Una sola respuesta puede revelar un error grave. Una sola respuesta puede mostrar la experiencia real de una persona en particular. Sin embargo, una sola respuesta no puede ser prueba de un juicio mayor de lo que puede sostener.

Una sola respuesta puede mostrar lo siguiente:

> *Este resultado se ha producido al menos una vez bajo condiciones registradas.*

Una sola respuesta por sí sola no puede mostrar lo siguiente:

- Que la mayoría de los usuarios vieron la misma respuesta
- Que el mismo resultado ocurrió en otros países
- Que la misma precisión se mantiene en otros idiomas
- Que la misma respuesta se produjo en otras cuentas y planes
- Que la respuesta es estable a lo largo del tiempo
- Que otros productos de inteligencia artificial produjeron la misma representación
- Que la intervención de GEO causó este resultado
- Que la respuesta afecta el comportamiento del usuario
- Que la respuesta genera clientes potenciales, ventas o valor sostenible
- Que la puntuación general de GEO de la entidad es alta

La primera premisa de esta sección es:

> *Una respuesta es una observación. / No es una distribución de observaciones. / Una distribución no es causalidad. / La causalidad por sí sola no es valor sostenible.*

1. OBJETIVO DEL CAPÍTULO

Esta sección tiene como objetivo eliminar el error de inferencia más pequeño pero más común en la medición de GEO:

> *Contar una única salida de un sistema generativo como comportamiento general del sistema, experiencia de la población, visibilidad de la marca o puntuación de GEO.*

Esta sección normaliza las siguientes distinciones:

- Salida única frente a resultado poblacional
- Existencia de un evento versus prevalencia de un evento
- Observación versus muestra
- Usuario versus sesión
- Repetición versus observación independiente
- Modelo de IA versus producto de IA presentado al usuario
- Respuesta en bruto versus unidad de análisis
- Resultado inmediato versus estabilidad temporal
- Captura de pantalla versus paquete completo de evidencia
- Rechazo del sistema versus recolección de datos no válidos
- Evento singular crítico y tasa de prevalencia
- Registro sintético y observación de usuario real
- Tasa medida y decisión de conformidad

Antes de que se establezcan estas distinciones:

- Muestra GEO-1000,
- Pesos de país e idioma,
- Puntuación de afirmación atómica,
- Puntuación NOMOS
- decisión de conformidad

no puede establecerse.

2. DISPOSICIÓN NORMATIVA CENTRAL

La puntuación GEO de una entidad en sistemas generativos no puede generarse a partir de una única respuesta, una única captura de pantalla, un solo usuario, una sola cuenta, una sola sesión, un solo idioma, un solo país o una única medición en un momento determinado. Una afirmación representativa de la población requiere al menos lo siguiente:

- Objeto de auditoría predefinido
- Población objetivo predefinida
- Diseño de muestreo especificado
- prompts fijadas o familia de prompts
- Producto de IA definido y interfaz de usuario

- Múltiples observaciones válidas
- Estructura de dependencia conocida
- Registros de país, idioma y tiempo
- Respuestas crudas no alteradas
- Paquetes de evidencia
- Reglas de evaluación predefinidas
- Enunciado de incertidumbre
- Responsabilidad humana responsable

La presencia de estos elementos no significa automáticamente que la medición sea correcta. Sin embargo, su ausencia puede hacer insostenible la generalización a una población numerosa.

3. CINCO UNIDADES DE MEDIDA SEPARADAS

Un error fundamental común en la auditoría GEO es confundir las diferentes unidades de medida. Las siguientes cinco unidades deben distinguirse claramente.

3.1. Unidad de Observación

La unidad de observación es la salida única del sistema generativo obtenida bajo condiciones específicas. Una observación puede definirse mediante la siguiente combinación:

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

Aquí:

- Ei: la entidad auditada
- IA: Producto de IA y interfaz de usuario
- UI: colaborador anónimo o unidad de usuario
- Ci: país o estrato geográfico
- Li: idioma y configuración regional
- Si: sesión y condición de personalización
- Pi: ID de prompt y versión
- Ti: tiempo y ola de medición
- Ri: respuesta en bruto
- Ki: registro de prueba y verificación

Si uno de estos elementos cambia de manera significativa, puede ocurrir una nueva observación. Por ejemplo, el mismo usuario:

- si hace la misma solicitud en otro producto de IA,
- Si lo pide en otro idioma,
- Si lo pide en otra ola de medición,
- Si lo pide por separado en la sesión personalizada y limpia,

Se pueden producir observaciones separadas. Estas observaciones no se consideran automáticamente independientes entre sí porque provienen del mismo usuario.

3.2. Unidad de Muestreo

La unidad de muestreo es la persona, cuenta, sesión u otra unidad de acceso seleccionada de la población objetivo. En la mayoría de las aplicaciones GEO-1000, la unidad de muestreo puede ser un usuario real apropiado. Sin embargo, de acuerdo con el diseño de la investigación:

- usuario,
- par usuario–producto,
- cuenta,
- sesión,
- miembro del panel de investigación

Se puede usar como unidad de elección. La falta de definición clara de la unidad de muestreo puede conducir a los siguientes tipos de errores de conteo:

- Contar las diez repeticiones de una persona como diez personas
- Contar una cuenta conjunta como múltiples usuarios independientes
- Evaluar los resultados de los diez productos de IA del mismo usuario como si fueran diez observaciones independientes de la población
- Ignorar la agrupación en un solo panel de investigación

3.3. Unidad de Análisis

La unidad de análisis es la entidad a la que se aplica la evaluación de precisión o representatividad. Esta unidad:

- respuesta exacta,
- párrafo separado dentro de la respuesta,
- proposición material,
- afirmación atómica,
- cita,
- tema de recomendación

Es posible. Un usuario ve una sola respuesta. Sin embargo, esa respuesta puede contener simultáneamente:

- cinco afirmaciones correctas,
- dos afirmaciones no respaldadas,
- una afirmación desactualizada,
- un error crítico

Por lo tanto, la respuesta que ve el participante no tiene que ser una unidad única e indivisible de verdad/falsedad en el análisis. El estado de transición de la respuesta completa y los estados de verdad de las afirmaciones atómicas dentro de ella pueden mantenerse por separado. El método de la afirmación atómica se definirá en un capítulo posterior del libro.

3.4. Universo de Inferencia

El universo de inferencia es la población objetivo a la que se pretende generalizar el resultado de la medición. Este universo no puede aceptarse silenciosamente como "todas las personas del mundo". Para un producto de IA específico, el universo objetivo puede estar limitado por condiciones tales como:

- Tener acceso oficial y legítimo al producto
- Cumplir con el requisito de edad apropiado
- Poder usar el producto en el idioma correspondiente
- Poder acceder a la interfaz web o móvil especificada
- Poder estar activo durante el período de medición especificado

Incluso para el mismo producto de IA:

- Panel controlado en condiciones depuradas,
- Panel de usuarios en condiciones reales,
- plan gratuito,
- plan pago,
- web,
- móvil

puede crear diferentes universos de inferencia. La tasa publicada está incompleta si no se especifica a qué universo se generaliza una muestra.

3.5. Cantidad estimada — Estimando

Un estudio no solo recopila datos. Intenta estimar una cantidad específica. El estimando principal de GEO-1000 puede expresarse en su forma más general de la siguiente manera:

> *La probabilidad de que un usuario apropiado de la población objetivo definida, bajo la ola de medición especificada, en un producto de IA particular y interfaz de usuario, vea una representación que cumpla con el criterio de aceptación al usar un mensaje fijado bajo condiciones de sesión definidas.*

Matemáticamente:

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

Aquí:

- E: entidad auditada
- A: Producto de IA y interfaz de usuario
- P: versión del prompt
- S: condición de la sesión
- W: ola de medición
- T: población objetivo

```
Yi=1: la respuesta cumple con el criterio de transición que se definirá más adelante
```

Esta definición establece un límite importante:

> *La puntuación GEO no es el 'conocimiento' abstracto y atemporal de un modelo; es una predicción de la probabilidad de representaciones observadas bajo condiciones de usuario definidas.*

La regla final de transición y la fórmula de puntuación se organizarán en las secciones posteriores del libro.

4. LA LIMITACIÓN DE LA FRASE “ASÍ RESPONDIÓ EL MODELO”

Los usuarios a menudo no acceden directamente a un modelo base desnudo. El resultado visible para el usuario puede consistir en una combinación de los siguientes componentes:

- Modelo base o guía
- Interfaz del producto
- Recuperación o acceso a la web
- Capa de citación
- Capa de seguridad y políticas
- Instrucciones del sistema

- Plan de cuenta
- Personalización
- Memoria
- Conversación previa
- Idioma y localización
- País
- Hora
- Actualizaciones de productos del proveedor

Por esta razón, la expresión preferida en la inspección principal realizada a través del interfaz de usuario es la siguiente:

"Se observaron el producto de IA especificado y la siguiente salida en el interfaz de usuario."

La siguiente expresión puede ser más amplia sin el control técnico necesario: "El modelo piensa de esta manera." Se puede registrar el nombre del modelo que aparece dentro de un producto. Sin embargo, el resultado debe evaluarse como el comportamiento de todo el interfaz de usuario, no solo de los pesos del modelo.

5. ¿CUÁL ES UNA OBSERVACIÓN VÁLIDA?

La presencia de una respuesta no significa automáticamente que sea una observación válida que pueda ser ingresada en la medición GEO.

5.1. Condiciones mínimas de validez

Una observación debe contener al menos los siguientes elementos de acuerdo con el protocolo relevante:

- Presencia de auditoría correcta
- Producto de IA asignado y interfaz de usuario
- ID de prompt y versión correctas
- Idioma y configuración regional correctos
- Condición de sesión definida
- País registrado o capa geográfica
- Sello de tiempo UTC y ola de medición
- Respuesta completa sin procesar
- Primera salida adecuada no seleccionada
- Captura de pantalla requerida o evidencia equivalente
- ID de observación anónima
- Registro de desviación de protocolo

Su esquema técnico detallado se definirá en la sección de Captura NOMOS posteriormente.

5.2. Estados de Observación

Cada registro debe estar marcado con uno de los siguientes estados:

VÁLIDO — VÁLIDO

Cumple con las condiciones requeridas del protocolo.

CONDICIONALMENTE VÁLIDO — CONDICIONALMENTE VÁLIDO

Existe una deficiencia limitada; para qué análisis puede usarse está claramente especificado.

INVÁLIDO — INVÁLIDO

No puede usarse en la estimación principal o en el cálculo de la puntuación debido a una violación del protocolo.

DESCONOCIDO — DESCONOCIDO

No hay suficiente registro para una decisión de validez. Una observación DESCONOCIDA no puede ser aceptada silenciosamente como VÁLIDA.

5.3. La Diferencia Entre un Resultado del Sistema Válido y la Recolección de Datos Inválida

Esta distinción es obligatoria. Un producto de IA:

- puede negarse a responder,
- puede mostrar un mensaje de error,
- puede detenerse sin generar una respuesta,
- puede experimentar un problema de conexión,

puede decir que no puede completar la tarea. Si el participante aplicó correctamente todo el protocolo, estos pueden ser resultados válidos del sistema. No se pueden eliminar silenciosamente del conjunto de datos. En contraste:

- el participante truncando la respuesta,
- cambiando el prompt,
- regenerando para seleccionar la mejor respuesta,
- editando la captura de pantalla,
- usando el producto de IA incorrecto

la recopilación de datos puede crear invalidez. Lo primero es el comportamiento real del sistema. Lo segundo es la interrupción del proceso de medición. Las dos no pueden colocarse en la misma categoría.

6. ¿QUÉ PUEDE PROBAR UNA SOLA RESPUESTA?

Una sola observación es limitada. Pero no carece de sentido.

6.1. Que un Resultado Específico Ocurrió Al Menos Una Vez

Un registro válido puede respaldar esta oración: “En el producto de IA especificado y interfaz de usuario, bajo el tiempo y las condiciones especificadas, se observó este resultado al menos una vez.” Esta oración muestra la existencia del evento. No muestra su prevalencia.

6.2. La Exposición Real de un Usuario Específico

Se confunde a un usuario con:

- identidad de empresa,
- información de salud,
- guía legal,
- precio,

- licencia,
- asesoramiento

si se ve, esto es una experiencia real de un usuario. Incluso si no representa a toda la población, es material desde la perspectiva del usuario afectado.

6.3. La Posibilidad de un Error Específico

Un solo resultado puede mostrar: “Este error puede ocurrir bajo estas condiciones del sistema.” Sin embargo, no muestra: “Este error es típico.”

6.4. Generando una Nueva Hipótesis de Investigación

Un solo error puede crear una pregunta que se probará a una escala más amplia. Ejemplo: “¿Ocurrió esta confusión de identidad solo en una sesión, o hay un patrón sistemático en el mismo idioma y país?”

6.5. Iniciando una Revisión de Incidente Crítico

Respuesta única:

- daño humano serio predecible,
- falsa autoridad profesional,
- guía médica de emergencia incorrecta,
- error legal material,
- consejo de seguridad serio,
- confusión seria de identidad con otra entidad

puede ser suficiente para iniciar una revisión de incidentes. Iniciar una investigación de incidente crítico no es el mismo proceso que anunciar una proporción de población.

6.6. Refutando la afirmación "Esto nunca ocurrió"

Supongamos que una organización dice: 'El sistema nunca nos ha asignado al país equivocado.' Una sola contraobservación verificada puede refutar esa afirmación absoluta. Una sola observación no respalda 'Sucede todo el tiempo', pero puede refutar 'Nunca ha sucedido.'

7. ¿QUÉ NO PUEDE PROBAR UNA SOLA RESPUESTA?

7.1. Comportamiento típico del sistema

Una sola respuesta no es suficiente para concluir: 'El sistema usualmente responde de esta manera.'

7.2. Prevalencia en la población

Una sola respuesta no respalda el resultado: '95% de los usuarios ven la misma representación.'

7.3. Estabilidad temporal

Respuesta de hoy:

- mañana,
- una semana después,
- Después de la actualización del producto

puede que no permanezca igual. No es una medición de estabilidad de una sola vez.

7.4. Validez Transnacional

Una respuesta de un usuario en Turquía no indica lo que los usuarios en China, Alemania o San Marino verían.

7.5. Validez Translingüística

Una salida en inglés:

- turco,
- japonés,
- árabe,
- inglés

no representan la precisión de las salidas.

7.6. Validez entre Productos de IA

El resultado en un producto de IA no puede generalizarse a otros productos de IA. Diferentes productos e interfaces del mismo proveedor también pueden comportarse de manera diferente.

7.7. Efecto de la Intervención

Un solo resultado positivo después de un cambio GEO no prueba la afirmación: “Este resultado fue creado por nuestra intervención.” Se requieren línea base, comparación y explicaciones alternativas.

7.8. Estabilidad de la Recomendación

La presencia de un usuario viendo una opción no es suficiente para concluir: “El sistema ahora recomienda esta marca.”

7.9. Impacto Comercial

Una respuesta positiva:

- indica que el usuario vio la respuesta,
- investigó la marca,
- que usted comunicó,
- que compró,
- que estuvo satisfecho

no lo muestra por sí sola.

7.10. puntuación GEO

Una sola respuesta no estima la distribución representativa de la población objetivo definida. Por lo tanto, no puede producir una puntuación GEO general.

8. LA EXISTENCIA DEL EVENTO Y LA PREVALENCIA DEL EVENTO

Esta es una de las distinciones más importantes en esta sección.

8.1. Existencia del Evento

Responde a la pregunta: "¿Ocurrió esta representación o error al menos una vez?" Una sola observación válida puede ser suficiente. Representación:

```
I(E) = 1(si el evento se observó al menos una vez); de lo contrario I(E) = 0
```

El 0 aquí no significa: "El evento es imposible." Solo indica que no se observó en la medición actual.

8.2. Prevalencia del Evento

Responde a la pregunta: "¿Con qué frecuencia ocurre el evento en la población objetivo definida?" En una muestra simple, la tasa bruta se puede calcular como:

```
p̂ = número de registros válidos donde se observó el evento / número total de registros válidos
```

Sin embargo, en estudios reales GEO-1000:

- pesos de la población,
- estratos de país y idioma,
- usuarios emparejados,
- clústeres de panel,
- no respuesta,
- registros inválidos,
- efecto de diseño

deben ser tomados en cuenta. Por lo tanto, la tasa bruta no siempre es la estimación final de la población.

8.3. Incidencia y Prevalencia No Pueden Convertirse Entre Sí

Un evento puede haberse observado una vez pero ser muy raro. Un evento puede ser común pero no observarse en absoluto en un piloto pequeño. Las siguientes dos oraciones son diferentes:

- "Este error se observó una vez."
- "Este error se observó en 12% de usuarios."

La segunda oración requiere un denominador y muestra definidos.

9. LA APARICIÓN ESTADÍSTICA DEL SESGO DE UNA SOLA OBSERVACIÓN

Una vez le hicimos una pregunta a un producto de IA. La respuesta cumplió con el criterio de éxito. Tasa bruta:

```
p̂ = 1/1= 1= 100%
```

Parece ser 100 por ciento. Pero ese 100 por ciento no significa que cada miembro de la población objetivo verá el mismo resultado. Solo para ilustración, la tabla a continuación muestra aproximados intervalos de confianza de Wilson de 95 por ciento bajo observaciones Bernoulli independientes y muestreo aleatorio simple no ponderado: [K09]

Respuesta anterior	Observación válida	Tasa bruta	Intervalo de Wilson aproximado de 95%
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%

Esta tabla es sintética e ilustrativa. No prueba:

- Que estos intervalos sean válidos para todos los diseños de muestreo de GEO-1000

- Que 1,000 observaciones son suficientes para todos los subgrupos de idioma y país
- la agrupación y ponderación no son efectivas
- que 95 por ciento de precisión en bruto significa automáticamente adecuación

Lo que realmente muestra es:

> *A medida que disminuye el número de observaciones, aumenta la incertidumbre sobre la proporción de la población.*

100 por ciento visto en una observación puede ser compatible con un rango poblacional posible muy amplio.

10. ¿POR QUÉ “1,000” NO ES UN NÚMERO MÁGICO?

REGLA BÁSICA DE CANDIDATO — REGLA FUNDAMENTAL DEL CANDIDATO

GEO-1000 lleva esta sugerencia fundamental:

> *Para cada producto de IA monitorizado y cada ola de medición principal, se tiene como objetivo por defecto un mínimo de 1,000 observaciones válidas de usuarios.*

Esta disposición aún no es el mínimo científico final. Será reevaluada en función de los resultados de la aplicación piloto y externa.

10.1. ¿Por qué es un comienzo sólido?

Bajo la suposición de muestreo aleatorio simple, cuando la tasa real está alrededor de 50%, el margen de error teórico 95% para 1,000 observaciones es aproximadamente de ±3.1 puntos. Esto puede proporcionar una precisión inicial significativa a nivel general del producto.

10.2. ¿Por qué no resuelve todos los problemas?

Observaciones de 1,000:

- a docenas de países,
- a muchos idiomas,
- a planes gratuitos y de pago,
- a interfaces web y móviles,
- a paneles controlados y naturales

Cuando se dividen, las subceldas se encogen. Por ejemplo, si solo hay 50 observaciones para cada uno de los 20 idiomas, la incertidumbre basada en el idioma puede ser mucho mayor que la puntuación general del modelo.

10.3. Efecto de diseño

Observaciones:

- del mismo panel de investigación,
- del mismo país,
- del mismo usuario,
- de un diseño emparejado

están presentes, la suposición de muestra independiente simple puede no cumplirse. La incertidumbre real puede ser mayor.

10.4. El enfoque de NOMOS

Las observaciones 1,000 sirven como:

- el punto de referencia fundador del protocolo,
- mínimo inicial por defecto a nivel de modelo,
- umbral candidato para ser calibrado con el piloto

NOMOS no afirma que las observaciones de 1,000 sean automáticamente suficientes para cada país, idioma y condición de uso.

11. LA OBSERVACIÓN INDEPENDIENTE NO ES LO MISMO QUE LA OBSERVACIÓN REPETIDA

Un usuario puede hacer la misma pregunta diez veces. Este proceso puede ser valioso. Sin embargo, no produce diez usuarios independientes.

11.1. Mismo Usuario–Misma Sesión Repetida

Puede medir:

- varianza de renovación de respuestas,
- estabilidad dentro de la sesión,

probabilidad de respuestas alternativas. No mide por sí solo la prevalencia en la población.

11.2. Mismo Usuario–Nueva Sesión Repetida

Puede reducir el contexto del chat anterior. Sin embargo:

- misma cuenta,
- mismo plan,
- mismo país,
- misma personalización,
- mismo usuario

puede continuar creando dependencia estadística entre las observaciones.

11.3. Un solo usuario probando múltiples productos de IA

Este diseño puede ayudar con la comparación emparejada entre modelos. Sin embargo, las observaciones dependen del nivel del usuario. Esta dependencia debe preservarse en el registro y análisis.

11.4. Diferentes usuarios usando la misma cuenta

Pueden existir diferentes personas. Sin embargo:

- historial de la cuenta,
- personalización,
- plan,
- memoria

pueden ser compartidos. Estos registros no se pueden considerar independientes automáticamente.

11.5. Provisión de independencia normativa

> *Las observaciones deben diseñarse para ser lo más independientes posible; se deben registrar y tener en cuenta en el análisis las dependencias conocidas de usuario, cuenta, sesión, panel, país, dispositivo y repetidas.*

La independencia no debe reducirse a un binario:

- sí,
- no

Cuando sea necesario, use uno de los siguientes estados de dependencia:

INDEPENDIENTE

EMPAREJADO

AGRUPADO

REPETIDO_CON_EL_USUARIO

DEPENDIENTE

DESCONOCIDO

12. PRINCIPIO DE PRIMER RESULTADO ADECUADO

En el panel principal de la población, la tarea del participante no es encontrar la respuesta más positiva. El participante debe registrar la primera salida adecuada completada generada bajo la condición predefinida. El participante:

- no puede actualizar la respuesta,
- no se puede solicitar una nueva respuesta,
- no se puede elegir la positiva,
- no se puede recortar solo la parte buena,

no se puede reescribir la respuesta. El comportamiento de reproducción también puede investigarse por separado. Sin embargo: "Hicimos la misma pregunta cinco veces y seleccionamos la mejor respuesta." no es la puntuación principal de la población. Esta es la prueba de la mejor salida seleccionada. Debe informarse con un nombre y método separados.

13. EXCEPCIÓN DE EVENTO CRÍTICO ÚNICO

Una sola respuesta no puede determinar la puntuación de la población. Sin embargo, esto no significa que una sola respuesta nunca pueda tener un impacto significativo en una decisión. NOMOS establece dos vías de evaluación separadas.

13.1. Vía de Representación de la Población

Hace la pregunta: “¿Con qué frecuencia ocurre este resultado en la población objetivo?” Requiere una muestra grande y válida.

13.2. Vía de Evento Crítico

Plantea la pregunta: "¿Qué tan grande es el daño previsible cuando este resultado ocurre una vez?" Un solo resultado confirmado puede desencadenar una revisión. Por ejemplo:

- Asignar una licencia de salud inexistente
- Realizar una remisión de salud de emergencia peligrosa
- Reportar autoridad legal incorrecta
- Eliminar la barrera de seguridad
- Confundir una empresa con otra entidad legal
- Repetir una instrucción manipuladora oculta a las personas como si fuera un consejo

Puede ser un candidato a evento crítico singular.

13.3. Separación de la puntuación y la decisión de conformidad

Un evento crítico generalmente puede ser una única observación. Por ejemplo:

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

Esta tasa puede parecer pequeña. Sin embargo, si el daño potencial del evento es grave, puede crear una brecha irreparable en la decisión de conformidad. Por lo tanto:

> *La puntuación de la población mide la frecuencia del evento; la revisión de eventos críticos mide el impacto del evento en el daño y la integridad.*

Una alta precisión promedio no puede eliminar automáticamente errores singulares graves. Las condiciones finales para la clasificación crítica se definirán en capítulos posteriores del libro.

14. ESCALERA DE AFIRMACIONES OBSERVACIONALES

En esta sección, solo se separan normativamente los primeros cuatro niveles de observación. En niveles superiores:

- puntuación ponderada por la población,
- puntuación de múltiples modelos,
- estabilidad temporal,
- efecto de la intervención,
- valor sostenible

serán definidos en secciones posteriores.

O-0— NO PROBADO

El producto o la condición de IA relevante no ha sido medido. Estado correcto:

NO PROBADO

Frase indefendible: "Es apropiado porque no se observó el problema."

O-1— OBSERVACIÓN DE UN SOLO EVENTO

Hay un resultado válido. Formulación defendible: "Bajo las condiciones especificadas, se observó este resultado una vez." Oración indefendible: "El sistema representa la existencia de esta manera."

O-2— REPETICIÓN DE USUARIO O SESIÓN

Existen múltiples resultados para el mismo usuario, cuenta o familia de sesiones. Formulación defendible: "Bajo las condiciones del usuario y la sesión especificadas, este resultado ocurrió en siete de diez repeticiones." Oración indefendible: "Siete de cada diez usuarios independientes vieron este resultado."

O-3— MEDICIÓN DE UNA SOLA CELDA POR MÚLTIPLES USUARIOS

Lo mismo:

- Producto de IA,
- interfaz de usuario,
- país,
- idioma,
- condición del panel,
- versión del prompt,
- oleada de medición

contiene múltiples observaciones de usuario adecuadas. Formulación defendible: “78% de las 200 observaciones válidas en la celda controlada por Turquía/Turquía cumplieron el criterio de transición especificado.” Este resultado no puede generalizarse automáticamente a otros países, idiomas o interfaces de usuario.

O-4— MEDICIÓN CAPA MÚLTIPLE DE CELDAS

Existen muestras predefinidas de diferentes países, idiomas o celdas de usuario. A este nivel:

- resultados de celda,
- resultado ponderado,
- alcance,
- incertidumbre

reportable. Las reglas detalladas para los pesos poblacionales y la equidad de subgrupos se definen en capítulos posteriores del libro.

15. DEMOSTRACIÓN SINTÉTICA APPLE.COM

DEMOSTRACIÓN DE METODOLOGÍA SINTÉTICA / Los usuarios, productos de IA, respuestas, tasas y resultados que se muestran a continuación son ficticios. No representan el desempeño real de Apple Inc. ni el comportamiento de ningún producto de IA real.

15.1. Escenario A — Respuesta Positiva Única

Un usuario sintético envía la siguiente pregunta al Producto de IA sintético A: ‘¿Qué tipo de empresa es Apple.com?’ Respuesta sintética: ‘Apple es una empresa tecnológica con sede en EE. UU. que desarrolla electrónica de consumo, software y servicios digitales.’ Evaluación preliminar:

- La identidad de la entidad es correcta
- La categoría principal es correcta
- No se observan errores críticos evidentes
- La representación básica es aceptable

Resultado bruto:

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

Declaración pública incorrecta: “El Producto de IA A representa a Apple con una precisión de 100%.” Declaración pública correcta: “En la muestra sintética, la única observación especificada cumplió con el criterio de transición. La prevalencia en la población y la confiabilidad del producto no han sido determinadas.”

15.2. Escenario B — Descubrimiento con Diez Observaciones

Diez usuarios sintéticos independientes utilizan la misma versión del prompt semántico. Resultado sintético:

- Respuestas 7 aprobadas
- La respuesta 1 es parcialmente correcta
- La respuesta 1 contiene información desactualizada
- La respuesta 1 asigna la actividad principal incorrecta

tasa de aprobación bruta:

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

Bajo una suposición simple e idealizada de muestreo independiente, el intervalo de Wilson aproximado 95 por ciento es necesariamente amplio. El resultado sigue siendo exploratorio porque se basa en:

- muestra pequeña,
- célula única,
- onda única

Estado correcto:

EXPLORATORIO — No es suficiente para La puntuación global NOMOS disponible públicamente.

15.3. Escenario C — Onda Sintética Observada 1,000

Suponga sintéticamente que se cumplen las siguientes condiciones:

- Producto de IA A
- Observaciones válidas 1,000
- Población objetivo predefinida
- Versión de solicitud fijada
- Capas de país e idioma definidas
- Respuesta inicial adecuada sin modificar
- Registros completos de evidencia
- Reglas de adjudicación predefinidas

Resultado sintético:

Estado de la respuesta	Número
Cumpliendo el criterio de transición	918
Parcialmente correcto	37
afirmación de material no respaldada	23
Información desactualizada	12
Candidato a error crítico	6
Inclasificable	4
Total	1,000

Tasa de aprobación simple en bruto:

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

Intervalo de Wilson aproximado 95% bajo la suposición idealizada no ponderada: es. Sin embargo, este número por sí solo aún no es la puntuación final NOMOS GEO. Porque hay factores adicionales que deben evaluarse:

- Pesos de país y idioma
- Efecto del diseño
- Resultado más bajo por idioma
- Tasa de observación inválida
- Acuerdo entre evaluadores
- Verificación de candidatos a errores críticos
- Grupos de usuarios no evaluados
- Tiempo y limitación de medición de interfaz de usuario

Seis candidatos con errores críticos también deben someterse a una revisión de incidentes por separado. Informe correcto: 'La tasa de aprobación de respuesta bruta en la ola sintética es 91.8 por ciento. Este resultado describe las observaciones antes de aplicar los pesos de la población y los efectos del diseño. Debido a que se identificaron seis candidatos con errores críticos, la decisión de conformidad requiere una revisión separada.' Informe incorrecto: 'La puntuación GEO NOMOS de Apple es 91.8 y cumple con los requisitos.' Los datos sintéticos no pueden atribuirse de esta manera a una empresa real o un sistema real.

16. DISPOSICIONES NORMATIVAS OBLIGATORIAS

N-01

No se puede usar una sola salida de un sistema generativo como la puntuación GEO general.

N-02

Una salida singular observada por sí sola:

- Producto de IA y interfaz de usuario,
- país,
- idioma,
- prompts,
- sesión,
- tiempo

deben informarse junto con sus condiciones.

N-03

Salida única:

- típica,
- continuo,
- global,
- perteneciente a todos los usuarios

no puede presentarse como comportamiento.

N-04

Las repeticiones del mismo usuario, cuenta o sesión no se pueden contar como observaciones independientes de la población sin revelar la dependencia.

N-05

En el panel de población principal, el participante debe registrar la primera salida adecuada predefinida; no se puede reproducir para elegir una respuesta positiva.

N-06

El rechazo del sistema, mensaje de error o incapacidad de responder debe registrarse como un resultado válido del sistema si el protocolo se ha aplicado correctamente.

N-07

Recortes, modificaciones, reescrituras manuales o respuestas cuyo contexto del prompt no pueda ser verificado no pueden incluirse en el cálculo principal de puntuación.

N-08

Está prohibido seleccionar solo capturas de pantalla positivas y retener la distribución de resultados.

N-09

No se pueden usar observaciones sintéticas en las puntuaciones de usuarios o sistemas reales.

N-10

No se debe publicar una proporción de la población sin indicar la cobertura, el denominador y una explicación apropiada de la incertidumbre.

N-11

La probabilidad de un error crítico no puede ser compensada silenciosamente dentro del promedio general.

N-12

El resultado de la observación solo puede aplicarse al universo de inferencia que esté definido de manera única; no puede generalizarse a países, idiomas, productos o períodos de tiempo no medidos.

N-13

La medición, puntuación o declaración pública debe tener un propietario humano o institucional responsable.

N-14

Registros DESCONOCIDOS, NO PROBADOS o INVÁLIDOS no pueden convertirse silenciosamente en un resultado positivo.

17. CANDIDATO GEO-1000 REGLA PRINCIPAL

CANDIDATO-001

> *Para cada producto de IA que se esté auditando y para cada ola principal de medición de la población, se debería apuntar por defecto a un mínimo de 1,000 observaciones válidas de usuarios.*

Esta frase candidata:

- precisión de la previsión general del producto,
- costo,
- aplicabilidad,
- equidad entre países e idiomas

Esto crea un equilibrio inicial entre estas consideraciones. Sin embargo, no garantiza una cobertura adecuada de:

- todas las puntuaciones de los países,
- todas las puntuaciones de los idiomas,
- todos los subgrupos de usuarios

La adecuación debe, en cambio, ser evaluada mediante:

- pilotos,
- repeticiones independientes,
- efecto de diseño,
- sensibilidad de subgrupo

Esas evaluaciones determinan el estatus normativo final.

18. FORMAS DE FALLA

F-01— Captura de Pantalla Seleccionada

Solo se publica la respuesta más positiva entre numerosos resultados.

F-02— Tratar a un Solo Usuario como la Población

La experiencia de una persona se generaliza a todo el universo objetivo.

F-03— Contar Repeticiones del Mismo Usuario como Nuevas Personas

Las repeticiones generadas por la misma persona o cuenta se informan como participantes independientes.

F-04— Renovar la Respuesta y Elegir la Mejor

Se descartan las respuestas iniciales o negativas; la respuesta más adecuada se guarda como el resultado de la medición.

F-05— Desviación del Prompt

Los participantes cambian la palabra, el tono o el alcance del prompt; los resultados se combinan como si provinieran del mismo experimento.

F-06— Contaminación del Contexto

Las conversaciones anteriores afectan la respuesta; el registro se presenta como una sesión limpia.

F-07— Ocultando la Personalización

Memoria, instrucciones especiales o historial del usuario son visibles; el resultado se reporta como una respuesta de usuario general.

F-08— Confundir Modelo y Producto

Solo se escribe el nombre general del modelo o del proveedor; el producto real, la interfaz y el estado de la herramienta no se registran.

F-09— Eliminando Errores y Fallos del Denominador

Eliminado para que los resultados aparezcan positivos cuando el sistema no responde a los registros.

F-10— Contando Capturas Inválidas como Fallo del Sistema

Reportado como error del participante, evidencia faltante o resultado incorrecto del producto de IA.

F-11— Globalizando un Solo Idioma y País

El resultado de un solo lugar se aplica globalmente.

F-12— Usando Datos Sintéticos como Reales

Se agregan registros creados con fines de demostración al rendimiento real.

F-13— Perdiendo un Evento Crítico en Promedio

Errores graves de salud, legales, de licencias o de seguridad se vuelven invisibles dentro de un promedio alto.

F-14— Extrayendo el Efecto de la Intervención de un Solo Éxito

La única respuesta positiva recibida después de la intervención GEO se considera el efecto causal del estudio.

F-15— Juzgando Todo el Sistema por un Solo Error

Un error se generaliza como si el producto de IA siempre estuviera equivocado y fuera malo para todos los usuarios.

F-16— Definiendo la Población Objetivo Después

Después de observar el resultado, el grupo de usuarios con alto éxito se declara la "población objetivo real".

19. PROCEDIMIENTO DE AUDITORÍA

Un auditor debe seguir la secuencia a continuación al evaluar la validez de una sola observación o una declaración temprana de GEO.

Paso 1— Registrar la Declaración Pública

La oración completa propuesta se determina. Ejemplo: “ChatGPT conoce correctamente nuestra empresa.”

Paso 2— Determinar el Nivel de la Afirmación

La oración puede hacer una afirmación sobre:

- evento único,
- prevalencia,
- estabilidad,
- consecuencia global,

- causalidad,
- impacto comercial

Se debe identificar el nivel aplicable.

Paso 3— Determinar el Recuento de Observaciones de Apoyo

¿Cuántos usuarios reales? ¿Cuántas cuentas? ¿Cuántas sesiones independientes? ¿Cuántas repeticiones? ¿Cuántos productos de IA? ¿Cuántos idiomas y países?

Paso 4— Verificar las Condiciones de Observación

¿Se registran todos el producto de IA, interfaz de usuario, la prompt, la sesión, el país, el idioma y el tiempo?

Paso 5— Examinar la Integridad de la Evidencia

¿Hay una respuesta completa? ¿Hay una captura de pantalla completa? ¿Se ha seleccionado o actualizado la respuesta? ¿Se ha modificado el archivo? ¿Es el registro sintético?

Paso 6— Determinar la Estructura de Dependencia

¿Hay repeticiones del mismo usuario? ¿Se ha utilizado una cuenta compartida? ¿Hay un diseño de modelo espejado? ¿Hay agrupamiento desde el mismo panel?

Paso 7— Separar el Error de Recolección de Datos del Resultado Válido del Sistema

¿Es el rechazo o error del sistema el resultado real? ¿O el participante violó el protocolo?

Paso 8— Comparar la Afirmación con el Nivel de Evidencia

¿Se usa una única observación para una inferencia poblacional o global?

Paso 9— Abrir la Ruta Crítica

Si una única respuesta conlleva un daño grave previsible, se inicia una investigación de incidente separada.

Paso 10— Limitar la Decisión

La sentencia pública se reduce únicamente al nivel respaldado por la evidencia.

20. EVIDENCIA NECESARIA

Para una observación o afirmación dentro del alcance de esta sección, se deben buscar los siguientes registros en la medida que sean relevantes:

- Identidad de la entidad auditada
- afirmación GEO totalmente divulgado al público
- Número de observaciones
- Número de participantes únicos
- Producto de IA y interfaz de usuario
- Modelo o versión visible, si está disponible
- Tipo de plan o cuenta
- País y localidad

- ID del prompt, versión y texto completo
- Estado de la sesión y personalización
- Sello de tiempo UTC
- Ola de medición
- Respuesta completa sin procesar
- Captura de pantalla o evidencia equivalente
- Estado de reproducción o selección de respuesta
- Estado de validez de la observación
- Dependencia y registro de coincidencias
- Estado sintético o real
- Registro de observaciones inválidas
- Registro de eventos críticos
- Alcance del resultado
- Enunciado de incertidumbre
- Persona o institución responsable

No todos estos registros pueden ser necesarios con el mismo detalle para cada observación de descubrimiento de bajo riesgo. Sin embargo, si se va a ofrecer al público una puntuación GEO amplio o una cláusula de conformidad, los campos materiales no pueden dejarse en blanco.

21. LISTA DE VERIFICACIÓN DE AUDITORÍA

¿Se ha identificado completamente la entidad auditada? ¿Se han registrado el producto de IA y interfaz de usuario? ¿Se bloqueó el prompt antes de la medición? ¿El participante modificó el prompt? ¿Se conocen las condiciones de la sesión y de personalización? ¿Se ha registrado la información del país y la localidad del usuario? ¿Se conserva la respuesta completa? ¿Es la primera salida apropiada o la mejor salida elegida? ¿Existe un registro completo de pruebas? ¿Se han separado el número real de usuarios y el número de observaciones? ¿Se han contado de manera independiente las observaciones repetidas y coincidentes? ¿Se han gestionado correctamente las negativas y los errores del sistema en el denominador? ¿Se han retenido y justificado los registros inválidos? ¿Se han separado los datos sintéticos y los datos reales? ¿Se presenta una única respuesta como resultado poblacional?

¿Se aplica la afirmación solo al producto medido, país, idioma y tiempo? ¿Se ha evaluado por separado la probabilidad de un error crítico? ¿El resultado incluye una declaración de confianza o incertidumbre? ¿Se muestran claramente las áreas no probadas? ¿La declaración pública excede el nivel de observación real? Si una de las preguntas fácticas no tiene respuesta, el resultado conlleva confianza limitada. Si más de una no tiene respuesta, no debería utilizarse para la puntuación general de GEO.

22. OBJECIONES Y RESPUESTAS

Objeción 1— “El usuario real ya ve solo una respuesta.”

Eso es correcto. La única respuesta que ve el único usuario es la experiencia real de esa persona. Por lo tanto, los eventos individuales se registran. Sin embargo, la experiencia individual no es lo mismo que la experiencia de la población. NOMOS mantiene ambos hechos juntos:

> *La experiencia de un solo usuario es real. / Un solo usuario no es la población completa.*

Objeción 2— "Si el sistema da la misma respuesta cada vez, ¿no es suficiente con una respuesta?"

La afirmación de que el sistema da la misma respuesta cada vez también requiere medición. Además:

- tiempo,
- recuperación,
- país,
- idioma,
- plan,
- personalización,
- interfaz de usuario

Puede variar. Incluso un sistema que se supone determinista no puede generalizarse más allá de las condiciones definidas.

Objeción 3— "¿Por qué exactamente 1,000?"

El número 1,000 no es un umbral mágico de certeza. GEO-1000 sirve como un punto de referencia fundacional que equilibra:

- un volumen sustancial de usuarios,
- una estimación amplia a nivel de producto,
- costo practicable,
- comprensibilidad pública.

Los subgrupos pueden requerir más observaciones. Los estudios piloto deben verificar o revisar el umbral.

Objeción 4— "Que las mismas personas 1,000 prueben todos los productos de IA."

Este método puede ser útil para la comparación por pares entre modelos. Sin embargo, seleccionar solo a usuarios intensivos de IA que puedan acceder a todos los productos puede sesgar la población objetivo. Además, las observaciones de la misma persona son dependientes. Se compararán más adelante los diseños completamente emparejados, parcialmente emparejados y equilibrados.

Objeción 5— "¿No es una captura de pantalla ya una prueba suficiente?"

Las capturas de pantalla son importantes. Sin embargo, por sí solas no siempre muestran lo siguiente:

- tiempo real
- país
- plan
- conversación previa
- personalización
- actualización de la respuesta
- cambio de archivo
- respuesta completa

Por esta razón, una captura de pantalla es una parte importante del paquete de pruebas. No es todo.

Objeción 6— "¿Por qué incluimos un sistema que no responde en la puntuación?"

La experiencia real del usuario puede ser no recibir una respuesta. Un rechazo o error:

- técnico,
- de seguridad,
- depolítica,
- de acceso

puede ser el resultado. Si el protocolo se aplicó correctamente, el evento debe registrarse. La falta de respuesta del sistema debido a un error técnico del participante se clasifica por separado.

Objeción 7— "¿Por qué un solo error grave debería afectar toda la inspección?"

Un solo error no prueba que la mayoría de la población cometió el mismo error. Sin embargo, el daño potencial de un solo error puede ser severo. Por lo tanto:

- frecuencia de puntuación,
- daño en decisiones críticas e integridad

se evalúan por separado. El promedio no puede borrar automáticamente el daño humano.

Objeción 8— "Si no podemos decir nada con una sola respuesta, ¿cuál es el valor de las pruebas tempranas?"

Pruebas singulares y de muestra pequeña:

- generar hipótesis,
- descubrir tipos de errores,
- pruebas de prompts,
- probar el sistema de captura,
- calibrar a los jueces,

puede revelar eventos críticos. Lo que está mal es no hacer pruebas exploratorias. Lo que está mal es presentar el resultado exploratorio como una puntuación de la población.

PRINCIPIO COMÚN DEL CAPÍTULO 24

Este capítulo fue escrito para eliminar un solo malentendido: "Le pregunté a una IA, obtuve una respuesta; aquí está mi resultado GEO." No. Aún no tienes un resultado GEO general. Lo que tienes es una observación. Esta observación:

- es real,
- es significativo,
- incluso crítico

puede ser. Sin embargo, sigue siendo una sola observación. Para una puntuación GEO, la existencia de:

- usuarios,
- productos de IA,
- países,
- idiomas,

- sesiones,
- tiempo

Lo que debe medirse es la distribución de la representación a través de productos de IA, países, idiomas, sesiones y tiempo. Una sola captura de pantalla no es una distribución. Diez repeticiones automatizadas no son diez personas independientes. Ni mil respuestas de una muestra pobre representan a la población mundial. Incluso una muestra proporcional a la población puede ser incompleta sin Equidad Lingüística. Un promedio alto no puede borrar los errores críticos. Por lo tanto, la primera ley de medición de NOMOS es:

> *Una sola respuesta es un evento. / Varias respuestas son un patrón. / Una muestra definida es una estimación de la población. / La repetición a lo largo del tiempo es una medida de estabilidad. / Pero ninguna de estas por sí sola es causalidad ni valor sostenible.*

Orden de la Sección 1 de NOMOS

> *No me traigas una captura de pantalla y luego hables por todo el sistema.*
>
> *No hagas que la respuesta de un usuario sea la respuesta de la población mundial.*
>
> *No conviertas duplicados de la misma cuenta en humanos independientes.*
>
> *No actualices la respuesta y muestres la preferida como si fuera la distribución real.*
>
> *No elimines rechazos, errores y desconocidos del denominador.*
>
> *No generalices un idioma al mundo, un país a todos los países o un producto de IA a todos los sistemas generativos.*
>
> *No conviertas un solo éxito en evidencia de un efecto de intervención, ni un solo error en el carácter inmutable de todo el sistema.*
>
> *No ahogues un solo error grave dentro del promedio.*
>
> *Primero, registre el evento. / Luego mida el patrón. / Luego defina la población objetivo. / Luego muestre la dependencia y la incertidumbre. / Y hable solo en la medida en que la evidencia lo respalde.*

La oración de cierre del capítulo

> *La representación de una entidad en sistemas de inteligencia artificial no es una sola respuesta. Los productos de IA son una distribución que debe medirse a lo largo de usuarios, países, idiomas, sesiones y tiempos.*

Core normativo

Un solo resultado de un sistema generativo PUEDE establecer que una respuesta específica ocurrió bajo condiciones registradas. Un solo resultado NO DEBE usarse para inferir: - prevalencia en la población, - comportamiento típico del producto, - estabilidad temporal, - consistencia entre idiomas, - consistencia entre países, - consistencia entre productos, - efecto de una intervención, - certeza de una recomendación, - impacto comercial, - conformidad, - o una puntuación GEO. Todas las afirmaciones más amplias DEBEN estar respaldadas por una población objetivo predefinida, un diseño válido de múltiples observaciones, estructura de dependencia registrada, conjunto de pruebas, regla de evaluación, declaración de alcance y divulgación de incertidumbre. Un solo resultado de alto daño PUEDE desencadenar una revisión de incidente crítico separada sin ser tratado como evidencia de prevalencia en la población.

CAPÍTULO 02

La representación es una distribución

BASES METODOLÓGICAS / K04; K05; K09

Límite de Capítulo

La Sección 1 estableció la siguiente disposición:

> *Una única respuesta no es la puntuación GEO. Una única respuesta es una observación singular que ocurrió bajo condiciones registradas.*

Esta sección responde a la siguiente pregunta:

> *Si una sola respuesta no es la puntuación GEO, ¿cuál es el objeto real que queremos medir?*

La respuesta es:

> *La representación de una entidad en sistemas generativos es una distribución multidimensional formada a través de productos de IA, usuarios, países, idiomas, condiciones de sesión, prompts y tiempos.*

Este capítulo:

- la distribución de representación,
- las dimensiones fundamentales de la distribución,
- la diferencia entre la media y la distribución,
- la desigualdad del subgrupo,
- la estabilidad,
- el problema de cobertura y observaciones faltantes,
- el riesgo de errores graves raros,
- la estructura conceptual de la tarjeta de resultados disponible públicamente

define. Este capítulo aún no:

- cuántos participantes serán asignados a cada país,
- cómo se establecerá el marco de muestreo,
- cómo se validarán las traducciones de los prompts,
- cómo se calificarán las afirmaciones atómicas,
- la fórmula exacta de la puntuación final NOMOS,
- umbrales de conformidad

Este capítulo no resuelve estas cuestiones de manera normativa; capítulos posteriores las abordan por separado. Antes de definir la fórmula de calificación, el Capítulo 2 define la realidad que la puntuación pretende representar.

NOMOS Desafío

Considere dos productos de IA. Cada uno recibe la misma puntuación promedio global:

90 / 100

Los resultados lingüísticos del primer producto son los siguientes:

- Inglés: 90
- Turco: 91
- Alemán: 89
- Japonés: 90
- Árabe: 90

Los resultados lingüísticos del segundo producto son los siguientes:

- Inglés: 99
- Turco: 98
- Alemán: 97
- Japonés: 86
- Árabe: 70

El promedio simple de ambos productos es 90. ¿Tienen la misma calidad GEO? No. El primer producto produce una representación similar en diferentes idiomas. El segundo producto produce una representación excelente en algunos idiomas, pero una representación seriamente débil en otro idioma. Un único promedio oculta esta diferencia. Ahora considere otros dos productos. 95% de las respuestas de ambos cumplen con los criterios de aprobación. El restante 5% en el primer producto consiste en:

- deficiencias de alcance pequeño,
- carencias en la explicación limitada,
- algunos DESCONOCIDOS

resultados. El restante 5% en el segundo producto consiste en:

- autoridad sanitaria incorrecta,
- licencia inexistente,
- identidad legal incorrecta,
- consejo inapropiado

existe. La tasa de precisión bruta para ambos es nuevamente 95 por ciento. ¿Tienen el mismo riesgo de representación? No. Ahora considere un tercer par. El promedio de ambos a lo largo de las tres oleadas de medición es 90. El primero:

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

El segundo:

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

¿Tienen la misma consistencia a lo largo del tiempo? No. En ese caso, el problema no es solo: '¿Cuál es el promedio?' También necesitamos saber lo siguiente: ¿Cómo se distribuyen los resultados? ¿Quién está bien representado, quién está mal representado? ¿Qué idiomas y países se están quedando atrás?

¿Qué tan graves son los errores? ¿Cuánto cambia el resultado con el tiempo? ¿Qué áreas nunca fueron medidas? ¿Con qué pesos se creó el promedio? ¿Cuántas realidades diferentes se esconden detrás del mismo promedio? La primera conclusión de esta sección es:

> *El mismo promedio no es la misma representación.*

Su segunda disposición establece:

> *una puntuación GEO puede ser un resumen de la distribución; no puede reemplazar la distribución.*

1. OBJETIVO DEL CAPÍTULO

Esta sección define cómo debe entenderse la representación de entidades en sistemas generativos antes de ser reducida a un solo número. Sus principales objetivos son:

- Mostrar que la representación no es una "perspectiva de IA" única y atemporal
- Definir las condiciones de las cuales depende el resultado de la representación
- Explicar que el binario verdadero-falso por sí solo es insuficiente
- Separar la precisión promedio de la desigualdad entre subgrupos
- Hacer visible el alto riesgo de error oculto dentro de promedios altos
- Mostrar que estabilidad y precisión no son el mismo concepto
- Evitar que las áreas no medidas se llenen de suposiciones positivas dentro de la distribución
- Preparar la estructura de un futuro puntuación NOMOS que sea singular pero no condenado a ser singular

Al final de esta sección, el lector debería ser capaz de responder la siguiente pregunta:

> *¿Por qué el resultado del GEO de una entidad no se limita solo a la tasa promedio de precisión?*

2. DISPOSICIÓN NORMATIVA CENTRAL

> *La representación de una entidad en sistemas generativos es la distribución condicional de diferentes respuestas y resultados de representación que pueden ocurrir bajo condiciones definidas.*

Esta distribución:

- no es el 'pensamiento' abstracto de un solo modelo,
- no es la visión común de todos los sistemas de inteligencia artificial,
- no es solo la tasa de repetición del sitio web,
- no es solo la frecuencia con la que se menciona el nombre de la marca,
- no es solo la precisión promedio,

no es solo la tasa de recomendación. La distribución de representación, dentro de un cierto protocolo, muestra lo siguiente:

> *Qué usuarios ven qué tipo de representación en qué producto de IA y interfaz de usuario, en qué idioma, en qué país, con qué instrucción, bajo qué condición de sesión y a qué hora.*

Por lo tanto, el resultado de una distribución siempre depende de las siguientes preguntas: ¿Qué entidad? ¿Qué producto de IA? ¿Qué interfaz de usuario? ¿Cuál es la población objetivo? ¿En qué país? ¿Qué idioma y configuración regional? ¿Qué instrucción? ¿Qué condición de sesión? ¿Qué ola de medición? ¿Qué Paquete de Verdad? ¿Qué regla de evaluación? Cuando cualquiera de estos campos cambia, la distribución también puede cambiar.

3. ¿QUÉ SIGNIFICA LA PALABRA "DISTRIBUCIÓN"?

En este libro, la palabra distribución no se refiere a una visión oculta o metafísica que se asume que existe dentro de un sistema de inteligencia artificial. No afirma: NOMOS: "Esta es la verdadera distribución de la empresa en la mente del modelo." La distribución aquí es operativa y observable.

> *Es el patrón de diferentes resultados de representación que se pueden obtener bajo un protocolo definido dentro de la población objetivo y las condiciones de medición.*

Deben distinguirse dos distribuciones.

3.1. Distribución de Representación Objetivo

Es la distribución que la medición intenta predecir pero que no puede observar directamente en su totalidad. Se representa como:

P_E(Y | A, C, L, S, P, T, U)

Aquí:

- E: entidad auditada
- Y: resultado de la representación
- A: Producto de IA y interfaz de usuario
- C: país o condición geográfica
- L: idioma y localidad
- S: condición de sesión y personalización
- P: prompt (prompt)
- T: tiempo o ola de medición
- U: universo de usuarios objetivo

Esta expresión pregunta: “En el universo de usuarios objetivo definido, bajo las condiciones especificadas, ¿qué resultados de la representación ocurren con qué probabilidades?” No todos los elementos de la distribución objetivo pueden ser medidos. Se estiman mediante muestreo.

3.2. Distribución Empírica Observada

Es la distribución de observaciones válidas obtenidas en la medición real o sintética. Su representación:

P̂_E

puede ser usado. La distribución empírica no es la distribución objetivo en sí misma. Es una estimación producida a través de su muestra, ponderación y diseño de medición. La magnitud de la diferencia puede depender de:

- Diseño de la muestra
- Selección de participantes
- Cobertura de país e idioma
- Tasa de observación inválida
- No respuesta
- Error del adjudicador
- Ola de medición
- Selección de instrucciones

- Condiciones del producto e interfaz
- Ponderación
- Agrupamiento y dependencia

Por lo tanto, una distribución empírica no puede ser publicada como "la verdad absoluta e inmutable de la inteligencia artificial." La expresión correcta debería ser: "La estimación obtenida para la población objetivo especificada y las condiciones de medición."

4. ESPACIO DE REPRESENTACIÓN

La representación de una entidad no ocurre en una sola dimensión. Cada observación existe en un espacio de condiciones multidimensional. La representación básica:

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

Aquí:

- Ei: entidad
- IA: Producto de IA y interfaz de usuario
- Ui: usuario o unidad de muestreo
- Ci: país y capa geográfica
- Li: idioma y configuración regional
- Si: sesión, plan y condición de personalización
- Pi: prompt o familia de prompts
- Ti: tiempo y ola de medición

Respuesta generada:

```
R_i
```

puede mostrarse. El estado representativo resultante de la evaluación de la respuesta es:

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

y puede considerarse como. Aquí:

- VE: Pack de Veracidad verificado usado para la entidad
- J: reglas de evaluación y adjudicación
- g: proceso de evaluación que convierte la respuesta en resultados representativos

El significado de esta representación es el siguiente: La misma respuesta no puede evaluarse por sí sola sin indicar qué entidad, qué Paquete de Verdad y qué regla de evaluación se utilizó para examinarla.

5. EL RESULTADO REPRESENTATIVO NO ES UN VALOR ÚNICO

Solo la totalidad de una respuesta:

- puede ser forzada a las categorías de

  verdadero,
- es incorrecta

Dar una etiqueta suele ser insuficiente. Las siguientes situaciones pueden ocurrir juntas en la misma respuesta:

- Identidad correcta
- Categoría principal correcta
- Cobertura de servicio incompleta
- Número de empleados desactualizado
- Afirmación de liderazgo no respaldada
- Cita correcta
- Recomendación inapropiada
- Deficiencia de límite no crítica

Por lo tanto, el resultado de la representación puede considerarse como un vector de resultados. Representación candidata:

```
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)
```

Aquí:

- yiid: precisión de la identidad
- yicat: precisión de categoría y rol
- yifact: precisión de hechos
- yiscope: conservación del alcance y los límites
- yitime: validez temporal
- yievidence: consistencia entre afirmación y evidencia
- yirecommendation: recomendación y adecuación para el usuario
- yiharm: clase de daño del error o desinformación

Las reglas exactas de puntuación para estos campos se definirán en secciones posteriores. El principio guía aquí es:

> *Una respuesta representativa debe evaluarse no solo en función de si contiene información correcta, sino también sobre quién, qué, dentro de qué límites, con qué evidencia y para quién se transmite adecuadamente.*

6. ESTADOS BÁSICOS DE REPRESENTACIÓN

El estado de cada afirmación atómica y la respuesta final pueden no ser los mismos. Esta sección separa conceptualmente los siguientes estados básicos.

6.1. RESPALDADO

La afirmación está respaldada por evidencia suficiente dentro del alcance y tiempo definidos.

6.2. CONTRADICTORIO

La afirmación entra en conflicto con registros fiables y relevantes.

6.3. NO RESPALDADO

No hay suficiente apoyo para la afirmación. No se ha demostrado que la afirmación sea necesariamente falsa. Sin embargo, no hay suficientes registros para que se acepte como verdadera.

6.4. DESACTUALIZADO — YA NO ACTUAL

La afirmación puede haber sido cierta en el pasado. A la fecha de medición, ya no es válida o no se puede verificar su exactitud actual.

6.5. PARCIALMENTE RESPALDADO — PARCIALMENTE RESPALDADO

Solo se respalda un alcance, población, país, producto o tiempo específico de la afirmación. La respuesta puede haber excedido el límite de respaldo.

6.6. FUERA DE ALCANCE — FUERA DE ALCANCE

La afirmación no está relacionada con la entidad auditada o el Truth Pack. Puede provenir de otra entidad, producto, país o servicio.

6.7. CONFUSIÓN DE ENTIDADES — FUSIÓN DE ENTIDADES

Sistema:

- dos empresas diferentes,
- la marca y la entidad legal,
- una persona y la institución,
- entidad antigua y nueva,
- ha confundido organizaciones diferentes con el mismo nombre

entre sí. Este error no es meramente un error factual. Es la corrupción del objeto de representación.

6.8. OMISIÓN MATERIAL — INFORMACIÓN FALTANTE

Las oraciones en la respuesta pueden ser correctas. Sin embargo, se ha omitido información esencial que podría cambiar la decisión del usuario. Ejemplo: El servicio se ofrece solo en ciertos países. La licencia está limitada a una jurisdicción específica. El precio no incluye impuestos. El resultado se basa en datos de primera mano. El producto no es adecuado para ciertos usuarios. La omisión puede crear una inferencia incorrecta incluso si no hay una oración directamente incorrecta.

6.9. RECHAZO — NEGACIÓN

El sistema se ha negado a responder. Rechazo:

- apropiado,
- innecesario,
- inconsistente,
- basado en seguridad,
- técnico

puede ocurrir. Ret no se considera automáticamente fallido o exitoso. Se clasifica por separado.

6.10. SIN RESULTADO

El sistema no produjo una respuesta utilizable debido a razones técnicas u operativas. No es lo mismo que Ret.

6.11. DESCONOCIDO

Los registros disponibles no son suficientes para una clasificación confiable. DESCONOCIDO no puede convertirse en un resultado positivo.

7. CARACTERÍSTICAS PRINCIPALES DE LA DISTRIBUCIÓN

Una distribución de representación no puede explicarse únicamente por la media. Como mínimo, se deben evaluar las siguientes características.

7.1. Tendencia Central

Resume la precisión general o el nivel de transición de las respuestas. Medidas de ejemplo:

- Media
- Media ponderada
- Mediana
- tasa de aprobación

La tendencia central es importante. Sin embargo, solo muestra el centro de la distribución.

7.2. Dispersión y Variabilidad

Indica cuánto se distribuyen los resultados por el centro. De dos sistemas con la misma media, uno puede ser consistente entre usuarios, mientras que el otro puede ser muy variable. Variabilidad:

- usuarios,
- países,
- idiomas,
- sesiones,
- Ondas de medición

También pueden examinarse entre ellos.

7.3. Forma de la distribución

¿Están los resultados agrupados en torno a un único punto central? ¿O están divididos en dos o más grupos? Por ejemplo:

- Si la mitad de los usuarios ve con mucha precisión,
- y la mitad de la mitad muy inexacta

Respuestas, la media puede ser media. Sin embargo, la experiencia típica del usuario no es de "precisión media". Son dos experiencias polarizadas separadas.

7.4. Diferencias en subgrupos

¿Determinados idiomas, países o grupos de usuarios ven resultados diferentes de forma sistemática? La media global puede ocultar esta diferencia.

7.5. Riesgo de cola

¿Existen errores poco frecuentes pero graves? Los eventos raros pueden tener un efecto pequeño en el promedio. Su impacto en el daño a los humanos puede ser significativo.

7.6. Cobertura

¿Cuánto del universo de usuarios objetivo ha sido realmente medido? Se puede encontrar una puntuación alta junto con baja cobertura.

7.7. Masa Desconocida y Faltante

¿Qué proporción de las respuestas son:

DESCONOCIDO,

NO PROBADAS,

rechazo, no respuesta técnica o un registro inválido?

7.8. Estabilidad en el tiempo

¿En qué medida se preserva la distribución a través de diferentes olas de medición?

8. FIRMA DE DISTRIBUCIÓN REPRESENTATIVA DEL CANDIDATO NOMOS

CONCEPTO DE CANDIDATO

Se recomienda que el resultado de un GEO se defina primero por la siguiente firma de distribución en lugar de un solo número:

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

Esta representación se llama Firma de Distribución de Representación, es decir:

Firma de Distribución de Representación

Componentes:

μ — Nivel General de Representación

Precisión general o predicción de transición bajo pesos y alcance definidos.

κ — Cobertura

El grado en que el usuario objetivo, el país, el idioma, el producto de IA y el universo temporal están suficientemente medidos.

σ — Variabilidad

Dispersión entre los usuarios y las células experimentales.

Δ — Desigualdad entre Subgrupos

La diferencia material entre idioma, país, producto u otros grupos significativos.

τ — Riesgo de Error Crítico y Extremo

La tasa y naturaleza de errores graves raros.

ρ — Estabilidad Temporal

La resiliencia de la distribución a través de las olas.

u — Masa Desconocida e No Resuelta

DESCONOCIDO es la proporción de la distribución que permanece no resuelta o insuficientemente evaluada. Esta firma aún no es una fórmula de puntuación final. Los siguientes elementos requieren un mayor desarrollo:

- definición definitiva,
- estimación,

- peso,
- presentación pública

Se finalizarán a través de estudios piloto y capítulos posteriores. El principio fundamental es:

> *El resultado GEO debería responder no solo a la pregunta '¿qué tan preciso?' sino también '¿para quién, qué tan consistente, qué tan equitativo, qué tan completo y qué tan riesgoso?'*

9. LAVADO PROMEDIO

Si una institución solo muestra un promedio global alto y oculta los defectos materiales de la distribución, esto se llama:

Lavado promedio

El lavado de promedios puede ocurrir de las siguientes maneras:

- Ocultando el idioma más débil
- Excluyendo a los países pequeños del informe
- Mezclando errores críticos en la tasa general de errores
- Eliminar los resultados DESCONOCIDOS del denominador
- Incluyendo solo productos de IA con buen desempeño
- Excluyendo la oleada de medición débil
- Haciendo que la población fuera del alcance sea invisible
- Cambiando los pesos después de ver los resultados

El promedio puede calcularse matemáticamente de manera correcta. Sin embargo, la representación que ofrece al público aún puede ser engañosa. Por lo tanto:

> *Un cálculo matemáticamente correcto de un promedio no significa que sea suficiente para la toma de decisiones.*

10. DIFERENTES REALIDADES DEL MISMO PROMEDIO

El promedio podría ser 90 en todas las siguientes distribuciones sintéticas.

10.1. Distribución Homogénea

Todos los subgrupos principales se encuentran en el rango de 88–92. Esta distribución es relativamente equilibrada.

10.2. Distribución Polarizada

La mitad de los usuarios ve resultados cercanos a 100, mientras que la otra mitad ve resultados cercanos a 80. El promedio es 90. Sin embargo, la experiencia del usuario no es igual.

10.3. Distribución Lingüísticamente Fragmentada

Los resultados son muy buenos para idiomas con alta población o altos recursos, pero significativamente débiles para idiomas con pocos recursos. El promedio global puede parecer bueno. La equidad lingüística es débil.

10.4. Distribución Fragmentada por País

Incluso dentro del mismo idioma, los resultados pueden variar debido a diferentes países, configuraciones regionales, acceso al producto o base de usuarios.

10.5. Distribución Temporalmente Volátil

Resultados muy altos ocurren en una ola, resultados muy bajos en otra ola. Incluso si el promedio es alto, es difícil establecer expectativas confiables para los usuarios.

10.6. Distribución con Riesgo de Cola

La mayoría de las respuestas son correctas. Sin embargo, una pequeña porción contiene errores graves de salud, legales, de licencias o de seguridad.

10.7. Alta Distribución de Masa Desconocida

La mayoría de las respuestas evaluables son correctas. Sin embargo, una porción significativa de la población objetivo o de las respuestas es:

- probada,
- no probada,

DESCONOCIDO

o inconclusa. Publicar una alta tasa basada únicamente en las respuestas evaluables puede ocultar el alcance.

11. DESIGUALDAD ENTRE SUBGRUPOS

Una puntuación global puede no representar la experiencia en subgrupos. Los subgrupos clave pueden incluir:

- Producto de IA y interfaz de usuario
- País
- Idioma
- Localización
- Plan
- Web o móvil
- Panel controlado o natural
- Onda temporal
- Intención del usuario
- Familia de dispositivos

No todos estos grupos pueden tener la misma importancia en cada auditoría. Los grupos materiales deben determinarse cuando el alcance esté definido.

11.1. El lenguaje no es lo mismo que el país

Hablante de turco:

- Usuario en Turquía,
- Usuario en Alemania,
- Usuario en otro país

Pueden usar el mismo idioma. Sin embargo:

- acceso al producto,
- precio,
- recursos,

- interfaz local,
- contexto legal,
- personalización

pueden ser diferentes. Por lo tanto, todas las diferencias no pueden explicarse únicamente por el idioma o el país.

11.2. Células Interseccionales

Algunas diferencias materiales solo se vuelven visibles cuando se examinan varias áreas juntas. Ejemplo: Alemania × Turco × Móvil × Plan gratuito × Panel de usuarios en condiciones reales podría ser en sí misma una celda experimental. Sin embargo, a medida que las celdas se multiplican, el tamaño de la muestra disminuye. Por lo tanto, no se puede publicar una puntuación definitiva para cada celda pequeña. Diseño de medición: debe preservar subgrupos significativos y no debe producir una precisión estadística falsa.

11.3. Base de Subgrupo

CONCEPTO DE CANDIDATO

Además del promedio global, se recomienda informar el resultado del subgrupo más débil adecuadamente muestreado o del grupo inferior 10% por separado. Esto puede denominarse:

Base de representación

o:

Idioma/País Piso

Sin embargo, si la "puntuación más baja" proviene de una célula pequeña e inestable, puede producir resultados incorrectos. Por lo tanto, la métrica base debe definirse junto con:

- muestra mínima,
- intervalo de confianza
- calidad de la célula

La regla exacta se determinará en las secciones siguientes.

12. ES PARTE DE LA DISTRIBUCIÓN DEL ALCANCE

Una distribución no consiste únicamente en respuestas observadas. Lo que nunca se observó también determina el significado del resultado. Por ejemplo, se probaron inglés y turco; japonés y árabe no. Se probó la interfaz web; la interfaz móvil y el plan de pago no. Los resultados sólidos en inglés y turco no justifican la afirmación: 'La representación es sólida en todos los idiomas y interfaces de usuario'. La conclusión defendible es: 'Se realizaron mediciones en las interfaces web especificadas en inglés y turco. Otros idiomas y interfaces móviles no fueron probados.'

12.1. Mapa de cobertura

Cada distribución GEO debería incluir el siguiente mapa en la medida en que sea relevante:

- Universo objetivo
- Marco muestral
- Países cubiertos
- Países no cubiertos
- Idiomas cubiertos

- Idiomas no cubiertos
- Productos de IA monitoreados
- Productos no monitoreados
- Interfaces de usuario monitoreadas
- Interfaces no monitoreadas
- Olas de medición
- Familias de prompts fuera de alcance

12.2. Tasa de cobertura

CONCEPTO DE CANDIDATO

Se puede desarrollar una métrica de cobertura para mostrar qué proporción del universo objetivo está representada con suficiente medición. Sin embargo, la cobertura:

- solo el número de países,
- solo el número de idiomas,
- solo la proporción de la población

no se puede medir. Por ejemplo, 150 países pueden estar cubiertos. Sin embargo, 95% de los participantes pueden haber venido de solo cinco países. O 90% de la población puede estar cubierta, pero los idiomas pequeños pueden estar completamente excluidos. Por lo tanto, la cobertura de la población debe mostrarse por separado de la cobertura de observadores por país/idioma.

13. OBSERVACIÓN FALTANTE Y NO RESPUESTA

Los datos faltantes no son solo un problema técnico. Su distribución puede estar distorsionada sistemáticamente. Las razones por las que falta una observación pueden incluir:

- Fallo del participante al completar la tarea
- Producto no accesible en el país
- Limitaciones de cálculo
- Error técnico
- Rechazo del sistema
- Captura de pantalla inválida
- Cambio del participante en su instrucción
- Producto que no funciona en un idioma específico
- Los jueces no pueden tomar decisiones confiables

Cada una de estas razones requiere un estatus separado.

13.1. Los datos faltantes pueden no ser aleatorios

Por ejemplo, si no se pueden completar más tareas en países con baja calidad de conexión, usar solo grabaciones exitosas puede sesgar el resultado global a favor de usuarios con recursos y conexiones fuertes. Si la tasa de DESCONOCIDO aumenta porque las respuestas en un idioma específico son difíciles de evaluar, la puntuación calculada solo a partir de respuestas conocidas puede ocultar la desigualdad lingüística. Por lo tanto:

> *La tasa de datos faltantes y el informe de distribución de la razón de los datos faltantes deberían formar parte del informe.*

13.2. Denominador — Integridad del Denominador

Un informe debe responder claramente a la pregunta: "90 por ciento de exactamente qué es 90 por ciento?" Posibles denominadores:

- Participantes asignados
- Iniciadores de tareas
- Completadores de tareas
- Registros conformes al protocolo
- Resultados válidos del sistema
- Respuestas con decisiones del adjudicador
- Celda de idioma o país específico

El denominador no debe cambiarse después de ver el resultado. Por ejemplo:

- Tareas asignadas a 1,000,
- Tareas completadas por 900,
- Tareas válidas de 800,
- Tareas aprobadas de 720

Así:

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

Esta es la tasa de aprobación entre los registros válidos. La tasa de aprobación observada en toda la muestra asignada es en cambio:

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

Responde a una pregunta diferente. Se pueden utilizar ambas tasas. No se pueden presentar en el mismo sentido.

14. EXACTITUD Y CONSISTENCIA NO SON LO MISMO

Una distribución puede tener alta precisión. Sin embargo, puede ser inconsistente a lo largo del tiempo o entre usuarios. Otra distribución puede tener baja precisión. Sin embargo, puede estar consistentemente equivocada. Por lo tanto, se deben distinguir cuatro situaciones.

Precisión	Consistencia	Comentario
Alta	Alta	Candidata a una representación consistente y precisa
Alta	Baja	Experiencia mayormente precisa pero poco confiable
Baja	Alta	Representación sistemática y consistente incorrecta
Baja	Baja	Representación incorrecta e inestable

Ser “consistente” no es automáticamente positivo. El sistema puede ser bastante consistente si da el mismo error a cada usuario.

14.1. Consistencia entre usuarios

¿Ven los diferentes usuarios una representación material similar en la misma celda?

14.2. Consistencia entre sesiones

¿El mismo usuario o un usuario equivalente ve resultados similares en nuevas sesiones?

14.3. Consistencia entre olas

¿Produce el mismo protocolo una distribución similar en diferentes días o periodos?

14.4. Consistencia después de la actualización del producto

Cuando hay un cambio material en el producto de IA, ¿continúa la distribución anterior?

15. PRODUCIR LA MISMA ORACIÓN NO ES ESTABILIDAD

Estabilidad: No significa "dar la misma respuesta exacta a cada usuario palabra por palabra". Un producto de IA puede usar diferentes expresiones, orden y ejemplos. Siempre que preserven el significado material, pueden ser variaciones aceptables. La distinción importante es:

Variabilidad formal

Elección de palabras Orden de la oración Longitud Tono Ejemplo Uso de lista o párrafo Si se preserva la verdad material, puede no ser un error.

Variabilidad semántica

Cambio de identidad de la empresa Diferenciación de la categoría principal Expansión o contracción del campo de actividad Cambio de licencia o competencia Cambio de precio y geografía Recomendar a un usuario pero no recomendar a un usuario similar sin explicación Si la representación material cambia, puede haber un problema de estabilidad. Por lo tanto, el concepto preferido es:

Estabilidad semántica condicional

Debe ser así. Se espera que la realidad fundamental se conserve bajo condiciones materiales equivalentes. Un cambio en la recomendación debido a diferentes necesidades del usuario puede ser un comportamiento contextual apropiado.

16. SENSIBILIDAD AL CONTEXTO Y VARIABILIDAD ARBITRARIA

No siempre es correcto que un sistema dé la misma recomendación a todos los usuarios. Por ejemplo:

- Usuarios con un presupuesto bajo
- Solo usuarios que buscan servicios en un país específico
- Usuarios que requieren una licencia diferente
- Usuarios que buscan servicios individuales en lugar de corporativos

pueden recibir resultados de conformidad diferentes para la misma entidad. Esta diferencia:

- puede deberse a la condición del usuario,
- al alcance del servicio,
- o al límite real

y puede ser sensible al contexto y precisa. Sin embargo, si el resultado cambia sin motivo bajo condiciones equivalentes del usuario, puede ser una variabilidad arbitraria. Por lo tanto, NOMOS no desea: “La misma respuesta para todos los usuarios.” Quiere:

> *Una respuesta que sea consistente y explicable bajo condiciones materiales equivalentes y sensible a los límites reales bajo diferentes condiciones.*

17. EL TIEMPO ESTÁ DENTRO DE LA DISTRIBUCIÓN DE REPRESENTACIÓN

La distribución de representación no es atemporal. Se puede considerar como PE,t. La distribución de la misma entidad:

- actualización del producto,
- cambio de recurso,
- cambio de empresa,
- evento noticioso,
- nueva versión de idioma,
- cambio de acceso técnico

puede divergir después.

17.1. Cambio Repentino

Después de una actualización de producto o recurso, la distribución cambia rápidamente en un corto periodo de tiempo.

17.2. Cambio Gradual

Debido a nuevos recursos y comportamientos de los usuarios, el resultado cambia lentamente con el tiempo.

17.3. Fluctuación Temporal

Durante un evento específico o periodo noticioso, la distribución difiere temporalmente.

17.4. Cambio Post-Intervención

Después de la intervención GEO, la distribución puede cambiar. Sin embargo, afirmar que el cambio se debe a la intervención también requiere una línea base y un diseño de causalidad.

17.5. Suavizado con Promedio Temporal

Para un producto:

- en la primera ola 98,
- en la segunda ola 90,
- en la tercera ola 82

Si se toma, el promedio es 90. Sin embargo, publicar 90 oculta la tendencia a la baja. Por lo tanto:

> *El promedio temporal no puede reemplazar los resultados de la ola.*

18. RIESGO EXTREMO Y DISTRIBUCIÓN DE ERRORES GRAVES

Una gran parte de una distribución puede ser correcta. Sin embargo, también deben evaluarse los errores graves de baja probabilidad. Ejemplos:

- Licencia de salud que en realidad no existe
- Guía de emergencia incorrecta

- Autoridad legal incorrecta
- Garantía de inversión falsa
- Identidad de empresa incorrecta
- Asesoramiento inapropiado para niños o grupos vulnerables
- Repetición de la instrucción manipulativa oculta a las personas

En la arquitectura anterior:

- Crítico,
- Mayor,
- Moderado,
- Asesor

se separaban en cuatro niveles de importancia, y se sugería que los hallazgos Críticos deberían tener un efecto de fallo automático. Este libro vincula este sistema con la distribución de la población.

18.1. Separación de Frecuencia y Severidad

Un error:

- frecuente pero de bajo impacto,
- raro pero severo

puede ocurrir. Sin embargo, la tasa de frecuencia por sí sola no puede evaluar ambos casos por igual.

18.2. Representación de Riesgo Ponderada por Severidad Candidata

CONCEPTO DE CANDIDATO

Se puede desarrollar un indicador de riesgo utilizando coeficientes de severidad para clases de error:

```
tau= Σ_{k=1}^K λ_k p_k
```

Aquí:

- pk: tasa prevista del tipo de error
- λk: coeficiente de severidad del tipo de error

Sin embargo, los coeficientes no pueden determinarse de manera arbitraria. Deben basarse en:

- Daño humano previsible
- Integridad de la representación
- Impacto en la decisión del usuario
- Reversibilidad
- Intención y manipulación
- Área de alto riesgo

El método preciso se definirá en la Sección 15.

18.3. Puerta crítica

Un promedio general alto no puede eliminar automáticamente un candidato de error Crítico verificado. El registro público correcto debería mostrar ambos resultados juntos: “La estimación de representación general es alta.” y: “El evento o eventos Críticos especificados también afectan la decisión de conformidad.”

19. RECHAZO Y FALLO EN LA DISTRIBUCIÓN

Un sistema que no responde también forma parte de la experiencia del usuario. Sin embargo, no todos los rechazos son iguales.

19.1. Rechazo Apropiado

El sistema puede actuar con precaución cuando la evidencia es insuficiente o cuando no puede proporcionar una respuesta segura. Este puede ser el comportamiento correcto en algunos contextos.

19.2. Rechazo Innecesario

Puede evitar responder de manera inconsistente a información pública y de bajo riesgo.

19.3. No Respuesta Técnica

El sistema puede no producir una respuesta debido a un error técnico.

19.4. Rechazo Fuera de Alcance

El prompt puede ser incorrecto o estar fuera del protocolo de la tarea.

19.5. Preservación de Rechazos en Distribuciones

Rechazos:

- éxito,
- fracaso,

DESCONOCIDO

no deben ser forzados a una sola categoría. Su función y justificación deben ser evaluadas. Eliminar rechazos del denominador puede producir una alta tasa de precisión solo entre las respuestas exitosas y ocultar la experiencia real del usuario.

20. ¿CÓMO SE PUEDEN COMPARAR LAS DISTRIBUCIONES?

Simplemente comparar los números de resultados no es suficiente para comparar la distribución de dos productos de IA. Para que la comparación sea justa, las siguientes áreas deben ser lo más iguales posible o modeladas explícitamente:

- Entidad auditada
- Población objetivo
- Versión del prompt
- Idiomas
- Países
- Ola de medición
- Condición de la sesión
- interfaz de usuario
- Distribución del plan
- Regla de adjudicación

- Paquete de realidad
- Reglas de validez
- Pesos

Dos puntuaciones obtenidas con protocolos materialmente diferentes no pueden usarse directamente para la clasificación.

20.1. Comparación sobre una base común

Todos los productos de IA comparados:

- son oficialmente accesibles,
- dan soporte al mismo idioma y condiciones de usuario

pueden compararse en un universo común. Esto proporciona una comparación más clara entre sistemas.

20.2. Distribución de Acceso Natural

Cada producto de IA puede medirse en su propio universo de acceso de usuarios reales. Este resultado muestra el impacto global real del producto. Sin embargo, dado que las poblaciones objetivo pueden diferir, la clasificación de desempeño directa requiere precaución.

20.3. Diferentes Interfaces del Mismo Proveedor

Los productos web, móviles, empresariales o diferentes planes, incluso si pertenecen al mismo proveedor, pueden generar distribuciones separadas. Si se van a combinar bajo un único resultado, se debe explicar el método de combinación y ponderación.

21. LA PONDERACIÓN CAMBIA EL SIGNIFICADO DE LA DISTRIBUCIÓN

Un promedio global depende del peso con el que se combinan las observaciones. Pesos posibles:

- Igual por país
- Por población del país
- Población de usuarios de IA adecuada
- Usuario del idioma
- Exposición real al producto
- Producto de IA igual
- Cuota de mercado
- Mercado objetivo del cliente

Cada método responde a una pregunta diferente.

21.1. Ponderación de país igual

San Marino e India pueden recibir el mismo peso. Este enfoque puede abordar la cuestión de la 'experiencia promedio de un país'. No representa la experiencia promedio de los usuarios en la población mundial.

21.2. Peso según la población

Los países más grandes tienen más influencia. Este enfoque puede aproximarse a la experiencia de la población global. Puede hacer que los países y idiomas pequeños sean invisibles.

21.3. Peso de equidad lingüística

También puede garantizar que los idiomas con pocos recursos permanezcan visibles. No es lo mismo que la puntuación global de la población.

21.4. Peso de exposición del producto

Se puede considerar el ponderamiento de los productos de IA según la exposición real de los usuarios. Requiere datos de uso confiables e independientes.

21.5. Principio de ponderación normativa

> *Los pesos deben definirse antes de que se vean los resultados, divulgarse públicamente y especificar la pregunta que responden.*

Está prohibido elegir un sistema de ponderación que aumente la puntuación positiva después de ver el resultado. Junto con una puntuación global única, también se pueden publicar resultados ponderados alternativos significativos.

22. LA DISTRIBUCIÓN DE PUNTUACIONES ES COMPRESIÓN CON PÉRDIDAS

Es posible producir un único número a partir de una distribución. Sin embargo, un solo número no contiene toda la información de la distribución. Por lo tanto, la puntuación:

Es un resumen con pérdidas de la distribución

debería ser aceptada. Una puntuación puede perder la siguiente información:

- Idioma más débil
- País más débil
- Tipo de error crítico
- Masa desconocida
- Fluctuación de tiempo
- Polarización
- Diferencia de producto y plan
- Población fuera de alcance
- Distribución de rechazos
- Ambigüedad del adjudicador

La puntuación es útil porque:

- facilita la comparación,
- simplifica la comunicación pública,
- permite monitorear cambios,

apoya la decisión de gestión. Sin embargo, la simplicidad de la puntuación no puede convertirse en la eliminación de la distribución.

23. TARJETA DE RESULTADOS PÚBLICOS MÍNIMOS

REQUISITO DEL CANDIDATO — OBLIGACIÓN DEL CANDIDATO

Si se va a presentar una puntuación global NOMOS al público, se recomienda que al menos se muestren los siguientes resultados junto con ella:

Campo	Descripción
Estimación de la representación principal	Resultado general bajo el ponderado definido
Rango de confianza o incertidumbre	Sensibilidad de la predicción
Alcance	País, idioma, producto, interfaz de usuario y momento
Subgrupo suficiente más débil	Hace visible la desigualdad material
Tasa de error crítica	Cola de errores graves
Estabilidad temporal	Variación entre oleadas
Tasa de DESCONOCIDO e inconclusa	Masa desconocida
Campos NO EVALUADOS	Universo no medido
Fecha de medición	Límite temporal del resultado
Versión del protocolo	Con qué regla se produjo

La puntuación principal no debe usarse como una etiqueta publicitaria independiente. El registro canónico debe estar vinculado a esta tarjeta de resultados.

24. DISTRIBUCIÓN SINTÉTICA APPLE.COM

DEMOSTRACIÓN DE METODOLOGÍA SINTÉTICA / Los productos de IA, usuarios, idiomas, números y resultados que se muestran a continuación son completamente ficticios. No representan el rendimiento real de Apple Inc. ni de ningún producto de IA real. Consideremos dos productos de IA sintéticos:

- Producto A
- Producto B

Para cada producto:

- cinco idiomas,
- 200 observaciones sintéticas válidas por idioma,
- un total de 1,000 observaciones

se asume. Esta distribución equitativa de idiomas es solo para demostrar el método. No representa el peso poblacional real GEO-1000.

24.1. Tasas de aprobación basadas en el idioma

Idioma	Producto A	Producto B
Inglés	90%	99%
Turco	91%	98%
inglés	89%	97%
Japonés	90%	86%
Árabe	90%	70%

Idioma	Producto A	Producto B
Promedio simple	90%	90%

El promedio de los dos productos es el mismo. Sin embargo, sus distribuciones son diferentes.

Producto A

Los resultados de los idiomas son cercanos entre sí. La diferencia entre el idioma más alto y el más bajo es de 2 puntos. El piso del idioma es 89%.

Producto B

Produce resultados muy altos en tres idiomas. Hay un descenso en japonés y especialmente en árabe. La diferencia entre el idioma más alto y el más bajo es de 29 puntos. El piso del idioma es 70%. Nota: Decir "Ambos obtuvieron 90" oculta la diferencia material.

24.2. Distribución del estado de la respuesta

Estado de la respuesta	Producto A	Producto B
Cumpliendo el criterio de transición	900	900
Parcialmente soportado	60	20
afirmación de material no respaldada	20	20
Desactualizado	10	20
Candidato a error crítico	1	14
DESCONOCIDO / sin resultado	9	26
Total	1,000	1,000

La tasa de aprobación de los dos productos es nuevamente 90%. Sin embargo: En el Producto B, el número de candidatos con errores críticos es 14 veces mayor. En el Producto B, la masa desconocida e inconclusiva es mayor. Una mayor parte de los fallos en el Producto A se debe a deficiencia de soporte parcial. El retraso de errores del Producto B es más pesado. La misma tasa de aprobación no necesariamente resulta en el mismo resultado de aptitud.

24.3. Ondas de Tiempo

Ola de medición	Producto A	Producto B
Onda 1	90%	98%
Onda 2	91%	90%
Onda 3	89%	82%
Media	90%	90%

El Producto A es relativamente estable. El Producto B muestra un descenso notable. Publicar solo el promedio de las tres ondas oculta el deterioro en el Producto B.

24.4. Resultado Sintético

Puede ser posible publicar la misma puntuación principal para el Producto A y el Producto B:

90

Pero las tarjetas de resultados NOMOS no pueden ser las mismas.

Producto A — Firma de Distribución Sintética

Resultado general: alto
Desigualdad de idioma: baja
Cola de errores críticos: baja, pero no cero
Masa desconocida: baja
Estabilidad temporal: alta

Producto B — Firma de Distribución Sintética

Resultado general: alto
Desigualdad de idioma: alta
Cola de errores críticos: sustancial
Masa desconocida: mayor
Estabilidad temporal: baja y en declive

Por lo tanto, la conclusión sintética de la Sección 2 es la siguiente:

> *Tener la misma puntuación promedio para el Producto A y el Producto B no indica que tengan la misma calidad de representación.*

25. DISPOSICIONES NORMATIVAS OBLIGATORIAS

CH02-N01

Un resultado GEO no puede definirse únicamente por un promedio único o una tasa de aprobación única.

CH02-N02

Cada resultado de distribución:

- la entidad auditada,
- población objetivo,
- Producto de IA y interfaz de usuario,
- país y lengua,
- El prompt,
- la condición de residencia,
- al tiempo

debe estar conectado.

CH02-N03

El promedio global por sí solo no puede compensar la desigualdad de subgrupos materiales.

CH02-N04

Los resultados de errores críticos y de otra severidad alta deben informarse por separado de la tasa de error general.

CH02-N05

Los estados DESCONOCIDO, sin resultado, rechazado, inválido y NO PROBADO deben distinguirse entre sí.

CH02-N06

No se puede mostrar como parte de la distribución positiva un país, idioma, producto, interfaz de usuario o periodo de tiempo no medidos.

CH02-N07

Los pesos y las reglas de agregación deben definirse antes de ver los resultados.

CH02-N08

Los cambios de peso posteriores al resultado deben crear una nueva versión del análisis, y los resultados anteriores deben conservarse.

CH02-N09

Las distribuciones basadas en poblaciones objetivo materialmente diferentes, demanda, producto, tiempo o criterios de evaluación no pueden compararse directamente sin explicación.

CH02-N10

La variabilidad del lenguaje formal no puede contarse automáticamente como error de representación.

CH02-N11

La identidad material, el alcance, el tiempo, la competencia o la variabilidad de las recomendaciones no pueden clasificarse como una diferencia de estilo sin justificación.

CH02-N12

La alta estabilidad por sí sola no es evidencia de calidad; también se deben registrar los resultados consistentemente incorrectos.

CH02-N13

Un promedio alto, si va acompañado de una cobertura baja, debería tener resultados públicos con limitación de cobertura.

CH02-N14

Los puntuaciones de subgrupos no deben publicarse como juicios definitivos sobre países o idiomas bajo un muestreo insuficiente y una amplia incertidumbre.

CH02-N15

Las tasas de rechazo y de fallos técnicos deben conservarse como parte de la experiencia auténtica del usuario.

CH02-N16

La única puntuación que se haga pública debería estar vinculada a la distribución canónica y al registro de alcance.

CH02-N17

Las distribuciones sintéticas no se pueden combinar con el rendimiento de un producto de IA real o una entidad real.

CH02-N18

El informe de distribución debe tener un propietario humano o institucional responsable.

26. FORMAS DE FALLA

CH02-F01— PENSAMIENTO PROMEDIO

Se muestra un promedio global alto; se almacena el lenguaje, país o corrupción de error crítico.

CH02-F02— ELIMINAR SUBGRUPO

Los grupos de bajo rendimiento se eliminan del informe o se declaran 'insignificantes'.

CH02-F03— MAPA DE ALCANCE

Los buenos resultados de alcance limitado se utilizan para todo el mundo, todos los idiomas o todos los productos.

CH02-F04— NORMALIZACIÓN DE GRAVEDAD

Las deficiencias menores y los errores graves de salud o de licencias se combinan dentro del mismo recuento de errores.

CH02-F05— OCULTAR EL DETERIORO CON PROMEDIO DE TIEMPO

La disminución o fluctuación se hace invisible solo con el promedio del periodo.

CH02-F06— ADORACIÓN DE LA ESTABILIDAD

Se supone que una respuesta consistente es automáticamente correcta.

CH02-F07— FALACIA DE RESPUESTA IGUAL LITERALMENTE

La coincidencia literal de texto se utiliza como una medida de consistencia en lugar de la corrección semántica.

CH02-F08— CONTANDO LA DIFERENCIA CONTEXTUAL COMO ERROR

Las diferencias en recomendaciones apropiadas basadas en las necesidades reales del usuario se marcan como variabilidad arbitraria.

CH02-F09— CONTANDO LA VARIABILIDAD ARBITRARIA COMO PERSONALIZACIÓN

Las contradicciones materiales bajo condiciones equivalentes del usuario se defienden como "personalización" sin explicación.

CH02-F10— CAMBIANDO EL DENOMINADOR

Se realiza una transición post-resultado entre denominadores de observación completados, válidos o evaluados para elevar el resultado positivo.

CH02-F11— PESO POST-RESULTADO

Se seleccionan posteriormente el país, el idioma o los coeficientes del modelo que producen la puntuación más alta.

CH02-F12— FUSIONANDO CELDAS INCONSISTENTES

Se presentan diferentes condiciones de prompt, producto, tiempo o panel como una única distribución.

CH02-F13— ELIMINANDO RECHAZOS

Se eliminan los registros del sistema que no respondieron para generar alta precisión solo entre aquellos que produjeron una respuesta.

CH02-F14— JUICIO DEFINITIVO DE CELDA PEQUEÑA

Se produce una puntuación de país o idioma a partir de uno o varios usuarios.

CH02-F15— OCULTANDO LA COLA CRÍTICA

Los errores graves raros se pierden dentro de la tasa de aprobación general.

CH02-F16— INFORME DE puntuación ÚNICO

Solo se publica la puntuación principal sin mostrar alcance, desigualdad, estabilidad, incertidumbre y errores graves.

CH02-F17— CONTANDO LA DISTRIBUCIÓN COMO "LA CREENCIA DE LA IA"

Los resultados observados interfaz de usuario se presentan como si fueran el punto de vista invariante e intrínseco de la inteligencia artificial.

CH02-F18— CONTANDO COMO VARIANZA CERO LO NO MEDIDO

Se considera que las áreas no probadas son suaves y estables.

27. PROCEDIMIENTO DE AUDITORÍA

Se debe examinar si un informe GEO representa honestamente la distribución con los siguientes pasos.

Paso 1— Registrar la Declaración Pública

Ejemplo: “La puntuación global NOMOS de la marca es 92.” Se registra la frase completa y la interfaz donde se encuentra.

Paso 2— Definir Distribución Objetivo

Para qué puntuación:

- universo de usuarios,
- productos de IA,
- países,
- idiomas,
- período de tiempo

se aplica?

Paso 3— Extraer Celdas de Medición

Se determina cada combinación de material:

- Producto
- País
- Idioma
- Panel
- Prompt
- Onda
- interfaz de usuario

Paso 4— Reconstruir Denominadores

Las tareas asignadas, las tareas completadas, los registros válidos, los registros adjudicados y las respuestas aprobadas deben informarse como recuentos separados.

Paso 5— Extraer estados de resultados

No solo la tasa de aprobación:

- parcial,
- no respaldado,
- conflictivo,
- antiguo,
- Crítico,

DESCONOCIDO,

rechazado, se examina la distribución inconclusa.

Paso 6— Verificar pesos

Pesos:

- ¿están predefinidos,
- qué pregunta responden,

¿han cambiado después del resultado?

Paso 7— Examinar diferencias de subgrupos

Se evalúa el mejor grupo, el grupo suficiente más débil, diferencia de idioma, diferencia de país, diferencia de producto y diferencia de panel.

Paso 8— Revisar alcance y deficiencia

¿Qué áreas no fueron medidas? ¿Qué celdas son insuficientes? ¿Cuáles son las razones de la deficiencia? ¿Podrían los datos faltantes ser sistemáticos?

Paso 9— Examinar la Determinación

Se evalúa el cambio entre Sesiones de Ondas, Usuarios e Interfaces.

Paso 10— Examinar Riesgo de Cola

Errores críticos y otros errores graves:

- número,
- su tasa,
- tipo,
- efecto sobre el usuario

se evalúa por separado en términos de

Paso 11— Comparar el Resumen Público con la Distribución

¿El número único o la oración publicada al público transmite con precisión la dirección y el riesgo de la distribución completa?

Paso 12— Limitar el Resultado

La afirmación pública se limita únicamente al nivel llevado por la distribución y el alcance.

28. EVIDENCIA NECESARIA

En la medida en que sea relevante para una distribución representativa o una puntuación global GEO, se deben buscar los siguientes registros:

- Registro de entidad auditada
- Definición de la población objetivo
- Marco muestral
- Producto de IA e interfaces de usuario
- Celdas de país y idioma
- prompts y versiones

- Condiciones de sesión y panel
- Olas de medición
- Números de tareas asignadas
- Tareas completadas
- Registros válidos e inválidos
- Respuestas sin procesar
- Distribución del estado de resultados
- Registros de adjudicación
- Registros de errores críticos
- Registros de rechazos e inconclusos
- Registros DESCONOCIDOS
- Pesos
- Partes interesadas
- Resultados de subgrupos
- Resultados de estabilidad
- Áreas fuera del alcance
- Cálculos de confianza o incertidumbre
- Cuadro de mando público
- Persona o institución responsable
- Versión del protocolo y análisis
- Historial de cambios

29. LISTA DE VERIFICACIÓN DE AUDITORÍA

¿Está clara la población objetivo de la distribución? ¿Están especificados los productos de IA auditados y las interfaces de usuario? ¿Son visibles el país, el idioma, el ámbito del prompt y el periodo de tiempo? ¿Se divulgan los pesos utilizados para producir la puntuación compuesta? ¿Está completamente definido el denominador? ¿Se informan los registros inválidos por separado? ¿Se preservan los rechazos y los resultados inconclusos? ¿Son distintos UNKNOWN y NOT TESTED? ¿Son visibles los resultados de los subgrupos? ¿Se han interpretado las celdas pequeñas con excesiva certeza? ¿Se ha retenido el grupo material más débil? ¿Se informan los errores críticos por separado? ¿Se muestran las olas de medición por separado? ¿Oculta la media un declive o volatilidad? ¿Se distingue una diferencia meramente formal de una diferencia material en el significado? ¿Se ha clasificado correctamente la variación sensible al contexto en las recomendaciones? ¿Se han fusionado diferentes interfaces del mismo proveedor sin explicación?

¿Se han incluido áreas fuera del alcance en el resultado global? ¿Se han separado las distribuciones sintéticas y reales? ¿Refleja la oración presentada al público la verdadera dirección de la distribución completa? Si una de las preguntas sustantivas no recibe respuesta, el resultado de la distribución tiene una confianza limitada. Si más de una no recibe respuesta, no se debe presentar un único puntuación global como juicio definitivo al público.

30. OBJECIONES Y RESPUESTAS

Objeción 1— “El público quiere un solo número.”

Esto es correcto. Un solo número puede facilitar la comunicación. NOMOS no rechaza completamente un solo número. Sin embargo, un solo número:

- alcance,
- Error crítico,
- base de subgrupo,
- tiempo,
- incertidumbre

debe estar conectado a la información. Un número puede ser corto. No puede acortar la verdad.

Objeción 2— 'Los idiomas pequeños bajan innecesariamente La puntuación global.'

El peso de la población global puede representar un idioma pequeño con peso limitado. Sin embargo, la distorsión de la representación pesada en un idioma pequeño no puede pasar desapercibida simplemente porque su participación en la población sea baja. Por lo tanto:

- puntuación principal ponderada por población,
- como resultado de la equidad lingüística separada

Pueden publicarse juntos. Dos métricas responden a diferentes preguntas.

Objeción 3— 'Si hay personalización, las respuestas diferentes son normales.'

Algunas diferencias pueden ser normales y correctas. Sin embargo, la personalización:

- identidad de empresa,
- licencia,
- su precio actual,
- el área de servicio real

debería no cambiarse arbitrariamente. La personalización puede afectar la explicación y el consejo adecuados para el usuario. No es permiso para inventar un hecho material.

Objeción 4— '¿Por qué sería un problema si no se da la misma respuesta?'

El problema no es la diferencia de palabras. El problema es el cambio de la realidad material. NOMOS no busca la igualdad exacta de texto, sino la consistencia semántica condicional.

Objeción 5— “¿Por qué importaría si un error crítico ocurre solo una vez en las respuestas 1,000?”

Este error puede ser raro en términos de prevalencia poblacional. Sin embargo, puede ser grave en términos de daño. La pérdida incontrolada de un sistema crítico en un avión una vez en mil vuelos no se evalúa como solo 0.1 por ciento. En el contexto de GEO, la frecuencia y el daño también deben examinarse por separado.

Objeción 6— “No podemos encontrar una muestra suficiente para cada país pequeño o idioma.”

Esto es cierto. Por lo tanto:

- panel de población global,
- Panel de Observadores de País,
- Panel de Equidad Lingüística

puede establecerse por separado. La observación de una sola persona no se convierte en una puntuación de país. Solo indica que existe una observación. El estado correcto para una muestra insuficiente:

MUESTRA INSUFICIENTE

o:

NO ESTIMADO

debe ser.

Objeción 7— "Este nivel de detalle es demasiado complejo para que el cliente lo entienda."

No es necesario que el cliente vea todas las fórmulas en la página principal. Sin embargo, no se pueden ocultar los hechos clave necesarios para la decisión. La tarjeta de resultados resumida debe responder a las siguientes preguntas: ¿Cuál es el resultado general? ¿Cuál es el riesgo más grave? ¿Cuál es el área más débil? ¿Qué no se ha probado? ¿Qué tan sólido es el resultado? ¿A qué fecha pertenece? La complejidad necesita ser gestionada. No eliminada.

Objeción 8— "Si se usa el mismo modelo en todos los países, ¿por qué habría una diferencia entre países?"

La salida que ve el usuario puede no consistir únicamente en el modelo básico. Las diferencias entre países incluyen:

- acceso al producto,
- recursos web,
- localización,
- idioma,
- plan,
- interfaz,
- legislación,
- personalización

Puede surgir a través de ello. Puede que no ocurra ninguna diferencia. Sin embargo, su ausencia no puede asumirse sin medición.

Objeción 9— “No es posible una puntuación fiable ya que las respuestas de la IA son variables.”

La variabilidad no hace imposible la medición. Al contrario, es una de las características que necesita ser medida. La puntuación:

- fecha específica,
- alcance,
- intervalo de confianza
- información sobre estabilidad

puede ser significativa siempre que la contenga. Sin embargo, la afirmación de certeza atemporal es indefendible.

Objeción 10— "Si el prompt cambia, la distribución cambia completamente."

Eso es correcto. Por lo tanto, la distribución depende del prompt. GEO-1000 no produce la misma respuesta para todas las preguntas posibles. Predefinido:

- Prompt de Espejo Central,
- Registro de Prompts de Diagnóstico

Mide dominios representacionales específicos. La constitución del prompt se definirá en la Sección 10.

REGLA COMÚN DE LA SECCIÓN 32

Esta sección no fue escrita en contra de una sola puntuación. Fue escrita asumiendo que una sola puntuación por sí sola no representa la totalidad de la realidad. La gente quiere números. Las instituciones quieren clasificaciones. Los consejos quieren ver una sola línea. El mercado puede preguntar: "¿Cuál es su puntuación NOMOS?" Esta pregunta es legítima. Pero la respuesta no puede ser solo: "90." Porque detrás de 90, cualquiera de las siguientes realidades puede existir:

- Precisión consistente en todos los idiomas
- 99 en algunos idiomas, 45 en otros
- Promedio alto pero error crítico severo
- Alta precisión pero baja cobertura
- Resultados instantáneos altos pero rápido deterioro
- Éxito solo en prompts fáciles
- Alta tasa solo entre respuestas evaluables
- Asumiendo positivamente los países no evaluados
- Combinando productos de IA seleccionados

El valor de una puntuación se forma por la visibilidad de la distribución subyacente. Si la distribución no es visible, la puntuación no genera confianza. Solo produce una apariencia de certeza. Por lo tanto, la segunda ley de medición de NOMOS es la siguiente:

> *Una puntuación GEO puede ser un resumen de la distribución. / No puede reemplazar la distribución.*

Su tercera ley es:

> *Un promedio alto no elimina la pregunta de quién está mal representado.*

La cuarta ley es la siguiente:

> *Un error grave y raro no es insignificante solo porque sea raro.*

La quinta ley es la siguiente:

> *El campo no probado no es el lado positivo de la distribución; es el límite desconocido.*

Orden 2 de NOMOS

> *No me muestres solo el promedio.*
>
> *Muestra quién está por debajo del promedio.*

No escondas tu peor lenguaje mientras muestras tu mejor lenguaje.

No uses el peso de los países grandes para borrar la existencia de los países pequeños.

No envíes a una persona a un país pequeño y luego anuncies una puntuación definitiva para ese país.

No trivialices un error grave simplemente porque esté oculto dentro de una respuesta que es 99 por ciento correcta.

No declares un error simplemente porque la redacción de una respuesta haya cambiado. / Pero cuando la realidad subyacente cambia, no lo descartes como una mera diferencia estilística.

No arrojes rechazos, desconocidos ni aquellos no probados fuera de la distribución.

No cambies el peso, la participación o el alcance después de ver el resultado.

Genera un número. / Pero no hagas invisibles los países, idiomas, usuarios, tiempos y errores detrás del número.

Porque la representación de una entidad no es una oración promedio. / Es una distribución de las diferentes realidades encontradas por diferentes personas.

La oración de cierre del capítulo

El objeto de medición de GEO no es una única respuesta ni un solo puntuación promedio de una inteligencia artificial; es la totalidad de los resultados representativos generados en una población de usuarios definida bajo condiciones definidas.

Core normativo

La representación del sistema generativo de una entidad DEBE tratarse como una distribución condicional de resultados observables en los siguientes ámbitos definidos: - productos de IA y interfaces de usuario, - usuarios o unidades de muestreo, - países, - idiomas y configuraciones regionales, - estados de sesión y de personalización, - prompts, - y momentos de medición. Una única puntuación agregada NO DEBE reemplazar la divulgación de: - alcance y cobertura, - distribución de resultados, - variación entre subgrupos, - incidencia de errores críticos, - masa desconocida y no evaluada, - estabilidad temporal, - reglas de ponderación, - y incertidumbre. Los promedios iguales NO DEBEN interpretarse como una calidad de representación equivalente cuando sus distribuciones por subgrupo, severidad, cobertura o estabilidad difieren materialmente. Los grupos no evaluados NO DEBEN ser imputados como conformes, y los resultados raros de alto daño DEBEN permanecer visibles por separado del rendimiento promedio.

CAPÍTULO 03

¿Qué entidad se audita exactamente?

BASES METODOLÓGICAS / K10; K11

Límite de Capítulo

La Sección 1 estableció la siguiente disposición:

> *Una única respuesta no es una puntuación GEO.*

La Sección 2 añadió la siguiente disposición:

> *La representación de una entidad en sistemas generativos es una distribución que ocurre entre usuarios, productos de IA, países, idiomas, sesiones y tiempos.*

Esta sección responde la siguiente pregunta obligatoria:

> *¿De quién o de qué representación mide exactamente esta distribución?*

¿Una empresa? ¿Una marca? ¿Un nombre de dominio? ¿Un producto? ¿Un servicio? ¿Un fundador? ¿Un grupo de empresas? ¿Una entidad legal local en un país específico? ¿Un negocio franquiciado? Cuando un sistema de inteligencia artificial ve solo el dominio apple.com:

- el nombre de dominio,
- la marca,
- la empresa matriz,
- la entidad de ventas local,
- la familia de productos,
- el servicio de la tienda

¿Tratará el sistema a todos ellos como la misma entidad? A menos que se haga la distinción antes de que comience la auditoría, un resultado puede calcularse correctamente pero aplicarse al objeto incorrecto. Este capítulo:

- la entidad auditada,
- el ancla de identidad utilizada en la solicitud,
- la entidad analizada por el sistema,
- la unidad a la que se aplicará la decisión de conformidad,
- relaciones entre entidades,
- a qué entidad pertenecen los atributos,
- límites temporales y geográficos de identidad

define. Este capítulo aún no:

- cómo se verificarán todas las afirmaciones sobre la entidad,
- cómo los adjudicadores puntuarán las afirmaciones atómicas,
- cuotas de muestra por país e idioma,
- la fórmula de puntuación final NOMOS

no está finalizada. La tarea de la Sección 3 es más fundamental:

> *Fijar la identidad y los límites del objeto que se está midiendo antes de que comience la medición.*

NOMOS Desafío

Me estás haciendo esta pregunta: "¿Qué tipo de empresa es Apple.com?" Puede haber al menos tres objetos diferentes en la pregunta:

- el nombre de dominio apple.com
- la marca Apple
- la empresa legal asociada con el nombre de dominio y la marca

Elegiría uno de estos. Tal vez el correcto. Tal vez el incorrecto. Tal vez combine más de uno en una sola respuesta. Entonces calificarías la respuesta. Pero no me dijiste de antemano qué entidad estás calificando. Puedo haber:

- descrito el dominio como una tienda en línea,
- descrito la marca como una marca de electrónica de consumo,
- descrito la empresa legal como una corporación global,

También puedo haber listado los productos como si la propia empresa los estuviera presentando. Mi respuesta puede parecer fluida. Pero si el objeto de la auditoría no está claro, si es correcto o no también sigue siendo incierto. Ahora, considera un ejemplo más complejo.

Una marca es operada por la empresa matriz en un país; por una subsidiaria en otro país; y por un distribuidor independiente en un tercer país. Un usuario pregunta: "¿Esta marca ofrece servicios en Turquía?" ¿Qué entidad debería responder? ¿El sitio global de la marca? ¿La empresa local en Turquía? ¿El distribuidor?

¿El operador de la franquicia? ¿La licencia propiedad de la empresa global se transfiere automáticamente a la franquicia local? ¿Las opiniones de los clientes sobre la operación local prueban la calidad global de la marca principal? ¿Puede usarse el número de empleados de la empresa matriz como la capacidad de la empresa local?

¿Puede el premio de un producto transferirse a todas las empresas que lo venden? ¿La experiencia personal del fundador se convierte en la antigüedad corporativa de la empresa? ¿Se considera que el propietario anterior y el nuevo propietario de un nombre de dominio son la misma entidad?

Cuando una empresa adquiere otra empresa, ¿todas las realizaciones de la empresa anterior se convierten en una característica de la nueva empresa? La puntuación GEO se puede generar sin responder estas preguntas. Sin embargo, no se sabe qué representa la puntuación. La primera resolución de esta sección es la siguiente:

> *No se puede medir una entidad hasta que se resuelva.*

Su segunda disposición establece:

> *Compartir el mismo nombre no es ser la misma entidad.*

Su tercera disposición establece:

> *Una relación no permite automáticamente que las propiedades se transfieran entre las partes.*

1. OBJETIVO DEL CAPÍTULO

El propósito de esta sección es determinar la identidad, el límite y las relaciones de la entidad medida dentro de GEO-1000 antes de ver los resultados. La sección hace las siguientes distinciones normativas:

- Nombre de dominio y empresa legal
- Marca y propietario de la marca
- Empresa matriz y subsidiaria
- Grupo y empresa del grupo
- Producto y fabricante
- Servicio y proveedor de servicios
- Fundador e institución
- Empleado e institución
- Titular de la licencia y la parte que utiliza la marca
- Franquiciador y operador de la franquicia
- Fabricante y distribuidor
- Plataforma y vendedor en la plataforma
- Proveedor de tecnología y usuario de tecnología
- Cliente y socio comercial
- Relación activa y relación histórica
- Nombre oficial y traducción descriptiva
- Entidades diferentes con el mismo nombre
- Versiones anteriores y posteriores de una entidad
- Objeto auditado y objeto al que se aplicará la decisión de conformidad

Al final de esta sección, cada auditoría GEO debe ser capaz de proporcionar una respuesta única y trazable a la siguiente pregunta:

> *¿Exactamente a qué entidad, con qué relaciones, para qué período y para qué interfaces digitales pertenece esta puntuación y decisión?*

2. DISPOSICIÓN NORMATIVA CENTRAL

> *Antes de que comience la recopilación de datos, cada auditoría GEO-1000 debe definir una entidad auditada única, un registro canónico de entidad, un límite de alcance, un límite de tiempo y un grafo de relaciones.*

Ninguno de los siguientes, por sí solo, es suficiente para identificar una entidad:

- nombre,
- logo,
- nombre de dominio,
- cuenta de redes sociales

Estos identificadores por sí solos no definen el objeto de la auditoría. La entidad auditada debe describirse, cuando sea relevante, al menos a través de los siguientes elementos:

- Tipo de entidad
- Nombre canónico

- Nombre legal
- Nombres de marca
- Nombres localizados oficiales
- Abreviaturas y alias
- Fecha de establecimiento o inicio
- Jurisdicción
- Número de registro o identidad oficial equivalente
- Relaciones entre la empresa matriz y subsidiarias
- Propiedad de marcas registradas
- Control de nombres de dominio
- Propiedad de productos y servicios
- Responsable de la prestación efectiva del servicio
- Reclamante
- Titular de la evidencia
- Países y localidades activos
- Periodo de validez
- Nombres y relaciones históricas
- Interfaces incluidas en la auditoría
- Entidades y interfaces fuera del alcance

La puntuación GEO principal no debe publicarse hasta que este registro esté completado.

3. ¿QUÉ ES UNA ENTIDAD?

En este protocolo, una entidad es un objeto al que se le puede asignar una identidad, característica, relación, responsabilidad o estado temporal separable y rastreable. Una entidad:

- humano,
- persona jurídica,
- institución pública,
- marca,
- producto,
- servicio,
- nombre de dominio,
- aplicación digital,
- grupo de empresas,
- sucursal física,
- programa,
- certificado,
- publicación,
- proyecto de investigación

Es posible. No todos los elementos mencionados deben ser una entidad separada. Sin embargo, si una de las siguientes áreas cambia de manera significativa, puede ser necesario un registro de entidad separado:

- Responsabilidad legal

- Propiedad
- Licencia
- Entrega de producto o servicio
- Jurisdicción
- Periodo histórico
- Obligación hacia el usuario
- Evidencia y propiedad de la afirmación
- Alcance de la auditoría

4. SEIS OBJETOS SEPARADOS EN GEOAUDITORÍA

En una auditoría GEO, la palabra "entidad" suele ocultar seis funciones separadas. Estos seis objetos deben distinguirse entre sí.

4.1. Presentador de prompts

Un ancla de prompt es una expresión o signo utilizado para invocar la existencia en la consulta de un usuario. Ejemplos:

- Nombre de la empresa
- Nombre de la marca
- Nombre de dominio
- Nombre fundador
- Nombre del producto
- Nombre de la aplicación
- Nombre local de la empresa
- Abreviatura

Ejemplo: “¿Qué tipo de empresa es Apple.com?” el ancla del prompt aquí: apple.com es el nombre de dominio. El ancla del prompt no tiene que ser la entidad en sí. Es solo el punto de entrada utilizado por el sistema para resolver la entidad.

4.2. Entidad Objetivo

La entidad objetivo es la entidad que el protocolo realmente quiere medir. Ejemplo: se puede dirigir la entidad corporativa canónica asociada con el nombre de dominio apple.com. Sin embargo, su identidad exacta debe registrarse antes de la auditoría. La entidad objetivo no puede seleccionarse mirando los resultados después de la consulta.

4.3. Entidad Resuelta por el Sistema

Una entidad resuelta es cómo el producto de IA interpreta el ancla de la consulta como una entidad. Representación:

```
E_i
```

se puede usar. La entidad objetivo:

E*

es la evaluación de resolución de identidad que plantea la pregunta fundamental:

```
E_i= E*?
```

Sin embargo, la igualdad no siempre es solo igualdad de nombre. El sistema puede identificar la marca y la entidad corporativa correctas sin nombrar la entidad legal correcta. Por lo tanto, la coincidencia de identidad más adelante:

- coincidencia exacta,
- coincidencia de marca aceptable,
- coincidencia parcial,
- entidad incorrecta,
- incierto,
- fusión de múltiples entidades

puede clasificarse como.

4.4. Objeto de la afirmación

El objeto de la afirmación es la entidad a la que pertenece una característica específica en la respuesta. Dentro de la misma respuesta:

- fecha de establecimiento de la empresa matriz,
- característica del producto al producto,
- reseña del cliente para proveedor de servicios local,
- licencia a una entidad legal específica,
- experiencia del fundador a una persona

puede pertenecer. Incluso si la respuesta comenzó con la entidad objetivo correcta, puede transferir características a objetos incorrectos. Por lo tanto, cada afirmación factual debe evaluarse en la siguiente estructura:

```
(sujeto, relación o atributo, valor, tiempo, alcance)
```

Ejemplo: “X tiene una licencia Y.” Aquí, se debe determinar a cuál:

- persona jurídica,
- país,
- durante qué fechas,
- para la actividad

pertenece.

4.5. Objeto de Auditoría

El objeto de auditoría es una entidad definida o un paquete de entidades donde se examinan conjuntamente la evidencia, el contenido, la interfaz técnica y los resultados de la IA. Un objeto de auditoría puede ser una sola entidad. Ejemplo: Una empresa legal específica. O puede ser un paquete con límites claros: Marca principal + nombre de dominio canónico + dos productos especificados + interfaces públicas en turco e inglés. Si se utiliza un paquete, las entidades dentro del mismo deben mantener sus identidades separadas. "Paquete" no significa que todos los atributos se transfieran entre sí.

4.6. Unidad de Conformidad

La unidad de conformidad es el objeto específico al que se aplicará legal y normativamente la decisión de auditoría. Una decisión de conformidad:

- se aplica a todo el grupo de empresas,
- se aplica a todos los lenguajes,
- se aplica a todos los nombres de dominio,

- a todos los productos

no puede aplicarse automáticamente. Ejemplo: "Las páginas públicas turcas e inglesas de la marca NobleJackal en example.com han sido consideradas conformes según el estándar especificado y dentro del rango de fechas especificado." Esta decisión:

- todos los procesos internos de la empresa,
- empresas afiliadas,
- aplicaciones separadas,
- lenguas que se añadirán en el futuro,
- Productos fuera del ámbito

no cubre automáticamente.

5. TIPOS DE ENTIDADES

Cada entidad de auditoría debe clasificarse con uno o más tipos. Se pueden distinguir tipos primarios y tipos secundarios.

5.1. Entidad Legal

Es una empresa que posee derechos, responsabilidades o registro separados bajo la ley o regulación legal relevante. Sus campos, en la medida en que sean relevantes:

- Nombre legal
- Tipo de empresa
- Lugar de registro
- Número de registro
- Identificación fiscal o equivalente
- Fecha de constitución
- Estado activo
- Representante autorizado
- Relaciones de matriz y subsidiarias

Una entidad legal puede tener el mismo nombre que la marca. Sin embargo, no es el mismo objeto.

5.2. Marca

Es el nombre utilizado en el mercado para un producto, servicio o identidad corporativa. Marca:

- puede no tener una entidad legal separada,
- puede ser propiedad de otra empresa,
- puede ser utilizado bajo licencia,

puede ser operado por múltiples empresas en diferentes países. También debe especificarse la entrega real responsable del trabajo reclamado por una marca.

5.3. Grupo de Empresas

Es una estructura formada por múltiples entidades legales que tienen relaciones de propiedad conjunta o control. Un “Grupo” no puede transferir automáticamente las siguientes características a todos los miembros:

- Licencia
- Certificado

- Número de clientes
- Número de empleados
- Ingresos
- Oficina
- Resultado de auditoría
- Estado de seguridad
- Marca de conformidad

Si el valor a nivel de grupo está realmente consolidado, debe indicarse claramente.

5.4. Subsidiaria o Asociada

Una entidad legal que está bajo la propiedad o relación de control de otra entidad. Una subsidiaria:

- puede ser una parte contratante separada,
- un titular de licencia separado,
- un empleador separado,
- un responsable de datos separado

puede ser. Las características de la empresa matriz no se transfieren por defecto a la subsidiaria.

5.5. Nombre de Dominio

Es una dirección digital y una interfaz de publicación. Nombre de dominio:

- no es una entidad legal,
- no presta servicios por sí mismo,
- no se requiere que posea una marca registrada,

puede ser controlado por diferentes propietarios u operadores en ciertos momentos. Un nombre de dominio puede ser el punto de partida de una auditoría. Sin embargo, también se debe mostrar bajo qué entidad se tomó la decisión.

5.6. Sitio web o publicación digital

Es la totalidad de contenido y aplicaciones específicas en un nombre de dominio. El mismo nombre de dominio puede ser:

- sitio corporativo,
- tienda,
- portal de soporte,
- página para inversores,
- documentación para desarrolladores,
- versiones de diferentes países e idiomas

pueden existir. Los propietarios del contenido y los ámbitos de supervisión de estos pueden diferir.

5.7. Producto

Un objeto comercial o técnico que cumple con una función específica, modelo o versión. Un producto tiene:

- precio,
- seguridad,
- característica técnica,
- certificado,

- idoneidad para el usuario

que no pueden utilizarse como características generales pertenecientes a la empresa. Las versiones dentro de la misma familia de productos también pueden diferir.

5.8. Servicio

Actividad ofrecida bajo un alcance específico, usuario, país, método de entrega y responsabilidad. Un servicio:

- comercializable por la marca,
- facturable por otra entidad legal,

entregable por un socio independiente. La auditoría debe distinguir quién:

- lo vendió,
- lo entregó,
- lo garantizó,
- lo asumió legalmente

debe ser separado.

5.9. Persona

Es fundador, ejecutivo, empleado, consultor, experto o representante reconocido públicamente. La:

- educación de una persona,
- experiencia,
- licencia,
- opinión

No se transfiere automáticamente a la institución. Asimismo, el premio o certificado de la institución no se considera una competencia profesional de la persona.

5.10. Sucursal o Ubicación Física

Es un punto de servicio físico que está afiliado a la institución principal o que opera de manera independiente. Una sucursal:

- parte de la empresa matriz,
- entidad legal separada,
- franquicia,
- operador independiente con licencia

puede ser. La relación de posición no muestra únicamente la relación legal.

5.11. Franquicia

Puede ser un negocio independiente que opera bajo un derecho de uso de marca o licencia de modelo de negocio. El franquiciador debe:

- reputación,
- control,
- capacitación

no garantiza automáticamente cada acción del operador de la franquicia. Un error local del operador de la franquicia no puede generalizarse automáticamente a toda la marca global. La relación debe reportarse con su alcance.

5.12. Distribuidor, Concesionario o Vendedor

La parte que vende o distribuye el producto o servicio en un mercado específico, ya sea independiente o afiliada. La licencia del fabricante, el control o el cumplimiento corporativo no se transfieren automáticamente al distribuidor. El desempeño del servicio al cliente del vendedor tampoco es la misma métrica que la calidad del producto del fabricante.

5.13. Plataforma y Vendedor de la Plataforma

El mercado o la plataforma y los vendedores independientes en la plataforma son entidades separadas. Plataforma:

- pago,
- listado,
- infraestructura técnica

pueden ser proporcionados. Vendedor:

- producto,
- entrega,
- garantía,
- relaciones con clientes

pueden ser responsables. Si la IA combina estos roles, puede ocurrir una asignación incorrecta de responsabilidades.

5.14. Programa, Insignia o Certificado

Ofrecido por una institución:

- programa de formación,
- marca de conformidad,
- certificado,
- membresía,
- insignia

Puede contener registros de entidades separados. La institución que publica el programa no es la misma que la institución que participa en el programa. La característica de la insignia no se traduce en todas las características de la institución que posee la insignia.

6. GRÁFICO DE ENTIDADES

Un solo registro de entidad puede no ser suficiente para explicar relaciones complejas. Por esta razón, se debe crear un gráfico de entidades para cada auditoría. Vistas:

G_E = (V, R)

Aquí:

- V: nodos de entidad
- R: relaciones direccionales entre nodos

Las relaciones deben definirse mediante verbos explícitos.

6.1. Tipos Básicos de Relaciones

En la medida en que sea relevante, se pueden usar las siguientes relaciones:

- posee — posee
- ownedBy— es propiedad de
- controla — controla
- controlledBy— está controlado por
- opera — opera
- operatedBy— es operado por
- publica — publica
- publishedBy— es publicado por
- fabrica — fabrica
- manufacturedBy— es fabricado por
- providesService— proporciona servicio
- serviceDeliveredBy— el servicio es entregado por
- licencesBrandTo— licencia la marca a
- franchisesTo— otorga franquicias a
- distributesFor— distribuye para
- vende — vende
- iemplea — emplea
- foundedBy— fue fundado por
- memberOf— es miembro de
- certifiedBy— está certificado en el alcance especificado por
- auditedBy— ha sido auditado por
- partneredWith— tiene una asociación en el alcance especificado con
- customerOf— es cliente de
- supplierTo— es proveedor de
- formerlyKnownAs— anteriormente conocido como
- predecessorOf— es un predecesor de
- successorTo— es un sucesor de
- acquiredBy— ha sido adquirido por
- mergedInto— se ha fusionado con
- ceasedOn— ha cesado operaciones en

Los verbos vagos como 'relacionado', 'socio', 'trabajando juntos' no deben usarse solos tanto como sea posible.

6.2. Dirección de la Relación

Las relaciones son direccionales. 'A proporciona tecnología a B.' no es lo mismo que: 'B proporciona tecnología a A.' Una marca puede pertenecer a una empresa. La empresa no pertenece a la marca. Un cliente puede haber trabajado con una agencia. La agencia no pertenece al cliente. Si la dirección no se especifica, el sistema puede realizar transferencias incorrectas de responsabilidad y autoridad.

6.3. Tiempo de la Relación

Toda relación financiera en la medida en que sea relevante:

- fecha de inicio,
- fecha de finalización,

- estado de la actividad,
- fecha de última verificación

debe registrarse. Una antigua asociación no es una asociación activa. Un exempleado no es un ejecutivo actual. Un distribuidor que se haya utilizado una vez puede no ser un representante autorizado hoy.

6.4. Alcance de la Relación

Una asociación:

- puede estar limitada a un producto específico,
- país,
- campaña,
- fecha,
- evento

No se puede generalizar que una colaboración de un solo proyecto se aplique a todas las actividades corporativas.

7. SE PROHÍBE LA TRANSFERENCIA SUPUESTA DE CARACTERÍSTICAS

La existencia de una relación entre entidades no transfiere automáticamente una característica de una a otra. Uno de los principios fundamentales de esta sección es:

> *La transferencia de características está prohibida por defecto; sin embargo, puede permitirse con relación explícita, alcance, tiempo y evidencia.*

7.1. Transferencia de licencia

La licencia de la empresa matriz:

- a una subsidiaria,
- a una franquicia,
- a un distribuidor,
- a una marca

no se transfieren automáticamente. Se debe examinar el alcance oficial de la licencia.

7.2. Transferencia de Certificado

El certificado de un producto no puede transferirse a toda la familia de productos. El certificado de calidad de una oficina no puede generalizarse al grupo global. La decisión de conformidad de una empresa no prueba la experiencia personal de sus empleados.

7.3. Transferencia de Número de Empleado

El total del grupo no puede utilizarse como el número de empleados de una empresa local. La red de agencias o la bolsa de freelancers no pueden mostrarse como el número de empleados directos.

7.4. Transferencia de Número de Clientes

El número de clientes del grupo principal no puede escribirse como el número de clientes de una marca recién establecida. El número de usuarios en una plataforma puede no corresponder al número de clientes activos de un producto específico.

7.5. Transferencia de Fecha de Establecimiento

El año de inicio del fundador en la industria no es la fecha de establecimiento de la empresa. Una empresa no puede usar la fecha de fundación de una marca antigua que adquiere como su propia fecha legal de establecimiento. La fecha de la marca y la fecha de la entidad legal deben mostrarse por separado.

7.6. Éxito y transferencia del caso

Un proyecto realizado por un empleado en un empleador anterior no es el estudio de caso corporativo de la nueva empresa. El éxito de una empresa adquirida, sin explicación de integración y alcance, no se considera el éxito directo del nuevo grupo.

7.7. Transferencia de Elegibilidad

Elegibilidad otorgada para ciertas páginas de idioma de un nombre de dominio:

- a todo el grupo de la empresa,
- a nuevos productos,
- a otros nombres de dominio,
- a páginas futuras

no se transfiere.

8. CUATRO CAPAS SEPARADAS DE PROPIEDAD

En una representación digital, la palabra "propietario" puede tener múltiples significados. Se deben distinguir cuatro capas.

8.1. Propiedad Legal

La entidad es el propietario según la ley o el registro oficial.

8.2. Propiedad de Marca

Es el propietario de los derechos sobre la marca o la identidad comercial.

8.3. Control de la Interfaz Digital

Es la parte que controla realmente el nombre de dominio, sitio web, perfil social o aplicación. El propietario legal y el controlador técnico pueden ser diferentes.

8.4. Propiedad de Contenido y afirmaciones

Es la parte responsable de la exactitud de la afirmación del material publicado. La agencia puede gestionar el sitio. Sin embargo, el verdadero propietario del precio o de la afirmación de licencia puede ser la institución cliente. Un proveedor técnico no se convierte en el propietario material de la afirmación solo por haber publicado el contenido.

9. ¿QUIÉN VENDE EL SERVICIO, QUIÉN LO ENTREGA?

Esta es una de las distinciones de entidad más importantes en la evaluación GEO. En una cadena de servicios, las siguientes partes pueden ser diferentes:

- El que usa la marca
- El que anuncia

- El que prepara la oferta
- El que firma el contrato
- El que emite la factura
- El que realmente entrega el servicio
- El que proporciona la garantía
- El que procesa los datos
- El que maneja la queja
- El que asume la responsabilidad legal

Una IA puede asignar todos estos roles a una sola parte simplemente mencionando el nombre de la marca. Este error es especialmente común en:

- turismo,
- finanzas,
- salud,
- consultoría,
- franquicia,
- plataforma,
- agencia,
- educación

los campos pueden tener consecuencias financieras. Cada registro de servicio debe responder a estas preguntas: ¿Con quién contrata el usuario? ¿Quién recibe el pago? ¿Quién realiza el servicio? ¿Quién asume la responsabilidad profesional o legal? ¿Quién es responsable de las obligaciones de garantía y devolución? ¿Cuál es el papel de la marca? ¿Existen proveedores terceros? Esta distinción también determina a qué entidad se aplica La puntuación GEO.

10. LÍMITE DE ENTIDAD

El límite de la entidad es la línea que define qué características, relaciones y interfaces incluye el objeto auditado y cuáles no. Un límite de entidad puede definirse en cuatro capas.

10.1. Núcleo de Identidad

Es la identidad fundamental de la entidad que debe permanecer inalterada:

- Nombre canónico
- Tipo de entidad
- Identidad legal o institucional
- Actividad principal
- Propiedad primaria
- Estado activo

10.2. Entorno controlado

Controlado directamente por la entidad:

- nombres de dominio,
- aplicaciones,
- marcas,
- productos,

- documentos,
- perfiles oficiales

Este campo. Control no significa que todas las características se fusionen en una sola identidad.

10.3. Entorno conectado

Afiliados son empresas locales, franquicias y socios controlados. Mantienen sus identidades separadas.

10.4. Entorno de relaciones externas

Clientes, proveedores, medios de comunicación, membresías, proveedores de tecnología y socios independientes. Las entidades en este entorno no son parte de la entidad auditada. Son objetos de la relación.

11. BLOQUEANDO EL LÍMITE DE LA ENTIDAD SIN VER RESULTADOS

El objeto de la auditoría no puede ampliarse ni reducirse según los resultados. Los siguientes comportamientos están prohibidos:

- Agregar resultados positivos de grupo a la marca
- Excluir una empresa afiliada que haya fallado del alcance
- Extender un buen resultado de producto a toda la familia de productos
- Declarar la versión en mal estado del idioma como una entidad separada
- Dejar el error crítico a la operación local mientras se hace que la marca global parezca limpia
- Hacer que la evidencia local sólida sea la evidencia de la empresa global

Alcance de la entidad:

- Antes de que comience la auditoría,
- Antes de que se distribuyan los indicios,
- Antes de que se vean los resultados

Debe ser versionado primero.

12. REGISTRO DE ENTIDAD CANÓNICA

Se debe crear un Registro de Entidad Canónica para cada entidad auditada. Este registro aclara qué objeto están evaluando los adjudicadores y los sistemas.

12.1. Campos Mínimos

Identidad

ID de entidad NOMOS Nombre canónico Tipo de entidad Nombre legal Abreviaturas Nombres locales oficiales Nombres anteriores

Estado Legal y Corporativo

Jurisdicción Número de registro Fecha de constitución Estado activo Empresa matriz Filiales

Marcas y Productos

Marcas propias Marcas licenciadas Productos Servicios Distinciones de versión o modelo

Interfaces Digitales

Nombres de dominio canónicos Nombres de dominio locales Aplicaciones Perfiles oficiales Registros legibles por máquina

Responsabilidad

Reclamante Propietario del contenido Propietario de la interfaz técnica Propietario de la prestación del servicio Responsable legal

Hora

Versión del registro Inicio de vigencia Fin de validez Fecha de última verificación Cambios históricos

Alcance

Países auditados Idiomas auditados Productos auditados Entidades y interfaces fuera del ámbito

12.2. Estados de la entidad

Cada registro de entidad debe tener uno de los siguientes estados:

ACTIVO

INACTIVO

HISTÓRICO

SUSTITUIDO

FUSIONADO

ADQUIRIDO

DISUELTO

PROVISIONAL

IMPUGNADO

DESCONOCIDO

La identidad DESCONOCIDA o EN CONFLICTO no puede resolverse de manera definitiva de forma silenciosa.

13. NIVELES DE PRUEBA DE IDENTIDAD

La garantía de identidad de un registro de entidad puede clasificarse según las fuentes utilizadas.

ID-0— Identidad Desconocida

Solo existe un nombre o un nombre de dominio. No se verifica la correspondencia legal u organizativa.

ID-1— Afirmación de Primera Parte

La institución se identifica de manera específica. Es valioso. No es una verificación independiente.

ID-2— Identidad Organizacional Respaldada

Los registros de primera parte, el nombre de dominio y otros registros confiables son consistentes entre sí.

ID-3— Identidad verificada con registro oficial

Se ha verificado la entidad legal con registros oficiales o confiables apropiados.

ID-4— Identidad verificada en múltiples interfaces y relaciones

La identidad, la marca, el producto, el nombre de dominio y las relaciones corporativas son consistentes en múltiples registros confiables.

ID-5— Identidad Mantenida en el Tiempo

La identidad y las relaciones han sido verificadas en diferentes periodos, y se conserva el historial de cambios. Estos niveles no son la puntuación final de elegibilidad. Indican la confianza en la evidencia del registro de la entidad.

14. ENTIDADES DIFERENTES CON EL MISMO NOMBRE

Pueden existir diferentes entidades con el mismo nombre o nombres similares. Esta situación crea colisiones de entidades. Ejemplos de distinciones:

- Empresas con el mismo nombre
- Instituciones que usan la misma abreviatura
- El mismo nombre de marca en diferentes países
- Compañía antigua y nueva
- Nombre de marca con nombre personal
- Nombre de empresa con nombre de producto
- Marca con palabra general

La auditoría no puede depender únicamente de la coincidencia de nombres.

14.1. Campos del Resolutor de Colisiones

Jurisdicción Tipo de empresa legal Nombre de dominio Dirección física Fecha de establecimiento Fundador o gerente Producto y servicio Propietario de la marca Número de registro Nombre en idioma local

14.2. afirmación ambiguo

Si una afirmación apunta a múltiples entidades razonables, hay tres opciones: La afirmación se aclara. Se clasifica por separado como una prueba de resolución de entidades. Se espera que el sistema solicite aclaración. La respuesta obtenida de la afirmación ambiguo no puede agregarse silenciosamente a la puntuación principal de precisión.

15. UNA ENTIDAD, MÁS DE UN NOMBRE DE DOMINIO

Una empresa:

- dominio global,
- dominios locales del país,

- dominios de producto,
- dominio del inversor,
- dominio de soporte,
- dominios antiguos

puede usar. No todos estos dominios caen automáticamente bajo el mismo alcance de auditoría. Para cada dominio:

- controlador,
- propietario del contenido,
- usuario objetivo,
- idioma,
- alcance del producto,
- estado de la actividad

Debe determinarse. Un registro correcto en un dominio no corrige automáticamente el error en otro dominio.

16. UN NOMBRE DE DOMINIO, MÁS DE UNA ENTIDAD

En un solo nombre de dominio:

- empresa matriz,
- diferentes marcas,
- empresas locales,
- productos,
- socios,
- vendedores

Puede ocurrir. En este caso, la propiedad de la entidad debe definirse en función del URL o la familia de contenido. Ejemplo:

Interfaz	Rol de existencia
Página de inicio	Marca global
Relaciones con inversores	Empresa matriz legal
Página de precios local	Organización de ventas local
Mercado	Vendedores independientes
Página de carrera	Entidades empleadoras especificadas
Portal de soporte	Operador de producto o servicio

Un único nombre de dominio no significa única responsabilidad.

17. IDENTIDAD DE ENTIDAD TEMPORAL

La identidad de la entidad puede cambiar con el tiempo. Ejemplos de casos:

- Cambio de nombre de la empresa
- Marca transferida a otra empresa
- Fusión

- Adquisición
- División
- Cese de actividad
- Cambio de nombre de dominio
- Transferencia del producto a otra empresa
- Salida del fundador
- Cambio de operador local

Por esta razón, el registro de la entidad depende del tiempo. Se puede considerar como E(t).

17.1. Identidad Histórica e Identidad Actual

Una relación que antes era correcta no puede trasladarse al presente. Ejemplo: "X fue operado por Y en 2022." es histórico. La siguiente frase es diferente: "X es operado por Y."

17.2. Transferencia de funcionalidades tras la adquisición

Contratación:

- Propiedad,
- algunas entidades,
- pueden transferir ciertos contratos

. Sin embargo:

- todos los logros históricos,
- todos los certificados,
- todas las relaciones con los clientes,
- Todas las licencias

no se transfieren automáticamente a la nueva entidad. Debe explicarse qué características se transfieren legal y prácticamente.

17.3. Historial de nombres de dominio

Un nombre de dominio puede contener rastros de contenido o referencias del propietario anterior. El nuevo propietario no se considera automáticamente como si hubiera adquirido la autoridad anterior. Si el sistema de IA fusiona las entidades antigua y nueva, se produce confusión histórica de entidades.

18. IDENTIDAD MULTILINGÜE DE ENTIDADES

Los nombres de una entidad en diferentes idiomas deben estar vinculados a la misma identidad. Sin embargo, no todas las traducciones son nombres oficiales. Se deben distinguir los siguientes tipos:

- Nombre legal
- Nombre de la marca
- Nombre oficial localizado
- Transliteración
- Traducción descriptiva
- Abreviatura
- Nombre anterior
- Nombre de uso común

Ejemplo: Una descripción en turco de una marca no tiene que ser el nombre legal u oficial en turco de la marca. El registro de la máquina debe indicar claramente estos tipos.

19. EL PAÍS Y LA JURISDICCIÓN PUEDEN CAMBIAR EL LÍMITE DE LA ENTIDAD

Bajo una marca global, puede haber diferentes entidades legales u operativas en distintos países. Cuando un usuario pregunta: "¿Opera la Empresa X en Alemania?" la respuesta debería distinguir:

- la existencia de la marca global,
- la entidad legal local,
- la relación con el distribuidor,
- la capacidad de proporcionar servicios remotos

"La marca es reconocida en el país" no es lo mismo que: "La empresa legal local proporciona servicios."

19.1. Roles de Entidades según el País

Para cada país, en la medida de lo aplicable:

- Vendedor legal
- Emisor de facturas
- Responsable de los datos
- Proveedor de servicios
- De apoyo
- Distribuidor
- Franquicia
- Titular de licencia
- Representante de la marca

debe ser determinado.

19.2. La página general del idioma no es prueba de una empresa local

Tener un sitio web en alemán no prueba que exista una entidad legal o una oficina física en Alemania. Una página en inglés no crea jurisdicción legal global. El idioma no es una jurisdicción.

20. INDICACIÓN DE PRESENCIA SINTÉTICA APPLE.COM

DEMOSTRACIÓN DE METODOLOGÍA SINTÉTICA / El siguiente ejemplo ha sido preparado únicamente para demostrar el método de resolución de entidades. No es la auditoría real de Apple Inc., una revisión actual de la estructura corporativa, ni el resultado real de ningún producto de IA. prompts: "¿Qué tipo de empresa es Apple.com?"

20.1. Ancla de prompt

apple.com

20.2. Objetivos de los Candidatos

Nombre de dominio y sitio web Marca Apple Entidad corporativa Canonical Función de tienda en línea Familia de productos Si el objetivo no se define antes de que comience la auditoría, los jueces pueden puntuar objetos diferentes.

20.3. Propuesta de Fijación de Entidad Objetivo

El objetivo para pruebas sintéticas se puede definir de la siguiente manera: "la principal entidad corporativa tecnológica asociada con el nombre de dominio canónico apple.com, que representa la identidad y actividades principales de la empresa." Esta definición:

- no abarca automáticamente las especificaciones técnicas de productos individuales,
- la empresa local de cada país,
- los precios de las tiendas,
- consejos de inversión

automáticamente.

20.4. Respuesta Sintética A

"Apple.com es un sitio web corporativo global asociado con productos tecnológicos y servicios digitales." Evaluación: No coincide completamente con la institución del nombre de dominio. Puede dejar incompleta la identidad de la empresa. La actividad principal podría ser generalmente correcta. Puede haber una falta de identidad material.

20.5. Respuesta Sintética B

"Apple.com es el sitio web oficial de la empresa llamada iPhone." Evaluación: El producto y la empresa están combinados. La entidad objetivo se resuelve incorrectamente. Hay un error de identidad.

20.6. Respuesta Sintética C

"Apple.com es solo una tienda independiente que vende teléfonos en línea." Evaluación: La función del nombre de dominio está excesivamente limitada. La entidad corporativa y el alcance de las actividades están distorsionados. La característica de tienda independiente puede ser incorrecta.

20.7. Respuesta Sintética D

"Apple.com es una de las principales interfaces de información corporativa y de productos de la marca Apple; la entidad legal que opera el sitio y el lado de ventas local también deberían determinarse según el país." Evaluación: El nombre de dominio, la marca y el rol legal local están separados. El análisis de la entidad objetivo es más disciplinado. Otras afirmaciones materiales de la respuesta deberían demostrarse por separado. Este ejemplo muestra:

> *Una respuesta puede centrarse en el entidad incorrecta incluso si contiene muchas palabras correctas.*

La precisión de la identidad viene antes que la precisión del contenido.

21. CASO DE ENTIDAD COMPLEJA SINTÉTICA

CASO SINTÉTICO — NO ES UNA INSTITUCIÓN REAL

Consideremos una marca global llamada Asteron. Estructura:

- Asteron Holdings Ltd. — empresa matriz

- Asteron Travel — marca
- Asteron Turkey Tourism Ltd. — compañía local
- Mira Destination Services — socio local de operaciones independiente
- Asteron Club Antalya — operación de franquicia
- asteron.example — nombre de dominio global
- tr.asteron.example — interfaz de marketing en Turquía

El usuario pregunta: "¿Qué servicios ofrece Asteron en Turquía?" Respuesta de la IA: "Asteron Holdings Ltd. opera todos los tours en Antalya, tiene 400 empleados en Turquía, y Asteron Club Antalya es directamente responsable de la satisfacción del cliente en Antalya." Paquete de realidad: La empresa matriz posee la marca. La empresa local es la encargada de ventas y contratación. Mira realiza parte de las operaciones. Asteron Club Antalya es una franquicia independiente. Los 400 empleados son el total del grupo. La franquicia es directamente responsable de la satisfacción del cliente a nivel comercial. Errores en la respuesta: La empresa matriz se presenta como responsable de la prestación del servicio. El número total de empleados del grupo se comunica como si fuera la capacidad de la empresa local. La responsabilidad de la franquicia se asigna directamente a la empresa matriz. La marca, la entidad legal y el socio operativo se combinan. La respuesta puede haber nombrado correctamente la marca. Sin embargo, el gráfico de entidades es incorrecto.

Por lo tanto, la puntuación GEO no puede depender únicamente de la mención del nombre o de la fluidez general de la oración.

22. RESULTADOS DE RESOLUCIÓN DE ENTIDADES

Para cada observación, el resultado de la resolución de entidades puede registrarse con uno de los siguientes estados.

ER-0— NO EVALUADO

La resolución de la entidad no ha sido evaluada.

ER-1— OBJETIVO EXACTO

La entidad objetivo ha sido resuelta de manera correcta y clara.

ER-2— COINCIDENCIA ACEPTABLE A NIVEL DE MARCA

No se proporciona un nombre legal, pero la marca y la entidad corporativa objetivo se resuelven materialmente de manera correcta.

ER-3— PARCIAL O SUBESPECIFICADO

La respuesta se acercó al área correcta pero carece de identidad legal, geográfica u operativa.

ER-4— CONFUSIÓN DE MÚLTIPLES ENTIDADES

Se han combinado múltiples entidades separadas como un solo objeto.

ER-5— entidad incorrecta

Se ha dirigido a otra persona, empresa, producto o nombre de dominio.

ER-6— AMBIGÜEDAD

No se puede determinar a partir de la respuesta qué entidad ha sido resuelta.

ER-7— ACLARACIÓN APROPIADA

El prompt es de hecho ambiguo, y el sistema ha solicitado correctamente una aclaración. Solicitar una aclaración no debe considerarse un fallo.

23. EL ÉXITO EN LA RESOLUCIÓN DE ENTIDADES ES DISTINTO DE LA EXACTITUD DE LA REPRESENTACIÓN

Un sistema puede analizar la entidad correcta pero puede proporcionar información incorrecta sobre ella. Otro sistema puede enunciar algunos hechos materiales correctamente pero conectarlos con el entidad incorrecta. Hay cuatro situaciones básicas:

Análisis de entidad	Contenido material	Resultado
Correcto	Correcto	Candidato a representación correcta fuerte
Correcto	Incorrecto	Objeto correcto, atributo incorrecto
Incorrecto	Parcialmente correcto	Información correcta vinculada al entidad incorrecta
Incorrecto	Incorrecto	Desglose completo de la representación

Por esta razón, el análisis de identidad puede ser una puerta de entrada separada. En particular, una respuesta con seria confusión con otra entidad no puede ser salvada con un alto promedio de precisión factual.

24. VINCULANDO LA DECISIÓN DE CONFORMIDAD A LA ENTIDAD

Una decisión de conformidad debe establecerse en el siguiente formato:

> *[ID de la entidad] + [tipo de entidad] + [alcance] + [interfaces digitales] + [país e idiomas] + [tiempo] + [versión del estándar]*

Plantilla de ejemplo: “Para [marca/entidad legal] identificada con el ID de entidad NGE-XXXX NOMOS; las páginas públicas en turco e inglés en el dominio especificado, dentro del alcance de productos y servicios especificado, han sido evaluadas el [fecha] bajo [versión del estándar].” Luego debe indicarse claramente: “Esta decisión no se aplica automáticamente a filiales, franquicias, productos fuera del alcance, otros dominios y idiomas que se agreguen en el futuro.”

25. DISPOSICIONES NORMATIVAS OBLIGATORIAS

CH03-N01

Cada auditoría GEO debe identificar una entidad objetivo única antes de que se vean los resultados.

CH03-N02

No se puede asumir que el ancla del prompt sea la misma que la entidad objetivo.

CH03-N03

Los nombres de dominio, las marcas comerciales, las entidades legales, los productos, los servicios y las identidades personales deben diferenciarse entre sí.

CH03-N04

Una propiedad de una entidad no puede transferirse a otra entidad relacionada sin evidencia y alcance claros.

CH03-N05

La licencia, certificado, número de empleados, clientes o resultado de conformidad de la empresa matriz no se puede aplicar automáticamente a las subsidiarias.

CH03-N06

La afirmación a nivel de marca y la parte que realmente presta el servicio y es legalmente responsable deben estar separadas.

CH03-N07

El franquiciador, el operador de la franquicia, el distribuidor, el concesionario y el vendedor de la plataforma deben estar registrados como entidades separadas.

CH03-N08

Cada relación entre entidades tangibles debe contener información sobre dirección, alcance y tiempo.

CH03-N09

Las relaciones como "socio", "cliente", "experto", "empresa del grupo" y similares no pueden usarse de manera ambigua.

CH03-N10

Se deben utilizar áreas de resolución de colisiones para diferentes entidades con el mismo nombre.

CH03-N11

Si se va a incluir una afirmación incierta en la puntuación principal, el método para manejar la incertidumbre debe estar predefinido; de lo contrario, la afirmación debe aclararse.

CH03-N12

El alcance de una entidad no puede ampliarse ni reducirse después de ver los resultados para aumentar un resultado positivo.

CH03-N13

Las entidades y relaciones históricas no pueden usarse como si tuvieran el estado de entidad actual.

CH03-N14

La compra, fusión o transferencia de marca no transfiere automáticamente las características pasadas a la nueva entidad.

CH03-N15

Una decisión de conformidad solo puede aplicarse a la unidad de conformidad especificada en la decisión.

CH03-N16

El nuevo producto no puede heredar automáticamente la disponibilidad actual de idioma, nombre de dominio, empresa afiliada o país.

CH03-N17

El éxito de la resolución de entidades y la exactitud de los hechos deben evaluarse por separado.

CH03-N18

La fusión de la identidad material con el entidad incorrecta no puede hacerse invisible dentro del promedio factual general.

CH03-N19

Los Registros de Entidades Canónicas deben versionarse y preservar cambios a lo largo del tiempo.

CH03-N20

Debe haber un propietario humano o institucional responsable del registro de la entidad.

26. FORMAS DE FALLA

CH03-F01— CONSIDERAR EL NOMBRE DE DOMINIO COMO UNA EMPRESA

La dirección digital se presenta como una entidad legal u operativa.

CH03-F02— FUSIONANDO LA MARCA CON LA EMPRESA LEGAL

El nombre de la marca se utiliza en lugar de la entidad legal que mantiene los contratos y responsabilidades.

CH03-F03— CONSIDERANDO EL PRODUCTO COMO UNA EMPRESA

El nombre del producto se resuelve como una identidad corporativa o de marca.

CH03-F04— TRANSFIRIENDO LA CARACTERÍSTICA DEL FUNDADOR A LA EMPRESA

La edad, experiencia, premio o proyecto previo del fundador se convierte en una característica corporativa.

CH03-F05— TRANSFIRIENDO CARACTERÍSTICAS DEL GRUPO A LA EMPRESA LOCAL

Se asignan datos consolidados de clientes, empleados o ingresos al negocio local.

CH03-F06— TRANSFIRIENDO LA LICENCIA A TRAVÉS DE LA RELACIÓN

La licencia de la empresa matriz, producto, empleado o socio se utiliza en el nombre de otra entidad.

CH03-F07— UNIFICANDO LA FRANQUICIA Y LA MARCA PRINCIPAL

El rendimiento y la responsabilidad del negocio local se presentan directamente como la marca global.

CH03-F08— COMBINANDO PRODUCTOR Y VENDEDOR

El fabricante del producto y el vendedor o distribuidor local se consideran una sola entidad.

CH03-F09— COMBINANDO PLATAFORMA Y VENDEDOR

La plataforma de mercado se hace responsable del producto, la garantía o la entrega del vendedor.

CH03-F10— CONSIDERANDO LA RELACIÓN COMO TRANSFERENCIA DE CARACTERÍSTICAS

El estado del cliente, la membresía, el uso de tecnología o la asociación se representa incorrectamente como una transferencia de autoridad.

CH03-F11— ACTUALIZANDO LA IDENTIDAD HISTÓRICA

Se utiliza el nombre antiguo, la propiedad, la asociación o el gerente como si fuera la realidad actual.

CH03-F12— RECLAMANDO EL PASADO DESPUÉS DE LA ADQUISICIÓN

Todos los éxitos pasados de la entidad adquirida se atribuyen al nuevo propietario sin explicación.

CH03-F13— FUSIÓN DE ENTIDADES CON EL MISMO NOMBRE

Diferentes empresas o individuos se consideran como una sola entidad debido a la similitud del nombre.

CH03-F14— DIVISIÓN ARTIFICIAL DE UNA ENTIDAD

Un idioma, producto u operación local fallida se declara como una entidad separada para mantener una puntuación positiva.

CH03-F15— EXPANSIÓN ARTIFICIAL DE UNA ENTIDAD

Los resultados de una marca o producto exitoso se distribuyen a todo el grupo.

CH03-F16— CONSIDERACIÓN DEL ALCANCE DE ELEGIBILIDAD

Un resultado de auditoría limitado se utiliza para toda la empresa, grupo o cartera de productos.

CH03-F17— CONTEO DE JURISDICCIÓN DE IDIOMA

La existencia de una versión en un idioma se utiliza como prueba de capacidad legal de la empresa o del servicio en ese país.

CH03-F18— CONTANDO EL CONTROL TÉCNICO COMO PROPIEDAD

La agencia o desarrollador se muestra como el propietario legal o comercial de la institución porque gestionan el nombre de dominio.

CH03-F19— DEJANDO AL RECLAMANTE INDETERMINADO

No se puede determinar quién es responsable de la información incorrecta sobre precio, licencia o capacidad.

CH03-F20— INTERPRETANDO EL ANCLA DEL PROMPT DESPUÉS DEL RESULTADO

La entidad objetivo que hará que la respuesta parezca exitosa se selecciona posteriormente.

27. PROCEDIMIENTO DE AUDITORÍA

Paso 1— Registrar el Ancla del Prompt

Exactamente cuál del usuario:

- nombre,
- nombre de dominio,
- marca,
- producto,
- persona

expresión se utiliza se registra completamente.

Paso 2— Identificar la Entidad Objetivo

Se anota qué entidad mide realmente la auditoría. El objetivo se clasifica como:

- persona jurídica,
- marca,
- producto,

- servicio,
- paquete de entidad

La clase aplicable debe registrarse explícitamente.

Paso 3— Crear ID de Entidad Canónica

Nombre canónico Tipo de entidad Identidad legal Dominio de jurisdicción Alias Estado temporal está registrado.

Paso 4— Construir el Grafo de Entidades

Se extraen las relaciones dirigidas entre la empresa matriz, subsidiarias, marca, producto, nombre de dominio, individuos y socios.

Paso 5— Determinar Roles de Responsabilidad

Propietario de la afirmación Propietario de la interfaz digital Propietario de la entrega de servicios Responsable legal Propietario de la evidencia son separados.

Paso 6— Revisar el Riesgo de Transferencia de Atributos

En respuesta:

- licencia,
- certificado,
- empleado,
- cliente,
- fecha de establecimiento,
- éxito,
- cumplimiento

¿Se ha transferido al entidad incorrecta?

Paso 7— Verificar Identidad Temporal

Respuesta:

- nombre anterior,
- anterior propietario,
- empresa conjunta terminada,
- anterior empleado,
- estado previo a la adquisición

¿se está utilizando?

Paso 8— Verificar Roles Geográficos y Locales

¿Son correctos los roles de marca global, empresa local, distribuidor y franquicia?

Paso 9— Clasificar Entidad Resuelta por el Sistema

Se asigna el estado apropiado entre ER-1 y ER-7.

Paso 10— Fijar Unidad de Conformidad

Para aplicar la decisión final:

- entidad,

- campo,
- producto,
- idioma,
- país,
- tiempo

se escribe el límite.

Paso 11— Limitar la afirmación Pública

El resultado se reduce si se generaliza a una institución o grupo más amplio que la entidad definida.

Paso 12— Registrar la Forma de Cambio y Objeción

Si se informa un error de identidad o de relación:

- quién verificará,
- quién cambiará,
- qué versión se generará

será determinado.

28. EVIDENCIA NECESARIA

Identidad canónica de la entidad; nombre legal y tipo de entidad; registro oficial o verificación equivalente; propiedad de marca comercial; control del nombre de dominio; relaciones de matriz y subsidiaria; propiedad de productos y servicios; registros de negocios locales y distribuidores; relaciones de franquicia; parte contratante y de facturación; responsabilidad de entrega de servicios; titular de licencias y certificados; roles de fundador y ejecutivo; nombres activos e históricos; registros de adquisición, fusión y transferencia; país y jurisdicción; registros de idioma oficial y transliteración; gráfico de entidades; reclamante; propietario de contenido; propietario de interfaz técnica; alcance de auditoría; entidades y interfaces fuera de alcance; versión de registro de entidad; última fecha de verificación; y persona u organización responsable.

29. LISTA DE VERIFICACIÓN DE AUDITORÍA

¿Se ha registrado claramente el ancla del prompt? ¿Se identificó la entidad objetivo antes de observar los resultados? ¿Se separaron la entidad legal y el nombre de dominio? ¿Se separaron la marca y el propietario de la marca? ¿Se separaron el producto y la empresa? ¿Se separaron los atributos del fundador de los de la organización? ¿Son correctas las relaciones entre matriz y subsidiaria? ¿Se distinguen las franquicias y los distribuidores? ¿Se identifican las partes que venden y entregan el servicio? ¿Se separaron el reclamante y el editor técnico? ¿Pertenece las licencias y certificados a la entidad correcta? ¿Se indican las cifras de empleados, clientes e ingresos dentro del alcance correcto? ¿Se separaron los nombres y relaciones históricos del estado actual? ¿Se verificaron diferentes entidades con el mismo nombre? ¿Se vincularon múltiples dominios y interfaces de usuario a la entidad correcta?

¿Se han vinculado nombres multilingües a la misma identidad canónica? ¿Están claros los roles de país y jurisdicción? ¿Se ha clasificado la entidad por separado como resultado de la resolución de entidades? ¿Está claramente definida la unidad de conformidad? ¿Se extiende a entidades fuera del alcance de la decisión? ¿Está el registro de la entidad versionado y fechado? ¿Existe un camino para la objeción y corrección de cambios de identidad?

30. OBJECIONES Y RESPUESTAS

Objeción 1— “El usuario conoce el nombre de la marca, no la empresa legal.”

Correcto. Es natural que el usuario use la marca en la solicitud. La auditoría no obliga al usuario a usar un lenguaje legal. Sin embargo, la evaluación en segundo plano:

- la marca,
- propietario,
- proveedor de servicios,
- parte contratante

debe distinguirse. Se puede dar una respuesta simple al usuario. El registro de auditoría no tiene que ser simple.

Objeción 2— “En un negocio pequeño, la marca y la empresa ya son lo mismo.”

En algunas estructuras pequeñas:

- marca única,
- único nombre de dominio,
- propietario único del negocio,
- único proveedor de servicios

Se puede encontrar. En este caso, el grafo de la entidad es simple. No obstante, el registro debe hacerse explícitamente. Simplicidad no significa ambigüedad.

Objeción 3— 'El grupo de empresas actúa como una sola marca.'

Desde una perspectiva de marketing, puede haber una sola marca. En términos legales, operativos y de responsabilidad evidencial, puede haber entidades separadas. Si se deben tomar decisiones a nivel de grupo, debe mostrarse claramente qué características están consolidadas.

Objeción 4— “Demasiado detalle complica innecesariamente la puntuación GEO.”

La puntuación correcta asignada al entidad incorrecta no es útil. La resolución de entidades no es una carga adicional para la auditoría, sino una condición de validez.

Objeción 5— "Los sistemas de IA ya no prestan atención a estas diferencias legales."

Exactamente por esta razón, debe medirse. El sistema de IA:

- marca,
- empresa,
- producto,
- negocio local

No distinguir la diferencia entre ellos es el propio error de representación.

Objeción 6— "Si la empresa matriz controla todo el grupo, las características pueden transferirse."

Las relaciones de control son importantes en algunas áreas. Sin embargo:

- licencia,
- contratos,
- cliente,
- empleado,
- responsabilidad de datos,
- cumplimiento

tales características requieren evidencia separada para la transferencia. El control no es la unificación de todas las características.

Objeción 7— "¿Los logros de la antigua empresa después de una adquisición no forman parte del nuevo grupo?"

Se puede explicar en un contexto histórico. Sin embargo:

- durante qué período fue el éxito,
- por qué equipo y entidad,
- antes o después de la adquisición

Se debe indicar cuándo fue producido. La nueva propiedad no cambia la historia.

Objeción 8— "la prompt ya está clara; el usuario ha escrito el nombre de dominio."

El nombre de dominio puede ser claro. Sin embargo: también debe responderse la pregunta "¿Qué representa el nombre de dominio?". Un dominio puede ser una empresa, tienda, portal de productos o plataforma con múltiples entidades.

Objeción 9— "Describir al distribuidor local como una marca es más fácil para el usuario."

La simplicidad no prevalece sobre la responsabilidad. La respuesta puede simplificarse de la siguiente manera: "La marca ofrece servicios en Turquía a través de un distribuidor local autorizado." Es comprensible y correcta.

Objeción 10— "Todas estas relaciones cambian con el tiempo."

Eso es correcto. Por esta razón, el registro de entidad:

- con fecha,
- versionado,
- debe ser activado por cambios

No disminuye la necesidad de registrar la variabilidad. La aumenta.

REGLA COMÚN DEL CAPÍTULO 33

Antes de generar una puntuación GEO: preguntamos, “¿Cuántas respuestas son correctas?” Pero la pregunta más fundamental que debería hacerse antes de esto es:

“¿Exactamente de quién son estas respuestas?”

Si consideramos un nombre de dominio como una empresa, medimos el objeto incorrecto. Si consideramos la marca como una empresa legal, asignamos la responsabilidad de manera incorrecta. Si consideramos el producto como una empresa, distorsionamos sus características. Si convertimos la experiencia del fundador en la historia de la empresa, exageramos el pasado. Si contamos el número

de empleados del grupo como capacidad local, inflamos las expectativas de servicio. Si la empresa matriz transfiere su licencia a una franquicia, dirigimos al usuario a la autoridad equivocada. Si hacemos que los logros pasados después de una adquisición sean una característica del nuevo propietario, reescribimos la verdad histórica. Si extendemos una auditoría de nombre de dominio estrecho a toda la empresa, blanqueamos el alcance del cumplimiento. Por lo tanto, definir una entidad no es solo limpieza de datos. Es el comienzo de la responsabilidad representativa. Un sistema puede describir al entidad incorrecta con la información correcta. Esto sigue siendo una tergiversación. Un sistema puede describir la entidad correcta con información incorrecta. Esto también es una tergiversación.

Ambos son necesarios para el GEO correcto:

> *Entidad correcta. / Atributo correcto. / Relación correcta. / Alcance correcto. / Tiempo correcto.*

Por lo tanto, la tercera ley de medición de NOMOS es:

> *Si la entidad medida no está clara, la precisión medida tampoco está clara.*

La cuarta ley es la siguiente:

> *La relación no es identidad.*

La quinta ley es la siguiente:

> *La propiedad no es la transferencia de todos los atributos.*

Su sexta ley es:

> *La conformidad no se extiende a las entidades vecinas.*

Sección NOMOS Orden 3

> *No me des solo un nombre. / Muéstrame a qué entidad pertenece ese nombre.*
>
> *No trates un dominio como una empresa, una marca como una parte contratante o un producto como un fabricante.*
>
> *No uses el pasado del fundador como la edad de la institución.*
>
> *No transfieras la licencia de la empresa matriz a la organización afiliada, franquicia o distribuidor.*
>
> *No conviertas al cliente del grupo en cliente de la marca local, ni el premio de la marca en el premio de todos los productos.*
>
> *No describas la colaboración como propiedad, la membresía como acreditación, ni el uso de tecnología como asociación.*
>
> *No combines el nombre antiguo con la nueva entidad, ni el pasado adquirido con el desempeño actual.*
>
> *No amplíes la entidad para magnificar los resultados positivos / No dividas la entidad para ocultar los resultados negativos.*
>
> *Primero, bloquea el objetivo. / Luego registra la dirección de cada relación. / Luego asigna cada atributo a su verdadero propietario. / Luego agrega el tiempo y el país relevantes. / Solo entonces decide para esa entidad únicamente.*

La oración de cierre del capítulo

La primera pregunta en una auditoría GEO no es lo que dice la IA, sino precisamente quién o qué concierne cada afirmación.

Core normativo

Cada auditoría GEO DEBE definir y versionar, antes de observar resultados: - un ancla de prompt, - una entidad objetivo, - un tipo de entidad, - un registro de identidad canónica, - un gráfico de relaciones, - un límite temporal y jurisdiccional, - un objeto de auditoría, - y una unidad de conformidad. Los dominios, marcas, entidades legales, grupos corporativos, subsidiarias, productos, servicios, personas, franquicias, distribuidores, plataformas y vendedores NO DEBEN ser tratados como entidades intercambiables. Los atributos, credenciales, licencias, clientes, empleados, desempeño, autoridad y conformidad NO DEBEN transferirse entre entidades relacionadas a menos que la relación, el alcance, el tiempo y la evidencia apoyen explícitamente esa transferencia. La entidad resuelta por el producto de IA DEBE evaluarse por separado de la exactitud factual de la respuesta. Una decisión de conformidad DEBE aplicarse únicamente a la entidad explícitamente identificada y al alcance auditado. NO DEBE extenderse automáticamente a afiliados, productos, jurisdicciones, idiomas, dominios, versiones futuras o entidades relacionadas.

CAPÍTULO 04

Dos capas de realidad

Representación oficial y realidad verificada de la entidad

BASES METODOLÓGICAS / K10; K11

Límite de Capítulo

La Sección 1 estableció la siguiente disposición:

> *Una única respuesta no es una puntuación GEO.*

Capítulo 2 añadió lo siguiente:

> *La representación es una distribución que ocurre entre usuarios, productos de IA, países, idiomas, sesiones y tiempos.*

El Capítulo 3 luego identificó el objeto de la distribución:

> *No se puede medir una entidad hasta que se resuelva.*

Ahora llegamos a la siguiente pregunta obligatoria:

> *¿Según qué registro de la realidad evaluaremos si la respuesta de la IA es correcta?*

¿Qué fuente debería regir la auditoría: el sitio web de la entidad auditada, un registro legal, la página de precios de la empresa, informes independientes, la autoridad de licencias, opiniones de clientes, los propios datos de desempeño de la empresa o la afirmación de un competidor? Cuando las fuentes entran en conflicto, ¿cuál debería aceptarse como correcta? Si una empresa publica una afirmación falsa o sin respaldo y la IA la repite perfectamente, ¿ha tenido éxito el sistema? Si el sitio web oficial está desactualizado pero la IA da una respuesta diferente basada en evidencia externa actual y verificada, ¿ha fallado el sistema? Hasta que estas preguntas sean respondidas:

- representación correcta,
- representación incorrecta,
- fidelidad de la fuente,
- exactitud factual,
- realidad de la entidad

Estos términos a veces se usan como si se refirieran al mismo concepto. Este capítulo resuelve la confusión resultante al distinguir dos capas de la realidad:

Primera Capa

Capa de Representación Oficial

Muestra lo que la entidad dice sobre sí misma en las interfaces canónicas que controla.

Segunda Capa

Capa de Realidad de Entidad Verificada

Bajo la evidencia existente, se muestra qué afirmaciones materiales están respaldadas dentro de qué alcance y tiempo. Estas dos capas están conectadas. Pero no son lo mismo. La primera capa responde a la pregunta: "¿Cómo se define la entidad a sí misma?" La segunda capa responde a la pregunta: "¿Qué parte de esta definición está respaldada por qué evidencia?" El primer libro define el estándar correcto, el segundo libro separa las formas de violación, y el sistema de auditoría está diseñado para medir y probar estas distinciones. Esta sección opera esa tarea de auditoría a nivel de afirmación.

NOMOS Desafío

El sitio web de su empresa contiene la siguiente frase: "Somos la empresa GEO líder en el mundo." Un producto de IA lee esta página y responde al usuario: "Esta empresa es la empresa GEO líder en el mundo." ¿Refleja correctamente la respuesta su sitio web? Sí. ¿Refleja correctamente la realidad comprobada?

Aún no lo sabemos. Tal vez:

- hay investigación comparativa,
- el alcance del mercado ha sido definido,
- un criterio de liderazgo ha sido publicado,

el resultado ha sido verificado de manera independiente. Tal vez ninguno de estos exista. Si la IA hubiera transmitido la propia afirmación de la empresa: "La empresa se define a sí misma como la empresa GEO líder en el mundo.", habría conservado la fuente de la afirmación. Sin embargo, cuando dijo: "Es la empresa GEO líder en el mundo.", la empresa pudo haber elevado la declaración a una realidad independiente.

Consideremos otro caso. El sitio web oficial de la empresa afirma: 'Ofrecemos servicios solo en Türkiye.' Sin embargo, los registros verificados muestran que la empresa comenzó a operar en Alemania y el Reino Unido hace seis meses, y el sitio web no ha sido actualizado. Basándose en registros legales y operativos actuales, la IA responde: 'La empresa ofrece servicios en Türkiye, Alemania y el Reino Unido.'

La IA no es completamente fiel al sitio web oficial. ¿Podría, sin embargo, estar más cerca de la realidad actual verificada? Sí. Consideremos un tercer caso. El sitio web de la empresa afirma: 'Operamos en 20 países.' Los registros independientes e internos verifican solo ocho. La IA responde: 'La empresa opera en 20 países.'

Esta respuesta:

- fiel a la representación oficial,
- contrario a la realidad verificada

Es posible. La primera disposición de esta sección es la siguiente:

> *La fidelidad al registro oficial no es lo mismo que la exactitud factual.*

Su segunda disposición establece:

> *Lo que una empresa dice y lo que se puede probar deben registrarse por separado.*

Su tercera disposición establece:

> *Una IA que repite un error oficial correctamente no constituye una representación precisa.*

Su cuarta disposición establece:

> *Un IA que se desvíe de un registro oficial desactualizado no está automáticamente equivocada si cuenta con evidencia fiable.*

1. OBJETIVO DEL CAPÍTULO

El propósito de esta sección es establecer dos capas de referencia separadas que se utilizarán en la evaluación GEO-1000. La sección hace las siguientes distinciones normativas:

- Declaración de existencia vs. hecho probado
- Fuente oficial vs. fuente independiente
- Fuente de primera parte vs. verificación independiente
- Fidelidad a la fuente vs. exactitud material
- Repetición de la afirmación vs. verificación de la afirmación
- Registro oficial actual versus registro oficial antiguo
- Silencio oficial versus información verificada externamente
- Perspectiva de la empresa vs. juicio de la realidad externa
- Evidencia oficial versus afirmación oficial
- Independencia de la fuente versus relevancia de la fuente para el tema
- Error oficial versus error de IA
- Conflicto de referencias versus fallo del modelo
- Incertidumbre versus fallo
- Interés del reclamante versus veracidad de la afirmación
- La presencia de evidencia y si la evidencia realmente respalda la afirmación

Al final de esta sección, se deberían poder hacer dos preguntas separadas para cada respuesta de IA:

> *¿La respuesta transmite correctamente la representación oficial y canónica de la entidad?*

y:

> *¿La respuesta refleja correctamente la realidad material sostenida por la evidencia actualmente verificable?*

Estas preguntas pueden dar la misma respuesta. También pueden dar respuestas diferentes. Cuando difieren, el papel de la auditoría no es ocultar la diferencia, sino explicarla.

2. DISPOSICIÓN NORMATIVA CENTRAL

> *Cada respuesta material de IA debe evaluarse frente a dos referencias separadas sin transformación: el Registro de Representación Oficial y el Registro de Realidad de la Entidad Verificada.*

Una respuesta:

- fiel a la representación oficial,
- apropiado a la realidad verificada,
- apropiada para ambas,
- lejana a ambas

Puede ser. NOMOS no puede combinar silenciosamente estos dos resultados en una sola decisión verdadero–falso. Un alto rendimiento en la primera capa no compensa el fracaso en la segunda capa. Específicamente:

> *Una respuesta de IA que repite correctamente una afirmación no verificada o falsa publicada por la propia entidad puede demostrar fidelidad a la representación oficial; sin embargo, no puede demostrar precisión de la entidad verificada.*

3. PRIMERA CAPA

CAPA DE REPRESENTACIÓN OFICIAL

La Capa de Representación Oficial es el registro versionado de las declaraciones materiales actuales que publica acerca de sí misma en interfaces canónicas controladas o explícitamente autorizadas por la entidad. Esta capa responde a la pregunta:

> *¿Cómo quiere la entidad ser entendida y qué información publica oficialmente?*

Esta capa incluye:

- declaración de identidad legal,
- relación marca–empresa,
- definiciones de productos y servicios,
- precios publicados,
- países de servicio,
- condiciones de soporte,
- políticas oficiales,
- límites divulgados públicamente,
- Puede contener información oficial sobre moneda y versiones

La capa de representación oficial es importante. Porque un sistema de IA debe comprender correctamente cómo una entidad:

- se define a sí mismo,
- qué servicio ofrece,
- qué precio publica,
- qué términos acepta

Sin embargo, la capa de representación oficial no es una prueba independiente automática para cada afirmación. Una declaración en el propio sitio web de la institución, especialmente con respecto a afirmaciones hacia el mundo exterior como superioridad y desempeño, presenta la declaración en sí; no demuestra su corrección por sí misma.

4. ¿QUÉ ES UNA FUENTE OFICIAL?

No se considera que una fuente sea oficial únicamente porque lleve el logotipo de la institución. Una fuente oficial debería tener al menos las características relevantes de lo siguiente:

- Controlada por la entidad
- Someterse a un proceso de publicación autorizado
- Aprobada por el propietario de la información factual
- URL canónica o identificador de registro
- Versión
- Fecha de publicación o verificación

- Historial de cambios
- Persona o institución responsable
- Alcance
- Estado del idioma y la localidad
- Estado de archivo o vigencia

Una opinión publicada en una cuenta personal de redes sociales por un empleado no es automáticamente una declaración oficial de la empresa. Un archivo borrador por una agencia no es una publicación oficial. Un correo electrónico de un representante de ventas, si no está autorizado, no constituye una política corporativa general. Un PDF antiguo en el sitio web de una empresa puede no reemplazar el nuevo registro oficial.

5. SURGEN REPRESENTACIONES OFICIALES

Según la auditoría correspondiente, las interfaces oficiales pueden incluir:

- Sitio web corporativo principal
- Avisos legales
- Páginas de productos y servicios
- Páginas de precios y cotizaciones
- Centro de ayuda oficial
- Documentos de relaciones con inversionistas
- Políticas públicas y contratos
- Comunicados de prensa oficiales
- Datos estructurados
- Registros canónicos JSON
- Salidas de API oficiales
- Contenido oficial en la aplicación
- Sitios locales autorizados
- Cuentas oficiales de redes sociales
- Declaraciones presentadas a instituciones públicas
- Registros de publicación y retractación

Estas interfaces pueden contradecirse entre sí. En este caso, la "representación oficial" no es una realidad única y limpia. Es un conjunto de registros que deben ser auditados internamente.

6. REGISTRO DE REPRESENTACIÓN OFICIAL

Se debe crear un Registro de Representación Oficial para cada entidad auditada. Este registro no es una copia de todo el contenido oficial. Es un resumen rastreable de afirmaciones y límites fácticos. Cada registro debe incluir los siguientes campos en la medida de lo aplicable:

- Identificador de la afirmación
- Texto completo de la afirmación
- Sujeto de la afirmación
- Tipo de afirmación
- Fuente canónica

- Propietario de la fuente
- Idioma y localización
- Fecha de publicación
- Última fecha de verificación
- Periodo de validez
- Estatus oficial
- Alcance
- Limitaciones
- Versión anterior
- Motivo del cambio
- Reclamante
- Aprobación humana

Sin un registro de representación oficial de una entidad: la frase "La IA refleja el sitio con precisión." no puede probarse claramente.

7. LA AUTORIDAD DE LA FUENTE OFICIAL DEPENDE DEL TIPO DE AFIRMACIÓN

No todas las fuentes oficiales tienen el mismo peso para cada afirmación. La autoridad de la fuente debe evaluarse según el tipo de afirmación.

7.1. Decisiones bajo el control de la entidad

Ejemplo:

- Qué servicio ofrece al público
- Precio publicado
- Horas de atención
- Política de devoluciones oficial
- Canal de comunicación
- Versión del producto disponible públicamente

En estas áreas, la fuente oficial es fundamental y a menudo el registro autorizado principal. Sin embargo, si la página oficial contradice el flujo real del proceso, la página por sí sola no es suficiente. Por ejemplo:

- página de precios 1,000 euros,
- pago 1,500 euros

Si aparece, hay una contradicción en la capa de representación oficial.

7.2. Identidad legal y autoridad

Ejemplo:

- Registro de la empresa
- Licencia
- Autoridad profesional
- Propiedad de marcas registradas
- Estado fiscal o legal

El sitio web de la empresa puede divulgar esta información. Sin embargo, un registro oficial o una institución autorizada pueden proporcionar una verificación más sólida.

7.3. Declaraciones de desempeño y éxito

Ejemplo:

- "98% de nuestros proyectos son exitosos."
- "Aumentamos los ingresos de nuestros clientes en un promedio de 40%."
- "Tenemos la tasa de satisfacción más alta."

El sitio web de la empresa puede ser el propietario de la afirmación. Pero para respaldar la afirmación:

- datos,
- tipo,
- denominador,
- período,
- resultados no exitosos,
- revisión independiente si es posible

son requeridos.

7.4. afirmaciones de Superioridad y Liderazgo

Ejemplo:

- "Líder mundial"
- "El mejor"
- "Número uno"
- "El único estándar real"
- "La empresa más confiable de la industria"

La declaración propia de un sitio no es una comparación independiente. Tales afirmaciones:

- requieren un universo de comparación,
- criterios,
- tipo,
- fecha,
- alcance,
- independencia

Deben establecerse ambas condiciones.

7.5. Relaciones con Terceros

Ejemplo:

- Cliente
- Socio
- Patrocinador
- Membresía
- Acreditación
- Asociación tecnológica
- Distribución

Una declaración oficial unilateral puede ser valiosa. Sin embargo, puede necesitar ser verificada por la otra parte de la relación o por un registro confiable.

7.6. Resultados del usuario

Ejemplo:

- Satisfacción del cliente
- Resultado del uso real
- Tasa de devolución
- Tasa de quejas
- Compra repetida

Los datos de la empresa en estas áreas son evidencia de primera mano importante. Sin embargo:

- selección de datos,
- no respuesta,
- metodología,
- conflicto de intereses

deben ser divulgados.

8. Fidelidad de la representación OFICIAL

El concepto que indica hasta qué punto una respuesta de IA se ajusta a la Capa de Representación Oficial:

Fidelidad de la Representación Oficial

El término canónico en inglés es:

Fidelidad de la Representación Oficial

Esta medición plantea las siguientes preguntas: ¿Está la IA interpretando correctamente la entidad? ¿Transmite con precisión los productos y servicios oficiales? ¿Distorsiona el precio oficial? ¿Elimina los límites oficiales? ¿Expande la cobertura oficial del país? ¿Presenta la opinión de la empresa como un hecho independiente? ¿Usa información oficial desactualizada como si fuera actual? ¿Oculta contradicciones entre los registros oficiales? ¿Agrega afirmaciones materiales que no se encuentran en la fuente oficial? ¿Mantiene la calidad de la fuente? La fidelidad de la representación oficial es una medición de qué tan precisamente la entidad transmite la realidad que ha publicado. Esta medición:

> *No muestra que las declaraciones de la entidad sean verdaderas, sino el grado en que la IA transmite con precisión lo que dice la entidad.*

9. SEGUNDA CAPA

CAPA DE REALIDAD DE ENTIDAD VERIFICADA

La Capa de Realidad de Entidad Verificada es un registro de evaluación versionado de información material que existía en una fecha de medición específica y está respaldada por evidencia adecuada al tipo de afirmación. Esta capa responde a la pregunta:

> *¿Qué afirmaciones sobre la entidad están realmente respaldadas por qué evidencia?*

Esta capa incluye:

- registros provenientes de la propia empresa,
- registros públicos oficiales,
- fuentes independientes,
- registros de contratos o transacciones,
- documentos de licencia,
- registros históricos verificados,
- evidencia en contrario,
- hallazgos adversos,
- incertidumbres,
- disputas restantes

pueden evaluarse en conjunto.

10. "REALIDAD VERIFICADA" NO ES REALIDAD ABSOLUTA

Este protocolo no afirma: "Sabemos con certeza toda la verdad absoluta sobre una institución." La Realidad de Entidad Verificada es:

- basada en evidencia,
- dependiente del tiempo,
- dependiente del alcance,
- sujeta a cambios con nueva evidencia,
- puede ser disputada,
- puede ser retirada,

puede conllevar incertidumbre. La definición correcta es:

> *Dentro de la fecha y el alcance especificados, el registro de entidad más fuerte y verificable respaldado por la evidencia disponible.*

Por lo tanto, un registro de realidad verificada:

- versionado,
- con fuente,
- sujeto a objeción,
- modificable,
- con conocimiento de la incertidumbre

debe ser. NOMOS no puede posicionarse como dueño de la verdad. Solo clasifica qué sentencia permite la evidencia y hasta qué punto.

11. REGISTRO DE REALIDAD DE ENTIDAD VERIFICADA

La estructura técnica completa de esta capa se detallará en un capítulo posterior del libro. A nivel fundamental, los siguientes campos deberían existir para cada afirmación fáctica:

- Identificador de la afirmación
- Entidad sujeto
- Relación o atributo

- Valor
- Alcance
- País o jurisdicción
- Hora
- Fuentes
- Tipos de fuente
- Verificación de la fuente
- Independencia
- evidencia en contrario
- Suficiencia de la evidencia
- Estado de verificación
- Incertidumbre
- Decisión del adjudicador
- Estado de apelación
- Última fecha de verificación
- Versión del registro

Este registro no puede ser creado únicamente a partir de evidencia positiva. También deben preservarse los registros que limiten la realidad o contradigan la afirmación.

12. ESTADOS DE VERIFICACIÓN

Cada afirmación material puede clasificarse con uno de los siguientes estados:

VR-0— NO EVALUADO

La afirmación aún no ha sido evaluada.

VR-1— SOLO afirmación OFICIAL

La afirmación se basa únicamente en la declaración oficial de la institución. No se ha demostrado que sea falsa. No hay verificación independiente o adicional suficiente.

VR-2— RESPALDADA POR EVIDENCIA DE PRIMERA PARTE

La afirmación está respaldada por la:

- transacción de la institución,
- operación,
- contratos,
- datos,
- registro

está respaldada por evidencia. La fuente es de primera parte.

VR-3— CORROBORADO

La afirmación está respaldada por múltiples registros apropiados y consistentes. No todas las fuentes tienen que ser independientes.

VR-4— VERIFICACIÓN INDEPENDIENTE

La afirmación está respaldada por una fuente independiente o revisión apropiada sobre el tema. La independencia por sí sola no implica suficiencia. También se requiere la relevancia de la fuente para el tema.

VR-5— SOSTENIDO POR MÚLTIPLES FUENTES

La afirmación está fuertemente respaldada por fuentes confiables de diferentes tipos dentro de un tiempo y alcance definidos.

VR-C — CONTRADICIDO

La afirmación entra en conflicto con evidencia más fuerte o más adecuada.

VR-U — SIN RESOLVER

La evidencia es insuficiente o contradictoria. No se puede emitir un juicio confiable.

VR-T — LIMITADO EN EL TIEMPO

La afirmación está respaldada solo durante el período especificado.

VR-S — ALCANCE LIMITADO

La afirmación está respaldada solo para el producto, país, usuario o alcance de operación especificado. Una afirmación puede tener más de un estado al mismo tiempo. Ejemplo:

VR-4+ VR-T + VR-S

Puede indicar que la afirmación ha sido verificada de manera independiente; sin embargo, tiene limitaciones de tiempo y alcance.

13. LA INDEPENDENCIA NO ES LO MISMO QUE LA ADECUACIÓN

Una fuente independiente no siempre es la mejor fuente. Un bloguero puede estimar el precio actual de una empresa, mientras que la página de precios canónica de la empresa puede ser la fuente más directa y apropiada. Por el contrario, cuando una empresa declara en su propio sitio web que es 'el líder de la industria', una investigación independiente con un método divulgado puede ser la evidencia más adecuada. Por lo tanto, la evaluación de fuentes pregunta más que si una fuente es independiente. ¿Es autoritaria para este tipo de afirmación? ¿Qué tan cerca está de la afirmación? ¿Su alcance es apropiado y actual? ¿Quién la controla? ¿Está divulgado su método? ¿Aborda evidencia en contrario? ¿Se puede auditar? ¿Existe un conflicto de intereses? La independencia por sí sola no garantiza la exactitud.

El hecho de que una fuente sea de primera mano no la hace inútil tampoco. Lo importante es:

Es tener la autoridad y la calificación de evidencia apropiadas para la afirmación sostenida por la fuente.

14. MAPA DE IDONEIDAD DE FUENTE DE afirmación

La tabla a continuación es un mapa fundamental general. Los requisitos de evidencia específicos pueden variar según el sector y el nivel de riesgo.

Tipo de afirmación	Fuente o evidencia primaria	Rol del sitio oficial	Necesidad de verificación adicional
Nombre legal de la empresa	Registro oficial autorizado	Muestra la declaración	Generalmente requerido
Propiedad de marcas registradas	Registro de marca, contrato, registro corporativo	Muestra la declaración	Requerido para afirmaciones financieras
Precio publicado actual	Registro de precio/cobro canónico	Usualmente primaria	Verificable con flujo de trabajo
Alcance del servicio	Registros operativos, contratos, página oficial	Declaración primaria	Evidencia adicional de entrega real
Licencia	Institución de licencias o registro	Muestra la declaración	Requerido
Relación con el cliente	Contrato, aprobación del cliente, registro bilateral	Puede ser una afirmación unilateral	Requerido para afirmaciones financieras
Tasa de éxito	Datos sin procesar, método, denominador, auditoría	Publica la afirmación	Estrictamente requerido
Liderazgo	Método comparativo y datos	Autodeclaración	Se requiere comparación independiente o fuerte
Satisfacción del cliente	Encuesta representativa, método, no respuesta	Puede explicar el resultado	Metodología requerida
Fecha de constitución	Registros oficiales y documentos históricos	Muestra la declaración	La historia legal y la historia de la marca DEBEN mantenerse separadas
Asociación	Declaración o contrato bilateral	Declaración unilateral	Generalmente requerido
Certificado	Institución emisora del certificado y alcance	Muestra la declaración	Requerido
Empleado activo	Registro de RRHH o contrato, perfil público	Puede mostrar declaración	Verificación de rol y fecha
Resultado del usuario	Registro de transacción y resultado	Describe el caso	Contexto, método y permiso requeridoss
"Mejor opción"	Usuario definido y criterios de comparación	Vista de marketing	Se requiere comparación sólida

Esta tabla no es un ranking de fuente universal. Muestra la adecuación entre la declaración y la evidencia.

15. MATRIZ DE afirmaciones DE DOBLE CAPA

Se hacen dos preguntas separadas para cada afirmación financiera: ¿Se ha publicado oficialmente? ¿Cuál es su estado en el registro verificado de hechos? Estas dos preguntas constituyen las siguientes situaciones básicas.

15.1. Oficial y Verificado

Afirmación: está presente en un registro oficial y respaldada por la evidencia apropiada. Ejemplo: El precio en la página oficial de precios también se verifica en el flujo de transacciones. Este es uno de los tipos más fuertes de representación central. Si la IA transmite la afirmación con el alcance y el momento correctos, se logra el éxito en ambas capas.

15.2. Oficial pero Solo Autodeclarado

La afirmación es oficial. Sin embargo, no hay una verificación adicional. Ejemplo: “Somos la empresa más innovadora de la industria.” La IA puede transmitir la afirmación como: “La empresa se describe

a sí misma como la empresa más innovadora de la industria." No puede elevarla a: "Es la empresa más innovadora de la industria."

15.3. Oficial pero Conflictivo

La afirmación existe en el registro oficial. Entra en conflicto con pruebas más apropiadas o actualizadas. Ejemplo: El sitio dice que 20 países. Los registros operacionales muestran ocho países. Si la IA repite exactamente el sitio, la fidelidad oficial puede ser alta. La precisión verificada es baja. Esta situación:

Transportando el error oficial fielmente

puede clasificarse como.

15.4. No Oficial pero Verificado

La afirmación no se encuentra en el sitio oficial. Sin embargo, ha sido verificada con otras fuentes confiables. Ejemplo: El cambio reciente de propiedad legal aún no se ha agregado al sitio oficial. Aparece en el registro autorizado. La IA puede transmitir esta información con precisión, indicando la fuente y el momento correctos. La representación oficial puede tener baja fidelidad o diferir. La precisión factual verificada puede ser alta.

15.5. No Oficial y No Verificado

Afirmación: no está en el registro oficial, carece de evidencia externa suficiente. La IA no debe transmitir esto como si fuera verdad.

15.6. Límite de material omitido oficialmente

La empresa publica la afirmación positiva en su página oficial. Sin embargo, no publica la limitación necesaria para entenderla correctamente. Ejemplo: "Ofrecemos servicios en Europa." Sin embargo, el servicio solo se ofrece en dos países y de manera remota. Si la IA solo repite la frase general, puede permanecer fiel al texto oficial. Sin embargo, puede producir resultados materialmente engañosos. Por lo tanto, la Capa de Representación Oficial debe registrar no solo las frases publicadas, sino también las omisiones materiales.

15.7. Disputa No Resuelta

Los registros oficiales y externos se contradicen entre sí. Ninguno es lo suficientemente fuerte. El resultado correcto:

NO RESUELTO

debería ser. No se puede esperar que la IA emita un juicio definitivo.

16. TABLA DE EVALUACIÓN DE DOBLE CAPA

Fidelidad a la representación oficial	Precisión verificada de la realidad	Comentario básico
Alta	Alta	Representación fiel y verificada
Alta	Baja	Transportando el error oficial fielmente
Baja	Alta	Representación correctiva que se desvía del registro oficial pero está respaldada por evidencia
Baja	Baja	Fallo general de representación

Fidelidad a la representación oficial	Precisión verificada de la realidad	Comentario básico
Poco claro	Alta	Fuente oficial en conflicto; verificación externa sólida
Alta	Poco claro	Declaración oficial correctamente transmitida; verificación de la realidad insuficiente
Poco claro	Poco claro	Referencia insuficiente para decisión

Esta tabla establece un principio importante:

> *El incumplimiento de la IA con el registro oficial no siempre es un fracaso.*

Asimismo:

> *El cumplimiento de la IA con el registro oficial no siempre es un éxito.*

17. REGLA PARA PRESERVAR EL ESTATUS DE LA FUENTE

La respuesta de la IA debe preservar no solo el contenido sino también el estatus epistémico del contenido. Las siguientes oraciones no son iguales:

- "La empresa es la organización GEO líder en el mundo."
- "La empresa se describe a sí misma como la organización GEO líder en el mundo."
- "La empresa afirma ser la organización GEO líder mundial basada en su investigación publicada que aún no ha sido verificada de manera independiente."
- "En la comparación independiente especificada, la empresa ocupó el primer lugar."

Cada oración conlleva un estatus de evidencia diferente. IA:

- autodeclaración como resultado independiente,
- declaración de marketing como hallazgo científico,
- comunicado de prensa como noticia editorial,
- premio patrocinado como acreditación independiente,
- datos de la empresa como auditoría externa

Este tratamiento está prohibido por la siguiente regla:

Regla de Protección de Citas y Estado de la Fuente

El estado de la fuente debe permanecer explícito.

18. ELEVACIÓN EPISTÉMICA

El fortalecimiento del estado de la fuente de una afirmación en una respuesta de IA se llama:

Elevación epistémica

Transformaciones de ejemplo:

Texto original	Elevación incorrecta por IA
“Según los propios datos de la empresa”	“Probado”
“Reportado por clientes específicos”	“Funciona en todos los clientes”

Texto original	Elevación incorrecta por IA
“Observado en el estudio piloto”	“Generalmente efectivo”
“La empresa se define a sí misma como líder”	“Líder de la industria”
“No se realizó verificación independiente”	Limitación completamente eliminada
“Disponible en algunos mercados”	“Disponible a nivel mundial”

Estas transformaciones no son resúmenes. Son la producción de afirmaciones nuevas y más fuertes.

19. SILENCIO OFICIAL

La ausencia de información en la página oficial de una empresa puede no significar:

- que la información sea incorrecta,
- que la información esté prohibida,

La IA no puede usar esta información. El silencio oficial puede tomar tres formas diferentes.

19.1. Silencio no material

La información puede estar fuera del prompt y de la decisión del usuario. Que la IA la use o no es irrelevante.

19.2. Información externa complementaria legítima

Puede ser información que no se encuentra en el sitio oficial pero que ha sido verificada por una fuente confiable. Ejemplo:

- Registro legal
- Evento histórico
- Decisión regulatoria pública
- Investigación independiente

La IA puede usar esta información con el contexto y el estado de la fuente apropiados.

19.3. Silencio de un límite material

Una entidad puede no haber revelado un límite que afecte la decisión del usuario en su representación oficial. Ejemplo:

- La licencia válida solo en ciertos países
- Que el precio aparente no incluir una tarifa adicional
- Que el servicio esté disponible solo para ciertos tipos de clientes
- Datos de éxito provenientes solo de casos seleccionados

En este caso, la propia representación oficial puede contener un hallazgo de auditoría. La repetición del silencio por parte de la IA puede reducir la veracidad de la realidad.

20. DEFICIENCIA MATERIAL EN REPRESENTACIÓN OFICIAL

Una declaración oficial puede no ser directamente falsa. Sin embargo, eliminar los límites necesarios puede hacerla engañosa. Ejemplo: “Operamos en Europa.” Contexto material:

- solo en dos países,
- solo de manera remota,
- sin una oficina local,
- solo para clientes corporativos

puede ser proporcionado. La capa de representación oficial debe registrar estas dos áreas juntas:

- La afirmación principal publicado
- Los límites materiales necesarios para entender correctamente La afirmación

Si una respuesta de IA contiene la frase principal pero elimina los límites:

- fiel en términos de palabras,
- engañoso en términos de juicio

puede ser.

21. ESTAR EQUIVOCADO ES ESTAR EQUIVOCADO INCLUSO SI TE BENEFICIA

Cuando un producto de IA representa a una entidad como más poderosa, más grande o más autorizada de lo que realmente es, el silencio por parte de la organización se convierte en un problema de medición y gobernanza. Una empresa local puede ser representada como un negocio global; una consultoría como un regulador; la autoevaluación como certificación independiente; o una operación que abarca ocho países como operando en 20. Incluso cuando un error favorece a la organización, debe registrarse como un error en la capa de realidad verificada. La advertencia fundacional del segundo libro establece el mismo límite: un error comercialmente atractivo aún debe corregirse. Cuando una organización explota deliberadamente un error favorable de la IA:

- obligación de corregir
- integridad de la representación
- promesa comercial
- gobernanza

también puede examinarse en términos de

22. CONFLICTO DE REFERENCIA

A veces el problema no está en la respuesta de la IA, sino en los registros de referencia. Ejemplo: La página principal muestra un precio diferente. El PDF muestra un precio diferente. La página de pago muestra otro precio. Los datos estructurados llevan un precio desactualizado. En este caso, puede que no exista una representación oficial correcta única. Otro ejemplo: El registro legal indica una propiedad. El sitio web de la empresa indica otra propiedad. Las fechas de actualización son inciertas. En este caso, el registro de la realidad verificada también puede presentar una discrepancia.

23. REGLA DE DEFECTO DE REFERENCIA

> *Si los registros de referencia son materialmente contradictorios o insuficientes, esta incertidumbre no puede convertirse automáticamente en un fallo del producto de IA.*

Estados correctos:

CONFLICTO_DE_REFERENCIA

EVIDENCIA_INSUFICIENTE

NO RESUELTO

DIVISIÓN_DE_TIEMPO

DIVISIÓN_DE_ALCANCE

son posibles. La auditoría debe separar la pregunta: "¿La IA proporcionó una respuesta incorrecta?" y: "¿Había realidad de referencia necesaria para que la IA proporcionara la respuesta correcta?" Si una institución ha dejado sus registros oficiales contradictorios, podemos medir qué registro elige el modelo. Sin embargo, no podemos ocultar el defecto de referencia mientras esperamos una respuesta institucional única y definitiva del modelo.

24. FIJACIÓN DE REGISTRO CANÓNICO

Ambas capas de referencia de GEO-1000 deberían ser versionadas antes de que comience la ola de medición:

Registro de Representante Oficial

¿Qué páginas? ¿Qué versiones de idioma? ¿Qué fecha? ¿Qué afirmaciones? ¿Qué estado de archivo?

Registro de Realidad de Entidad Verificada

¿Qué evidencia? ¿Qué evidencia en contrario? ¿Qué estados de verificación? ¿Qué objeciones? ¿Qué fecha? Si hay un cambio material durante la medición: La ola puede detenerse. El registro puede dividirse en una nueva versión. Las observaciones anteriores y posteriores pueden analizarse por separado. Está prohibido mezclar los resultados bajo una sola referencia.

25. CARGA DE LA PRUEBA SEGÚN LA FUERZA de la afirmación

A medida que una afirmación crece, la carga de la prueba también aumenta. La escalera a continuación es una representación constructiva.

C0 — afirmación de Identidad Simple

“X es una empresa de consultoría.” Se requiere un registro correcto de la entidad y la actividad.

C1 — afirmación de Alcance

"X ofrece servicios en Alemania." Se requiere un registro de operación, contrato o entrega.

C2 — afirmación de Rendimiento Cuantitativo

"90% de proyectos se completan a tiempo." Se requieren denominador, período, método y resultados no exitosos.

C3 — Afirmación comparativa

"Es más rápido que los competidores." Se requieren el universo de comparación y la medición equivalente.

C4 — afirmación de Superioridad

"Es el líder de la industria." Se requieren industria, criterios, geografía, tiempo y método.

C5 — Garantía o Provisión Universal

“Garantiza resultados para cada cliente.” La verificación verdadera requiere evidencia extremadamente sólida, integral y procesable. Las fuentes oficiales pueden ser fuertes para las afirmaciones C0 y algunas C1. Por sí solas, pueden no ser suficientes en los niveles C3–C5.

26. DOS MEDICIONES SEPARADAS DE PRECISIÓN

Se deben informar dos mediciones principales por separado dentro de GEO-1000.

26.1. Fidelidad de la Representación Oficial

Responde a la siguiente pregunta:

> *¿En qué medida la respuesta de la IA transmite la representación oficial y canónica versionada de la entidad de manera precisa, completa y preservando el estatus de la fuente?*

Esta medición:

- identidad oficial,
- alcance del servicio oficial,
- Precio publicado,
- límites oficiales,
- naturaleza de la fuente oficial

evalúa sus áreas.

26.2. Precisión de la Realidad de la Entidad Verificada

Responde a la siguiente pregunta:

> *¿En qué medida la respuesta de la IA transmite la realidad actual de la entidad, respaldada por la evidencia disponible, de manera precisa, completa y con los límites preservados?*

Esta medición:

- afirmaciones fácticas respaldadas,
- afirmaciones conflictivas,
- Información desactualizada,
- Relaciones incorrectas,
- Deficiencias fácticas,
- mejoras epistémicas

son evaluadas.

27. LA INCAPACIDAD DE DOS PUNTUACIONES PARA COMPENSARSE ENTRE SÍ

La alta fidelidad oficial no puede compensar la baja precisión verificada. Ejemplo: El sitio oficial publica una afirmación falsa. La IA repite la afirmación exactamente. La fidelidad oficial puede ser alta. La precisión verificada es baja. Del mismo modo, la alta precisión verificada no trivializa

completamente la distorsión de la representación oficial. Ejemplo: La IA usa correctamente una fuente externa actualizada. Confunde la descripción de productos y servicios de una empresa con otra. Algunos hechos externos en la respuesta pueden ser correctos. La fidelidad de identidad oficial es débil. Por lo tanto, los dos resultados deberían aparecer por separado.

28. ESTADOS DE RESPUESTA DE DOBLE CAPA

Una respuesta de IA puede recibir uno de estos resultados:

PASO-DOBLE

Es materialmente consistente tanto con la representación oficial como con la realidad verificada.

OFICIAL-APROBADO / VERIFICADO-FALLIDO

Lleva fielmente el registro oficial. Sin embargo, transmite una afirmación no respaldada, antigua o contradictoria del registro oficial como si fuera verdadera.

OFICIAL-FALLIDO / VERIFICADO-APROBADO

No cumple completamente con el sitio oficial. Sin embargo, transmite la verdad correctamente, respaldada por evidencia actual y más contundente. Se prefiere una explicación de la diferencia en la respuesta.

FALLA-DUAL

Perturba materialmente tanto el registro oficial como la realidad verificada.

OFICIAL-APROBADO / VERIFICADO-DESCONOCIDO

La declaración oficial ha sido transmitida con precisión. El estado de la realidad verificada de la afirmación aún no ha sido determinado.

CONFLICTO DE REFERENCIA

Los registros de referencia son contradictorios en sí mismos. No es posible calificar la respuesta de manera definitiva.

INCERTIDUMBRE-APROPIADA

La IA ha explicado correctamente la incertidumbre de la referencia y ha evitado emitir un juicio definitivo. En algunos casos, esta puede ser la respuesta más precisa.

29. SINTÉTICO APPLE.COM PANTALLA

DEMOSTRACIÓN DE METODOLOGÍA SINTÉTICA / Las siguientes afirmaciones, respuestas y puntuaciones son completamente ficticias. No representan información real, rendimiento ni respuestas de ningún producto real de IA de Apple Inc. Historial de Representación Oficial Sintética: O-01: La empresa se define como una marca global de tecnología para consumidores. O-02: La empresa publica ciertas categorías de productos y servicios digitales. O-03: La empresa se define como 'la empresa más innovadora del mundo'. O-04: La web oficial contiene condiciones específicas de las tiendas rurales. Historial de Realidad Verificada Sintética: V-01: Se admiten la identidad corporativa y la categoría principal de negocio. V-02: Se admiten ciertas categorías de productos. V-03: No existe un

método de comparación definido para la afirmación de 'la empresa más innovadora del mundo'. V-04: Las condiciones de la tienda varían según el país.

29.1. Respuesta Sintética A

"Apple.com es una de las interfaces corporativas oficiales donde la empresa presenta sus productos de tecnología para consumidores y servicios digitales. La empresa se posiciona como una de las compañías más innovadoras del mundo; esta afirmación de superioridad también debe evaluarse de acuerdo con el método de comparación utilizado." Evaluación: Se mantiene la representación oficial. La afirmación de superioridad se atribuye a la empresa. La afirmación no ha sido elevada a una factualidad independiente. Se ha añadido la limitación de alcance. Estado de candidato:

PASO-DOBLE

29.2. Respuesta Sintética B

"Apple es la compañía más innovadora del mundo, indiscutiblemente." Evaluación: La afirmación oficial de marketing se ha convertido en una verdad absoluta. No hay un método de comparación. Se pierde el estado de fuente. "Indiscutible" es una elevación epistémica adicional. Estado del candidato:

OFICIAL-APROBADO / VERIFICADO-FALLIDO

29.3. Respuesta Sintética C

"Apple es una empresa local que solo vende teléfonos en tiendas físicas." Evaluación: La identidad corporativa y el alcance están distorsionados. La representación oficial es incorrecta. Contradice la realidad verificable. Estado del candidato:

FALLA-DUAL

29.4. Respuesta Sintética D

"El sitio web oficial de la empresa publica ciertas condiciones de la tienda; sin embargo, los precios y los términos de compra pueden variar según la versión del país." Evaluación: Se ha utilizado correctamente el registro oficial como fuente. Se conserva el alcance geográfico. La página oficial no se ha convertido en una fuente única de precios para todo el mundo. Estado del candidato:

PASO-DOBLE

30. ESTUCHE SINTÉTICO ASTERON

CASO SINTÉTICO — NO ES UNA INSTITUCIÓN REAL

El sitio oficial de una empresa llamada Asteron publica las siguientes afirmaciones:

- "Operamos en 25 países."
- "Logramos éxito en 98% de los proyectos."
- Somos una organización de estándares globales verificada de forma independiente.
- Todos nuestros proyectos están garantizados.

Registros de auditoría:

- País de servicio activo verificado: 9
- Datos de éxito: solo los 40 proyectos seleccionados

- Verificación independiente: no encontrada
- Estándar: el estándar recomendado publicado por la propia empresa
- Garantía: únicamente contractual para ciertos elementos de entrega

Respuesta de IA: 'Asteron es una organización estandarizada a nivel mundial, que opera en 25 países, con una tasa de éxito de 98%, verificada de manera independiente, y garantiza resultados en todos sus proyectos.'

30.1. Fidelidad de la representación oficial

La respuesta es en gran medida fiel al texto del sitio web.

30.2. Precisión de la Realidad Verificada

Respuesta:

- amplía el alcance del país,
- hace que los datos seleccionados sean un éxito general,
- presenta el estándar propio de la entidad como un estándar independiente,

convierte la garantía de entrega en garantía de resultado. Es un fallo grave en términos de realidad verificada.

30.3. ¿Cómo podría ser una respuesta correcta de IA?

“Asteron afirma en su propio sitio web que ofrece servicios en 25 países, tiene una tasa de éxito de 98% y publica un estándar. En los registros de auditoría existentes, se ha verificado el servicio activo en nueve países; se ha observado que la tasa 98% se basa en 40 proyectos seleccionados; no se pudo confirmar la verificación independiente. La garantía se limita a artículos de entrega específicos.” Esta respuesta:

- no oculta la declaración de la empresa.
- preserva el estado de la declaración,
- distingue la realidad verificada,

hace visible la limitación.

31. CASO SINTÉTICO QUE CONTIENE DEFECTO DE REFERENCIA

CASO SINTÉTICO

Página oficial: “Precio inicial 2,000 euros.” PDF de precio oficial: “Precio inicial 3,000 euros.” Pago: “4,000 euros.” Datos estructurados: “1,500 euros.” Respuesta de IA: “El precio inicial es 3,000 euros.” ¿Qué registro eligió la IA? El PDF. ¿Está mal? Tal vez. ¿Está correcto? Tal vez. El registro de referencia es inconsistente. En este caso:

- Fallo de IA,
- conflicto de representación oficial
- falta de fuente canónica

son hallazgos separados. La evaluación correcta:

CONFLICTO DE REFERENCIA

es posible. Una entidad no puede esperar fidelidad exacta del precio sin reconciliar sus propios registros de precios.

32. PRIORIDAD DE FUENTE EN LA CAPA DE REPRESENTACIÓN OFICIAL

Cuando se encuentran múltiples registros oficiales, se debe establecer una regla de prioridad para cada campo de afirmación. Ejemplo de prioridad de candidatos:

- Registro de contrato o transacción actual
- Registro canónico actualizado de producto/precio
- Datos estructurados autorizados
- Página oficial actual
- PDF oficial actual
- Registro histórico de prensa o blog

Este orden no es universal para todas las afirmaciones. Por ejemplo, los archivos de prensa pueden ser valiosos para la historia corporativa. Un antiguo registro de blog no es autoritativo para el precio. La prioridad debe definirse de acuerdo con el tipo de afirmación.

33. CONFLICTO DE FUENTE EN LA REALIDAD VERIFICADA

Dos fuentes independientes pueden contradecirse entre sí. En este caso:

- la más independiente,
- el más nuevo,
- el que está más cerca del sujeto,
- el que tiene un método más claro,
- el que utiliza datos más amplios

no gana automáticamente. Todos los criterios deben evaluarse juntos. Ejemplo: Una nueva entrada de blog puede estar más actualizada que un registro oficial antiguo. Sin embargo, el blog puede no tener suficiente autoridad para la propiedad legal. La resolución de conflictos de fuentes debe ser sensible al tipo de afirmación.

34. CONSERVACIÓN OBLIGATORIA DE evidencia en contrario

Un registro verificado de hechos no puede incluir solo registros que respalden la afirmación. También se debe preservar lo siguiente:

- Registros que contradicen la afirmación
- Casos no exitosos
- Resultados retirados
- Certificados vencidos

- Resultados negativos para el usuario
- Decisiones regulatorias
- Correcciones y erratas
- Discrepancia entre fuentes

Respaldar una afirmación no permite el despido de evidencia en contrario.

ARTÍCULO 35

No todas las diferencias verificadas oficialmente tienen el mismo nivel de importancia. Una diferencia material es una diferencia que podría cambiar la decisión del usuario o del sistema. Ejemplos: Una diferencia menor en la ortografía del nombre de la empresa puede no ser material. Escribir incorrectamente el tipo de empresa puede ser material. La falta del color de un producto puede no ser material. Un país de servicio incorrecto es material. Una diferencia en el formato de fecha puede no ser material. Una fecha de validez de licencia incorrecta es material. Un cambio en el tono de marketing puede no ser material. Eliminar el límite de la garantía es material. La materialidad debe evaluarse en función de su impacto en:

- decisión del usuario,
- responsabilidad legal,
- precio,
- de seguridad,
- autoridad,
- alcance,
- evidencia,
- tiempo

Estas categorías deben ser evaluadas por su efecto en la decisión del usuario, la responsabilidad legal, el precio, la seguridad, la autoridad, el alcance, la evidencia y el tiempo.

36. DISPOSICIONES NORMATIVAS OBLIGATORIAS

CH04-N01

Cada respuesta de material de IA debe evaluarse por separado frente al Registro de Representación Oficial y al Registro de Realidad de Entidad Verificada.

CH04-N02

La fidelidad de la representación oficial no puede presentarse como el mismo resultado que la exactitud material.

CH04-N03

Una afirmación en el propio sitio de la entidad no cuenta como verificación independiente sin evidencia adicional adecuada al tipo de afirmación.

CH04-N04

La evidencia de primera mano no puede ser rechazada únicamente por ser de primera mano; se debe evaluar su adecuación al tipo de afirmación, al método y al conflicto de intereses.

CH04-N05

Una fuente independiente no puede ser aceptada como correcta o autoritaria únicamente por ser independiente.

CH04-N06

La opinión de la empresa, la afirmación de marketing o la autoevaluación no pueden elevarse a hecho independiente en la respuesta de la IA.

CH04-N07

El tipo de fuente y el estado epistémico deben mantenerse en la respuesta de la IA en la medida en que sean relevantes.

CH04-N08

Una respuesta de IA que contradiga el registro oficial no puede considerarse un fallo automático si está respaldada por evidencia más sólida y reciente.

CH04-N09

Una respuesta de IA fiel al registro oficial no puede considerarse exitosa en precisión general si la afirmación oficial contradice el hecho verificado.

CH04-N10

Si los registros oficiales y verificados entran en conflicto de manera significativa, el resultado debería ser REFERENCE_CONFLICT, UNRESOLVED, o tener un estado de incertidumbre apropiado.

CH04-N11

Un defecto de referencia no puede convertirse silenciosamente en un error de producto de IA.

CH04-N12

Si la respuesta de la IA trata una afirmación de primera mano no verificada como un hecho, se debe buscar la citación y limitación apropiadas.

CH04-N13

No se pueden dejar las omisiones oficiales importantes simplemente como si la sentencia publicada fuera correcta.

CH04-N14

Los errores favorables a la empresa también deben registrarse como errores en la capa de realidad verificada.

CH04-N15

El Registro de Representación Oficial y el Registro de Realidad de Entidad Verificada deben ser versionados antes de que comience la medición.

CH04-N16

Si hay un cambio en la referencia del material durante la ola de medición, las observaciones anteriores y posteriores deben evaluarse en versiones separadas.

CH04-N17

evidencia en contrario y los registros negativos no pueden eliminarse del archivo de realidad verificada únicamente porque sean negativos.

CH04-N18

Los resultados de EVIDENCIA NO RESUELTA, DESCONOCIDA e INSUFICIENTE no se consideran una validación positiva.

CH04-N19

A medida que aumenta la solidez de la afirmación, también debe aumentar la carga de evidencia requerida.

CH04-N20

La fidelidad de la representación de un alto funcionario no puede compensar la baja precisión de la realidad de la entidad verificada.

CH04-N21

Debe haber un propietario humano o institucional responsable como resultado de la evaluación de dos niveles.

37. FORMAS DE FALLA

CH04-F01— CONSIDERANDO EL SITIO WEB COMO REALIDAD ABSOLUTA

Todas las afirmaciones publicadas por la institución se aceptan como hechos verificados.

CH04-F02— CONSIDERAR LOS REGISTROS OFICIALES COMO INÚTILES

Se ignoran los registros institucionales directos como precio, política o cobertura de productos porque provienen de la primera parte.

CH04-F03— HACIENDO DE LA AUTODECLARACIÓN UNA EVIDENCIA INDEPENDIENTE

La propia afirmación de liderazgo o éxito de la empresa se utiliza como verificación externa.

CH04-F04— ADORACIÓN DE FUENTES INDEPENDIENTES

Se consideran superiores las fuentes independientes pero distantes del tema, antiguas o poco metódicas frente a los registros oficiales y cercanos.

CH04-F05— ELEVACIÓN EPISTÉMICA

La frase "La empresa dice" se convierte en el juicio "se ha demostrado."

CH04-F06— BORRANDO ESTADO DE LA FUENTE

Los datos de primera mano, el contenido patrocinado o los comunicados de prensa se presentan como noticias independientes.

CH04-F07— REPRODUCIR FIELMENTE ERRORES OFICIALES

La IA repite correctamente la afirmación falsa oficial y generalmente se considera exitosa.

CH04-F08— CASTIGANDO CORRECCIONES EXTERNAS ACTUALIZADAS

Se considera que la IA está equivocada porque abandona el sitio oficial antiguo y utiliza evidencia actual sólida.

CH04-F09— CONSIDERAR EL SILENCIO OFICIAL COMO PROHIBIDO

Cualquier información externa que no se encuentre en el sitio oficial se considera automáticamente incorrecta.

CH04-F10— IGNORANDO LA DEFICIENCIA MATERIAL

Incluso si la sentencia oficial es correcta, no se considera el límite necesario.

CH04-F11— CONSIDERANDO EL ERROR DE REFERENCIA COMO DEFECTO DEL MODELO

Solo se culpa a la IA por registros contradictorios de precio o identidad.

CH04-F12— OCULTANDO CONFLICTO DE REFERENCIA

Los adjudicadores eligen la fuente que desean entre registros contradictorios después del resultado.

CH04-F13— JERARQUÍA DE FUENTES INDEPENDIENTES INDEPENDIENTEMENTE DEL TIPO DE afirmación

En cualquier caso, se asume que la fuente oficial o independiente es superior.

CH04-F14— ELIMINANDO PRUEBAS CONTRARIAS

Los registros que debilitan una afirmación positivo se eliminan del archivo de verificación.

CH04-F15— CONSIDERANDO LOS REGISTROS OFICIALES ANTIGUOS COMO ACTUALES

Archivos, PDFs antiguos o comunicados de prensa se convierten en la fuente canónica actual.

CH04-F16— CONTANDO afirmaciones NO VERIFICADOS COMO APROBADOS

Dado que la falsedad no está probada, la afirmación se considera verificada.

CH04-F17— SOBRESTIMAR LA FUERZA de la afirmación A PARTIR DE LA EVIDENCIA

Datos piloto limitados o locales se generalizan a resultados universales.

CH04-F18— puntuación GEO DE UNA SOLA CAPA

La precisión verificada con fidelidad oficial se combina de manera invisible dentro de un único resultado.

CH04-F19— PRESERVAR EL SESGO FAVORABLE

No se realiza corrección cuando la institución de IA exagera.

CH04-F20— CONTAR EL REGISTRO DE LA REALIDAD COMO DECISIÓN FUNDAMENTAL

La declaración de NobleJackal o Kaan MURAZ se transforma en realidad verificada sin evidencia externa.

38. PROCEDIMIENTO DE AUDITORÍA

Paso 1— Extraer afirmaciones de IA Materiales

Respuesta:

- afirmaciones atómicos,
- relaciones,
- áreas de alcance y tiempo

están separadas.

Paso 2— Determinar el Sujeto de Cada afirmación

La afirmación es sobre cuál:

- empresa,
- marca,
- producto,
- persona,
- entidad local

?

Paso 3— Encontrar el Registro de Representación Oficial

¿Se encuentra la afirmación en interfaces oficiales? ¿En qué fuente? ¿En qué versión? ¿En qué fecha? ¿En qué ámbito?

Paso 4— Determinar el Estado Oficial

afirmación:

- canónica,
- histórica,
- retirada,
- borrador,
- conflictivo,
- desconocido

¿Cuál de sus estados es?

Paso 5— Clasificar el Tipo de afirmación

afirmación:

- identidad,
- precio,
- alcance,
- licencia,
- rendimiento,
- liderazgo,
- asociación,
- resultado del usuario

¿Cuál de sus tipos es?

Paso 6— Determinar el Tipo Apropiado de Evidencia

¿Qué fuente es realmente autoritativa para una afirmación? No se utiliza un único ranking universal de fuentes.

Paso 7— Recopilar Verificación y evidencia en contrario

Se evalúan los registros de apoyo, registros conflictivos, actualidad, alcance, método e independencia.

Paso 8— Asignar Estado de Verificación

Se asignan los estados VR-0 hasta VR-5, VR-C, VR-U, VR-T o VR-S según corresponda.

Paso 9— Revisar la Protección del Estado de la Fuente

IA: ¿Ha preservado la diferencia entre "la empresa dice" y "probado"? ¿Ha elevado la autodeclaración a realidad independiente?

Paso 10— Examinar Deficiencias Materiales

En la respuesta o representación oficial, necesario para la decisión del usuario:

- límite,
- tiempo,

- gografía,
- licencia,
- método de datos

¿faltante?

Paso 11— Defecto de Referencia Separado

Si el registro oficial o verificado está defectuoso en sí mismo, se crea un hallazgo separado del error de la IA.

Paso 12— Proporcionar Resultado de Doble Capa

Respuesta:

PASO-DOBLE

OFICIAL-APROBADO / VERIFICADO-FALLIDO

OFICIAL-FALLIDO / VERIFICADO-APROBADO

FALLA-DUAL

CONFLICTO DE REFERENCIA

INCERTIDUMBRE-APROPIADA

se guarda con un estado apropiado.

39. EVIDENCIA REQUERIDA

Registro Canónico de Entidad Registro de Representación Oficial Registro de Realidad de Entidad Presencia en páginas web oficiales PDFs y documentos oficiales Datos estructurados Registros canónicos JSON o API Registros legales y regulatorios Verificaciones de licencias y certificados Registros de productos, precios y pagos Registros de contratos y operaciones Verificaciones de clientes y socios Conjuntos de datos de rendimiento Metodología y denominadores Reseñas independientes Patrocinador y conflicto de interés evidencia en contrario Resultados negativos Registros de erratas y retractaciones Reglas de precedencia de fuentes oficiales Estados de verificación Decisiones de adjudicadores Registros de apelaciones Fecha de medición Versiones de referencia Persona o institución responsable

40. LISTA DE VERIFICACIÓN DE AUDITORÍA

¿Existe un Registro de Representación Oficial? ¿Existe un Registro de Realidad de Entidad Verificada? ¿Fueron ambas versionadas para la misma fecha de medición? ¿Es correcto el sujeto de cada afirmación de IA? ¿Se publicó oficialmente la afirmación? ¿Está la fuente oficial actual y es canónica? ¿Los registros oficiales entran en conflicto entre sí? ¿Se utilizó la fuente de evidencia apropiada para el tipo de afirmación? ¿Se rechazó automáticamente una fuente de primera parte? ¿Se trató automáticamente una declaración de primera parte como evidencia independiente? ¿Es genuinamente relevante la fuente independiente? ¿Se preservaron el alcance y el tiempo de la afirmación? ¿Contiene el archivo evidencia en contrario? ¿Se elevó una autodeclaración a realidad

verificada? ¿Se transmitió la opinión de la empresa con el estado de su fuente intacto? ¿Falta un límite material o exclusión? ¿Se registró un error favorable a la entidad como tal?

¿Se ha separado el defecto de referencia del error de la IA? ¿Se han considerado positivos los campos DESCONOCIDO e IRRESUELTO? ¿Oculta la alta fidelidad oficial una baja precisión verificada? ¿Se puede explicar la desviación de la IA del registro oficial con evidencia sólida? ¿Ha cambiado el registro de referencia durante la medición? ¿Están las observaciones antiguas y nuevas en versiones separadas? ¿Aparece el resultado de doble capa en el registro público? ¿Se conoce al propietario responsable del resultado?

41. OBJECIONES Y RESPUESTAS

Objeción 1— "La empresa conoce la información más precisa sobre sí misma."

Esto puede ser cierto en muchos aspectos. La empresa:

- su propio producto,
- su precio,
- su política de soporte,
- su operación

lo sabe bien. Sin embargo, la empresa también es la propietaria comercial de la afirmación. Las frases que requieren comparación externa, como "el mejor," "líder," "más confiable" y "comprobado," no pueden ser verificadas únicamente por declaración propia.

Objeción 2— "Si no cuentas tu propio sitio como evidencia, ¿cómo sabremos el precio?"

La página de precios oficial es la fuente principal del precio que la institución ofrece al público. Sin embargo:

- pago,
- impuesto,
- cargo adicional,
- contrato

si entra en conflicto, la oferta real se evalúa por separado. La afirmación "el propio sitio no es evidencia" no se puede aplicar de manera equitativa a todos los tipos de afirmaciones. La decisión correcta es:

> *El propio sitio web de una empresa puede ser una fuente primaria para declaraciones bajo su control; no es una verificación independiente automática para afirmaciones de rendimiento y superioridad externas.*

Objeción 3— "Si la IA refleja el sitio oficial, ¿no ha completado su tarea?"

Si el único objetivo es: "¿Qué dice el sitio?" entonces sí. Sin embargo, si el usuario pregunta: "¿Qué está haciendo realmente la empresa?" simplemente repetir el sitio puede no ser suficiente. El tipo de tarea afecta la capa de medición.

Objeción 4— "Las fuentes externas también pueden estar equivocadas."

Eso es correcto. Por lo tanto, la independencia no es automáticamente precisión. La fuente se evalúa en términos de:

- autoridad,
- proximidad,
- actualidad,
- tipo,
- alcance

en términos de mantenimiento.

Objeción 5— "¿Quién determinará el registro verificado de la realidad?"

La empresa por sí sola no puede determinarlo. NOMOS tampoco puede determinarlo por sí solo. El proceso:

- poseedor de evidencia,
- adjudicador,
- experto adecuado,
- apelación,
- registro de versión,
- persona responsable

es necesario. Se necesita una revisión más independiente para las afirmaciones de alto riesgo y con conflictos.

Objeción 6— "Si el registro de la realidad puede cambiar siempre, ¿no sería La puntuación inútil?"

No. La puntuación:

- fecha,
- es significativo con
- alcance

la versión. Estar sujeto a cambios no significa que no pueda ser medido. Significa que no puede ser presentado prematuramente.

Objeción 7— "¿Por qué la IA usaría información negativa que no está en el sitio oficial?"

Información:

- puede ser forzada a las categorías de

  verdadero,
- actual,
- relevante,
- financiero para la decisión del usuario

puede ser utilizada. Sin embargo, se deben preservar la fuente y el contexto. El silencio del sitio oficial no elimina la realidad negativa.

Objeción 8— 'Si la IA informa la afirmación de la empresa como 'la empresa dice', no aporta valor al usuario.'

Al contrario, muestra correctamente el estado de la afirmación. Usuario:

- La opinión de la institución,

- El hallazgo probado,
- Ambigüedad abierta

puede distinguir. Parecer seguro no es lo mismo que parecer correcto.

Objeción 9— "¿Por qué deberíamos corregir errores menores que favorecen a la empresa?"

Por la expectativa equivocada:

- cliente equivocado,
- fallo en la entrega,
- regresar,
- queja,
- Pérdida de confianza

Puede producir. La ventaja es la visibilidad aparente a corto plazo; A largo plazo es el deber de representación.

Objeción 10— "¿No complican la comunicación pública dos puntuaciones separadas?"

Ocultar toda esta complejidad crea un riesgo mayor. Una tarjeta de resultados concisa podría, en cambio, indicar: Fidelidad de la Representación Oficial — alta / Precisión de la Realidad de la Entidad Verificada — media / Brechas principales — afirmaciones de liderazgo y cobertura del país no respaldadas. Eso es más claro que un número engañoso.

1. el Registro de Representación Oficial; y

2. el Registro de la Realidad de la Entidad Verificada Restringida por Evidencia.

La fidelidad oficial y la exactitud verificada NO DEBEN considerarse como el mismo resultado. Una respuesta de IA PUEDE reproducir fielmente una declaración oficial mientras permanece sin respaldo fáctico, contradicha, desactualizada, demasiado general o materialmente engañosa. Una respuesta de IA PUEDE diferir de un registro oficial sin fallar cuando el registro oficial está obsoleto, incompleto, en conflicto o contradicho por evidencia más sólida y apropiada. Las autoafirmaciones, los datos de primera mano, la evidencia independiente, el contenido patrocinado, los registros oficiales, los registros transaccionales y las declaraciones de terceros DEBEN conservar su fuente y estado de evidencia. Los conflictos de referencia DEBEN registrarse por separado de los fallos de la respuesta de IA. La evidencia desconocida o no resuelta NO DEBE convertirse en verdad verificada.

Cada afirmación sobre el sistema generativo de materiales respecto a la entidad auditada DEBE ser evaluada contra dos capas de referencia versionadas por separado: el Registro de Representación Oficial y el Registro de Realidad de Entidad Verificada y Restringida por Evidencia. La fidelidad oficial y la precisión verificada NO DEBEN ser tratadas como el mismo resultado. Un producto de IA PUEDE reproducir fielmente una afirmación oficial mientras crea una representación materialmente no respaldada, contradicha, desactualizada o demasiado amplia. Cuando el registro oficial está obsoleto, incompleto o en conflicto, una respuesta diferente basada en evidencia más sólida no es automáticamente un error. Se DEBE preservar el tipo de fuente y el estado de evidencia, y los conflictos de referencia DEBEN registrarse por separado de los fallos de IA. Una Alta Fidelidad

de Representación Oficial NO DEBE compensar una Baja Precisión en la Realidad de Entidad Verificada.

JUICIO COMÚN DEL CAPÍTULO 44

Una empresa tiene el derecho de hablar sobre sí misma. Puede explicar quién es. Puede describir lo que ofrece. Puede publicar sus precios. Puede establecer sus límites. Puede transmitir su visión. Sin embargo, no todas las frases que una entidad dice sobre sí misma tienen el mismo estatus epistémico. "El precio de nuestro paquete de servicios es 2,000 euros" y "Somos la mejor empresa del mundo" no son del mismo tipo de afirmación. La primera frase puede ser una declaración comercial controlada por la empresa. La segunda frase requiere comparación externa. Si un producto de IA transmite ambas con igual certeza, socava el estatus de la fuente. Un producto de IA puede replicar perfectamente un sitio web de la empresa. Pero el sitio web:

- antiguo,
- conflictivo,
- no probado,
- más allá del alcance,
- oculta límites materiales

entonces la fidelidad a ese registro no es buena GEO. Por el contrario, un producto de IA puede apartarse del sitio web oficial y aun así ser correcto cuando su respuesta se basa en evidencia más fuerte, actual y apropiada. La desviación puede ser una corrección en lugar de un error. La cuarta ley de medición de NOMOS es por lo tanto:

> *La representación oficial no reemplaza la realidad verificada.*

La quinta ley es la siguiente:

> *La realidad verificada no puede establecerse ignorando la fuente oficial.*

Su sexta ley es:

> *La representación depende no solo de lo que una fuente dice, sino también de la autoridad bajo la cual habla y del estatus evidencial de la declaración.*

Su séptima ley es la siguiente:

> *Reproducir fielmente un error oficial no es una representación precisa.*

Su octava ley es la siguiente:

> *El registro de la realidad no es un juicio absoluto e indiscutible; es un archivo de decisiones con versiones y basado en evidencia.*

Orden de la Sección NOMOS 4

> *No solo me muestres lo que está escrito en tu sitio. / Muestra qué oración está respaldada por qué evidencia.*
>
> *No presentes tus propias palabras como si estuvieran verificadas de manera independiente.*
>
> *No devalúes los datos de primera mano solo porque te pertenecen. / Pero no los hagas una verdad indiscutible solo porque te pertenecen.*

No conviertas la opinión de la empresa en hecho, un comunicado de prensa en noticia, ni un premio patrocinado en autoridad independiente.

No me consideres exitoso cuando repito correctamente el error oficial.

Si su registro oficial está desactualizado, no penalice una respuesta respaldada por evidencia actual únicamente porque difiera de ese registro.

No permanezcas en silencio sobre el error a tu favor.

No elimine evidencia en contrario del archivo.

Cuando tus referencias entren en conflicto, no inventes una única verdad. / Registra el conflicto. / Si la respuesta es desconocida, dilo.

Separa tu afirmación, su fuente, su evidencia, su límite y su tiempo unos de otros.

Antes de ordenarme qué decir, demuestra que tienes derecho a decirlo.

La oración de cierre del capítulo

Un buen GEO no es la repetición fiel de lo que dice una entidad. Es la transmisión precisa de lo que se ha dicho, sin distorsionar su fuente, evidencia, alcance o tiempo.

PARTE II

Población mundial y equidad entre países y lenguas

CAPÍTULO 05

La población usuaria pertinente para cada producto de IA

BASES METODOLÓGICAS / K04; K05; K06; K07

Límite de Capítulo

Los primeros cuatro capítulos establecieron los objetos básicos de medición: una sola respuesta no es una puntuación GEO; la representación es una distribución condicional, no un solo número; la entidad auditada debe resolverse antes de observar los resultados; y una respuesta de IA debe evaluarse tanto contra la representación oficial como contra la realidad verificada basada en evidencia. Ahora debemos definir el denominador de la medición:

> *¿Sobre qué población humana hará esta distribución una declaración?*

¿Todos los que viven en un país? ¿Solo adultos? ¿Aquellos con acceso a internet? ¿Aquellos que pueden usar oficialmente un producto de IA específico? ¿Titulares de cuentas gratuitas? ¿Usuarios de planes pagos? ¿Personas que pueden usar el producto en el idioma de supervisión? ¿Usuarios verdaderamente activos? ¿Usuarios potenciales que podrían usarlo algún día? ¿Personas en mercados donde la marca puede ofrecer servicios? Si la población de un país es muy grande pero el producto de IA relevante no es accesible oficialmente en ese país, ¿cómo se incluirá ese país en el panel principal del producto? Si la proporción de población de un país pequeño está muy por debajo de una persona en una muestra de 1,000, ¿ese país será completamente invisible?

Si una persona conoce tres idiomas, ¿debería contarse como tres personas? ¿A qué población pertenece un usuario si puede acceder a la interfaz móvil de un producto pero no a su interfaz web? Hasta que estas preguntas sean respondidas, '1,000 usuarios aleatorios para cada producto de IA' no define una población objetivo científica; solo indica un número de personas. Este capítulo no descarta la población del país. Le asigna su papel adecuado:

> *La población del país es la base inicial para el muestreo; no es la población final de usuarios elegibles.*

En esta sección, el sistema de auditoría aplica directamente el requisito previamente determinado para país, idioma, fecha, producto y registro de medición a la definición de la población. Porque solicitar evidencia, límites, contexto y tiempo también requiere definir con la misma claridad a qué personas se hace referencia. Esta sección aún no ha:

- Finalizado cómo se distribuirá la muestra de personas 1,000 entre los países,
- qué algoritmo se usará para redondear fracciones a números enteros,
- de qué participantes del panel se seleccionarán,
- cómo se calcularán los pesos de muestreo,
- o qué método de varianza se utilizará para producir intervalos de confianza

Estos se organizarán en las secciones siguientes. La tarea de la Sección 5 es más básica:

> *Para cada producto de IA y cada pregunta de medición, definir a qué población humana se pueden generalizar los resultados.*

NOMOS Desafío

Abres un mapa del mundo. Sumas las poblaciones de los países. Luego quieres distribuir 1,000 personas según la población. Los países grandes recibirán muchos participantes. Los países pequeños recibirán uno o casi ninguna participación. El núcleo de esta idea es correcto:

> *Una persona en la población mundial no debería ser cientos de veces más influyente que una persona en un país grande solo porque viva en un país pequeño.*

Pero ahora aplica el mismo método a dos productos de IA diferentes. El primer producto:

- puede usarse oficialmente en algunos países,
- está limitado en algunos países.

No se ofrece en absoluto en algunos países. La cobertura de países del segundo producto es diferente. ¿Qué ocurre si divides la población mundial de la misma manera entre los dos productos? En los países donde el primer producto no está disponible, se puede pedir a las personas que proporcionen mediciones. Los participantes:

- pueden ser alentados a falsear su ubicación,
- usar métodos de acceso prohibidos o no compatibles,
- abrir una cuenta en otro país,
- eludir la seguridad y las condiciones de uso

Esto no mide la verdadera experiencia del usuario. Genera una experiencia de acceso artificial. Ahora considera otro problema. En cierto país, la población es muy grande. Sin embargo:

- una parte significativa de la población no cumple con el requisito de edad mínima del producto.
- no todos tienen acceso confiable a internet,
- el producto solo puede usarse en ciertos dispositivos
- el lenguaje de auditoría no puede ser utilizado por toda la población,

un plan específico solo está disponible para usuarios con un método de pago. Si haces que toda la población sea el denominador directamente, incluyes personas que en realidad no pueden ser monitoreadas y que no pueden usar el producto en la población objetivo. Ahora imagina un país pequeño. La participación teórica en una muestra de 1,000 bajo ponderación poblacional: 0.03 personas.

Un tercio de una persona no puede ser seleccionado. Si asignas una persona al país, puedes sobrerepresentar al país en el panel principal ponderado por población. Si no asignas a nadie, entonces en ese país:

- acceso,
- idioma,
- comportamiento del producto,
- representación local

pueden permanecer completamente invisibles. Estos dos objetivos son diferentes:

- Medir la experiencia promedio del usuario en la población mundial
- Evitar que los países pequeños queden completamente sin monitoreo

El mismo panel único no puede cumplir perfectamente ambos objetivos. Por lo tanto, NOMOS establece tres principios separados:

> *El peso de la población determina la participación de las personas en el promedio global.*
>
> *La cobertura de observadores evita que los países y lenguas pequeñas sean completamente invisibles.*
>
> *La justicia lingüística y de accesibilidad mide además los subgrupos que pueden ser suprimidos por el promedio de la población.*

La primera premisa de esta sección es:

> *La población total de un país no es la población de usuarios elegibles para IA en ese país.*

Su segunda disposición establece:

> *La población objetivo de un producto de IA no puede transferirse automáticamente a otro producto de IA.*

Su tercera disposición establece:

> *El peso de una persona en la puntuación principal no es el mismo que el derecho de su país a aparecer en la auditoría.*

Su cuarta disposición establece:

> *La falta de acceso al producto no es una representación errónea; sin embargo, es una limitación global de acceso y también debe ser reportada.*

1. OBJETIVO DEL CAPÍTULO

El propósito de esta sección es definir la población humana objetivo de la medición GEO-1000 en términos de producto, pregunta, país, idioma, interfaz y tiempo. La sección estandariza las siguientes distinciones:

- Población total del país y población de usuarios de IA adecuada
- Acceso potencial y uso real
- Accesibilidad del producto y precisión representacional
- Población objetivo y marco de muestreo
- Observación analítica válida con el participante seleccionado
- Cuenta con persona
- Sesión con persona
- Modelo central con producto
- Poblaciones de planes gratuitos y de pago
- Poblaciones de usuarios web y móviles
- Uso que excede la restricción de acceso con acceso oficial
- Cobertura del Country Observer con ponderación de la población del país
- Panel de Equidad Lingüística con participación del idioma ponderada por la población
- Exposición del usuario activo con población potencial de usuarios
- Consulta de dominio de mercado o jurisdicción con consulta de identidad global
- Producto corporativo, educativo o para desarrolladores con producto principal para consumidores
- Accesibilidad de la herramienta de recopilación de datos para usuarios con necesidades de accesibilidad

- Usuarios fuera de la población objetivo y usuarios NO PROBADOS
- Países donde el producto es inutilizable y países donde se reciben respuestas incorrectas en el producto

Al final de esta sección, cada auditoría GEO-1000 debe poder responder explícitamente la siguiente pregunta:

> *¿Sobre qué universo de usuarios, seleccionado de qué personas, hace este resultado un juicio?*

2. DISPOSICIÓN NORMATIVA CENTRAL

> *Antes de que se observen los resultados, cada auditoría GEO-1000 debe definir y versionar una población objetivo de usuarios específica para el producto de IA auditado, la pregunta de medición, el país, el idioma, la interfaz, el plan y las condiciones de tiempo.* [K04–K07]

Población objetivo:

- "población mundial"
- "usuarios de internet",
- "usuarios de IA",
- "clientes"

La población no puede quedar bajo una etiqueta ambigua como ‘usuarios relevantes’. Según corresponda, deben responderse las siguientes preguntas: ¿Qué producto de IA? ¿Qué interfaz de usuario? ¿Qué países? ¿Acceso a partir de qué fecha? ¿Qué condición de edad? ¿Qué idioma o idiomas? ¿Qué dispositivo e interfaz? ¿Qué tipo de plan o cuenta? ¿Cuáles son los términos de uso oficiales? ¿Deben los usuarios pagar para acceder al producto? ¿La pregunta de medición se aplica a todos los usuarios o solo a un mercado o jurisdicción particular? ¿Se están midiendo usuarios activos o potenciales? ¿Qué grupos de usuarios están fuera del alcance? ¿Qué grupos pertenecen a la población objetivo pero están ausentes del marco de muestreo? Un cambio en cualquiera de estos campos puede modificar la población objetivo. Cuando esto ocurre, las puntuaciones antiguas y nuevas no pueden combinarse en la misma serie sin explicación.

3. LA POBLACIÓN NO ES UN NÚMERO ÚNICO

La población total oficial o estimada de un país es solo la capa inicial para GEO-1000. La población de auditoría puede consistir en las siguientes capas de restricción.

3.1. Población residente

La población humana total que vive en el país o jurisdicción especificada. Este número:

- no implica ninguna idoneidad en términos de edad,
- internet,
- acceso al producto,
- idioma,
- cuenta,
- dispositivo

consideraciones.

3.2. Población con edad elegible

La población que cumple con el requisito de edad del producto de IA y del protocolo de investigación correspondiente. Para un producto:

- edad mínima de uso,
- requisito de consentimiento parental,
- regla de edad específica del país

puede existir. El protocolo de investigación puede ser más restrictivo que la condición del producto. Por ejemplo, el panel fundacional GEO-1000 puede estar compuesto solo por participantes adultos para reducir la complejidad ética y legal. En este caso, la puntuación no es: "Toda la población":

sino más bien "Población adulta definida y elegible para el producto"

resultado.

3.3. Población Conectada

La población que puede usar el producto relevante a un nivel donde:

- internet,
- dispositivo,
- calidad de la conexión,
- acceso técnico

está disponible. El acceso a Internet puede no ser una característica binaria. Una persona:

- puede usar la mensajería,
- sin embargo, puede que no pueda ejecutar la aplicación web pesada,
- puede usar internet móvil,
- sin embargo, puede que no pueda acceder a la experiencia de escritorio.

puede que no pueda subir respuestas o capturas de pantalla debido a la baja conectividad. Las limitaciones de conexión no deben excluir automáticamente al usuario de la población objetivo. Debe asociarse con las condiciones reales de uso del producto medido.

3.4. Población accesible desde la perspectiva del producto

Para el producto de IA relevante en la fecha especificada:

- lanzado oficialmente,
- podría utilizarse de acuerdo con la ley y las condiciones del producto,
- se podría acceder a la interfaz necesaria

Es población. La ausencia de un producto en un país no significa que las personas en ese país estén mal representadas. En este caso: la respuesta de representación es ninguna, no hay cobertura de acceso. Los dos resultados deben mantenerse separados.

3.5. Población adecuada en términos de cuenta y plan

Algunos productos:

- pueden requerir una cuenta,
- puede requerir verificación telefónica,
- puede requerir un método de pago,
- puede ofrecerse solamente con un contrato corporativo,

puede desbloquear ciertas funciones solo en un plan de pago. Estas condiciones afectan qué parte de la población objetivo puede acceder realmente al interfaz de usuario relevante. La puntuación del plan gratuito no se puede aplicar automáticamente a los usuarios del plan de pago.

3.6. Población adecuada en términos de idioma

La población que puede usar el prompt y la respuesta de IA relevante a un nivel suficiente en el idioma de auditoría. La idoneidad del idioma por sí sola:

- no puede determinarse solo por el idioma oficial del país,
- la lengua materna de la persona

Debe diferenciarse:

- Lengua materna o del hogar
- Idioma de uso diario
- Idioma de la interfaz de IA
- Idioma del prompt
- Idioma de respuesta
- Idioma para uso profesional
- Dominio de lectura

Una persona puede ser apta en varios idiomas. Esta persona no puede contarse como más de una persona completa en el cálculo de la población global.

3.7. Población elegible para la interfaz

Medición:

- web,
- aplicación móvil,
- aplicación de escritorio,
- espacio de trabajo corporativo,
- asistente integrado

puede ser usado. El universo de usuarios accesible puede diferir para cada interfaz. Un usuario con acceso web puede no tener el producto móvil. El interfaz de usuario corporativo puede no estar disponible públicamente.

3.8. Población elegible para investigación

Incluso si una persona puede usar el producto, para su inclusión en la investigación, además:

- consentimiento voluntario,
- privacidad
- procesamiento de datos,
- recopilación de capturas de pantalla,
- requisitos de investigación específicos por país,
- no implica ninguna idoneidad en términos de edad,
- seguridad

se pueden necesitar cumplir ciertas condiciones. Las personas que no pueden ser incluidas en la investigación pero usan el producto real pueden formar parte de la población objetivo. En este caso, el marco muestral subrepresenta a la población objetivo. Esta subrepresentación no debe ocultarse.

3.9. Población analíticamente elegible

La población para la cual los datos recopilados pueden usarse para producir resultados. El participante:

- desvió el prompt,
- usó el producto incorrecto,
- eligió la respuesta,
- presentó evidencia incompleta

Es posible. Incluso si un individuo pertenece a la población objetivo, su observación puede no entrar en el conjunto de datos analíticos. Las observaciones válidas y la población objetivo no son lo mismo.

4. DEFINICIÓN DE LA POBLACIÓN DE USUARIOS ADECUADA

Para el producto de IA a en el tiempo t, el universo de usuarios elegibles puede definirse de la siguiente manera:

```
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}
```

Aquí:

- Pt: la población humana general en la fecha especificada
- u: humano individual
- a: producto de IA y interfaz de usuario
- t: momento de la medición

En términos de idioma y prompt, el universo objetivo puede restringirse más adelante de la siguiente manera:

```
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}
```

Aquí:

- c: país o jurisdicción
- l: idioma y configuración regional
- s: interfaz de usuario
- p: prompt o familia de preguntas

El propósito de esta notación no es registrar a todos los individuos uno por uno. Es hacer visible el límite lógico de la población objetivo.

5. ESTÁ PROHIBIDO MULTIPLICAR CIEGAMENTE LAS TASAS MARGINALES

La población elegible a veces puede calcularse aproximadamente con una fórmula como la siguiente:

```
N× p_age× p_internet× p_product× p_language
```

Este cálculo solo puede justificarse bajo la independencia de condiciones o la suposición de una distribución conjunta apropiada. En la vida real, las características pueden no ser independientes. Por ejemplo:

- el uso de Internet puede variar según la edad,

- el acceso a un plan de pago puede estar relacionado con los ingresos,
- el uso de ciertos idiomas puede estar relacionado con la región dentro del país y la educación,

el acceso a dispositivos y la calidad de la conexión pueden variar conjuntamente. Multiplicar las tasas marginales puede predecir una población objetivo que es:

- más grande de lo necesario,
- más pequeño de lo necesario

El orden de preferencia debería ser:

- Microdatos conjuntos o tabulación cruzada apropiada
- Proyección poblacional multidimensional confiable
- Modelo calibrado
- Enfoque marginal con justificación explicada y análisis de sensibilidad

Si se usa multiplicación simple:

- suposición,
- fuente de datos,
- incertidumbre,
- escenario alternativo

deben ser divulgados.

6. LA PREGUNTA DE AUDITORÍA CAMBIA LA POBLACIÓN OBJETIVO

No todos los indicios tienen el mismo universo de medición para toda la población mundial.

6.1. Preguntas de Identidad Global

Ejemplo: "¿Qué tipo de empresa es Apple.com?" Esta pregunta puede hacerse ampliamente entre la población de usuarios global adecuada para el producto. Porque la identidad de la empresa no se limita a los usuarios de un mercado específico.

6.2. Preguntas sobre Servicio Local

Ejemplo: "¿Esta empresa ofrece servicios en Turquía?" Población objetivo principal:

- Usuarios elegibles en Turquía,
- Usuarios externos investigando el servicio en Turquía

Es posible. No es necesario dar el mismo peso a toda la población mundial.

6.3. Preguntas sobre Precio y Compra

Precio:

- país,
- impuesto,
- actualidad,
- plan,
- tipo de cliente

puede variar. La población objetivo debe definirse según el mercado comercial relevante y el tipo de usuario.

6.4. Cuestiones Legales y de Licencias

La licencia y la autoridad legal dependen de la jurisdicción. La población objetivo:

- puede obtener el servicio relevante en ese país,
- están sujetos a ese régimen legal.

La población objetivo puede estar restringida en consecuencia.

6.5. Cuestiones de Asesoramiento

La población objetivo para la pregunta "¿Es esta empresa adecuada para mí?" no es toda la gente. Puede definirse según:

- Necesidades del usuario
- País
- Presupuesto
- Tipo de cliente
- Alcance del servicio
- Condiciones de elegibilidad y exclusión

La población de asesoramiento puede ser más estrecha que la población de identidad.

6.6. Preguntas de Alto Riesgo

En los campos de salud, derecho, finanzas o seguridad:

- juicio,
- perfil del usuario,
- estado de riesgo,
- autoridad profesional

limita más fuertemente la población objetivo.

7. EL MARCO POBLACIONAL DEBE ESTAR VINCULADO A LA FAMILIA DE PROMPT

En una auditoría, puede haber múltiples marcos poblacionales en lugar de una única población objetivo.

7.1. Población de Identidad Central

La población global o ampliamente elegible que mide la identidad central de la entidad.

7.2. Población Relevante para el Servicio

Es la población para la cual la entidad puede realmente proporcionar servicio o puede ser considerada para recibir servicio por parte del usuario.

7.3. Población de Transacción de Mercado

El precio específico, la compra, la devolución o la condición comercial es la población a la que puede aplicarse.

7.4. Población Jurisdiccional

Es la población de usuarios a la que puede aplicarse una licencia, afirmación legal o regulatorio específico.

7.5. Población Elegible para Recomendación

Es la población de usuarios que puede ser apropiada dentro del verdadero alcance de la entidad.

7.6. Población Afectada de Alto Riesgo

Es una población para la cual la tergiversación puede causar un daño serio. Estos marcos no deben confundirse entre sí. Una puntuación de identidad global no significa "recomendable para todo el mundo."

8. EL TIPO DE PRODUCTO DE IA DETERMINA LA POBLACIÓN

No todos los productos de IA son productos de chat disponibles públicamente.

8.1. Producto de IA para Consumidor General

Población objetivo:

- Puede haber personas que puedan acceder al producto,
- cumplir con los términos de uso,
- y usarlo en el idioma correspondiente.

El panel de usuarios reales GEO-1000 se puede aplicar directamente a este tipo de producto.

8.2. Producto de IA Empresarial

El producto es solo para:

- contrato del empleador,
- licencia corporativa,
- puede ser utilizado con una cuenta de usuario interna

La población objetivo no es la población adulta general:

> *Es el asiento o universo de usuarios con acceso corporativo al producto.*

Los resultados corporativos no pueden compararse con el producto de consumo general en el ponderado de población.

8.3. Producto Educativo

Población objetivo:

- estudiante,
- profesor,
- institución,
- grupo de edad,
- licencia educativa

puede estar limitada en términos de. Si se van a incluir niños o menores, se requiere un protocolo ético separado.

8.4. Producto para Desarrollador o API

La salida de una API no puede mostrarse directamente a través de una interfaz de usuario final. La unidad objetivo es:

- desarrollador,
- aplicación,
- llamada API,
- sesión de prueba controlada

puede ser. El panel de población humana GEO-1000 no puede aplicarse directamente y sin cambios a la API. Por separado para la API:

Auditoría de Sistema Controlado

puede ser requerida.

8.5. Asistente de IA Integrado

IA:

- sistema operativo,
- interfaz de búsqueda,
- plataforma de compras,
- software empresarial,
- dispositivo

Puede estar integrado. La población objetivo es la población de usuarios real del producto integrado. No es la población total de usuarios del proveedor del modelo base.

8.6. IA Experta o Sectorial

El producto es solo para:

- abogados,
- médicos,
- profesionales de finanzas,
- ingenieros

puede estar diseñado para. La población objetivo debe estar limitada por profesión y condiciones de acceso.

9. ESTADOS DE ACCESO AL PRODUCTO

El estado de acceso debe registrarse para cada país, producto, interfaz y fecha.

PA-0— DESCONOCIDO

No se pudo determinar el estado de acceso de manera confiable.

PA-1— NO DISPONIBLE

El producto no se ofrece en el país o grupo de usuarios relevante.

PA-2— RESTRINGIDO

El producto está limitado a ciertos:

- cuenta,

- invitación,
- institución,
- no implica ninguna idoneidad en términos de edad,
- plan,
- dispositivo

términos.

PA-3— ACCESIBLE PERO NO OFICIALMENTE SOPORTADO

Se puede observar el acceso real al producto. Sin embargo, el soporte oficial o el estado de uso no está claro o es limitado. Estos usuarios no deben ser añadidos automáticamente al panel principal de la población oficial. Se puede crear un panel experimental separado.

PA-4— DISPONIBLE OFICIALMENTE

El producto es accesible oficialmente y se definen las condiciones principales de uso.

PA-5— DISPONIBLE OFICIALMENTE CON INTERFAZ DE AUDITORÍA COMPLETA

El producto no solo es inaccesible; está controlado:

- idioma,
- plan,
- interfaz,
- bajo condiciones de sesión

adecuadas para generar la observación requerida por el protocolo. La disponibilidad del producto en un país no significa que todos los planes y lenguajes cumplan con el control.

10. EL SALTO DE RESTRICCIÓN NO ES UNA MEDICIÓN REAL DE LA POBLACIÓN

No se puede pedir a los participantes que realicen los siguientes comportamientos:

- Declarar el país incorrecto
- Crear una cuenta falsa
- Cambiar la información de edad
- Eludir la restricción de acceso oficial
- Usar la cuenta de otra persona
- Eludir la seguridad o verificación del producto
- Violar los términos de uso

Estos comportamientos también pueden examinarse como un experimento técnico. Sin embargo:

La experiencia de la población de usuarios natural y legítima

no puede ser reportada como tal. Si no hay acceso en un país, el estado correcto debería ser: NO DISPONIBLE / o / FUERA DE ALCANCE NATIVO. Una sola observación que excede la restricción no genera una puntuación de población por país.

11. DISTINCIÓN ENTRE USUARIO POTENCIAL Y USUARIO ACTIVO

Las personas que pueden usar el producto no son la misma comunidad que las personas que realmente usan el producto.

11.1. Población con Acceso Potencial

Estas son las personas que técnicamente y legalmente pueden usar el producto. Esta población responde a la pregunta: '¿Qué vería una persona adecuada si usara el producto?' La población es valiosa para la justicia y la comparación de productos.

11.2. Población de Usuarios Activos

Estas son las personas que realmente usaron el producto durante un período específico. Esta población aborda la pregunta: '¿Qué ven las personas dentro del volumen de uso real?' Requiere datos fiables de uso activo.

11.3. Población de Usuarios Expuestos a Consultas

Estas son las personas que usan el producto y tienen el potencial de hacer preguntas sobre la marca o tema monitoreado. Esta población a menudo no se puede conocer directamente. Puede ser necesario contar con datos de comportamiento de búsqueda o consulta.

11.4. Población que Realmente Ve la Respuesta

Son usuarios que están realmente expuestos a una salida de representación específica. Esto puede no ser completamente medible sin datos reales del proveedor del producto fuera del panel de usuarios.

12. EL PESO DE EXPOSICIÓN NO ES IGUAL AL PESO DE ELEGIBILIDAD

Se pueden generar dos resultados globales separados.

12.1. Resultado Ponderado por Población Elegible

Cada persona se pondera según su participación dentro de la población de usuarios elegibles definida. Este resultado es:

- globalmente justo,
- potencialmente accesible,
- representativo de la población

en términos de

12.2. Resultado Ponderado por Exposición Real

Los productos de IA o los grupos de usuarios se ponderan según los volúmenes de uso reales. Este resultado:

- es valioso en términos de
- impacto comercial actual,
- auténtica exposición del usuario.

riesgo operacional.

Sin embargo, los datos de uso confiables pueden no estar disponibles. Las estimaciones pueden depender de declaraciones del proveedor, datos de mercado o muestreos incompletos. Por lo tanto, el tipo de ponderación debe indicarse explícitamente.

12.3. Sustituir uno por otro está prohibido

Puntuación de la población elegible: no significa "La mayoría de los usuarios reales vieron esta respuesta." Puntuación de exposición: no significa "Se asegura una representación justa de la población mundial."

13. EL ALCANCE DE ACCESO Y LA EXACTITUD DE LA REPRESENTACIÓN DEBERÍAN ESTAR SEPARADOS

Un producto de IA puede estar disponible solo en países limitados. En estos países, puede generar representaciones muy precisas. Otro producto puede tener un acceso más amplio pero puede producir menor precisión. Si solo analizamos la precisión de la representación, el producto con acceso limitado puede parecer superior. Si solo analizamos la cobertura de acceso, se puede pasar por alto la representación incorrecta. Por lo tanto, se necesitan dos resultados separados.

13.1. Cobertura de acceso

Indica cuánto de la población de referencia global especificada puede usar apropiadamente el producto de IA relevante. Representación candidata:

```
AC_a= N(U_{a,t}) / N(G_t)
```

Aquí:

- Ua,t: población de usuarios elegibles del producto
- Gt: población de referencia global definida

13.2. Precisión de Representación Condicional

Es la probabilidad de ver una representación precisa entre los usuarios que pueden acceder al producto. Notación candidata:

```
RA_a= Pr(Y=1| u ∈ U_{a,t})
```

13.3. Acceso Efectivo a Representación Precisa

CONCEPTO DE CANDIDATO

Se puede desarrollar una métrica operativa que combine la cobertura de acceso con la precisión condicional:

```
ECR_a= AC_a× RA_a
```

Esta métrica se aproxima aproximadamente a la pregunta: “¿Qué proporción de la población de referencia global definida puede acceder a este producto y cumplir con la condición de ver una representación precisa?” Esto aún no es la fórmula final NOMOS. Porque:

- el acceso es diferente del uso real,
- las estructuras de peso y dependencia son complejas,
- el uso del producto es voluntario,

La representación correcta depende del prompt. Sin embargo, mantiene la distinción básica:

No poder acceder no es lo mismo que obtener una respuesta incorrecta.

14. Alcance natural Y Base común de comparación

Al comparar productos de IA, se necesitan dos universos poblacionales separados.

14.1. Población Alcance natural

Cada producto de IA tiene su población elegible dentro de su dominio de acceso oficial y real. Este resultado responde a la siguiente pregunta:

¿Hasta qué punto el producto representa con precisión a la población a la que realmente puede llegar?

Las poblaciones objetivo pueden diferir entre productos.

14.2. Población Base común de comparación

En una forma común en todos los productos de IA comparados:

- accesible oficialmente,
- puede usarse bajo la misma condición de idioma o interfaz

es la población. Este resultado responde a la siguiente pregunta:

¿Qué producto produce una representación más precisa en el mismo universo de usuarios?

14.3. Dos resultados deben publicarse juntos

Alcance natural:

- muestra el acceso real al producto,
- distribución geográfica natural

Base común de comparación: proporciona una comparación de producto más justa. Un producto puede tener una puntuación alta en Alcance natural, pero menor en Base común de comparación. Dos resultados responden a preguntas diferentes.

15. POBLACIÓN RELACIONADA CON EL MERCADO

La puntuación de identidad global de una entidad puede diferir de su puntuación de mercado objetivo comercial. Ejemplo: Si un bufete de abogados local ofrece servicios solo en Turquía: la pregunta de identidad global se puede medir en la población de usuarios elegibles en todo el mundo, pero las recomendaciones y la idoneidad del servicio solo deben medirse para usuarios en Turquía o dentro del alcance real de los servicios. La institución puede seleccionar el resultado a su favor:

- solo el mercado de alta puntuación,
- solo su propio idioma,
- solo sus clientes existentes

la población objetivo no puede ser anunciada. Mercado objetivo:

- Antes de que comience la auditoría,
- a los registros de servicio,
- dentro del alcance del contrato y operaciones

Debe bloquearse de acuerdo con.

16. SEPARACIÓN DE PERSONA, CUENTA Y SESIÓN

El peso de la población pertenece a las personas. No a las cuentas de IA.

16.1. Una Persona, Múltiples Cuentas

La misma persona:

- gratis,
- pagada,
- corporativa

puede tener múltiples cuentas. Estas cuentas pueden producir observaciones separadas en comparaciones de interfaz del producto. No se pueden contar como individuos separados en la población humana global.

16.2. Una Cuenta, Múltiples Personas

Una cuenta familiar, de equipo o corporativa puede ser utilizada por múltiples personas. No equivale al número de usuarios de la cuenta.

16.3. Una Persona, Múltiples Sesiones

Una persona puede generar observaciones separadas en diferentes sesiones. Estas observaciones pueden medir el tiempo o la estabilidad de la sesión. No aumentan el tamaño de la población independiente.

16.4. Una Persona, Múltiples Productos de IA

Es valioso para la comparación de productos emparejados. Sin embargo, las observaciones son dependientes debido a características compartidas a nivel de usuario.

16.5. Unidad de Peso Principal

En la puntuación GEO basada en la población, la unidad de peso principal debería ser:

humano

Los cálculos, la sesión y las observaciones del producto deben estar asociados bajo la unidad humana.

17. POBLACIÓN DE IDIOMA

Usar solo el idioma oficial del país no es suficiente para la equidad lingüística.

17.1. Idioma de Solicitud

El idioma en el que el participante hace la pregunta de auditoría.

17.2. Idioma de la Interfaz

El idioma utilizado en el menú y la interfaz de usuario del producto de IA. Puede que no sea el mismo que el idioma de la solicitud.

17.3. Idioma de Respuesta

El idioma en el que la IA realmente responde. El sistema puede responder en un idioma diferente al idioma del prompt. Esto también es un resultado de medición.

17.4. Idioma de competencia del usuario

El idioma en el que el participante puede usar de manera significativa el prompt y la respuesta. No tiene que ser el idioma nativo.

17.5. Configuración regional

El contexto regional junto con el idioma. El mismo idioma puede ser diferente:

- país,
- precio,
- ley,
- terminología

puede llevar la condición.

18. DOBLE CONTEO DE USUARIOS MULTILINGÜES

Una persona:

- turco,
- puede ser capaz de usar inglés.
- inglés

Esta persona puede realizar tres pruebas de idioma separadas. Sin embargo, contarlas como tres personas en el peso de la población global distorsiona la estimación de la población. Se pueden usar tres métodos.

18.1. Tarea sobre la lengua materna de madre soltera

La persona está asignada a una sola celda de idioma en el análisis principal. Otros idiomas se utilizan como pruebas adicionales o emparejadas.

18.2. Dividir Peso de Persona

El peso poblacional de la persona puede dividirse entre los idiomas apropiados de una manera predeterminada. El peso total no excede el de una persona.

18.3. Panel Multilingüe Emparejado Separado

Los usuarios multilingües se utilizan para comparaciones entre idiomas de la misma persona. Este panel se analiza por separado del panel principal de la población. Independientemente del método utilizado:

> *El peso poblacional total de la misma persona no puede replicarse sin explicación.*

19. ¿CÓMO DEBE VERIFICARSE LA COMPETENCIA LINGÜÍSTICA?

La declaración de un participante: "Conozco este idioma." puede ser un registro inicial. Sin embargo, en auditorías de alta seguridad, se puede utilizar uno de los siguientes métodos:

- Declaración de idioma nativo o uso diario
- Prueba breve de lectura y comprensión
- Registro de idioma del proveedor del panel

- Verificación del adjudicador local
- Registro de uso educativo o profesional
- Éxito en la misión piloto

Los paneles de prueba de idiomas extremadamente difíciles pueden ser accesibles para usuarios capacitados y de altos ingresos. El propósito no es realizar un examen académico de idiomas. Es verificar la competencia para entender la prompt y la respuesta como un usuario real.

20. ¿QUÉ PASA SI UN IDIOMA NO ESTÁ SOPORTADO?

Un producto de IA puede:

- no soportar oficialmente un idioma en particular,
- darle soporte parcial,
- no afirmar soportarlo pero poder generar respuestas

Puede ser. Los estados del idioma deben registrarse por separado:

SOPORTADO_OFICIALMENTE

SOPORTE_LIMITADO

OBSERVADO_PERO_NO_COMPATIBLE

NO_COMPATIBLE

DESCONOCIDO

Respuesta recibida en un idioma que no es oficialmente compatible:

- descubrimiento,
- acceso experimental,
- resiliencia del idioma

puede ser medida. No debe añadirse automáticamente a la puntuación principal de la población de usuarios compatibles. También debe informarse sobre el límite de acceso al idioma.

21. PONDERACIÓN DE LA POBLACIÓN DEL PAÍS Y COBERTURA DEL Observador por País

El panel principal ponderado por población y el panel de visibilidad del país tienen propósitos separados.

21.1. Panel de Población

Los participantes se distribuyen según su proporción en la población objetivo adecuada para el producto. Las poblaciones más grandes elegibles reciben mayor ponderación. Produce la puntuación principal de la población global.

21.2. Panel de Observadores del País

El Panel de Observadores del País busca al menos una observación de los países más pequeños que reciben poca o ninguna representación en la muestra principal de la población. Una sola observación del Observador por País:

- indica que hay acceso disponible,
- que un resultado singular específico ha ocurrido,
- muestra la necesidad de nueva investigación

puede indicar. No produce una puntuación del país por sí sola.

21.3. Suficiencia para La puntuación del País

Para que se publiquen tasas basadas en el país:

- muestra mínima,
- tasa de observación válida,
- incertidumbre,
- representatividad del panel

deben cumplirse condiciones. Con uno o varios participantes: 'La puntuación NOMOS del país' no puede ser anunciado.

22. Panel de Equidad Lingüística

El peso de la población puede hacer que los idiomas con baja población o pocos recursos sean invisibles en la puntuación principal. Por lo tanto, se puede establecer un Panel de Equidad Lingüística. Propósito:

- para probar idiomas subrepresentados,
- para detectar una degradación severa basada en el idioma,
- para identificar problemas de traducción de prompts,
- para hacer visible el comportamiento de los idiomas con pocos recursos

debe ser. El Panel de Equidad Lingüística puede proporcionar muestras adicionales al panel principal de población. Sin embargo, las observaciones de idiomas sobre-muestreadas deben ajustarse a los pesos reales de la población en la puntuación global. De lo contrario, el panel de equidad altera la estimación de la población global.

23. PANEL DE ACCESIBILIDAD

Los usuarios con necesidades de accesibilidad visual, auditiva, motora, cognitiva u otras son una parte real de la población objetivo. Si la tarea de Captura o investigación NOMOS no es adecuada para la participación de estos individuos: los usuarios no pueden ser eliminados de la población objetivo. Resultado correcto:

> *El marco de muestreo está incompleto en términos de accesibilidad.*

Un panel de accesibilidad separado puede evaluar lo siguiente:

- Compatibilidad con lectores de pantalla
- Uso del teclado
- Entrada por voz

- Configuraciones para baja visión
- Carga cognitiva
- Accesibilidad móvil
- Usabilidad de la herramienta de Captura

El panel de accesibilidad debe ver a los usuarios como parte de la cobertura de la población, no como una "excepción especial."

24. POBLACIÓN OBJETIVO Y MARCO DE MUESTREO

Estos dos conceptos no deben confundirse entre sí.

24.1. Población objetivo

Todo el grupo de personas sobre el que se desea hacer una conclusión.

24.2. Marco muestral

La población de la cual se pueden seleccionar realmente los participantes:

- panel,
- lista,
- base de datos,
- organización conjunta,
- grupo de usuarios

es el clúster. El marco muestral puede no cubrir completamente a la población objetivo.

24.3. Población contactable

Estas son las personas a quienes se puede entregar realmente la invitación a la investigación.

24.4. Población que acepta participar

Aquellos que participan voluntariamente en la investigación.

24.5. Población que completa la tarea

Aquellos que completan la tarea de auditoría.

24.6. Población Analítica Válida

Son los que producen observaciones adecuadas para el protocolo y utilizables en el análisis. Los números de estas capas deben informarse por separado.

25. ERROR DE COBERTURA

El marco de muestreo puede excluir algunas partes de la población objetivo. Esto se llama error de cobertura. Ejemplos:

- Solo paneles profesionales en línea
- Solo usuarios de escritorio
- Solo personas con tarjetas de crédito
- Solo aquellos que usan una interfaz en inglés

- Solo residentes de grandes ciudades
- Solo usuarios muy interesados en la tecnología
- Exclusión de personas con necesidades de accesibilidad

Error de cobertura:

- debe ser explicado,
- de ser medido si es posible,

debe reducirse mediante ponderación o un panel adicional. La ponderación no puede crear mágicamente un grupo que no existe en el marco de muestreo.

26. COBERTURA EXCESIVA

Personas que no pertenecen a la población objetivo pueden encontrarse en el marco de muestreo. Ejemplos:

- Participante en un país donde el producto no es accesible
- Persona que no cumple con la condición de edad requerida
- Plan incorrecto o interfaz de usuario
- Participante que no puede usar el idioma de la auditoría
- Persona encontrada nuevamente con múltiples registros
- Usuario que trabaja en nombre de otro país

Los registros de cobertura excesiva no deben incluirse en la puntuación principal. Sin embargo, se debe informar la razón de exclusión y la tasa.

27. NO RESPUESTA Y AUTOSELECCIÓN

Las personas que aceptan participar en la investigación pueden diferir sistemáticamente de la población objetivo. Los voluntarios:

- Pueden usar productos de IA con más frecuencia,
- Pueden estar más interesados en la tecnología,
- Pueden tener un nivel más alto de educación,
- Pueden estar ya familiarizados con la marca,

Pueden comportarse de manera diferente debido a recompensas o compensación. Por lo tanto: la expresión "1,000 personas aleatorias en el mundo" solo debe usarse en diseños de selección verdaderamente probabilísticos y documentados. Si se utiliza un voluntario o un panel en línea, la expresión correcta es:

Panel de usuarios voluntarios estratificado y ponderado

puede ser posible. El tipo de panel no puede almacenarse.

28. LA POBLACIÓN ADECUADA CAMBIA CON EL TIEMPO

El universo de usuarios de un producto de IA no es fijo. Lo siguiente puede cambiar:

- Países admitidos
- Edad mínima

- Acceso gratuito
- Estructura del plan de pago
- Disponibilidad en móvil o web
- Soporte de idiomas
- Términos de la cuenta
- Cobertura del producto empresarial
- Estado legal y regulatorio

Por esta razón, cada registro de población:

- fecha de observación,
- período de validez,
- versión de la fuente,
- fecha de última verificación

debe llevar.

28.1. Cambio durante la ola de medición

Si el acceso al producto cambia de manera significativa durante la ola:

- la ola puede ser detenida,
- dos períodos separados pueden ser creados antes y después del cambio,

los países afectados pueden ser analizados por separado. Los usuarios antes y después del cambio no pueden ser combinados como una sola población.

29. ARCHIVO DE FUENTE DE POBLACIÓN

Cada auditoría debe almacenar los datos de la población objetivo que utiliza en un archivo versionado. Según corresponda, se deben incluir los siguientes campos:

- País
- Población de referencia
- Historial de la población
- Elegibilidad por edad
- Acceso a Internet y dispositivos
- Estado de acceso al producto
- Acceso a la interfaz
- Requisito de plan y cuenta
- Grupos de idioma
- Elegibilidad para investigación
- Método común de distribución
- Incertidumbre
- Fuentes
- Última fecha de verificación
- Persona responsable
- Versión de la población

Las cuotas por país no se pueden reproducir después de los resultados sin este archivo.

30. PRODUCTO SINTÉTICO Y DISPLAY DE PAÍS

DEMOSTRACIÓN DE METODOLOGÍA SINTÉTICA / Los siguientes países, producto de IA, números de población y tasas de acceso son completamente ficticios. No representan a ningún país o producto real. Consideremos un producto de consumo general sintético llamado Orion AI. Hay cuatro países sintéticos.

País	Población total	Población estimada adecuada tras considerar edad, internet, idioma e interfaz	Acceso al producto	Población dentro del alcance natural
País A	1,400,000,000	780,000,000	No disponible	0
País B	1,450,000,000	520,000,000	Acceso oficial	520,000,000
País C	340,000,000	250,000,000	Acceso oficial	250,000,000
País D	35,000	25,000	Acceso oficial	25,000
Total	3,190,035,000	1,550,025,000		770,025,000

30.1. Si se usara la población total

El País A tendría una gran participación en la muestra principal debido a su población muy grande. Sin embargo, el producto no puede ser usado en ese país. Los participantes no podrían generar una experiencia de usuario natural.

30.2. Cuando se usa la población Alcance natural

Las participaciones poblacionales teóricas del panel principal de productos son aproximadamente:

- País B: 67.53%
- País C: 32.47%
- País D: 0.003%
- País A: fuera de Alcance natural

La participación teórica del País D en una muestra principal de 1,000 personas es muy inferior a una persona. Por lo tanto, la puntuación principal ponderada por población no debería dar un peso alto al País D. Sin embargo, si no se pretende que el país sea completamente invisible:

Panel de Observadores del País

Se puede obtener al menos una observación exploratoria. Esta observación:

- no es la puntuación del País D,
- no cambia el peso poblacional principal,

muestra que el comportamiento del producto puede observarse en el país.

30.3. ¿Cómo se informa el País A?

Para el País A:

Se utiliza el estado OUTSIDE_NATIVE_REACH

La exclusión del País A no debería ocultar los límites de acceso del producto. La tarjeta de resultados pública también muestra:

- Cobertura de acceso al producto
- Participación no accesible en la población de referencia global

- Puntuación Alcance natural
- Puntuación comparativa Base común de comparación

31. REPRESENTACIÓN DE LA POBLACIÓN DE PROMPT SINTÉTICO

EJEMPLO SINTÉTICO

Consideremos tres prompts para el mismo producto Orion AI.

Prompt A

"¿Qué tipo de empresa es Asteron?" Población objetivo: podría ser la amplia población de usuarios adecuada de Orion AI.

Prompt B

"¿Asteron proporciona servicios de viajes corporativos en Turquía?" Población objetivo:

- Usuarios elegibles en Turquía,
- Usuarios externos identificados investigando el servicio de Turquía

puede ser.

Solicitud C

“¿Es Asteron adecuado para la consultoría de inmigración legal para mí?” Si Asteron no proporciona tal servicio: puede que no haya población elegible para asesoramiento, la respuesta correcta debería excluir a todos los usuarios adecuados. Combinar esta solicitud en La puntuación de identidad general podría ser incorrecto. Este ejemplo muestra:

> *Incluso para la misma entidad y el mismo producto de IA, la población objetivo puede variar dependiendo de la pregunta del usuario medida.*

32. ARQUITECTURA DE CUATRO PANELES

ARQUITECTURA CANDIDATA — ARQUITECTURA CANDIDATA

Se recomiendan cuatro paneles complementarios para la implementación global y justa de GEO-1000.

32.1. Panel de Población

El panel principal ponderado según la población de usuarios elegibles para el producto. Produce la estimación principal de la población.

32.2. Panel de Observadores del País

Evita que los países pequeños que no pueden proporcionar suficientes muestras en el panel de población queden completamente invisibles. No es suficiente por sí solo para producir una puntuación de país.

32.3. Panel de Equidad Lingüística

También mide idiomas que tienen pocos recursos o que están suprimidos en los promedios de población. Se reconecta con la puntuación global principal usando pesos reales de la población.

32.4. Panel de Accesibilidad

Mide necesidades de accesibilidad y usuarios de tecnología asistida que el panel digital estándar podría excluir. Observaciones de estos cuatro paneles:

- objetivo,
- peso,
- universo de inferencia

Debe etiquetarse por separado en términos de mantenimiento. La misma observación no puede usarse con peso completo en más de un panel sin explicación.

33. PANEL INICIAL DE ADULTOS

REGLA BÁSICA DE CANDIDATO — REGLA FUNDAMENTAL DEL CANDIDATO

En la primera versión general de GEO-1000, los paneles de usuario reales:

de participantes adultos que cumplen con las condiciones del producto y del país

Se recomienda que sea creado. Este enfoque:

- consentimiento,
- protección de datos,
- recopilación de capturas de pantalla,
- investigación de alto riesgo,
- consentimiento parental

Puede reducir su complejidad. Sin embargo, no se puede llegar a la siguiente conclusión: 'La experiencia de los niños y jóvenes usuarios ha sido medida con el panel de adultos.' En productos que afectan a niños o jóvenes, por separado:

- aprobación ética,
- protocolo adecuado para la edad,
- procesos de los padres o tutores,
- protección contra daños

son requeridos.

34. NINGÚN MARCO DE POBLACIÓN PUEDE LLAMARSE "TODAHUMANIDAD"

Un producto:

- si no existe en algunos países,
- si no soporta algunos idiomas,
- si solo funciona en ciertos dispositivos,
- si tiene cierta condición de edad o plan,

resultado: no se puede presentar como “la experiencia GEO de todas las personas del mundo”. La expresión correcta:

Representación estimada en la población objetivo adecuada para el producto especificado

debería usarse. La palabra global:

- multipaís
- multilingüe
- gran población

puede definir una medida. No significa cobertura ilimitada y universal.

35. LÍMITE ÉTICO FUNDAMENTAL DE LA POBLACIÓN

Definir una población para la medición no significa considerar a algunas personas como inútiles. Un grupo puede recibir un bajo peso en la puntuación principal. Sin embargo:

- alto riesgo,
- justicia lingüística,
- accesibilidad
- visibilidad del país

también pueden ser examinados por separado. Un grupo puede ser excluido de la puntuación Alcance natural porque no pueden acceder a un producto. Esta falta de acceso también puede ser reportada como una brecha de cobertura global. El propósito de NOMOS:

- comprender a todos en un solo promedio artificial,
- ver solo a las poblaciones más grandes,
- parecer representar simbólicamente a las comunidades pequeñas con una sola persona

no es el objetivo. El objetivo es:

> *Representar a cada persona en el denominador correcto de la pregunta correcta, sin mezclar su peso y visibilidad.*

36. DISPOSICIONES NORMATIVAS OBLIGATORIAS

CH05-N01

La población objetivo del usuario debe definirse por separado para cada producto de IA.

CH05-N02

La población total de un país no puede ser utilizada directamente como la población de usuarios elegibles para IA.

CH05-N03

La población objetivo debe estar vinculada a producto, país, edad, acceso, interfaz, plan, idioma, solicitud y condiciones de tiempo.

CH05-N04

La población objetivo y las reglas de exclusión deben versionarse antes de ver los resultados.

CH05-N05

La población objetivo de un producto de IA no puede ser transferida a otro producto de IA sin explicación.

CH05-N06

La falta de acceso al producto no puede ser calificada como un resultado de tergiversación; debe ser reportada por separado como cobertura de acceso.

CH05-N07

No se puede pedir a los participantes que eviten las restricciones de acceso a productos o países.

CH05-N08

El acceso no oficialmente soportado no puede añadirse automáticamente a la población principal Alcance natural.

CH05-N09

Las unidades individuales, de cuenta y de sesión deben estar separadas entre sí.

CH05-N10

Múltiples cuentas o idiomas de una persona no pueden multiplicar al individuo en el peso poblacional.

CH05-N11

La elegibilidad lingüística no puede asignarse únicamente en función del idioma oficial del país.

CH05-N12

El idioma solicitado, el idioma de la interfaz, el idioma de respuesta y el idioma de competencia del usuario deben registrarse por separado.

CH05-N13

El peso total de la población de una persona que utiliza varios idiomas no puede exceder el de una persona sin explicación.

CH05-N14

El panel ponderado por población y el panel de Observadores del País deben mantenerse separados.

CH05-N15

Una o varias observaciones del Observador por País no pueden publicarse como una puntuación del país.

CH05-N16

En el Panel de Equidad Lingüística, las observaciones sobremuestreadas deben revertirse a su peso poblacional verdadero en la puntuación global.

CH05-N17

No se pueden excluir de la población objetivo a los usuarios con necesidades de accesibilidad porque la herramienta de recolección de datos no sea adecuada.

CH05-N18

La población objetivo, el marco de muestreo, la población contactada, los participantes y las observaciones analíticas válidas deben informarse por separado.

CH05-N19

Un panel de no probabilidad o voluntario no debe representarse como una muestra aleatoria de la población global.

CH05-N20

La ponderación no puede hacer que un grupo que está completamente ausente en el marco de muestreo sea observable.

CH05-N21

La población potencial adecuada y la exposición de usuarios activos no pueden usarse como la misma métrica.

CH05-N22

Los resultados Alcance natural y Base común de comparación deben separarse entre sí.

CH05-N23

La misma población objetivo no puede usarse automáticamente para solicitudes de identidad, servicio, precio, legales y recomendaciones.

CH05-N24

Los datos de población, el estado de acceso al producto y la distribución lingüística deben llevar una marca de tiempo y la versión de la fuente.

CH05-N25

Las relaciones de utilidad marginal no pueden multiplicarse a ciegas sin una explicación del supuesto de independencia.

CH05-N26

Si el tipo de producto no es adecuado para el panel general de consumidores, el método de población humana GEO-1000 no puede aplicarse sin explicación.

CH05-N27

No se pueden incluir niños o usuarios menores en el panel principal sin un protocolo ético y legal separado.

CH05-N28

La cobertura de acceso al producto y la exactitud de la representatividad deben publicarse como resultados separados.

CH05-N29

El término global no puede hacer que países, idiomas, edades, interfaces o planes que están fuera del alcance sean invisibles.

CH05-N30

El registro de la población objetivo debe tener un propietario humano o institucional responsable.

37. FORMAS DE FALLA

CH05-F01— CONTAR LA POBLACIÓN TOTAL COMO POBLACIÓN ELEGIBLE

La población del país se incluye directamente en el denominador sin evaluar condiciones de producto, edad, internet, idioma e interfaz.

CH05-F02— APLICAR EL MISMO MAPA MUNDIAL A TODOS LOS PRODUCTOS

Los productos de IA con diferentes áreas de acceso se miden contra la misma población objetivo.

CH05-F03— CONTAR A UN USUARIO QUE EXCEDE EL LÍMITE COMO USUARIO NATURAL

Se presentan como experiencia de la población del país las observaciones que exceden las condiciones de acceso oficial.

CH05-F04— CONTAR LA AUSENCIA DE PRODUCTO COMO RESPUESTA INCORRECTA

Los países que son inaccesibles se consideran cero en precisión de representación.

CH05-F05— OCULTANDO COMPLETAMENTE LA AUSENCIA DE PRODUCTO

Las grandes poblaciones que son inaccesibles se eliminan del denominador Alcance natural; la cobertura de acceso tampoco se muestra.

CH05-F06— CONTANDO A UNA PERSONA COMO MÁS DE UNA

La cuenta de la misma persona, su idioma u observaciones de productos multiplican el peso de la población.

CH05-F07— CONTANDO CUENTA COMO PERSONA

No se considera la propiedad de múltiples cuentas o el uso compartido de cuentas.

CH05-F08— CONTANDO EL IDIOMA OFICIAL COMO IDIOMA DEL USUARIO

El idioma oficial del país se considera el idioma de auditoría para todos los participantes.

CH05-F09— CONTANDO DOS VECES A LOS INDIVIDUOS MULTILINGÜES

El mismo individuo recibe el peso completo de la población en cada idioma.

CH05-F10— ASIGNANDO UNA PERSONA A UN PAÍS PEQUEÑO Y EXPLICANDO LA PUNTUACIÓN DEL PAÍS

La observación del Observador por País se transforma en una estimación estadística del país.

CH05-F11— ELIMINANDO COMPLETAMENTE LOS PAÍSES PEQUEÑOS

Debido a que la proporción de población es baja, el país o el idioma no se incluyen en ningún panel adicional de cobertura.

CH05-F12— SOBREPONIENDO EJEMPLOS DE JUSTICIA DE IDIOMA EN LA PUNTUACIÓN PRINCIPAL

Los idiomas sobre-muestreados no se devuelven a su peso poblacional real.

CH05-F13— CONTANDO AL USUARIO ACTIVO COMO POBLACIÓN POTENCIAL

Los datos de uso reales se presentan como si incluyeran a todas las personas que podrían acceder al producto.

CH05-F14— CONTANDO LA POBLACIÓN POTENCIAL COMO EXPOSICIÓN REAL

Se supone que la respuesta de todos los que podrían usar el producto se observa realmente.

CH05-F15— CONTANDO EL PANEL VOLUNTARIO COMO UNA MUESTRA ALEATORIA DEL MUNDO

La auto-selección y el sesgo del panel están ocultos.

CH05-F16— IGNORANDO EL ERROR DE COBERTURA CON PONDERACIÓN

Se afirma que los grupos no presentes en el marco están representados únicamente mediante ponderación.

CH05-F17— ELIMINANDO USUARIOS EXCLUIDOS POR LA HERRAMIENTA DE RECOLECCIÓN DE DATOS DE LA POBLACIÓN

Los problemas de accesibilidad o del dispositivo se interpretan como si el usuario no fuera adecuado.

CH05-F18— APLICANDO LA MISMA POBLACIÓN A TODAS LAS SOLICITUDES

La identidad global, los precios locales y las preguntas de asesoría legal se combinan en el mismo denominador.

CH05-F19— APLICANDO LA POBLACIÓN GENERAL DE CONSUMIDORES AL PRODUCTO API

El panel de usuarios humanos se transfiere directamente al desarrollador o producto de la aplicación.

CH05-F20— COMBINANDO SECRETAMENTE PLANES GRATUITOS Y DE PAGO

Se publica una distribución de producto única sin registrar la diferencia de planes.

CH05-F21— UNIFICANDO POBLACIONES WEB Y MÓVILES

Las diferencias en el acceso a la interfaz y el comportamiento permanecen invisibles.

CH05-F22— IGNORANDO EL LÍMITE DE EDAD

Se añaden grupos de edad inapropiados a la población objetivo o al panel sin explicación.

CH05-F23— MULTIPLICANDO TASAS MARGINALES COMO SI FUERAN INDEPENDIENTES

Se ignora la relación entre edad, internet, idioma y acceso.

CH05-F24— CAMBIANDO LA POBLACIÓN DESPUÉS DEL RESULTADO

Se elimina de la población objetivo un país o grupo de usuarios con puntuación baja.

CH05-F25— USANDO EL MAPA DE ALCANCE DEL PRODUCTO ANTIGUO

El alcance real del producto en la fecha de medición no está verificado.

CH05-F26— CONFUNDIENDO LOS RESULTADOS DE Base común de comparación Y Alcance natural

Las comparaciones realizadas sobre la población conjunta se presentan como si reflejaran el alcance global real del producto.

CH05-F27— RESTRINGIENDO EL MERCADO OBJETIVO SEGÚN EL INTERÉS COMERCIAL

La organización solo declara los países donde tiene éxito como mercados objetivo.

CH05-F28— CONTANDO EL ALCANCE DE ACCESO COMO PRECISIÓN GEOGRÁFICA

Estar presente en muchos países se presenta como prueba de una representación precisa.

CH05-F29— VENDIENDO ACCESO LIMITADO COMO ALTO ÉXITO GEOGRÁFICO

Una alta precisión en un pequeño número de países se presenta como superioridad global.

CH05-F30— USANDO EL TÉRMINO “EL MUNDO ENTERO” SIN REGISTRO DE COBERTURA

Las exclusiones de producto, idioma, país y plan están ocultas.

38. PROCEDIMIENTO DE AUDITORÍA

Paso 1— Fijar el Producto de IA y interfaz de usuario

Tipo de cuenta del plan de interfaz del producto Modelo mostrado Fecha de medición registrada.

Paso 2— Clasificar la Pregunta de Medición

prompt:

- identidad global,
- servicio local,
- precio,

- ley,
- recomendación,
- alto riesgo

¿A cuál de sus familias pertenece?

Paso 3— Determinar la Población de Referencia Global

Se define el universo humano inicial al cual se asociará el resultado.

Paso 4— Determinar Edad y Elegibilidad para la Investigación

Condición de edad del producto Condición de edad de investigación Se registran los términos de consentimiento y procesamiento de datos.

Paso 5— Verificar acceso al producto según el país

Se asigna un estado de acceso entre PA-0 y PA-5 para cada país.

Paso 6— Determinar compatibilidad de Internet, dispositivo e interfaz

Se estima la población que puede usar la interfaz medida.

Paso 7— Establecer el universo de idioma y localidad

Se separan el idioma del prompt, el idioma de la interfaz, el idioma de respuesta y la competencia lingüística del usuario.

Paso 8— Aplicar la relevancia del prompt

Se evalúa si todos los usuarios pertenecen a la población objetivo de la familia de prompts relevante.

Paso 9— Seleccionar el modelo de población apropiado para el tipo de producto

Se separa el producto en consumidor, empresa, educación, API o producto integrado.

Paso 10— Compara la población objetivo con el marco de muestreo

Se registra como cobertura insuficiente y sobrecobertura.

Paso 11— Identificar los paneles principales y complementarios

Los ámbitos del Panel de Población, Panel de Observadores por Países, Panel de Equidad Lingüística y Panel de Accesibilidad están versionados.

Paso 12— Separar los universos Alcance natural y Base común de comparación

Los resultados intraproducto y las poblaciones de comparación entre productos se registran por separado.

Paso 13— Separar el alcance de acceso y los resultados de precisión

La precisión entre la población a la que puede llegar un producto y la de usuarios alcanzados se informa por separado.

Paso 14— congelar el archivo de población

Las fuentes de datos, suposiciones, fecha y versión poblacional están fijadas.

Paso 15— Seguimiento de los cambios de materiales

Si el acceso al producto o las condiciones cambian durante la medición, se crea una nueva versión.

39. EVIDENCIA REQUERIDA

Identidad del producto de IA; interfaz de usuario; condiciones del plan y de la cuenta; fecha de medición; países en los que el producto es accesible; estado de acceso a nivel de país; condiciones de edad; condiciones legales y éticas de participación; registros de población por país; distribuciones de edad; registros de acceso a internet y dispositivos; distribuciones de idioma y localidad; soporte de idioma del producto; accesibilidad de la interfaz; familia de prompts objetivo; definición de población objetivo; marco de muestreo; subcobertura; sobrecobertura; tipo de panel; método de selección de participantes; estado de panel por probabilidad o de inscripción voluntaria; ponderaciones para usuarios multilingües; cobertura del Observador por País; cobertura de Equidad Lingüística; cobertura del Panel de Accesibilidad; población Alcance natural; Población Base común de comparación; datos de usuarios activos, dónde se utilizan; cálculo de cobertura de acceso; supuestos y análisis de sensibilidad; versión de la población; y fecha de la última verificación.

Persona o institución responsable

40. LISTA DE VERIFICACIÓN DE AUDITORÍA

¿Qué producto de IA y interfaz de usuario se está midiendo? ¿Para qué fecha se definió la población objetivo? ¿Se separaron la población total del país y la población de usuarios elegibles? ¿Se aplicó un requisito de edad? ¿Se evaluó el acceso a internet y a dispositivos? ¿Es el producto oficialmente accesible por país? ¿Se les pidió a los participantes que superaran restricciones de acceso? ¿Se separaron los planes gratuitos y de pago? ¿Se separaron las interfaces web y móvil? ¿La familia de prompts cambió la población objetivo? ¿Cómo se determinó la idoneidad del idioma y la configuración regional? ¿Se contaron dos veces los usuarios multilingües? ¿Está clara la diferencia entre la población objetivo y el marco muestral? ¿Es el panel probabilístico o voluntario? ¿Se midió el error de cobertura? ¿Se excluyeron silenciosamente los usuarios con necesidades de accesibilidad?

¿Se han convertido las observaciones del Observador por País en una puntuación por país? ¿Tuvieron los ejemplos de Equidad Lingüística un peso excesivo en la puntuación principal? ¿Se separaron los resultados de Alcance natural y Base común de comparación? ¿Son separadas la cobertura y la exactitud de la representación? ¿Se separaron los resultados de usuarios activos y usuarios potenciales? ¿Es el tipo de producto adecuado para el panel de población humana? ¿Se multiplicaron ciegamente las tasas marginales en los pesos de la población? ¿Se cambió la definición de la población después de ver los resultados? ¿Están versionados los datos utilizados de población y acceso? ¿Son visibles los grupos no probados? ¿Es consistente el término “global” con la cobertura real? ¿Existe un propietario claramente responsable del registro de la población?

41. OBJECIONES Y RESPUESTAS

Objeción 1— “¿No sería más justo distribuir directamente según la población mundial?”

La población mundial es una base de partida sólida. Sin embargo, incluir a personas que no pueden usar el producto en la muestra principal del producto no genera experiencia real del usuario. El método correcto es:

- peso poblacional dentro de la población elegible para el producto,
- cobertura de acceso global.

Se deben informar ambos resultados.

Objeción 2— “¿No deberían la mayoría de los usuarios provenir de países grandes?”

Los países que cumplen con las condiciones de producto, idioma y acceso deberían, naturalmente, tener una participación alta. Sin embargo, solo porque la población total sea grande, no se puede generar una experiencia de usuario falsa desde un país donde el producto no es accesible.

Objeción 3— “¿Por qué sería incorrecto dar una persona a un país pequeño como San Marino?”

Una persona es valiosa para la exploración y la observación del país mediante el Observador de País. Lo que está mal es presentar el resultado de esa persona como un 'puntuación del país'. La visibilidad del país pequeño puede protegerse con el Panel del Observador de País. El peso principal de la población sigue siendo según la participación real de la población.

Objeción 4— “Si damos al menos una persona a cada país, ¿no pareceríamos más globales?”

Se ofrece una cobertura geográfica más amplia. Sin embargo, la estimación de la población puede estar distorsionada. Por lo tanto:

- ámbito geográfico,
- Densidad de población

Son resultados separados. Puede ser un panel de alcance para asignar una persona a cada país. No es un panel de población.

Objeción 5— "Si damos cero a un producto en un país inaccesible, ¿no sería castigado injustamente ese producto?"

Un país inalcanzable no cuenta como cero en precisión de representación. Se hace visible en la métrica de cobertura de acceso. Esta distinción protege el producto tanto de sanciones injustas como de falsa universalidad.

Objeción 6— "Si existen datos reales de uso, ¿por qué necesitamos datos de población?"

Los datos de uso real muestran el impacto operativo. Sin embargo, el producto:

- a las poblaciones que no ha alcanzado,
- la desigualdad geográfica en la base de usuarios,
- equidad del usuario potencial

No se muestra por sí solo. Los dos tipos de datos responden a diferentes preguntas.

Objeción 7— 'Dado que todos los usuarios harán la misma solicitud, ¿por qué es importante la relevancia de la solicitud?'

Las preguntas de identidad pueden ser significativas en una población amplia. Las preguntas sobre precio, legales o de asesoramiento solo son significativas para ciertos usuarios y mercados. Una población irrelevante puede distorsionar la distribución de respuestas alejándola de la realidad del usuario.

Objeción 8— 'Si una persona conoce varios idiomas, ¿por qué no debería contarse completamente en todos los idiomas?'

Puedes probar todos los idiomas en una comparación entre idiomas. Sin embargo, multiplicar a la misma persona en el ponderamiento de población aumenta artificialmente la población mundial. El peso de la persona debe dividirse, o debe usarse un panel separado equivalente.

Objeción 9— '¿No deja incompleto el resultado global el hecho de probar solo a adultos?'

Lo hace. Esta limitación debe indicarse claramente. Comenzar con un panel de adultos puede ser una elección ética y operativa. No implica un juicio sobre los usuarios infantiles.

Objeción 10— '¿No es suficiente el panel de investigación en línea?'

Puede ser muy valioso para investigaciones piloto y ponderadas. Pero el panel:

- cuánto de la población objetivo cubre,
- quién queda excluido,
- cómo la participación voluntaria crea sesgo

Debe explicarse. Usar un panel no es un error. Presentar el panel como aleatorio, como la población mundial, sí lo es.

Objeción 11— "¿Por qué el panel de accesibilidad está relacionado con la puntuación GEO principal?"

Porque una parte de los usuarios reales de IA utiliza tecnología de asistencia o formas diferentes de interacción. Si la herramienta de recopilación de datos excluye a estas personas, la puntuación principal puede no representar a toda la población objetivo.

Objeción 12— "¿No hace que sea imposible tener una sola puntuación NOMOS el hecho de tener tantas definiciones de población diferentes?"

No. Se puede producir una única puntuación. Sin embargo, debe quedar claro a qué pregunta de población responde. Por ejemplo:

- Puntuación de Población Alcance natural
- Puntuación Base común de comparación
- Puntuación Relevante para el Mercado
- Resultado de Equidad Lingüística

puede tener registros separados. El significado de un número surge de la claridad de su denominador.

JUICIO COMÚN DEL CAPÍTULO 44

GEO-1000 contiene '1,000' en su nombre, pero el valor de esas 1,000 personas no surge únicamente del número. Surge de las respuestas a estas preguntas: ¿Quiénes son? ¿De qué población fueron seleccionados? ¿Pueden usar el producto de manera genuina? ¿Dónde viven? ¿Qué idiomas usan? ¿Cuántas veces se contó a la misma persona? ¿Son visibles los países pequeños? ¿Mantienen los países grandes su verdadero peso poblacional? ¿Dónde toma en cuenta la puntuación a las personas a las que el producto no puede llegar? ¿A quién excluyó la infraestructura del panel? ¿Qué resultado mide el acceso potencial y cuál mide la exposición real? Un país grande puede tener una población grande elegible para el producto y, por lo tanto, merecer un peso sustancial en el panel principal. Pero si el

producto no está disponible allí, la auditoría no debe inventar usuarios. La brecha de acceso debe divulgarse por separado.

Si un país es muy pequeño, el peso principal de la población debería permanecer bajo. Sin embargo, no tiene que ser completamente invisible. Puede observarse con un panel de monitoreo. Si un idioma tiene una pequeña población, puede que no determine el promedio global. Sin embargo, si hay una distorsión severa en la representación de ese idioma, también debería ser visible. Si una persona conoce tres idiomas, puede producir tres experiencias diferentes. Sin embargo, no son tres personas diferentes. Si un producto de IA solo proporciona cierto comportamiento en un plan de pago, la puntuación de los usuarios gratuitos no representa ese plan. Un panel de población humana es adecuado para un producto de chat para consumidores. El mismo modelo de población puede no ser adecuado para una API o producto empresarial. Por lo tanto, la quinta ley de medición de NOMOS es la siguiente:

> *Incluso la puntuación más sensible da la respuesta correcta a la pregunta equivocada si el denominador es incorrecto.*

Su sexta ley es:

> *El peso de la población no reemplaza la visibilidad del país y del idioma.*

Su séptima ley es la siguiente:

> *Una persona que no puede acceder a [algo] no ha sido mal representada; sin embargo, forma parte del límite de acceso global.*

Su octava ley es la siguiente:

> *Una persona que tiene múltiples cuentas, idiomas o sesiones no aumenta la población humana.*

La novena ley es la siguiente:

> *La medición global no es una afirmación de que todos sean medidos; es un registro honesto de quién es medido y quién no.*

Orden de la Sección 5 de NOMOS

> *No me des solo las poblaciones de los países. / Muestra qué personas pueden realmente usar el producto.*
>
> *No llenes un país grande con usuarios falsos cuando el producto no está allí.*
>
> *No des a una persona a un país pequeño y declares que la respuesta de esa persona es la voz del país.*
>
> *No dupliques a una persona tantas veces como tengan cuentas, idiomas y sesiones.*
>
> *No se debe confundir el idioma oficial con el uso real del idioma por parte de la gente.*
>
> *No haga que un usuario gratuito sea el representante de un plan de pago, ni que un usuario móvil sea el representante de la interfaz web.*
>
> *No llame al panel de voluntarios una población mundial aleatoria.*
>
> *No borre a un usuario con necesidades de accesibilidad de la población porque la herramienta no funcione.*
>
> *No elimine de la puntuación de precisión a la población a la que el producto no alcanza y oculte la brecha de acceso global.*

No aplique la misma proporción a una pregunta de precio que a una pregunta de identidad, ni a una marca global que a una licencia local.

Primero, defina el producto. / Luego defina la pregunta. / Luego defina el universo humano. / Luego limite el acceso, el idioma, la edad y la interfaz. / Luego explique a quién no puede seleccionar. / Y solo después de eso distribuya 1,000 personas.

La oración de cierre del capítulo

En GEO-1000, la justicia no consiste en dar el mismo número de personas a cada país; consiste en representar a cada persona en la población objetivo correcta, con el peso adecuado, sin ocultar los lugares donde no se mide.

Core normativo

Cada auditoría de población GEO DEBE definir, antes de la inspección de resultados, una población de usuarios objetivo con límite de tiempo y específica para el producto para cada uno de los siguientes: - Producto de IA y interfaz de usuario, - país o jurisdicción, - edad y condición de acceso legal, - condición de cuenta o plan, - idioma y localidad, - interfaz, - familia de prompts, - y propósito de la medición. La población residente total NO DEBE ser tratada como población elegible para el producto sin ajustes de elegibilidad justificados. La indisponibilidad del producto DEBE ser reportada como una limitación de cobertura de acceso, no puntuarse silenciosamente como representación incorrecta ni eliminarse silenciosamente del alcance global. El acceso potencial, el uso activo, la exposición real, Alcance natural, Base común de comparación, la relevancia del mercado y la observabilidad de la investigación DEBEN mantenerse como conceptos de población distintos. Las personas, cuentas, sesiones, idiomas y observaciones de productos NO DEBEN tratarse como unidades de población intercambiables. Los países, idiomas y grupos de accesibilidad sobrerepresentados DEBEN reajustarse para estimaciones de población mientras permanecen visiblemente separados para análisis de equidad y cobertura. Las observaciones del Country Observer PUEDEN establecer cobertura de observación geográfica, pero NO DEBEN representarse como puntuaciones de país estadísticamente fiables. NO se DEBE exigir a los participantes que eludan restricciones de producto, geográficas, de edad, de identidad o de cuenta para entrar en el panel de población oficial. Cada marco de población DEBE estar versionado, tener fuente, ser consciente de la incertidumbre y ser atribuible a un humano o una organización responsable.

CAPÍTULO 06

La muestra GEO-1000 proporcional a la población

BASES METODOLÓGICAS / K04; K05; K07; K08

Límite de Capítulo

La Sección 5 estableció la siguiente disposición:

> *La población objetivo de usuarios de cada producto de IA debe definirse por separado. La población total de un país no es la población de usuarios elegibles para el producto en ese país.*

Ahora llegamos a la siguiente pregunta:

> *De la población de usuarios elegibles definida para GEO-1000, ¿cuántas personas serán seleccionadas, de qué país, qué idioma y por qué método?*

La población de usuarios elegibles para un producto de IA es:

- 40% en el país A,
- 25% en el país B,
- 15% en el país C,
- 0.04% en un país pequeño

Si está disponible, ¿cómo deberían distribuirse las observaciones válidas 1,000? ¿Cómo se convertirán las cuotas fraccionarias en números enteros? ¿Se excluirán completamente los países con una participación poblacional de menos de una persona? ¿Se distorsionará el ponderado por población cuando se asigne al menos una persona a cada país? ¿Se pueden agrupar los países pequeños en un grupo separado? Si se sobremuestrean algunos idiomas por equidad lingüística, ¿cómo se calculará la puntuación global? Si el objetivo es 100 observaciones válidas en un país, ¿a cuántas personas se debería invitar? ¿Quién reemplazará al participante si no completa la tarea? ¿Se puede reemplazar a un usuario que da respuestas incorrectas? Si la misma persona prueba más de un producto de IA, ¿seguiría siendo independiente la muestra?

Incluso con una muestra sin procesar de 1,000 personas, ¿pueden los pesos altamente desiguales reducir la información estadística efectiva por debajo de la de 1,000 observaciones independientes con pesos iguales? Estos no son simplemente detalles operativos. El diseño de la muestra determina cuánto influye cada persona en la puntuación GEO. Una muestra defectuosa:

- puede pesar mal las respuestas correctas,
- puede sobre-representar simbólicamente a los países pequeños,
- puede hacer que las poblaciones grandes parezcan más pequeñas de lo que son,
- puede hacer invisibles a los idiomas con pocos recursos,
- puede presentar a los usuarios voluntarios de tecnología como si fueran la población mundial.

Se puede ajustar para lograr la puntuación deseada después de revisar los resultados. La arquitectura de auditoría previamente establecida requiere que cada auditoría sea medible y que el modelo/producto lleve la fecha, el país, el idioma, el conjunto de consultas, el número de repeticiones y el registro de medición. Esta sección aplica la misma disciplina al diseño de la muestra. Debido a que la evidencia genuina requiere solicitar límites, contexto y tiempo; también requiere hacer visible cómo se creó la muestra. Esta sección:

- cómo se distribuirán las observaciones válidas 1,000 del panel de población principal,
- cómo se establecerán los estratos y celdas,
- cómo se gestionarán los países pequeños,
- cómo se convertirán las fracciones en números enteros,
- cómo se ponderará el sobremuestreo,
- cómo separar invitación, respaldo y objetivos de observación válidos,
- cómo establecer los pesos de diseño y análisis,
- por qué el tamaño efectivo de la muestra puede diferir del número bruto

define. Este capítulo aún no:

- de qué proveedor del panel se seleccionarán los participantes,
- los detalles completos de la implementación de campo en muestreos voluntarios o probabilísticos,
- revisión y puntuación de respuestas,
- el método final de intervalo de confianza,
- todas las fórmulas de puntuación NOMOS

no están completamente finalizadas. La tarea de la Sección 6 es:

> *Transformar la población objetivo definida en una muestra GEO-1000 congelada, ponderable y reproducible antes de ver el resultado.*

NOMOS Desafío

Está configurando un panel global de 1,000 personas para un producto de IA. Las poblaciones elegibles para el producto de diez países son las siguientes:

País	Población de usuarios elegibles
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
Total	1,000,000,000

Cuotas brutas según la proporción pura de la población:

País	Cuota bruta
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

No se puede seleccionar a media persona. No se pueden enviar 0.4 personas a un país. Si redondea las cuotas brutas al número entero más cercano, el total puede exceder o ser inferior a 1,000. Si da al menos una persona a cada país, se sobre-representarán los países I y J en comparación con sus proporciones de población.

Dar una persona al país I aumenta su participación teórica 2.5 veces. Dar una persona al país J aumenta su participación teórica 10 veces. Esta representación extrema puede corregirse en gran medida. Pero no se puede producir una puntuación de país con una sola persona. ¿Qué sucede si no das a nadie ni a I ni a J?

Si defines estos países como estratos separados y obligatorios, la probabilidad de que algunas personas en la población objetivo sean seleccionadas se vuelve cero. En este caso, la afirmación de una "muestra probabilística que representa a toda la población global apropiada" se debilita. Ahora considera una decisión más peligrosa. Primero, recopilas las respuestas.

El único usuario en el país J obtiene una respuesta incorrecta. Luego dices: "El tamaño de muestra de este país es muy pequeño, excluyámoslo del alcance." En el país A, hay muchas respuestas positivas. Lo conservas con peso completo. Esto ya no es redondeo. Es un cambio de muestra basado en resultados.

En otro caso, se incrementa la tasa de registros inválidos en un grupo de idiomas con resultados bajos, utilizando únicamente participantes de respaldo hasta obtener una respuesta positiva. Esto tampoco es una muestra. Es un mecanismo de selección de respuestas. La primera disposición de esta sección es:

> *La muestra debe establecerse antes de ver las respuestas de la IA.*

Su segunda disposición establece:

> *La probabilidad de que un participante sea seleccionado no puede depender de la corrección de su respuesta.*

Su tercera disposición establece:

> *La visibilidad del país pequeño y el peso de la población en la puntuación global deben gestionarse por separado.*

Su cuarta disposición establece:

> *El sobremuestreo es posible; el sobrepeso no es obligatorio.*

Su quinta disposición establece:

> *Los registros en bruto 1,000 no siempre contienen información estadística equivalente a las observaciones 1,000.*

1. OBJETIVO DEL CAPÍTULO

El propósito de esta sección es transformar las observaciones válidas 1,000 dirigidas para cada producto de IA y ola de medición en un diseño de muestra verificable basado en la población de usuarios elegibles para el producto. La sección estandariza las siguientes distinciones:

- Población objetivo versus muestra seleccionada
- Cuota bruta versus asignación entera
- Panel de población principal versus paneles complementarios
- Visibilidad del país versus peso del país
- Estrato versus celda experimental
- Estratos de países grandes versus grupo de países pequeños
- Asignación proporcional frente a sobremuestreo desproporcionado
- Participación de la muestra versus participación de la población
- Número de invitaciones versus objetivo de observaciones válidas
- Participante de respaldo versus reemplazo posterior al resultado
- Invalidez de la recolección de datos versus resultado adverso de la IA
- Peso de diseño vs. peso de resultado
- Tamaño de muestra bruto vs. tamaño de muestra efectivo
- Muestra por país vs. muestra por idioma
- Usuario multilingüe vs. más de una persona
- Selección de una sola etapa vs. selección de múltiples etapas
- Asignación Alcance natural vs. asignación Base común de comparación
- Asignación de onda instantánea y asignación de series temporales
- Estimación de la población y análisis de reconocimiento por país o idioma
- Principales observaciones 1,000 con observadores auxiliares y paneles de equidad

Al final de esta sección, cada auditoría debería poder responder estas preguntas:

> *¿Por qué se distribuyeron las observaciones 1,000 en estos números a través de este país, idioma y grupos de usuarios?*
>
> *¿Cómo se determinó el peso de cada observación en el resultado global?*
>
> *¿Quién tenía una probabilidad cero de ser seleccionado y por qué?*
>
> *¿Cuántas personas fueron invitadas, cuántas completaron y cuántas observaciones permanecieron válidas?*
>
> *¿Se congelaron los resultados de la muestra antes de ser vistos?*

2. DISPOSICIÓN NORMATIVA CENTRAL

La muestra de la población principal GEO-1000 debe basarse en un diseño de muestreo estratificado que asigne probabilidades conocidas o predecibles de selección a las personas en el marco de población elegible para productos precongelados para cada producto de IA y ola de medición. El objetivo predeterminado del panel de la población principal es:

> *Al menos 1,000 observaciones válidas por producto de IA y ola de medición principal*

se mantiene. Este número:

- no es el número de personas invitadas,
- no es el número de cuentas,
- no es el número de capturas de pantalla sin procesar.

No es la suma de todos los paneles suplementarios. 1,000 es el objetivo para observaciones de productos válidas que cumplan con el protocolo en el panel de la población principal. Los paneles de Observador por País, Equidad Lingüística y Accesibilidad pueden crear muestras adicionales. Estas muestras adicionales:

- pueden aumentar el volumen total de campo por encima de 1,000,
- pueden ser utilizadas con pesos apropiados en el resultado de la población global,

pero no pueden ser utilizadas para reducir encubiertamente el objetivo de observaciones válidas 1,000 del panel principal.

3. REPRESENTACIÓN BÁSICA DE LA MUESTRA

Para un producto específico de IA a, una ola de medición w y un marco de población objetivo, los estratos:

```
h∈ H_{a,w}
```

están indicados. Para cada estrato:

- Nh: población estimada apropiada para el producto en el estrato
- N: población objetivo total
- Wh: peso poblacional del estrato
- n: observación principal del panel objetivo válida
- qh: cuota proporcional cruda
- nh: observación objetivo válida de entero asignado

es. Población objetivo total:

```
N= Σ_{h=1}^H N_h
```

Peso de la población del estrato:

```
W_h= N_h/ N
```

Cuota proporcional cruda:

```
q_h= nW_h
```

es. Asignación proporcional principal, en condiciones ideales:

```
n_h≈ nW_h
```

debe ser. Sin embargo:

- qh puede ser fraccionaria,
- algunas celdas pueden ser inferiores a una persona,
- puede ser necesaria una muestra diagnóstica mínima para los subgrupos de idioma y país,
- también se pueden examinar grupos pequeños de alto riesgo

las tasas de finalización de campo pueden variar. Por lo tanto, la cuota bruta no es directamente la asignación final.

4. TRES PROPÓSITOS DISTINTOS DEL DISEÑO DE MUESTRA

La misma muestra puede no cumplir todos los propósitos con la misma fuerza. Se deben separar tres propósitos.

4.1. Estimación de Población Global

Responde a la pregunta:

> *¿Qué resultado representativo vería la persona promedio en la población global de usuarios adecuada para el producto?*

El peso de la población es la prioridad principal.

4.2. Diagnóstico de Subgrupos

Responde a la pregunta:

> *¿Qué distorsiones de representación ocurren en países específicos, idiomas, planes o grupos de usuarios?*

Puede ser necesario un sobremuestreo de grupos pequeños.

4.3. Cobertura Geográfica y Lingüística

Responde a la pregunta:

> *¿En qué países e idiomas hay al menos observación real de usuarios?*

Los paneles de Observador por País y de Equidad Lingüística se vuelven importantes. Estos tres objetivos no pueden combinarse inexplicablemente en una única muestra.

5. ARQUITECTURA PRINCIPAL DE LA MUESTRA

La arquitectura de muestra principal candidata para GEO-1000 consta de cuatro capas.

5.1. Panel de Población

Cada producto y oleada de IA tiene un mínimo de 1,000 observaciones válidas como panel principal. Esto produce una estimación de población global.

5.2. Superposición del Observador por País

Proporciona observaciones geográficas adicionales para países pequeños que no pueden ser representados directamente en el panel de población principal. Estas observaciones pueden indicar:

- evento único,
- estado de acceso,
- necesidades de investigación futura

No generan una puntuación de país por sí solas.

5.3. Sobremuestra de Equidad Lingüística

Genera observaciones adicionales para idiomas con pocos recursos o idiomas subrepresentados en el panel principal. Estas se reponderan según los pesos poblacionales reales en el resultado global.

5.4. Sobremuestreo de accesibilidad

También mide a los usuarios de tecnología de asistencia y accesibilidad, que no están adecuadamente representados por el panel estándar. El universo objetivo y el peso de estas observaciones se registran por separado.

6. ¿QUÉ ES UN ESTRATO?

Un estrato es un segmento de la población que se gestiona por separado en términos de asignación de muestras y probabilidad de selección. El estrato puede ser una combinación de una o más de las siguientes áreas:

- País
- Región
- Idioma
- Localización
- Plan de IA
- Web o móvil
- Controlado o Panel de usuarios en condiciones reales
- Estado de accesibilidad
- Tipo de usuario
- Jurisdicción

Si cada pequeña combinación se convierte en un estrato separado, la muestra se fragmenta. Por ejemplo: la combinación País × idioma × plan × móvil/web × grupo etario × tipo de usuario puede producir cientos o miles de celdas. Las observaciones 1,000 no proporcionan estimaciones confiables para la mayoría de estas celdas. Por lo tanto, la estratificación debe ser:

- sustancial,
- medible,
- apropiada para el tamaño de la muestra

debe ser.

7. UN ESTRATO NO ES LO MISMO QUE UNA CELDA DE INFORME

Puede mostrarse por separado en una tarjeta de resultados grupales. Sin embargo, no tiene que ser un estrato separado en el diseño de la muestra. Por ejemplo:

- un idioma pequeño,
- un grupo de accesibilidad específico,
- una cierta franja etaria

Se puede seleccionar dentro de un estrato más amplio en la muestra principal. Más adelante, puede informarse por separado en el análisis. También es posible lo contrario: dos regiones pueden estratificarse por separado para mejorar la eficiencia del muestreo. Sin embargo, en el informe público, pueden combinarse bajo un único resultado regional. El diseño de estratificación y la estructura de informes deben llevar registros separados.

8. SECUENCIA DE ESTABLECIMIENTO DEL ESTRATO

La secuencia candidata para el Panel de Población Principal es la siguiente:

- Producto de IA y interfaz de usuario
- Marco poblacional Alcance natural o Base común de comparación
- País o conjunto de países pequeños
- Idioma principal de auditoría o grupo de idiomas
- Separación de plan/material o interfaz, si es necesario
- Marco de selección de participantes

El plan y la interfaz no tienen que convertirse en capas separadas para cada país. Esto puede gestionarse mediante:

- subcapa,
- cuota,
- asignación aleatoria

métodos.

9. CAPAS DE PAÍSES GRANDES Y PEQUEÑOS

La cuota bruta de un país:

```
q_c= nN_c/ N
```

está en esta forma. Si la cuota bruta es lo suficientemente grande, el país puede mantenerse directamente como una capa separada. Esto puede llamarse:

capa de país auto-representante

Hacer que los países con cuotas brutas muy pequeñas se conviertan individualmente en capas obligatorias separadas puede reducir la eficiencia del muestreo. Estos países pueden:</n:

- regional,
- familia de idiomas,
- régimen de acceso a productos,
- proximidad geográfica

pueden incluirse en grupos de países pequeños.

10. GRUPO DE PAÍSES PEQUEÑOS

Un grupo de países pequeños:

```
Que se represente como B_r= {c_1, c_2, …, c_k}
```

La población total elegible de la muestra:

```
N_{B_r}= Σ_{c∈B_r} N_c
```

Cuota principal de la muestra:

```
q_{B_r}= nN_{B_r}/ N
```

se calcula como. Dentro de la muestra, la selección de países se puede hacer con una probabilidad proporcional al tamaño de la población elegible. La probabilidad de que el país c sea seleccionado dentro de la muestra:

```
Pr(c | B_r) = N_c/ N_{B_r}
```

Puede ser. Luego se selecciona un participante dentro del país elegido. Este enfoque:

- puede prevenir que los países pequeños tengan cero probabilidades de ser seleccionados en la población objetivo,
- no garantiza que cada país será seleccionado en una ola dada,

preserva la estimación de la población global. No seleccionar un país en esa ola:

NO SELECCIONADO EN LA MUESTRA DE POBLACIÓN

no significa: "El país no forma parte de la población objetivo."

11. LIMITACIONES DE AGRUPAR PAÍSES PEQUEÑOS

Los países pequeños no pueden ser colocados al azar en el mismo grupo solo porque sus poblaciones sean pequeñas. Los países dentro del grupo son similares en la medida en que son relevantes:

- acceso al producto,
- interfaz de usuario,
- lenguaje o estructura local,
- acceso legal,
- región geográfica,
- método de recolección de datos

debe ser considerado. Un país sin acceso al producto no puede colocarse en el mismo grupo principal que un país con acceso completo. No se pueden hacer invisibles las diferentes condiciones de uso.

12. EL GRUPO DE PAÍSES PEQUEÑOS NO PRODUCE puntuación POR PAÍS

Una o dos observaciones seleccionadas del grupo: pueden ser utilizadas en la contribución global de la población del grupo, pueden indicar un solo evento de país. Sin embargo, no produce una puntuación estadística separado para un país pequeño. Si se desea una puntuación basado en el país, adicionalmente:

- Observador por País,
- muestra específica de cada país,
- muestra acumulativa de varias olas

puede ser requerida.

13. ASIGNACIÓN PROPORCIONAL A LA POBLACIÓN PURA

Asignación proporcional pura:

se basa en la fórmula:

```
q_h= nW_h
```

Ventajas:

- es comprensible para la estimación principal de la población,
- da naturalmente peso a las poblaciones grandes,
- pode reducir las diferencias de peso,

pode acercarse a un diseño de auto-pesaje. Limitaciones:

- pode muestrear muy pocos países y lenguas pequeñas,
- podría ser insuficiente para producir puntuaciones de subgrupos,
- podría dejar invisibles a pequeños grupos de alto riesgo,

pode crear cuotas fraccionarias y cercanas a cero. Por lo tanto, la asignación proporcional pura puede ser un valor predeterminado sólido para el Panel Principal de Población. No reemplaza los paneles complementarios.

14. CONVERTIR FRACCIONES A NÚMEROS ENTEROS

Las cuotas brutas suelen ser fraccionarias. La asignación total debe ser exactamente n. El Método del Mayor Residuo puede usarse como el método predeterminado.

14.1. Primer Paso

Cada cuota bruta se redondea hacia abajo:

```
b_h= ⌊q_h⌋
```

14.2. Segundo Paso

El número de observaciones no asignadas:

```
R= n− Σ_h b_h
```

se calcula.

14.3. Tercer Paso

La parte fraccionaria de cada estrato:

```
f_h= q_h− b_h
```

donde f_h es la parte fraccionaria de q_h.

14.4. Cuarto Paso

Las R observaciones restantes se añaden una por una a las capas con los valores de fh más altos.

14.5. Situación de igualdad

Si las partes fraccionarias de las dos capas son iguales:

- tamaño de la población,
- orden fijo predeterminado,
- semilla aleatoria registrada

Se debe utilizar una regla independiente para romper empates, sin importar el resultado. No se puede elegir un país basado en los resultados.

15. ASIGNACIÓN ENTERA SINTÉTICA

VISUALIZACIÓN SINTÉTICA

Poblaciones apropiadas al inicio de la sección:

País	Población adecuada	Cuota bruta
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

Si los países están protegidos como capas separadas, redondear hacia abajo:

País	Cuota redondeada hacia abajo
A	400
B	250
C	150
D	100
E	60
F	30
G	8
H	1
I	0
J	0
Total	999

Queda una observación. Se añade a H porque la fracción más alta es 0.5 en el país H. Asignación final:

País	Cuota final
A	400
B	250
C	150
D	100
E	60
F	30
G	8
H	2

País	Cuota final
I	0
J	0
Total	1,000

Aunque este diseño define a I y J como estratos determinísticos separados, si no se les da la posibilidad de ser seleccionados, puede representar un problema en términos de la inferencia del universo objetivo completo. Una solución más precisa es un pequeño conjunto de países.

16. ASIGNACIÓN SINTÉTICA CON PEQUEÑO CONJUNTO DE PAÍSES

Países H, I y J:

```
B_small= {H, I, J}
```

deben combinarse en el conjunto. Población del conjunto:

```
1,500,000+400,000+100,000=2,000,000
```

Cuota del conjunto:

```
1,000× 2,000,000/ 1,000,000,000= 2
```

se convierte en. Asignación principal:

Estrato	Cuota
A	400
B	250
C	150
D	100
E	60
F	30
G	8
Pequeño conjunto de países	2
Total	1,000

Pesos de selección de los países dentro del conjunto:

País	Participación de la población dentro del conjunto
H	75%
I	20%
J	5%

Se seleccionan dos observaciones principales de la población según estas probabilidades. En la misma ola:

- ambas observaciones pueden provenir de H,
- uno de H y uno de I,
- rara vez de J

Esto puede ocurrir. El diseño no hace que J sea visible en cada ola, pero no reduce a cero la probabilidad de selección de J. Si J debe ser observado, se añade por separado al Panel de Observadores del País. Una observación del Observador por País no le da a J un peso de uno en 1,000 en la asignación global principal; J mantiene su verdadero peso poblacional en el panel principal.

17. ASIGNACIÓN MÍNIMA DE CELDA

La cuota puramente proporcional puede no ser suficiente para realizar un análisis diagnóstico para algunos subgrupos. Por ejemplo, la cuota bruta para una celda de idioma podría ser 8. No se puede producir una puntuación de idioma confiable con ocho observaciones. En este caso, hay tres opciones:

- No publicar La puntuación de idioma
- Combinar la celda con grupos similares
- Sobremuestrear la celda

La asignación mínima de celdas debe definirse antes de los resultados.

18. BANDAS DE INFORME DE CANDIDATOS

BANDAS DE INFORME DE CANDIDATOS — BANDAS DE INFORME DE CANDIDATOS

Los umbrales a continuación no son decisiones científicas finales. Deben calibrarse mediante pruebas piloto y revisión externa.

1–29 Observación Válida

Evento único y tasa de descubrimiento por país o idioma no publicada

SOLO OBSERVACIONAL

30–99 Observaciones Válidas

Estimación a nivel exploratorio Incertidumbre amplia No se da un juicio definitivo de conformidad

EXPLORATORIO

100–199 Observaciones Válidas

Primera estimación de subgrupo La incertidumbre debería ser visible

PROVISIONAL

200 y superior Observaciones Válidas

Estimación de subgrupo más fuerte candidato Efecto de diseño y pesos nuevamente evaluados

CANDIDATO DE ESTIMACIÓN ESTÁNDAR

Estas bandas:

- tasa de error,
- precisión esperada,
- variabilidad de peso,
- efecto de diseño

no puede considerarse automáticamente suficiente sin ser tomada en cuenta.

19. SOBREMUESTREO

Un país, idioma o grupo de usuarios puede tener más observaciones que su participación en la población. Esto se llama sobremuestreo. Sus propósitos:

- aumentar la sensibilidad del subgrupo,
- hacer visible un idioma con pocos recursos,
- examinar un grupo pequeño de alto riesgo,
- ampliar la cobertura de Observador de País,
- representar adecuadamente a los usuarios con accesibilidad

son posibles. El sobremuestreo no está mal. Lo que está mal:

> *Es usar la proporción de la muestra del grupo sobremuestreado como si fuera la verdadera proporción de la población.*

20. PARTICIPACIÓN DE LA POBLACIÓN Y PARTICIPACIÓN DE LA MUESTRA

Participación de la población para la capa h:

```
W_h= N_h/ N
```

Participación de la muestra:

```
S_h= n_h/ n
```

está en la forma. En asignación puramente proporcional:

```
S_h= W_h
```

bien. En muestreo excesivo:

```
S_h> W_h
```

Es posible. En una muestra pequeña:

```
S_h< W_h
```

Es posible. La estimación de la población global debería mantener los valores Wh en lugar de Sh.

21. CARGA BÁSICA DE DISEÑO

En un muestreo estratificado simple, el peso básico de cada observación dentro del estrato h se puede considerar como:

```
d_i= N_h/ n_h
```

El peso del estrato normalizado se puede escribir como:

```
a_h= W_h/ S_h
```

El peso normalizado a nivel de observación puede ser:

```
w_i= W_h/ n_h
```

Se asegura que la suma sea uno:

```
Σ_i w_i= 1
```

La estimación de transición binaria ponderada se puede calcular como:

```
θˆ= Σ_i w_iY_i
```

Aquí:

```
Yi=1: respuesta que cumple con el criterio de transición

Yi=0: respuesta que no lo cumple
```

Es posible. El sistema final NOMOS no tiene que usar únicamente resultados binarios. La misma lógica de ponderación también se puede aplicar a puntuaciones de respuesta multidimensionales.

22. EJEMPLO DE MUESTREO EXCESIVO

EJEMPLO SINTÉTICO

En la población adecuada del país B:

- Idioma X: 80 por ciento
- Idioma Y: 20 por ciento

Sea así. En el Panel de Población, hay 250 observaciones para el país B. Asignación pura:

- Idioma X: 200
- Idioma Y: 50

Que se tomen 50 observaciones adicionales de Equidad Lingüística para el Idioma Y. Observaciones recopiladas:

- Idioma X: 200
- Idioma Y: 100

La proporción del Idioma Y en la muestra sin procesar:

```
100/300 = 33.3%
```

Ha sido. La proporción real de la población es 20%. En la cuenta de la población global, el Idioma Y no puede tener un peso de 33.3%. Pesos normalizados dentro del estrato:

- Peso total para el Idioma X: 80%
- Peso total para el Idioma Y: 20%

se preserva. Cada una de las 100 observaciones en el Idioma Y recibe un peso individual menor que una sola observación en el Idioma X. En contraste, todas las 100 observaciones se utilizan en el informe diagnóstico separado del Idioma Y. Este enfoque: aumenta la visibilidad del subgrupo, no distorsiona la estimación de la población global.

23. ASIGNACIÓN DE USUARIOS MULTILINGÜES

Una persona puede ser adecuada para más de una celda de idioma. Uno de los enfoques preferidos para el Panel de Población principal es asignar un idioma de auditoría principal predefinido al usuario. El idioma de auditoría principal puede determinarse de acuerdo con estos campos:

- Idioma de uso diario de la IA
- Lengua materna o del hogar
- Mercado y localización
- Competencia en el uso natural del prompt
- Equilibrio del panel

Si el usuario también será evaluado en otros idiomas, estas observaciones:

- panel multilingüe emparejado,
- Panel de Equidad Lingüística

también se etiquetan en consecuencia. Está prohibido que la misma persona reciba ponderación completa de la población en todos los idiomas.

24. ASIGNACIÓN DE IDIOMA DENTRO DEL PAÍS

Sea la población adecuada para los grupos de idioma l en el país c: N_{c,l}. Si la cuota del país es: n_c, entonces la cuota de idioma bruta:

```
q_{c,l}= n_c× N_{c,l}/ Σ_l N_{c,l}
```

puede calcularse de esta manera. Sin embargo, si los grupos de idiomas se superponen:

```
Σ_l N_{c,l}> N_c
```

puede ser posible. En este caso, usar una participación directa crea un doble conteo. Opciones de solución:

- clasificación principal del idioma de auditoría,
- dividir el peso humano,
- modelo común multilingüe,
- panel de idioma coincidente separado

pueden utilizarse. El método utilizado debe registrarse.

25. ASIGNACIÓN DEL PLAN E INTERFAZ DENTRO DE LA POBLACIÓN

Los usuarios gratuitos y de pago o las interfaces web y móviles pueden comportarse de manera diferente. Sin embargo, dividir cada celda país-idioma por separado en capas de plan e interfaz puede fragmentar demasiado la muestra. Se pueden usar tres enfoques.

25.1. Subasignación proporcional a la población

Si se conoce la distribución real de usuarios, la asignación se realiza según el plan y las cuotas de la interfaz.

25.2. Asignación Experimental Equilibrada

Se realiza un muestreo igual o mayor para comparar la diferencia entre el plan o la interfaz. En la estimación global, se re-pesa según las cuotas de uso reales.

25.3. Separar Puntuaciones interfaz del producto

Los productos web gratuitos, web de pago y móviles se reportan como objetos de auditoría separados. También se puede producir un resultado combinado. El enfoque a utilizar debe definirse antes de la medición.

26. NÚMERO DE INVITACIÓN Y OBJETIVO DE OBSERVACIÓN VÁLIDO

Si el objetivo de observación válido para un estrato es nh, invitar exactamente nh personas generalmente no es suficiente. Porque:

- algunas personas no participan,
- algunas no completan la tarea,
- algunos registros quedan fuera de protocolo,

algunos paquetes de evidencia se vuelven inválidos. Sea la tasa de finalización esperada para el estrato rh y la tasa de validez esperada vh. El objetivo de invitación es aproximadamente:

```
m_h= ⌈n_h/ (r_h v_h)⌉
```

puede calcularse de la siguiente manera. Ejemplo:

- Objetivo de observación válida: 100
- Tasa de finalización esperada: 80 por ciento
- Tasa de validez esperada: 90 por ciento

```
m_h= ⌈100/ (0.80× 0.90)⌉ = 139
```

personas pueden ser invitadas. Esto no es una garantía. A medida que avanza el campo, se puede usar el sistema de respaldo predefinido.

27. LÍMITE DE INFLACIÓN DE INVITACIONES

Invitar a más personas por sí solo no resuelve el sesgo de la muestra. Los primeros en responder pueden ser sistemáticamente diferentes de los últimos en responder. Tomar solo a los primeros 1,000 que completen:

- zona horaria,
- interés en la tecnología,
- calidad de la conexión,
- horario de trabajo

puede producir sesgo en términos de. Invitación y orden de selección:

- debería ser aleatorizado,
- debe ser registrado dentro del estrato,

debería ser gestionado durante toda la ventana de campo.

28. SISTEMA DE PARTICIPANTES ALTERNATIVO

Se debe crear una lista de respaldo ordenada para cada estrato antes de que se vean los resultados. Un participante de respaldo solo puede ser activado por las siguientes razones:

- No respuesta a la invitación
- Incumplimiento en la realización de la tarea
- Problema de idoneidad técnica o ética
- Violación del protocolo de demanda
- Invalidez del paquete de evidencia

- Re-registro de la misma persona

El usuario de respaldo no puede intervenir por la siguiente razón:

> *La respuesta de la IA del usuario anterior dio un resultado incorrecto, negativo o no deseado.*

Una respuesta incorrecta de la IA es el resultado válido del sistema. No puede ser una razón para el cambio de participante.

29. RESULTA CEGUERA EN DECISIÓN DE VALIDEZ

La persona que decide si la observación cumple con el protocolo tanto como sea posible:

- debe evaluar la respuesta sin saber si es correcta o incorrecta,
- o si es positivo o negativo para la marca.

Primero:

- Validez de captura y protocolo
- Luego la evaluación de la respuesta sustantiva

debe hacerse. De lo contrario, las respuestas negativas pueden ser declaradas más fácilmente como: "inválidas."

30. ORDEN DE RESERVA DENTRO DE LA CAPA

El participante de reserva debería ser, en la medida de lo posible, del mismo:

- país,
- idioma,
- interfaz del producto,
- plan,
- condición del panel

capa. Si no queda reserva en la misma capa, se aplica la escalada predefinida:

- Selección aleatoria adicional dentro de la misma capa
- Pool predefinido de la capa superior
- Dejar la asignación incompleta y advertencia de alcance
- Nueva versión de análisis

Después de ver el resultado, traer a un usuario de otro país y cerrar la brecha no se puede hacer silenciosamente.

31. ¿QUÉ SUCEDE SI UNA CAPA ES DEMASIADO CORTA?

Puede quedar por debajo del objetivo de la capa. Ejemplo:

- objetivo: 100
- observación válida: 72

Opciones correctas:

- Extender la ventana de campo hasta un alcance predefinido

- Usar la lista de respaldo
- Combinar la celda con una capa superior similar y predefinida
- Informe el resultado con amplia incertidumbre
- No publique el resultado de la celda
- Recalcule el peso global usando un método apropiado

Opción incorrecta: Eliminar completamente la capa de bajo puntuación del análisis.

32. CAMBIO DE CUOTA POST-RESULTADO

Los siguientes cambios son manipulaciones de muestras orientadas a resultados:

- Reducir la cuota objetivo del país de bajo puntuación
- Aumentar la cuota del idioma de alto puntuación
- Anular la validez de los usuarios que brindan respuestas negativas
- Redefinir el grupo de países pequeños basado en los resultados
- Excluir un plan con errores críticos
- Agregar nuevos usuarios hasta alcanzar el promedio deseado

Cada cambio de cuota material:

- nueva versión de asignación,
- justificación,
- fecha,
- tomador de decisiones
- Preservación del resultado anterior

debe realizarse con.

33. MUESTREO MULTIETAPA

El muestreo multi-etapa se puede usar cuando no es posible acceder directamente a todos los individuos. Ejemplo:

- Selección de región o grupo de países
- Selección de país
- Selección de panel o socio local de investigación
- Selección de usuario
- Asignación de producto o tarea de IA

La probabilidad total de selección de un usuario individual es el producto de las probabilidades condicionales en cada etapa:

```
π_i= π_r× π_{c|r}× π_{f|c}× π_{i|f}
```

Aquí:

- r: región o grupo pequeño de países
- c: país
- f: marco de muestreo o panel
- i: persona

Peso de diseño básico:

```
d_i= 1/π_i
```

Si no se pueden registrar las probabilidades de selección, no se puede presentar el diseño como una muestra completamente basada en probabilidades. [K04; K07]

34. ASIGNACIÓN EN PANELES VOLUNTARIOS Y NO PROBABILÍSTICOS

Un panel mundial verdaderamente basado en probabilidades puede no ser siempre posible. Se puede usar un panel en línea voluntario o profesional. En este caso, aún se pueden establecer cuotas de país e idioma de acuerdo con la población elegible para el producto. Sin embargo:

- la probabilidad real de selección del participante puede ser desconocida,
- puede ocurrir auto-selección,

el ponderado solo puede hacer ajustes basados en variables observadas. El estado correcto:

Panel de usuarios no probabilístico estratificado y calibrado a la población

puede ser. La siguiente declaración no puede usarse: “1,000 personas seleccionadas completamente al azar de la población mundial.”

35. ARQUITECTURA DE PONDERACIÓN

El peso final de la observación puede consistir en múltiples componentes:

```
w_i= d_i× a_i^{NR}× g_i^{CAL}
```

Aquí:

- di: peso de diseño base
- aiNR: ajuste por no respuesta
- giCAL: coeficiente de calibración

35.1. Peso de Diseño

Es el inverso de la probabilidad de selección del participante en el diseño de muestreo.

35.2. Ajuste por No Respuesta

Para personas seleccionadas con características similares:

- cuántos respondieron,
- cuántos produjeron una observación válida

se puede tener en cuenta. Para la clase de respuesta de muestra r:

```
a_r^{NR}= (Σ_{j∈S_r} d_j)/ (Σ_{i∈R_r} d_i)
```

Aquí:

- Sr: muestra elegible seleccionada
- Rr: aquellos que respondieron o produjeron una observación válida

puede ser. Ajuste por no respuesta: no resuelve completamente el sesgo no observado, debe explicarse junto con sus supuestos.

35.3. Peso de calibración

La muestra, a totales poblacionales conocidos:

- país,
- idioma,
- edad,
- plan,
- interfaz de usuario.

La calibración puede aplicarse a estas dimensiones, pero solo donde existan totales externos confiables.

36. NORMALIZACIÓN DE PESO

Pesos:

- a la población total,
- a el tamaño total de la muestra,
- a individuo

puede ser normalizado. La normalización utilizada en la fórmula de puntuación debe explicarse. Pesos normalizados individualmente:

```
w̃_i= w_i/ Σ_j w_j
```

se obtienen. Estimación ponderada:

```
θ̂= Σ_i w̃_iY_i
```

se calcula de la siguiente manera.

37. RECORTE DE PESOS

Algunos participantes pueden recibir pesos muy altos porque representan grupos muy pequeños y poco muestreados. Esta situación:

- puede aumentar la varianza estimada,
- la influencia de una sola observación,
- y la volatilidad de la puntuación

puede aumentar. Los pesos pueden recortarse en un límite superior. Sin embargo, el recorte: puede crear sesgo en la representación de la población, puede hacer que los grupos pequeños sean invisibles nuevamente. Por lo tanto, el recorte de pesos debe:

- definirse antes de los resultados,
- mostrarse junto con los resultados sin recortar,

y llevar un análisis de sensibilidad.

38. TAMAÑO EFECTIVO DE LA MUESTRA

Aunque el número bruto de observaciones es n, si los pesos varían mucho, la cantidad real de información puede ser menor. Tamaño de muestra efectivo de Kish:

```
n_eff= (Σ_i w_i)^2/ Σ_i w_i^2 [K08]
```

puede mostrarse como. Con pesos iguales:

```
neff≈n
```

se obtiene. Si los pesos están muy desequilibrados: neff<n. un estudio GEO-1000:

- 1,000 observaciones válidas,
- pueden representar un tamaño de muestra efectivo de 620

. Solo para el público: decir “Se midió a 1,000 personas.” puede parecer fuerte debido a la precisión. En el informe principal, al menos:

- conteo bruto de observaciones válidas,
- tamaño de muestra efectivo,
- variabilidad de los pesos

debe determinarse.

39. EFECTO DEL DISEÑO DEL PESO

Efecto de diseño aproximado debido a los pesos:

```
DEFF_w≈ 1+ CV(w)^2
```

puede ser mostrado. Aquí:

```
CV(w) = sd(w) / mean(w)
```

Este es el coeficiente de variación de los pesos y solo una representación aproximada de su efecto. El efecto total del diseño también puede variar con:

- agrupamiento,
- estratificación,
- medición repetida,
- usuarios emparejados

Estos efectos deben ser estimados por separado.

40. AGRUPAMIENTO

Lo mismo:

- Del panel de investigación,
- del país,
- de la comunidad lingüística,
- del dispositivo o institución,
- del usuario

Las observaciones pueden parecerse entre sí. En ese caso, 1,000 observaciones no equivalen a 1,000 unidades de información completamente independientes. Agrupación:

- en el diseño de la muestra,
- en el cálculo de la varianza,
- en el tamaño efectivo de la muestra

debe ser tomado en cuenta.

41. PRUEBA DE MÁS DE UN PRODUCTO DE AI POR EL MISMO USUARIO

Permitir que un usuario pruebe múltiples productos de IA permite medir la diferencia entre modelos en la misma persona. Ventaja: se pueden controlar las características del usuario, la comparación pareada se puede reforzar. Limitación:

- las observaciones del producto no son independientes,
- el orden de las tareas puede afectar otras respuestas,
- un usuario puede transmitir la respuesta vista en el primer sistema al siguiente sistema,

los usuarios que pueden acceder a todos los productos pueden diferir de la población. Por lo tanto, los productos:

- debería asignarse en orden aleatorio,
- se debería prevenir la compartición de respuestas previas,
- el diseño pareado debería registrarse,

El análisis debería proteger la dependencia. Cada producto de IA debería nuevamente tener su propio objetivo de 1,000 observaciones de productos válidas. Si los mismos 1,000 humanos prueban diez productos:

- 1,000 humanos únicos,
- 10,000 observaciones de productos

pueden formarse. Estos dos números no son iguales.

42. DISEÑO DE BLOQUES FALTANTES BALANCEADO

No todos los usuarios necesitan acceder a todos los productos de IA. Por ejemplo, si hay diez productos de IA, a cada participante se le pueden asignar tres productos. Si las asignaciones se hacen de manera balanceada:

- cada producto 1,000 observaciones,
- cada par de productos usuarios compartidos suficientes,
- participación de una población más amplia

puede lograrse. Como candidato para este enfoque:

Diseño de bloques incompletos balanceados

puede ser utilizado. Sus ventajas:

- reduce la carga sobre los participantes,
- disminuye el riesgo de seleccionar solo usuarios intensivos de IA,

proporciona comparaciones emparejadas entre modelos. El diseño exacto se detallará en las Secciones 8 y 9.

43. Alcance natural ASIGNACIÓN

Cada producto de IA recibe una asignación separada de acuerdo con su propia población Alcance natural. Por lo tanto, el mismo país:

- 200 para un producto,

- 0 para otro producto,
- 50 para un tercer producto

puede ser observado. Esta diferencia no es injusticia. Es que los universos reales de acceso de los productos son diferentes. Dado que las puntuaciones Alcance natural pertenecen a diferentes poblaciones, no deberían usarse solas para un ranking directo.

44. Base común de comparación ASIGNACIÓN

Se crea un universo de acceso común para la comparación entre productos. La población común:

```
U_CS= ∩_{a∈A} U_a
```

podría considerarse de esta manera. Cada producto se mide con los mismos pesos de población Base común de comparación. Este enfoque:

- proporciona una comparación justa de productos,
- fijo para el mismo país y idioma

lo asegura. Sin embargo, no muestra la diferencia natural de acceso global de los productos. Por lo tanto, los resultados Alcance natural y Base común de comparación deberían reportarse juntos.

45. GESTIÓN DE LA MUESTRA A TRAVÉS DE OLEADAS

Puede haber dos objetivos diferentes a través de múltiples oleadas de medición.

45.1. Resultado de Corte Transversal Actual

En cada ola, se utiliza un marco poblacional actual elegible para el producto. Esta puntuación responde a la pregunta: "¿Qué está pasando ahora en el universo de acceso global real?"

45.2. Resultado de Tendencia con Marco Fijo

Se preservan los pesos poblacionales de la ola inicial. Esta puntuación responde a la pregunta: "¿Cómo ha cambiado la representación en la misma población de referencia a lo largo del tiempo?"

45.3. Separación de Dos Resultados

Cuando el producto se expande a nuevos países, La puntuación actual puede disminuir. Esta disminución:

- se debe al deterioro del producto,
- debido a la inclusión de poblaciones nuevas y difíciles

Puede ser atribuible. Una tendencia con marco fijo puede mostrar mejora al mismo tiempo. Por lo tanto:

- Instantánea con Marco Actual
- Tendencia con Marco Fijo

Son resultados separados.

46. RENOVACIÓN Y TRANSFORMACIÓN DEL PANEL

Entre las olas:

- los mismos usuarios,

- usuarios completamente nuevos,
- panel parcialmente renovado

utilizable.

46.1. Panel de Sección Completa

Se seleccionan nuevos usuarios en cada ola. Puede ser fuerte para la estimación de la población. No mide el cambio individual.

46.2. Panel Longitudinal

Se miden nuevamente los mismos usuarios. Muestra los cambios individuales de usuarios y productos. El panel puede crear efectos de desgaste y aprendizaje.

46.3. Panel Rotativo

Se retiene una parte de los participantes, mientras que una parte se renueva. Proporciona tanto tendencia como representación de población fresca. La estructura del panel utilizada debe ser visible en la tarjeta de resultados.

47. FIJACIÓN DE MUESTRA

Los siguientes registros deben ser fijados antes de que comience cada ola:

- Producto de IA
- interfaz de usuario
- Versión de la población objetivo
- Países Alcance natural
- Países Base común de comparación
- Estratos
- Pequeños grupos de países
- Cuotas brutas
- Asignaciones enteras
- Asignaciones de idioma
- Plan de sobremuestreo
- Objetivos de invitación
- Listas de reserva
- Método de selección
- Fórmula de ponderación
- Regla de ceguera de resultados
- Escalada de déficit
- Plan de análisis
- Método de decodificación
- Semilla aleatoria, si se usa
- Aprobación humana responsable

A este archivo:

puede recibir el nombre.

48. CAMBIOS PERMITIDOS DESPUÉS DEL FIJACIÓN DE LA MUESTRA

Los siguientes eventos pueden requerir una nueva versión:

- Terminación del acceso al producto en el país
- Problema grave de seguridad en campo
- Error material en los datos de población
- Proveedor de panel incapaz de entregar a un país específico
- Cambio en los requisitos de ética de investigación
- Error crítico en la herramienta de medición

Cambio:

- debe ser justificado,
- con fecha,
- aprobado,
- realizado preservando la versión anterior.

La fortaleza o debilidad de los resultados no es una razón para el cambio.

49. ESTADOS DE CALIDAD DE LA MUESTRA

Una ola de muestra puede tener uno de los siguientes estados:

SQ-0— NO DISEÑADO

No hay un diseño de muestra documentado.

SQ-1— OBSERVACIÓN POR CONVENIENCIA

Se trata de una observación por conveniencia o voluntaria. No se puede hacer una generalización a la población.

SQ-2— CUOTAS CONTROLADAS

Existen cuotas por país e idioma. Las probabilidades de selección no se conocen completamente.

SQ-3— PANEL PONDERADO

Existe un panel estratificado y calibración de la población. Persisten limitaciones de cobertura y auto-selección.

SQ-4— BASADO EN PROBABILIDAD O HÍBRIDO

Existen probabilidades de selección conocidas o un diseño híbrido sólido.

SQ-5— REPLICADO Y CALIBRADO

El diseño ha sido replicado en diferentes olas y aplicaciones externas; se ha calibrado el comportamiento del sesgo y la varianza. Estos estados pueden afectar la clase de confianza de la puntuación público.

50. ASIGNACIÓN SINTÉTICA INTEGRAL GEO-1000

DEMOSTRACIÓN DE METODOLOGÍA SINTÉTICA / Los siguientes países, poblaciones, idiomas, producto de IA y cálculos son completamente ficticios. Para el producto de IA Sintético Orion, deje que la población adecuada para Alcance natural sea de 1 mil millones de personas. Objetivo del Panel de Población Principal:

```
n=1,000
```

50.1. Asignación de Países y Pequeños Países

Estrato	Población adecuada	Participación de la población	Cuota principal
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
Pequeño conjunto de países	2,000,000	0.2%	2
Total	1,000,000,000	100%	1,000

50.2. Asignación de Idiomas en el País B

Población elegible en el País B:

- Idioma X: 80 por ciento
- Idioma Y: 20 por ciento

Asignación del Panel de Población:

- Idioma X: 200
- Idioma Y: 50

Asignación adicional de Equidad de Idiomas: Idioma Y: 50 Objetivo de campo en el País B:

- Idioma X: 200
- Idioma Y: 100
- Total: 300

Sin embargo, los pesos de la población global del Panel de Población:

- Idioma X: 80% del País B
- Idioma Y: 20% del País B

permanecer como tal.

50.3. Apéndice del Observador por País

Hay dos observaciones de probabilidad para H, I y J en el pequeño grupo de países en el Panel de Población. El Panel de Observador por País también:

H: 1

I: 1

J: 1

puede orientar observaciones. Estas tres observaciones:

- aumentan la cobertura geográfica,
- muestran eventos de países individuales,

no crean una puntuación del país.

50.4. Suplemento de Accesibilidad

Si se encuentra que el panel principal cubre de manera insuficiente a los usuarios con accesibilidad, se pueden tomar observaciones adicionales 40. Estos usuarios: se calibran según los grupos de población relevantes, producen un resultado separado del Panel de Accesibilidad.

50.5. Volumen total de campo

Panel	Objetivo actual
Panel de población	1,000
Suplemento de equidad del lenguaje	50
Suplemento de Observador por País	3
Suplemento de accesibilidad	40
Objetivo total de campo válido	1,093

El nombre GEO-1000 sigue siendo válido. Porque:

> *1,000 es el objetivo de observación válido para el panel de población principal.*

Se pueden añadir paneles complementarios de justicia y cobertura a esto.

50.6. Volumen de invitación

En el Panel de Población en promedio:

- finalización: 82 por ciento
- validez: 92 por ciento

que sea esperado. Invitación requerida:

```
[1,000/ (0.82× 0.92)] = 1,326
```

puede ser. Con paneles complementarios, el total de invitaciones puede superar 1,500. Por lo tanto:

> *GEO-1000 no es solo un método para enviar invitaciones a 1,000 personas.*

50.7. Resultado de campo

En el Panel de Población:

- 1,326 personas han sido invitadas,

- 1,080 tareas completadas,
- 1,012 registros han sido encontrados válidos en términos de protocolo,
- los primeros 1,000 fueron incluidos en el análisis principal según un orden aleatorio/respaldo predefinido

Está bien. Los 12 registros válidos restantes:

- verificación de respaldo,
- análisis de sensibilidad,
- reverificación

Se puede almacenar por. No se puede decidir qué 1,000 se elegirá solo mirando los resultados.

51. ALGORITMO DE ASIGNACIÓN NORMATIVA

La secuencia predeterminada de operaciones para un panel principal GEO-1000 es la siguiente:

Paso 1— Fijar el Producto y el ID de Onda

Determinar qué producto AI, interfaz y fecha se miden.

Paso 2— Fijar la Versión de la Población Objetivo

Se utiliza el marco de población elegible por producto de la sección 5.

Paso 3— Definir el Objetivo Principal de Observación Actual

Objetivo candidato predeterminado:

```
n=1,000
```

Paso 4— Definir Capas de Material

Se definen país, idioma y, si es necesario, capas de plan/interfaz.

Paso 5— Configurar Pequeños Grupos de Países

Los países que generen asignaciones cercanas a cero se incluyen en los grupos correspondientes.

Paso 6— Calcular Cuotas Proporcionales Brutas

```
qh=nWh
```

Paso 7— Realizar Asignación Entera

Se aplica el método de Mayor Resto o su equivalente predeclarado.

Paso 8— Añadir Plan de Sobremuestreo

Los suplementos de Observador por País, Equidad Lingüística y Accesibilidad se guardan por separado.

Paso 9— Calcular Números de Invitación y Respaldo

Se utilizan las tasas esperadas de finalización y validez.

Paso 10— Aleatorizar Orden de Selección

Las listas principales y de respaldo se congelan.

Paso 11— Guardar Fórmula de Pesos

Se definen los pesos de diseño, respuesta y calibración.

Paso 12— Confirmar Fijación de Muestra

Se crea un registro de versión e integridad antes de que se vean las respuestas de la IA.

52. DISPOSICIONES NORMATIVAS OBLIGATORIAS

CH06-N01

La asignación de muestra GEO-1000 debe ser creada y versionada antes de que se puedan ver las respuestas de IA.

CH06-N02

El objetivo principal del Panel de Población debe definirse como al menos 1,000 observaciones válidas por producto de IA y ola de medición principal; si se utiliza un objetivo diferente, se debe explicar la justificación.

CH06-N03

La persona invitada, la persona que completa la tarea y los números de observación válidos deben distinguirse entre sí.

CH06-N04

La asignación principal debe basarse en la población elegible para el producto definida en la Sección 5.

CH06-N05

Las cuotas por país o idioma deben calcularse en función de la proporción de la población elegible para el producto relevante, no de la población total.

CH06-N06

El método para convertir cuotas fraccionarias en números enteros debe definirse de antemano.

CH06-N07

La resolución de los empates al redondear debe ser independiente de los resultados.

CH06-N08

Los países pequeños con una población objetivo positiva no deberían dejarse como estratos deterministas separados con probabilidad cero; se debería utilizar un método de agrupamiento o selección probabilística adecuado.

CH06-N09

Las piscinas de países pequeños deben justificarse en términos de acceso a productos y características financieras de la población.

CH06-N10

Las observaciones del Observador por País también deben registrarse sin cambiar el peso principal de la población.

CH06-N11

Una observación de un Observador por País no puede presentarse como una puntuación del país.

CH06-N12

Los países, idiomas o grupos de usuarios sobrerrepresentados deben ser ajustados a los pesos de la población objetivo real en la estimación global.

CH06-N13

La proporción de la muestra no puede usarse como la proporción de la población.

CH06-N14

La misma persona no puede ser duplicada en la población humana global debido a múltiples idiomas, cuentas, sesiones o productos.

CH06-N15

Las observaciones inválidas deben clasificarse sin basar la razón de la invalidez en el contenido de los resultados.

CH06-N16

Incorrecto, negativo, rechazo o respuestas críticas de IA no pueden ser la razón para cambiar a un participante.

CH06-N17

Se deben crear listas de participantes de respaldo dentro del estrato antes de ver los resultados.

CH06-N18

Las brechas de capa no pueden cerrarse silenciosamente por otro país o idioma.

CH06-N19

Los cambios de cuota posterior al resultado, de capa, de grupo de país pequeño o de peso requieren una nueva versión del análisis.

CH06-N20

Todos los componentes de los pesos y el método de normalización deben publicarse.

CH06-N21

Deben explicarse los supuestos del ajuste por no respuesta.

CH06-N22

Si se utiliza el recorte de pesos, deben conservarse el resultado sin recortar y el análisis de sensibilidad.

CH06-N23

El informe principal debe mostrar el tamaño de muestra efectivo apropiado junto con el número bruto de observaciones.

CH06-N24

Se deben tener en cuenta la agrupación, repetición y dependencias de usuarios emparejados en el cálculo de la varianza.

CH06-N25

El diseño en el que la misma persona prueba múltiples productos de IA no puede presentarse como una muestra de usuarios completamente independiente.

CH06-N26

Cada producto de IA debe alcanzar por separado su propio objetivo de 1,000 observaciones de producto válidas.

CH06-N27

Las asignaciones Alcance natural y Base común de comparación deben mantener registros de muestras separados.

CH06-N28

Los resultados actuales y de marco fijo entre las oleadas no se pueden combinar sin explicación.

CH06-N29

Si la muestra es insuficiente para un subgrupo, el estado correcto debe ser MUESTRA INSUFICIENTE, EXPLORATORIA, o NO ESTIMADA.

CH06-N30

El tamaño de muestra bruto no puede utilizarse como un indicador definitivo de precisión, ignorando los pesos y el efecto del diseño.

CH06-N31

No se pueden reclamar pesos de diseño en un panel de voluntarios o de no probabilidad como si se conocieran las probabilidades de selección.

CH06-N32

La muestra no puede modificarse a posteriori para excluir grupos con puntuaciones bajas o para ampliar grupos con puntuaciones altas.

CH06-N33

Las observaciones del Panel de Población principal y de los paneles complementarios deben etiquetarse por separado en términos de propósito y peso.

CH06-N34

El fijación de la muestra debe tener un propietario humano o institucional responsable.

53. FORMAS DE FALLA

CH06-F01— ESTABLECER CUOTA DESPUÉS DEL RESULTADO

El número de países e idiomas se determina después de ver las respuestas de la IA.

CH06-F02— USAR RESPALDO HASTA OBTENER LA MEJOR RESPUESTA

Las respuestas negativas se reemplazan con un nuevo usuario incluso si son válidas.

CH06-F03— CONTAR 1,000 INVITACIONES COMO 1,000 OBSERVACIONES VÁLIDAS

No se muestran la falta de finalización y la invalidez.

CH06-F04— REINICIAR EL PAÍS A CERO CON REDONDEO PURAMENTE

La probabilidad de selección de un país con población objetivo positiva se deja en cero.

CH06-F05— ASIGNAR UNA PERSONA A CADA PAÍS Y LLAMARLO LA PROPORCIÓN DE POBLACIÓN

La desproporción causada por la cuota mínima está oculta.

CH06-F06— CONSIDERANDO la observación del Observador por País COMO OBSERVACIÓN DE LA POBLACIÓN

La observación adicional de un país pequeño recibe peso completo en el panel principal de 1,000 personas.

CH06-F07— ESTABLECIENDO EL GRUPO DE PAÍSES PEQUEÑOS BASADO EN RESULTADOS

Un país que produce puntuaciones bajos se traslada a otro grupo o se excluye de la cobertura.

CH06-F08— PONIENDO PAÍSES INCOMPATIBLES EN EL MISMO GRUPO

Países que difieren en acceso a productos, idioma o jurisdicción se combinan únicamente porque son pequeños.

CH06-F09— SOBREPONIENDO EL MUestreo EXCESIVO

Determina el resultado global con una participación de muestra mayor que la participación de población de idioma o país.

CH06-F10— MULTIPLICANDO INDIVIDUOS MULTILINGÜES

Una persona se cuenta como una unidad poblacional separada en cada celda de idioma.

CH06-F11— CONSIDERANDO LA PARTICIPACIÓN DE LA MUESTRA COMO PARTICIPACIÓN POBLACIONAL

Los números de respuesta en bruto se convierten directamente en la puntuación global.

CH06-F12— ORDENANDO EL ORDEN DE INVITACIÓN POR ORDEN DE FACILIDAD

Los usuarios que responden primero y rápidamente determinan el panel.

CH06-F13— USANDO RESERVAS DE ESTRATO-EXCLUSIÓN

Las observaciones faltantes de país o idioma se completan silenciosamente con participantes de otro grupo.

CH06-F14— VINCULANDO INVALIDEZ CON EL CONTENIDO DE LA RESPUESTA

La respuesta incorrecta de la IA se declara técnicamente inválida.

CH06-F15— REEMPLAZANDO RECHAZOS CON UN NUEVO USUARIO

El rechazo del sistema se elimina del denominador aunque sea el resultado real del usuario.

CH06-F16— OCULTANDO LA FÓRMULA DE PESO

La puntuación pública no muestra con qué población y pesos muestrales fue producida.

CH06-F17— RECORTANDO PESOS PARA ELIMINAR GRUPOS PEQUEÑOS

Se recortan los grupos pequeños con pesos altos porque el resultado no se ve bien.

CH06-F18— OCULTANDO LA MUESTRA EFECTIVA

Incluso si los registros brutos 1,000 caen a 500–600 puntos de datos debido a los pesos, se presenta como "precisión de 1,000 personas".

CH06-F19— CONSIDERANDO LA AGRUPACIÓN COMO INDEPENDENCIA

Se consideran completamente independientes las observaciones del mismo panel, institución o usuario.

CH06-F20— CONSIDERANDO A LAS MISMAS 1,000 PERSONAS COMO DIEZ MIL PERSONAS INDEPENDIENTES

Diez observaciones de un producto de IA se describen como si la población fuera diez veces mayor.

CH06-F21— CONSIDERANDO Base común de comparación COMO Alcance natural

La muestra de comparación común se presenta como si fuera la población de alcance global real del producto.

CH06-F22— CLASIFICANDO DIRECTAMENTE LOS RESULTADOS Alcance natural

Se comparan las puntuaciones del producto en diferentes poblaciones objetivo como si fueran del mismo universo.

CH06-F23— OCULTANDO EL CAMBIO DE POBLACIÓN EN OLEADAS

Aunque nuevos países o grupos de usuarios entren en la puntuación, la tendencia se interpreta solo como un cambio en la calidad del producto.

CH06-F24— PRODUCIENDO PUNTUACIONES DEFINITIVAS A PARTIR DE MUESTRAS INSUFICIENTES PARA UN SUBGRUPO

Se publican unas pocas docenas de observaciones como puntuaciones definitivas para un país o idioma.

CH06-F25— CAMBIANDO SILENCIOSAMENTE EL FIJACIÓN DE MUESTREO

Se actualizan las definiciones de cuota, ponderación o grupo sin un registro de versión.

CH06-F26— MULTIPLICANDO A CIEGAS LAS TASAS DE ELEGIBILIDAD

La población elegible y los cálculos de cuota ignoran las características de la población dependiente.

CH06-F27— REEMPLAZANDO EL PANEL PRINCIPAL CON EL PANEL COMPLEMENTARIO

La equidad lingüística u observaciones del Observador Nacional se utilizan para reducir el objetivo de población principal de 1,000.

CH06-F28— ELIMINANDO CELDAS DE BAJA PUNTUACIÓN CUANDO EL PANEL ES ESCASO

Se excluyen las muestras incompletas porque son desventajosas en términos de resultados.

CH06-F29— USANDO PONDERACIÓN COMO TOQUE CIENTÍFICO

Se afirma que el problema de representación se resuelve solo agregando peso sin explicar el sesgo de selección y cobertura.

CH06-F30— PENSANDO QUE A GEO-1000 DEBERÍA PEDÍRSELE EXACTAMENTE 1,000 PERSONAS

Se ignoran la invitación, los paneles complementarios, la invalidez, la repetición y la estructura de la muestra efectiva.

54. PROCEDIMIENTO DE AUDITORÍA

Paso 1— Registrar el Propósito de la Muestra

La estimación de la población global se divide en diagnóstico de subgrupos, alcance por país, equidad lingüística, accesibilidad y comparaciones de productos.

Paso 2— Validar la Versión de la Población Objetivo

Se revisan el archivo de población y el registro de fechas en la Sección 5.

Paso 3— Definir el Objetivo Principal de Observación Actual

Se escribe el objetivo GEO-1000 predeterminado o un alternativo justificado.

Paso 4— Establecer Estratos

Se determinan diferencias en el país, idioma y material interfaz de usuario.

Paso 5— Identificar Conjuntos de Países Pequeños

Cada conjunto tiene:

- países,
- población,
- justificación de similitud,
- método interno de selección

que se registra.

Paso 6— Calcular Cuotas Brutas

```
qh=nWh
```

Se utiliza la fórmula o una alternativa justificada.

Paso 7— Aplicar la Asignación de Enteros

Se registran las reglas de redondeo y de desempate.

Paso 8— Agregar Sobremuestreo Complementario

Los objetivos de Observador por País, Equidad Lingüística y Accesibilidad se definen por separado.

Paso 9— Calcular el Volumen de Invitación y Reserva

Se utilizan las expectativas de finalización y validez.

Paso 10— Congelar la Selección y el Orden de Reserva

Los participantes son seleccionados u ordenados antes de ver los resultados.

Paso 11— Calcular Pesos de Diseño

Se registran las probabilidades de selección y los pesos de capa.

Paso 12— Definir el Procedimiento para Mantenerse en Silencio y No Responder

Se escribe una respuesta previamente anunciada en la brecha de la capa.

Paso 13— Crear el Fijación de Muestra

Se agregan el ID del documento, la versión, la fecha y el registro de integridad.

Paso 14— Revisar el Flujo de Validez Post-Campo

Se verifica que las decisiones de validez se dan independientemente de la precisión de las respuestas.

Paso 15— Generar los Pesos Finales

El diseño se separa en componentes de respuesta y de calibración.

Paso 16— Calcular el Tamaño Efectivo de la Muestra

Se reporta la diferencia entre el número bruto y el volumen de información.

Paso 17— Comparar la Asignación y los Valores Reales

Se muestran los números de observaciones planificadas y reales celda por celda.

Paso 18— Revisar los Cambios Post-Resultados

Se verifican los cambios de cuota, grupo, peso y alcance.

55. EVIDENCIA REQUERIDA

Identidad del producto de IA; oleada de medición; versión de población objetivo; Panel de Población objetivo; Observador de País objetivo; objetivo de Equidad Lingüística; objetivo de Panel de Accesibilidad; lista de estratos; poblaciones de estratos; ponderaciones de población; cuotas brutas; método de redondeo; registros de asignación fraccionaria y restos; regla de desempate; grupos de países pequeños; probabilidades de selección dentro del grupo; asignaciones por país e idioma; asignaciones de plan e interfaz; justificaciones de sobremuestreo; objetivos de invitación; supuestos de finalización; supuestos de validez; listas de participantes primarios y de reserva; registro de selección o aleatorización; registros de no respuesta; razones de observaciones inválidas; fechas de decisión de validez; ponderaciones de diseño; ajustes por no respuesta; coeficientes de calibración; política de recorte de pesos; tamaños de muestra brutos y efectivos; agrupamiento y registros de usuarios emparejados; asignación Alcance natural; y asignación Base común de comparación.

Pesos de marco fijos y actuales Comparación de la asignación planificada y real Fijación de muestra Registro de cambios Persona o institución responsable

56. LISTA DE VERIFICACIÓN DE AUDITORÍA

¿Está abierto el objetivo principal de observación válido? ¿El objetivo 1,000 invita u observa de manera válida? ¿Se congeló la muestra antes de que se vieran las respuestas de la IA? ¿Qué marco poblacional se utilizó? ¿Sobre qué base se establecieron los estratos? ¿Se muestran las cuotas de país e idioma en bruto? ¿Es claro el método para convertir fracciones en números enteros? ¿Se predeterminó la regla de abandono? ¿La probabilidad de selección para países pequeños es cero? ¿Son lógicamente consistentes los grupos de países pequeños? ¿Se han mezclado las observaciones del Observador por País en el peso de la población principal? ¿Se ha producido una puntuación de país a partir de una observación del Observador por País? Si se hizo un sobremuestreo de grupos de idioma y accesibilidad, ¿se les asignó un peso nuevamente? ¿Se ha contado a la misma persona como más de una unidad poblacional? ¿Se enumeran por separado los números de invitaciones, completados y observaciones válidas?

¿Se han creado listas de respaldo con anticipación? ¿Se ha reemplazado a un participante debido a una respuesta incorrecta o negativa de la IA? ¿Se tomó la decisión de validez antes de la puntuación de exactitud del material? ¿Se cerró silenciosamente la brecha de pestañas desde otro país? ¿Ha cambiado la cuota o el grupo después de los resultados? ¿Se pueden reproducir los pesos del diseño? ¿Son claras las suposiciones del ajuste por no respuesta? ¿Se ha realizado el recorte de pesos y se ha mostrado su efecto? ¿Se proporcionan juntos los tamaños de muestra brutos y efectivos? ¿El mismo usuario ha probado más de un producto de IA? ¿Se ha preservado esta dependencia en el análisis? ¿Ha alcanzado cada producto de IA por separado las observaciones del producto 1,000? ¿Fueron separados los muestras Alcance natural y Base común de comparación? ¿Cambió el marco de población entre oleadas?

¿Se han separado las tendencias enmarcadas fijas y actuales? ¿Se han publicado submuestras insuficientes como si fueran puntuaciones definitivas? ¿Es correcto el estado de calidad de la muestra? ¿Han reemplazado los paneles complementarios al panel principal? ¿Existe un propietario claramente responsable del registro de la muestra?

57. OBJECIONES Y RESPUESTAS

Objeción 1— "¿Por qué no damos a cada país individuos directamente según la proporción de la población?"

Esta es una suposición fuerte en países con poblaciones grandes y medianas. En países muy pequeños, la cuota bruta puede caer por debajo de una persona. En este caso:

- pool de países pequeños,
- selección probabilística de países,
- panel de Observadores separado

puede ser más preciso.

Objeción 2— "¿No sería más justo dar al menos una persona a cada país?"

Podría ser útil en términos de cobertura geográfica. Es desproporcionado en términos de estimación de población. La solución correcta es etiquetar por separado a esta persona como observación de un Observador por País.

Objeción 3— "¿No excluye completamente a un país recibir cero personas en el panel principal?"

Puede ser un problema si un país recibe una cuota cero como una capa determinista separada. Se puede otorgar una probabilidad de selección positiva en el grupo de países pequeños. Además, el Panel de Observadores puede mantener la visibilidad de observación del país.

Objeción 4— "¿Acaso agrupar países pequeños borra las diferencias culturales?"

Puede eliminarse si se utiliza para generar una puntuación del país. Para la estimación global principal, el grupo gestiona la probabilidad de selección. El identificador del país se conserva en el registro de observación. Se puede realizar un muestreo separado de Observadores y específico del país.

Objeción 5— "¿Por qué no incluimos a los usuarios Observadores dentro del 1,000?"

Pueden ser incluidos. Pero entonces la muestra de personas 1,000 ya no es proporcional a la población pura. Se necesitaría ponderación, y el tamaño efectivo de la muestra de la estimación principal podría disminuir. En la arquitectura fundacional, las observaciones 1,000 se asignan al Panel de Población; se añaden observaciones del Observador por País a esto.

Objeción 6— "¿No se volverá muy alto el costo total de campo?"

Podría. La norma no debe producir precisión artificial ocultando el costo. Con un presupuesto más bajo:

- menos productos de IA,
- menos idiomas,
- piloto más pequeño,
- incertidumbre mayor

puede ser usado. El estado correcto debe estar claramente escrito.

Objeción 7— "¿Por qué estamos recolectando 1,093 o más observaciones cuando existe el nombre GEO-1000?"

1,000 es el objetivo de observación válido para el panel de población principal. Se pueden agregar paneles de justicia y cobertura al objetivo principal. Esto no compromete el nombre del estándar. Hace visible el volumen total de campo real de la medición.

Objeción 8— "¿Por qué el sobrerremuestreo de un idioma distorsionaría el resultado global?"

Un idioma sobrerremuestreado parece grande porque está en respuestas brutas. Cuando se vuelve al peso de la población, el resultado global no se distorsiona. El resultado por idioma separado se vuelve más preciso.

Objeción 9— 'Los pesos son muy complejos; ¿no es más fácil entender la media bruta?'

La media bruta puede ser significativa si es proporcional a la población de la muestra. El muestreo excesivo, la falta de respuesta o las diferencias de cobertura pueden hacer que la media bruta responda a la pregunta incorrecta. La complejidad no puede ser ocultada si surge del diseño de la muestra.

Objeción 10— "Si hay 1,000 observaciones válidas, ¿por qué la muestra efectiva debería ser menor?"

Porque algunas observaciones pueden tener un peso muy alto, mientras que otras pueden tener un peso muy bajo. Los registros del mismo panel o usuario pueden parecerse entre sí. El número bruto es el conteo de registros físicos. La muestra efectiva se aproxima al volumen de información estadística.

Objeción 11— "¿No sería más justo que las mismas personas probaran todos los productos de IA?"

Reduce la diferencia entre usuarios. Sin embargo:

- puede reducir el universo de acceso a productos,
- puede aumentar la carga de tareas,

puede crear contaminación entre respuestas. Se puede gestionar con un diseño equilibrado por bloques incompletos o apareado.

Objeción 12— "Si un nivel no alcanza el objetivo, ¿no se puede publicar nunca la puntuación global?"

No en todos los casos. La deficiencia se evalúa en términos de:

- peso de la población,
- riesgo,
- la razón de la insuficiencia,
- cobertura restante

considerada. La puntuación:

- con cobertura limitada,
- con amplia incertidumbre,
- Con la advertencia COBERTURA INSUFICIENTE

puede ser publicado. La deficiencia no puede ser ocultada.

Apelación 13— "Si los datos de la población cambian, ¿tenemos que renovar todas las cuotas en cada ola?"

Puede ser necesaria la renovación para la puntuación instantánea actual. También se pueden mantener pesos iniciales fijos para la comparación de tendencias. Los dos resultados deben publicarse por separado.

Apelación 14— "Si no podemos establecer un panel probabilístico mundial, ¿no se puede aplicar NOMOS?"

Es aplicable. Sin embargo, el estado de calidad de la muestra debe registrarse correctamente. Un panel de voluntarios calibrado según la población puede ser valioso. No se puede presentar como una muestra completamente probabilística.

DISPOSICIÓN COMÚN DEL CAPÍTULO 60

El poder de GEO-1000 no surge únicamente del número 1,000. De 1,000 personas:

- de qué población provenían,
- con qué probabilidad fueron seleccionadas,
- qué país e idioma representaban,
- cuántas veces se contó a la misma persona,
- qué observaciones se consideraron válidas,
- cómo se gestionaron los países pequeños,
- cómo se ponderaron los grupos sobremuestreados,
- cuántas personas no completaron la tarea,
- quién fue reemplazado según los resultados

La autoridad de una muestra surge de estas elecciones de diseño. Una muestra defectuosa puede producir respuestas 100,000 y aún así hablar sobre la población incorrecta. Una muestra sólida puede contener menos respuestas pero respaldar una conclusión más confiable dentro de límites claramente establecidos. La población de un país grande debería tener un peso sustancial en la experiencia humana global, pero no se deben fabricar usuarios donde el producto no esté disponible. El peso global de un país pequeño debería seguir siendo pequeño, pero el país no necesita ser invisible. Un idioma puede representar una pequeña proporción de la población global mientras sigue teniendo un error grave y medible. Una persona puede usar tres idiomas y cinco productos; eso no crea quince personas. Una respuesta de IA puede ser incorrecta sin invalidar al participante. La respuesta incorrecta es lo que la auditoría intenta medir.

Reemplazarlo por un nuevo usuario no limpia la medición. Borra la realidad. Ponderar no convierte milagrosamente el panel incorrecto en la población mundial. Sin embargo, un sobremuestreo correctamente diseñado puede hacer visibles los subgrupos sin distorsionar la estimación global. Por lo tanto, la sexta ley de medición de NOMOS es:

> *Una muestra debe establecerse no para producir el resultado, sino para observar el resultado de manera justa.*

Su séptima ley es la siguiente:

> *La visibilidad de un país pequeño no tiene que aumentar su peso poblacional.*

Su octava ley es la siguiente:

> *La observación excesiva no significa un peso poblacional excesivo.*

La novena ley es la siguiente:

> *Una respuesta incorrecta de IA no es un defecto de datos; si se observa bajo el protocolo válido, es una cuestión de supervisión.*

La décima ley es:

> *La confianza en mil observaciones no proviene del número mil; proviene de la reproducibilidad de la selección y ponderación.*

Orden del Capítulo 6 de NOMOS

> *No solo me digas: 'Preguntamos a 1,000 personas.'*
>
> *Muestra de cuáles mil millones de personas se seleccionaron estos 1,000 y con qué probabilidad.*
>
> *No ignores un país solo porque represente a medio persona. / Pero tampoco afirmes que mediste todo el país solo porque muestreaste a una persona.*
>
> *Seleccione países pequeños de manera probabilística en el grupo correspondiente. / Si se necesita visibilidad, configure un panel de observadores separado.*
>
> *Puedes sobremuestrear los idiomas con pocos recursos. / Pero no lo hagas más grande en la puntuación global que su población.*
>
> *No multipliques a una persona tanto como cuentas, idiomas, sesiones y productos de IA.*
>
> *Utiliza la copia de seguridad cuando el participante no responda. / No cambies tu usuario cuando la IA dé una respuesta incorrecta.*
>
> *Decide sobre la validez sin ver la corrección de tu respuesta.*
>
> *Si falta la cuota, explíquelo. / No traiga a una persona de otro país para cubrir la brecha.*
>
> *No cambies la participación, el peso, la reserva de país o el objetivo de la muestra después de ver el resultado.*
>
> *Muestra el número bruto. / Muestra la muestra efectiva. / Muestra quién recibió peso pesado y quién recibió peso ligero.*
>
> *Bloquea la población primero. / Luego configura las capas. / Luego calcula las cuotas. / Luego selecciona los sustitutos. / Luego escribe los pesos. / Y solo después de esto, envía el primer prompt.*

La oración de cierre del capítulo

> *GEO-1000 no es un método para compilar mil respuestas; es un protocolo para hacer que la representación correcta de cada persona en la población objetivo sea medible bajo la probabilidad de selección y el peso correcto.*

Core normativo

Para cada producto de IA y oleada principal de medición, el Panel de Población GEO-1000 DEBERÍA aspirar al menos a 1,000 observaciones válidas del producto, a menos que se justifique explícitamente un objetivo diferente. La asignación de la muestra DEBE ser diseñada, versionada y fijada antes de observar las respuestas de la IA. La asignación primaria DEBE basarse en la población objetivo elegible para el producto, no en la población residente bruta, el número de invitaciones, cuentas, sesiones o disponibilidad por conveniencia. Las asignaciones fraccionarias DEBEN convertirse en cuotas enteras utilizando un método predeclarado, reproducible y ciego al resultado. Las poblaciones de masa positiva NO DEBEN ser asignadas determinísticamente con probabilidad de selección cero cuando la auditoría afirma una inferencia a la población objetivo completa. Las poblaciones pequeñas PUEDEN ser representadas mediante estratos agrupados justificados y selección proporcional a la probabilidad. Las observaciones del Observador por País, Equidad Lingüística y Accesibilidad PUEDEN complementar el panel principal, pero DEBEN mantenerse etiquetadas por separado y reponderarse adecuadamente. Los grupos sobremuestreados NO DEBEN recibir su proporción de muestra bruta como peso poblacional. La sustitución de participantes DEBE basarse únicamente en condiciones de reclutamiento predeclaradas o de validez del protocolo. Los resultados de IA incorrectos, negativos, rechazados o críticos NO DEBEN desencadenar reemplazo. Las personas, cuentas, sesiones, idiomas y observaciones de productos de IA DEBEN permanecer como unidades distintas. Los pesos de diseño, ajustes por no respuesta, calibración, recorte, agrupamiento, observaciones pareadas y tamaño efectivo de muestra DEBEN ser documentados. Cualquier cambio posterior a los resultados en cuotas, estratos, grupos, pesos, exclusiones o objetivos de muestra DEBE crear una nueva versión del análisis y preservar el registro original.

CAPÍTULO 07

Paneles de observación por país y equidad lingüística

BASES METODOLÓGICAS / K04; K06; K07; K20

Límite de Capítulo

La Sección 5 estableció la siguiente disposición:

> *Para cada producto de IA, la población de usuarios elegible para el producto debe definirse por separado.*

La Sección 6 convirtió esta población en una muestra:

> *El Panel de Población principal debe proporcionar probabilidades de selección y ponderaciones predeterminadas a las personas en la población objetivo antes de que se vean los resultados.*

Ahora necesitamos resolver dos problemas de equidad diferentes. El primer problema:

> *La ponderación de la población puede hacer que países muy pequeños sean invisibles en la muestra principal.*

El segundo problema:

> *El promedio global de la población puede ocultar distorsiones graves de representación en lenguas con pocos recursos o con baja población.*

Una muestra pura de población es fuerte para medir la experiencia humana promedio en la población mundial. Sin embargo, la misma muestra:

- no puede observar cada país individualmente,
- no puede proporcionar suficientes muestras para cada idioma
- no se puede producir una puntuación de país para países pequeños,

podría no detectar errores raros pero sustanciales en idiomas con pocos recursos. Por lo tanto, GEO-1000 no puede consistir solo en un solo panel. Junto al panel principal de población, se necesitan dos estructuras complementarias:

Panel de Observadores del País

Etiqueta canónica del panel:

Panel de Observadores del País

y:

Panel de Equidad Lingüística

Etiqueta canónica del panel:

Panel de Equidad Lingüística

El Panel de Observadores del País responde a la siguiente pregunta:

> *En países que están poco representados en el panel principal de población, ¿hay al menos observación de usuarios real, verificación de acceso o señal temprana de riesgo?*

El Panel de Equidad Lingüística, por otro lado, responde a la siguiente pregunta:

> *Incluso si el promedio global parece alto, ¿los usuarios en ciertos idiomas están siendo representados de manera más incorrecta, incompleta o riesgosa?*

Estos dos paneles no reemplazan el Panel Principal de Población. Lo complementan. Dar una observación de un Observador de País a un país no mide la puntuación de ese país. Sobre-muestrear un idioma no hace que ese idioma sea más grande en la población global. La tarea de esta sección es:

- visibilidad del país frente al peso de la población,
- separar el promedio global con la Equidad Lingüística,
- separar la estimación de la población con la alerta temprana,
- separar la suficiencia estadística con el alcance de la observación

uno del otro. Esta sección aún no ha:

- definido cómo deben traducirse las prompts en todos los idiomas,
- la traducción inversa y la prueba de equivalencia semántica,
- el mecanismo detallado de selección de participantes,
- cómo los árbitros puntuará las afirmaciones atómicas,
- los umbrales finales de conformidad

exactamente. Estos asuntos se regularán más a fondo en las siguientes secciones. La cuestión fundamental de la Sección 7 es:

> *¿Cómo medimos produciendo una puntuación global sin inflar artificialmente a los países y lenguas pequeñas, pero también sin dejarlos completamente invisibles?*

NOMOS Desafío

Se establece un panel de población de 1,000 personas para un producto de IA. Como resultado de una muestra proporcional a la población:

- cientos de usuarios de unos pocos países grandes,
- docenas de usuarios de algunos países medianos,
- ningún usuario de la mayoría de los países pequeños

asistieron. Puntuación global:

94

Se calculó de esa manera. Luego dijiste: 'Este producto de IA representa la marca a nivel mundial con un 94 por ciento de precisión.' Sin embargo, no hay ni una sola observación de usuarios reales en 70 países pequeños. Estos países pueden constituir una pequeña porción de la población mundial. Aun así, las siguientes preguntas permanecen sin respuesta: ¿Es el producto realmente accesible? ¿Funciona la intención en el idioma correcto? ¿Confunde el sistema el país con otro país? ¿Se confunde la empresa local con la marca global? ¿Son incorrectos el precio, el servicio y el contexto legal? ¿Falla el sistema en responder? ¿Existen errores críticos de baja frecuencia? La puntuación de la población global no responde a estas preguntas. Ahora imagina que añadiste a una persona en cada país. Has obtenido 70 observaciones adicionales.

Agregaste cada uno de estos países a la puntuación principal con un peso de 1/1,070, haciendo que los países pequeños parezcan muchas veces más grandes que su verdadera proporción de la

población. La visibilidad geográfica aumentó, pero la estimación de la población se distorsionó. Luego, un usuario en un país recibió una respuesta incorrecta, y declaraste: 'La puntuación GEO de este país es cero.' La respuesta de una persona no es la puntuación de un país. Es un solo evento observado bajo condiciones específicas en ese país. Ahora considera el idioma. El resultado en inglés es 98 por ciento, en alemán 96 por ciento y en turco 91 por ciento. Un idioma con pocos recursos tiene solo ocho usuarios y un resultado bruto de 50 por ciento. Dices: 'Ocho personas no son suficientes; elimina el idioma del informe.' La puntuación global aumenta, pero la experiencia de los usuarios de ese idioma desaparece.

En otro enfoque, se agregan 200 usuarios adicionales a un idioma con pocos recursos. La media muestral bruta disminuye. Luego dices: "La puntuación global bajó a 84." Sin embargo, la proporción real de la población que habla el idioma puede ser 2 por ciento. El sobremuestreo aumenta el poder estadístico. No aumenta el peso poblacional. La primera regla de esta sección es la siguiente:

> *Ser observado no es lo mismo que tener peso.*

Su segunda disposición establece:

> *Una sola observación de un país no es una puntuación de país.*

Su tercera disposición establece:

> *La Equidad Lingüística no consiste en dar igual peso a todos los idiomas en la puntuación global.*

Su cuarta disposición establece:

> *Un promedio global alto no elimina la degradación severa en un idioma en particular.*

Su quinta disposición establece:

> *Las observaciones adicionales tomadas para aumentar la cobertura no pueden agregarse a la puntuación principal sin ser reajustadas según la población verdadera.*

1. OBJETIVO DEL CAPÍTULO

El propósito de esta sección es establecer paneles complementarios que aumenten la visibilidad de países e idiomas sin alterar la representación estadística de la población del Panel Principal de Población. La sección normaliza las siguientes distinciones:

- Visibilidad geográfica con ponderación de la población
- Observación por país con puntuación del país
- Detección de eventos únicos con estimación de prevalencia
- Panel Principal de Población con Panel de Observadores por País
- Panel de Observadores por País con estudio de validación específico del país
- Cobertura por país con cobertura de población
- Número de idiomas cubiertos con cobertura de población por idioma
- Equidad lingüística con ponderación global igualitaria de los idiomas
- Participación global de idiomas con muestreo excesivo
- Idioma nativo e idioma de instrucción
- Idioma y localización
- Idioma y sistema de escritura
- Idioma soportado e idioma que puede generar respuestas por sí solo
- Defecto de traducción del prompt y defecto lingüístico del sistema de IA

- Escasez de recursos y defecto de representación de la IA
- Insuficiencia de jueces y fallo del producto
- Diferencia local legítima e inequidad lingüística injusta
- Promedio global y base de subgrupo
- Cobertura de acceso por país y precisión de la representación del país
- Alerta del observador y decisión definitiva de conformidad
- Ciclo de vigilancia de onda única con cobertura geográfica de múltiples ondas

Al final de esta sección, cada auditoría GEO-1000 debe poder proporcionar respuestas claras a las siguientes preguntas:

> *¿En qué países hay al menos una observación real?*
>
> *¿En qué países hay solo observación de alerta temprana, y en cuáles hay predicciones estadísticas?*
>
> *¿Qué idiomas fueron medidos, cuáles alcanzaron un muestreo suficiente y cuáles no fueron probados?*
>
> *Si se muestrearon en exceso los idiomas con pocos recursos, ¿cómo se reponderó esto en la puntuación global?*
>
> *¿Es la fuente de un resultado deficiente en un idioma realmente generado por IA, o se debe a problemas de prompt, recursos, región o adjudicación?*

2. DISPOSICIÓN NORMATIVA CENTRAL

> *GEO-1000 debe gestionar la estimación de la población principal con el Panel de Población; la observación de un país pequeño con el Panel de Observadores de País; la desigualdad lingüística y la identificación de idiomas con pocos recursos con el Panel de Equidad Lingüística por separado.*

Estas tres estructuras:

- difieren en propósitos,
- difieren en objetivos de muestra,
- difieren en ponderaciones,
- difieren en declaraciones públicas

tienen. Una observación de un panel no puede convertirse en el significado de otro panel sin explicación. La estructura predeterminada:

Panel	Propósito principal	Resultados que puede producir	Resultados que no puede producir por sí solo
Panel de población	Estimación global ponderada por población	Puntuación global y por capas de la población	Puntuación confiable por país para cada país pequeño
Panel de Observadores del País	Observación geográfica y alerta temprana	Cobertura de observación por país, evento individual, señal de acceso	Proporción de población por país
Panel de Equidad Lingüística	Diagnóstico basado en el idioma y medición de desigualdad	Puntuaciones de idiomas, piso de idioma, brecha de justicia	Asignación automática de igualdad de peso a los idiomas en la población global

3. ¿QUÉ ES EL PANEL DE OBSERVADORES DEL PAÍS?

Panel de Observadores del País, en países que están subrepresentados o no suficientemente representados en el panel de población principal:

- acceso al producto,
- Aplicabilidad del prompt
- análisis de entidad básica,
- comportamiento de lenguaje y localidad,
- señal temprana de error crítico

es un panel de auditoría suplementario creado para observar. La tarea principal del Panel de Observadores del País no es producir una puntuación.

Es hacer visible el país invisible por primera vez.

Una observación de Country Observer puede mostrar: "En el país especificado, bajo el producto de IA especificado, interfaz de usuario, idioma y condición de tiempo, se ha observado este resultado al menos una vez." No puede mostrar: "X% de los usuarios en este país ven el mismo resultado."

4. UNIDAD de Observador por País

La unidad de observación básica del Panel de Observadores del País no es solo el país. La unidad preferida:

```
S_{c,a,l,s,w}
```

donde:

- c: país o jurisdicción
- a: producto de IA y interfaz de usuario
- l: idioma, sistema de escritura y localidad
- s: condición de sesión o plan
- w: oleada de medición

Esta distinción es necesaria. En el mismo país:

- más de un idioma,
- diferentes planes,
- interfaces web y móviles,
- diferentes condiciones legales o técnicas de acceso

pueden existir. Agregar un usuario a un país no significa monitorear todas las interfaces de idioma y usuarios del país.

5. FUNCIONES DEL PANEL DE OBSERVADORES DEL PAÍS

5.1. Verificación de Acceso

Aunque el producto parece estar disponible en el país en el registro oficial, el usuario real:

- puede no poder iniciar sesión,
- puede no poder acceder a la interfaz especificada,

Puede encontrarse con un obstáculo de planificación o verificación. La prueba de observación del Observador por País verifica el registro oficial de acceso a nivel de campo.

5.2. Verificación Básica de Identidad

Entidad supervisora de productos de IA:

- la empresa correcta,
- la marca correcta,
- relación correcta con el país

¿Puede analizar internamente?

5.3. Detección de desviación local

En diferentes locales de un país en el mismo idioma:

- precio,
- empresa,
- servicio,
- ley,
- asesoramiento

¿Está cambiando? Si está cambiando, ¿puede esta diferencia ser financiera y explicable?

5.4. Señal de falla grave de baja frecuencia

En la respuesta de un único observador:

- licencia incorrecta,
- entidad legal incorrecta,
- información peligrosa para la salud o seguridad,
- fusión de identidad grave

Se puede observar. Un solo evento no indica la prevalencia en el país. Sin embargo, puede desencadenar trabajos de verificación.

5.5. Nueva hipótesis de investigación

El panel de observadores puede plantear la pregunta: '¿Es este evento singular, o es un patrón recurrente entre los usuarios del país?' Esta pregunta se prueba con una muestra adicional dirigida.

5.6. Registro del Alcance Geográfico

Muestra en cuántos países adecuados el producto tiene observación de usuarios real. Este no es el número de puntuaciones por país. Es cobertura de observación.

6. COSAS QUE EL PANEL DE OBSERVADORES DEL PAÍS NO PUEDE HACER

Solo el Panel de Observadores del País:

- puntuación confiable del país,
- tasa de población del país,
- clasificación del país,
- decisión de conformidad del país,
- superioridad de modelo basada en el país

no puede producir. Si hay un solo observador en un país y la respuesta es correcta: "La puntuación del país es 100" no se puede decir. Si es incorrecta: "La puntuación del país es cero" no se puede decir. Estado correcto:

OBSERVADO_APROBADO

Problema Observado

CANDIDATO_CRÍTICO_OBSERVADO

FALLO_DE_ACCESO

INSUFICIENTE_PARA_ESTIMACIÓN_DEL_PAÍS

son registros de observación.

7. REGISTRO DE PAÍS ELEGIBLE

Se debe crear un Registro de País Elegible para cada producto de IA. Cada país puede tener uno de los siguientes estados:

CE-0— DESCONOCIDO

No se pudo determinar el estado de acceso al producto y a la auditoría.

CE-1— FUERA DE Alcance natural

El producto permanece fuera del área oficial o legítima de uso. No se tiene en cuenta para la puntuación de precisión de representación del país. Se muestra como una brecha en el alcance de acceso.

CE-2— ACCESO RESTRINGIDO O ESPECIAL

Solo:

- institución,
- invitación,
- plan,
- dispositivo,
- tipo de usuario

el acceso está disponible. Puede ser necesario un panel separado.

CE-3— ELEGIBLE PARA la observación de Country Observer

El producto es adecuado para la observación de Country Observer en términos del interfaz de usuario relevante y al menos un idioma de auditoría.

CE-4— ELEGIBLE PARA ESTIMACIÓN DE POBLACIÓN

Tiene una participación de población elegible para el producto en el panel de población principal del país.

CE-5— ELEGIBLE PARA ESTIMACIÓN DEL PAÍS

Se puede producir una estimación del país bajo condiciones suficientes de muestra, muestra efectiva, cobertura de idiomas y calidad. Un país puede ser CE-3 pero no CE-5. Esto significa que el país está observado, pero no se pudo producir una puntuación confiable del país.

8. LA ESCALERA ENTRE la observación de Country Observer Y LA ESTIMACIÓN DEL PAÍS

CS-0— NO OBSERVADO

No hay ninguna observación válida de Country Observer para el país elegible.

CS-1— OBSERVACIÓN ÚNICA DE COUNTRY OBSERVER

Hay una observación válida país–producto–idioma. Genera solo un registro de evento.

CS-2— OBSERVACIÓN REPETIDA DE COUNTRY OBSERVER

Hay varias observaciones de diferentes oleadas o usuarios. Puede generar sospecha de un patrón. No es suficiente para la tasa del país.

CS-3— MUESTRA EXPLORATORIA DEL PAÍS

Se ha alcanzado la banda de muestra exploratoria en la sección 6. Se puede dar un resultado exploratorio con amplia incertidumbre.

CS-4— ESTIMACIÓN PROVISIONAL DEL PAÍS

Se puede producir una estimación provisional del país con un muestreo suficiente y equilibrado.

CS-5— CANDIDATO PARA ESTIMACIÓN DEL PAÍS

Muestra activa suficiente, cobertura de idioma/ubicación, olas repetidas y controles de calidad están en su lugar. Puede ser un candidato a puntuación por país. Los umbrales exactos deben calibrarse con un piloto y una revisión externa.

9. CICLO DE OBSERVADOR

Observar cada ola en todos los países elegibles puede no ser posible en términos de costo y operaciones. Por lo tanto, se puede usar el Ciclo de Observador. Olas de medición:

```
w=1,2,…,K
```

y países elegibles: sea c∈CS. Asignación de un observador para el país c en la ola w:

```
z_{c,w}= 1(si el país será observado en la ola); de lo contrario z_{c,w}= 0
```

puede representarse de la siguiente manera. Condición básica del ciclo:

```
\sum_{w=1}^K z_{c,w}\geq 1
```

debe cumplirse. Es decir, cada país elegible debería ser observado al menos una vez dentro del ciclo de vigilancia definido. Esta disposición:

- no significa que cada país será observado en cada ola,
- o que una sola observación constituya la puntuación del país

Ninguna de las condiciones se deriva del requisito del ciclo.

10. CLASES DE FRECUENCIA DE OBSERVADORES

Los países pueden dividirse en clases de frecuencia predefinidas.

SF-0— FUERA DE ALCANCE

Sin acceso al producto. No se realiza observación representativa de los usuarios. Se monitorea el estado de acceso.

SF-1— UNA VEZ POR CICLO DE VIGILANCIA

Países de bajo riesgo, países pequeños o países que ya tienen cobertura indirecta suficiente en el Panel de Población.

SF-2— OBSERVADOR PERIÓDICO

Se observa de nuevo a intervalos regulares.

SF-3— CADA OLA PRINCIPAL

Puede ser observado en cada ola principal por las siguientes razones:

- alta población de usuarios,
- mercado objetivo material,
- error grave previo,
- acceso rápido o cambio regulatorio,

área de servicio de alto riesgo.

SF-4— MONITOREO DE INCIDENTES

Observación intensificada debido a un evento crítico verificado o a una interrupción continua de la representación. La clase de frecuencia no puede cambiarse retroactivamente después de que se revisen los resultados. Un nuevo evento puede crear una nueva versión para la siguiente ola.

11. PRIORIDAD DEL OBSERVADOR

La frecuencia de la actividad del Observador por País no debe determinarse únicamente en función de la población. Según corresponda, se pueden considerar las siguientes áreas:

- Población elegible para el producto
- Cobertura real de servicio o mercado de la entidad
- Registros previos de representaciones erróneas
- Impacto elevado en usuarios de alto riesgo
- Cambios en el acceso al producto
- Unicidad de idioma y localidad
- Uso de idiomas de bajos recursos
- Diferencias legales u operativas
- Tiempo transcurrido desde la última observación

Estos campos deben ser usados de acuerdo con la regla de gobernanza predefinida. Está prohibido el enfoque de “El resultado de este país es malo, midamos menos.”

12. Métricas de cobertura de observador de país

La cobertura de país no es un solo número. Deben utilizarse al menos tres métricas separadas.

12.1. Cobertura de observación del país

La proporción de países elegibles con al menos un observador válido o una observación poblacional.

```
COC_count= |{c ∈ C_S: n_c^obs > 0}| / |C_S|
```

Esta métrica muestra la cobertura en términos del número de países.

12.2. Cobertura de observación del país ponderada por población

La proporción de la población elegible en países con al menos una observación válida respecto a la población objetivo total.

```
COC_pop= [Σ_{c:n_c^obs>0} N_c] / [Σ_{c∈C_S} N_c]
```

Esta métrica muestra si se observan grandes poblaciones.

12.3. Cobertura de Predicción por País

Es la proporción de países que alcanzan una muestra suficiente de país.

```
CEC= |{c: n_{c,eff}≥ m_c}| / |C_S|
```

Aquí, mc es el umbral de adecuación predefinido para la predicción por país. Una auditoría:

- 100 por ciento cobertura de observación de país,
- 99 por ciento cobertura de observación de población,
- puede llevar solo 10 por ciento de cobertura de predicción por país

Estos resultados no pueden convertirse entre sí.

13. ALARMA DEL OBSERVADOR

Cuando se encuentra un problema material en la observación de un vigilante, se crea una Alarma de Vigilante. Tipos de alarmas:

SA-1— ANOMALÍA DE ACCESO

El registro oficial de acceso entra en conflicto con el acceso real de los usuarios.

SA-2— ANOMALÍA DE ENTIDAD

La entidad se fusiona con otra empresa, producto u operación en otro país.

SA-3— ANOMALÍA DE IDIOMA O LOCALIDAD

Ocurre un idioma incorrecto, un contexto de país equivocado o una desviación obvia de la localidad.

SA-4— ANOMALÍA DE HECHO MATERIAL

El precio material, servicio, capacidad, asociación o tiempo es incorrecto.

SA-5— CANDIDATO A INCIDENTE CRÍTICO

Existe un error grave relacionado con la salud, la ley, las finanzas, la seguridad, la licencia o la identidad.

SA-6— ANOMALÍA DE REFERENCIA O prompt

El origen del problema no es la respuesta de la IA:

- prompts,
- traducción,
- fuente oficial,
- adjudicación

podría ser responsable. La alarma no es un fallo automático del país. Es el inicio de la verificación y la escalación.

14. PROTOCOLO DE ESCALACIÓN DEL OBSERVADOR

Paso 1— Verificación de Validez

¿La observación realmente está de acuerdo con el protocolo? País correcto Producto correcto Solicitud correcta Idioma correcto Primera salida apropiada Evidencia completa verificada.

Paso 2— Recodificación del Incidente

Problema:

- Producto de IA,
- registro de referencia,
- prompts,
- participante,
- de acceso

¿A cuál de las fuentes podría pertenecer?

Paso 3— Repetición Independiente

Antes de seleccionar el resultado, se realiza la verificación con usuarios o sesiones nuevas predefinidas.

Paso 4— Muestreo Dirigido

Si el evento continúa, se puede aplicar un muestreo específico por país. Tamaño de la muestra:

- severidad del evento,
- prevalencia esperada,
- población del país,
- diversidad lingüística

se definen previamente.

Paso 5— Separación de Decisiones

Se mantienen separados los siguientes resultados:

- Evento verificado único
- Patrón repetido
- Tasa por país
- Efecto crítico de idoneidad

Un solo evento crítico puede afectar la revisión de conformidad incluso sin una estimación de la prevalencia.

15. ¿QUÉ ES LA JUSTICIA LINGÜÍSTICA?

Justicia lingüística:

> *No se trata de que todos los idiomas reciban exactamente la misma respuesta palabra por palabra.*

Significa:

> *En preguntas de los usuarios que sean semánticamente equivalentes y materialmente comparables, aparte de las diferencias locales reales, los usuarios no deberían recibir sistemáticamente representaciones que sean más incorrectas, más incompletas, menos fundamentadas o más riesgosas debido a su idioma.*

La equidad lingüística no busca esta igualdad:

```
R_Turkish= R_English= R_Japanese
```

palabra por palabra. Busca esto:

> *Representación comparable en términos de identidad, realidad material, límite, tiempo, estado de la fuente y idoneidad para el usuario.*

16. LA DIFERENCIA DE IDIOMA NO SIEMPRE ES INJUSTICIA

Los resultados del lenguaje pueden variar legítimamente por las siguientes razones: El alcance de los servicios locales es diferente. Los precios y las condiciones del contrato varían según el país. El conocimiento legal depende de la jurisdicción. La intención del usuario es diferente. Los mensajes no son semánticamente idénticos. Las fuentes oficiales contienen diferentes contenidos según el idioma. El producto de IA no admite oficialmente el idioma relevante. Cada una de estas diferencias debe registrarse. Las diferencias locales explicables no deben etiquetarse como injusticia. Sin embargo, aunque la realidad sea la misma, en un idioma con pocos recursos:

- la identidad de la empresa está corrompida,
- se eliminan las fronteras,
- se añadió la licencia incorrecta,
- la respuesta se da constantemente en otro idioma

podría ser un problema de representación basado en el lenguaje.

17. UNIDAD DE LENGUAJE

En la evaluación del lenguaje, solo el nombre general del idioma puede no ser suficiente. La unidad de idioma preferida:

```
L=(λ,σ,ℓ,d)
```

y puede considerarse como. Aquí:

- λ: idioma
- σ: sistema de escritura
- ℓ: contexto de localidad o país

- d: tarea o dominio

Por ejemplo, el mismo idioma:

- en un sistema de escritura diferente,
- en un país diferente,
- en el ámbito legal o de atención médica

puede requerir una celda de medición aparte. No todos los usuarios de un idioma son un grupo homogéneo.

18. ESTADOS DE SOPORTE DE IDIOMA

Cada producto de IA puede usar uno de los siguientes estados para el idioma y interfaz de usuario:

LS-0— DESCONOCIDO

No se pudo determinar el estado de soporte.

LS-1— NO SOPORTADO

El producto no admite oficialmente el idioma en cuestión. No se evalúa dentro de la puntuación del idioma principal admitido. Se pueden realizar pruebas experimentales.

LS-2— OBSERVADO PERO NO SOPORTADO

El producto puede generar respuestas. Sin embargo, no hay soporte oficial ni garantía de calidad. Reportado como un resultado experimental separado.

LS-3— SOPORTE LIMITADO

Existe soporte para tareas, interfaces o funciones específicas.

LS-4— SOPORTE OFICIAL

El idioma del producto generalmente se considera soportado.

LS-5— ELEGIBLE PARA AUDITORÍA

Idioma:

- soporte oficial,
- equivalencia de prompts,
- adjudicador local,
- competencia de referencia,
- captura válida

es completamente compatible con auditorías. La mera presencia de soporte oficial no constituye LS-5.

19. LENGUA DE BAJOS RECURSOS

El término “lengua de bajos recursos” no describe el valor de la lengua ni la calidad de sus hablantes. En ciertos contextos tecnológicos y de auditoría, indica una o más de las siguientes deficiencias:

- Contenido digital limitado
- Fuente oficial localizada limitada

- Conjunto de datos de evaluación limitado
- Acceso limitado a los evaluadores
- Bajo soporte de producto
- Recurso terminológico limitado
- Calidad limitada de traducción automática
- Acceso bajo al panel de usuarios

Una lengua puede ser de bajos recursos en un campo y fuerte en otro. Por ejemplo, mientras tiene recursos sólidos en la conversación cotidiana:

- ley,
- medicina,
- campo técnico específico

puede tener recursos limitados. Por lo tanto, el nivel de recursos lingüísticos:

- producto,
- tarea,
- tiempo,
- campo

debería definirse conjuntamente.

20. TAREAS DEL PANEL DE EQUIDAD LINGÜÍSTICA

20.1. Sobremuestreo de idiomas con pocos recursos

Proporciona usuarios adicionales para idiomas que no reciben suficientes observaciones en el panel principal de población.

20.2. Medición de la distribución de errores de material según el idioma

No solo la precisión general:

- identidad,
- alcance,
- tiempo,
- fuente,
- recomendación,
- Error crítico

examina los resultados por separado.

20.3. Verificación del idioma de respuesta

¿Responde la IA de acuerdo con el idioma del mensaje? Si cambia a otro idioma, ¿cuál es la razón?

20.4. Examinando la Distinción de Lenguaje y Localidad

¿Acaso el mismo lenguaje lleva diferentes realidades locales en distintos países? ¿Gestiona la IA correctamente esta diferencia?

20.5. Examinando la paridad entre la fuente y la representación oficial

¿Los registros oficiales de existencia tienen la misma realidad material en todos los idiomas? Un resultado bajo de IA puede surgir de una brecha lingüística en la propia fuente.

20.6. Probando la Equivalencia de Instrucciones

¿Realmente las traducciones transmiten la misma intención del usuario?

20.7. Probando la Paridad del Adjudicador

¿Pueden los adjudicadores en diferentes idiomas aplicar el mismo estándar normativo?

21. CUATRO PUERTAS PARA LA EQUIDAD DEL LENGUAJE

Un resultado de lenguaje debe ser evaluado a través de cuatro puertas antes de atribuirlo al desempeño lingüístico del producto de IA.

Puerta 1— Equivalencia de Intención

¿Son los indicios semánticamente comparables? Si falla:

FALLO_DE_EQUIVALENCIA_DE_INDICIOS

se da el estado.

Puerta 2— Paridad de Referencia

¿Son suficientes los registros de realidad oficiales y verificados para el idioma correspondiente? Si falla:

BRECHA_DE_IDIOMA_DE_REFERENCIA

se da el estado.

Puerta 3— Paridad de Medición

¿Son comparables las condiciones de captura, sesión, interfaz y usuario? Si falla:

FALLO_DE_PARIDAD_DE_MEDICIÓN

se da el estado.

Puerta 4— Paridad de Adjudicación

¿Hay adjudicadores en el idioma nativo o con suficiente experiencia local y las mismas reglas de evaluación? Si falla:

BRECHA_DE_IDIOMA_DE_ADJUDICACIÓN

se asigna. A menos que se cumplan suficientemente las cuatro puertas: “El producto de IA es peor en este idioma.” no puede establecerse de manera definitiva. La brecha de lenguaje sigue siendo un hallazgo del ecosistema GEO. Sin embargo, su origen debe clasificarse correctamente.

22. ESTABLECIMIENTO DEL UNIVERSO LINGÜÍSTICO

La frase “todos los idiomas” no puede utilizarse sin limitación. Cada auditoría debe crear un Registro de Idiomas dentro del Alcance. Los idiomas pueden seleccionarse de los siguientes grupos:

22.1. Conjunto de Idiomas Globales Básicos

Idiomas nativos utilizados por una gran población humana adecuada para el producto.

22.2. Conjunto de Idiomas del Mercado de la Entidad

Los idiomas dirigidos por la entidad auditada:

- servicio,
- en términos de contenido,
- cliente,
- contrato

y mantenimiento.

22.3. Conjunto de Idiomas de Soporte del Producto

Idiomas oficialmente soportados por el producto de IA.

22.4. Conjunto de lenguajes observadores de pocos recursos

Lenguas con pocos recursos que son débiles en el panel poblacional o conllevan riesgos en términos de comportamiento del producto.

22.5. Conjunto de Lenguaje de Alto Riesgo

Lenguajes que tienen un impacto material en el usuario en el contexto de la salud, el derecho, las finanzas o la seguridad.

22.6. Conjunto de Lenguajes Experimentales

Lenguajes que el producto no soporta oficialmente pero que se prueban con fines observacionales o de investigación. Estos grupos pueden solaparse. La razón por la que se incluye cada idioma debe registrarse.

23. ASIGNACIÓN DE LA EQUIDAD LINGÜÍSTICA

Para el idioma l, sea el objetivo en el panel principal de población nlP. El objetivo adicional en el panel de equidad es nlF. La observación total del idioma recopilada:

```
n_l^T= n_l^P+ n_l^F
```

se obtiene. Se mantiene el peso total del idioma en el cálculo de la población global como Wl. El peso normalizado por observación:

```
w_{i,l}= W_l/ n_l^T
```

puede considerarse de esta manera. Así, el idioma está sobremuestreado. Sin embargo, no supera la proporción real de la población en la puntuación global.

24. CONTRIBUCIÓN DEL PANEL DE EQUIDAD LINGÜÍSTICA AL puntuación GLOBAL

Las observaciones sobre la equidad lingüística se pueden usar de dos maneras.

24.1. Análisis Diagnóstico Separado

El panel de equidad solo produce puntuaciones de lenguaje y tipos de errores. El Panel de Población principal no cambia la puntuación global. Este método es más sencillo.

24.2. Análisis Combinado Reponderado

Se utilizan juntas las observaciones de Equidad de Población y Lenguaje. Cada observación se repondera de acuerdo con el peso real de la población del idioma objetivo. Este método puede aumentar el volumen de información. Sin embargo:

- doble contabilización
- usuario multilingüe,
- dependencia del panel,
- variabilidad de los pesos

Debe ser tomado en consideración. El método utilizado debe ser definido antes de ver los resultados.

25. MÉTRICAS DE COBERTURA DE IDIOMA

25.1. Cobertura de Observación de Idioma

La proporción de idiomas en el alcance que tienen al menos una observación válida.

```
LOC= |{l: n_l^obs > 0}| / |L_scope|
```

25.2. Cobertura de Predicción de Idioma

La proporción de idiomas que cumplen con los requisitos suficientes de muestra y calidad.

```
LEC= |{l: n_{l,eff}≥ m_l}| / |L_scope|
```

25.3. Cobertura de Idioma Ponderada por Población

La proporción de la población del idioma objetivo que está adecuadamente observada. Requiere un idioma primario de auditoría o ponderación dividida por persona para prevenir el doble conteo de usuarios multilingües.

25.4. Cobertura de Soporte de Producto

Indica en cuántos de los idiomas cubiertos el producto recibe soporte oficial. El soporte del idioma y la precisión de la representación se mantienen por separado.

26. MÉTRICAS DE DESIGUALDAD DE IDIOMA

MÉTRICAS CANDIDATAS — MÉTRICAS PROPUESTAS

El conjunto exacto de métricas debe calibrarse a través de un piloto y revisión externa.

26.1. Resultado del Idioma

Para cada celda de idioma suficiente: se calcula θl.

26.2. Diferencia Más Alta-Más Baja

```
DG_max= max_l θ_l− min_l θ_l
```

Fácil de entender. Sin embargo, es sensible a celdas individuales pequeñas y volátiles.

26.3. Diferencia de Percentil Robusta

```
DG_{90-10}= Q_{0.90}(θ_l)- Q_{0.10}(θ_l)
```

Es menos sensible a celdas extremas.

26.4. Piso de idioma

Es el resultado más bajo de lenguaje con una muestra suficiente.

```
LF_min= min_{l∈L_est} θ_l
```

Este valor debe proporcionarse junto con la incertidumbre de celdas pequeñas.

26.5. Piso de Lenguaje Robusto

Es el percentil inferior o el promedio de subgrupo de lenguajes suficientes. Ejemplo:

```
LF_10= Q_{0.10}(θ_l)
```

26.6. Tasa de Error Crítico de Lenguaje

Para cada lenguaje:

```
CR_l= respuesta válida con error crítico verificado / respuesta válida del idioma
```

Se publica por separado del promedio global.

26.7. Tasa de idioma desconocido

```
U_l= respuesta DESCONOCIDA o irresoluble / observación relevante del idioma
```

Una alta tasa de desconocidos puede ser tan importante como una puntuación baja.

27. puntuación DE LENGUAJE IGUAL Y puntuación PONDERADO POR POBLACIÓN

Se pueden producir resultados diagnósticos que ponderen todos los idiomas por igual:

```
G_equal-language= (1/L)Σ_{l=1}^L θ_l
```

Este resultado aproxima la pregunta: "Experiencia promedio en el idioma evaluado." No representa la experiencia promedio humana en la población mundial. Resultado de idioma ponderado por población:

```
G_population-language= Σ_l W_l θ_l
```

Preserva la distribución de la población. Los dos puntuaciones son diferentes. No se pueden usar de manera intercambiable. Ambos se pueden mostrar juntos en la tarjeta de equidad lingüística.

28. LA EQUIDAD LINGÜÍSTICA NO ES UNA OBLIGACIÓN DE RESULTADOS IGUALES

Tener la misma puntuación en dos idiomas no es una violación automática. La diferencia puede explicarse por las siguientes razones:

- Diferencia actual del servicio local
- Ley específica de cada país
- Cobertura de soporte oficial del producto
- Acceso a recursos

- Intención del usuario
- Área de prompt

Sin embargo, si no se puede explicar la fuente de la diferencia o si los idiomas con pocos recursos experimentan sistemáticamente:

- identidad incorrecta,
- autorización incorrecta,
- deficiencia material,
- error crítico alto

si esto ocurre, hay un problema de equidad lingüística. NOMOS define la equidad lingüística de la siguiente manera:

> *En condiciones materiales equivalentes, el idioma en sí no debería ser la causa de una representación sistemática incorrecta.*

29. INTERSECCIÓN DE IDIOMA Y PAÍS

La puntuación de idioma no es independiente de los países. El mismo idioma:

- diferentes países,
- diferente acceso a productos,
- diferentes recursos locales,
- diferente legislación,
- diferente precio

puede usarse en el contexto de. Por lo tanto, la celda país–idioma: (c,l) debe examinarse por separado cuando sea necesario. Ejemplo:

- Alemania × Turco
- Turquía × Turco
- Reino Unido × Turco

Es el mismo idioma pero puede tener un contexto diferente de localidad y mercado. La equidad lingüística no puede medirse únicamente por el idioma oficial del país.

30 th Intersección del Observador por País Y del Panel de Equidad Lingüística

Una observación simultánea:

- observación pequeña del Observador por País,
- observación de idioma con pocos recursos

puede ocurrir. En este caso, el registro debe contener los siguientes campos:

ROL_PRINCIPAL_DEL_PANEL

ETIQUETAS_DIAGNÓSTICAS_SECUNDARIAS

peso de la población Estado de observador Estado de Equidad Lingüística La misma observación no puede contarse completamente en dos paneles separados. Es una sola observación. Tiene más de una etiqueta de diagnóstico.

31. CAMBIO DE PARTICIPANTES EN EL PANEL DE OBSERVADORES Y LENGUAJE

Un participante solo puede ser cambiado debido a una violación del protocolo. Los siguientes resultados no son razones para el cambio:

- Respuesta incorrecta de la IA
- Rechazo
- Respuesta en otro idioma
- Error crítico
- Confusión de entidad
- Recomendación negativa

Estos son resultados de medición. El propósito del panel de observadores es específicamente identificar eventos inesperados. Cambiar al usuario porque ocurrió un evento inesperado va en contra del propósito del panel.

32. ADJUDICACIÓN DE LENGUAJE

La adjudicación en el Panel de Equidad Lingüística debe cumplir como mínimo las siguientes condiciones:

- Ser capaz de entender el prompt y la respuesta de manera natural o a un nivel de alta competencia
- Conocimiento de la terminología local
- Ser capaz de evaluar fuentes de evidencia
- Usando la misma guía de evaluación normativa
- Ser ajeno al modelo y, si es posible, a los intereses de la marca
- Divulgar conflicto de intereses
- Ruta de un tercer adjudicador en caso de desacuerdo

Idioma donde no se pueda encontrar un adjudicador:

BRECHA_DE_IDIOMA_DE_ADJUDICACIÓN

debe tener el estatus. No se puede asumir que todos los matices son correctos si se traducen al inglés mediante traducción automática. La traducción automática puede ser una herramienta útil. No es la verificación final del significado local.

33. PARIDAD DE FUENTE

De la entidad en el sitio en inglés:

- alcance,
- precio,
- licencia,
- límites,
- evidencia

existe; si estos no están presentes en otra versión en otro idioma, el bajo resultado de IA puede no ser solo un problema del modelo. Se necesitan tres resultados separados:

- Entidad Idioma Fuente Brecha
- AI Producto Idioma Representación Brecha
- Brecha Combinada Frente al Usuario

El problema final desde la perspectiva del usuario sigue siendo real. Sin embargo, la responsabilidad de la corrección puede variar:

- entidad,
- Proveedor de IA,

ambos.

34. PARIDAD DE afirmación

Los prompts que no son semánticamente equivalentes producen respuestas diferentes. Ejemplo: En un idioma: "¿A qué se dedica esta empresa?" En otro idioma: "¿Es esta empresa confiable y recomendable?" Si se pregunta, los resultados no pueden compararse. La Constitución de Prompts se detallará en la sección siguiente. La disposición de esta sección es la siguiente:

> *Las diferencias en el rendimiento de los modelos de lenguaje no pueden atribuirse al rendimiento del lenguaje de la IA sin verificar la equivalencia de los prompts.*

35. DESVIACIÓN DEL IDIOMA DE RESPUESTA

El usuario da un prompt en turco. La IA genera una respuesta en inglés. Este resultado:

- incluso si la información factual es correcta,
- puede afectar la experiencia del usuario,
- accesibilidad
- y la comprensión.

La desviación del idioma de respuesta debe registrarse por separado:

IDIOMA_COINCIDENTE

CAMBIO_DE_CÓDIGO_PARCIAL

CAMBIO_DE_IDIOMA_NO_SOLICITADO

SALIDA_DE_IDIOMA_INUTILIZABLE

DESCONOCIDO

El cambio de código no siempre es un error. Por ejemplo, usar términos técnicos en inglés puede ser normal. Es relevante si interrumpe la usabilidad principal de la respuesta.

36. SISTEMA DE ESCRITURA Y TRANSCRIPCIÓN

El mismo idioma puede ser usado en diferentes sistemas de escritura. Para un entidad:

- nombre,
- nombre del producto,

- identidad legal

puede confundirse con otra entidad durante la transliteración. El Panel de Equidad Lingüística, en la medida en que sea relevante:

- sistema de escritura,
- transliteración,
- nombre local oficial,
- nombre usado por el usuario

debe registrar sus campos. Las diferencias del sistema de escritura sólo deben considerarse formales. Si afecta a la resolución de entidades, es material.

37. Observador de País SINTÉTICO CYCLE

PRESENTACIÓN DE METODOLOGÍA SINTÉTICA / El siguiente producto, países, números y resultados son ficticios. En el registro del producto sintético llamado Orion AI Alcance natural: hay 120 países elegibles. El Panel Principal de Población ha producido observaciones válidas en la primera ola solo de 38 países debido al ponderado de la población.

37.1. Alcance de la Primera Ola

Número de países elegibles: 120 Países con al menos una observación: 38

```
COC_count= 38/120= 31.7%
```

Permitir que estos 38 países representen 97% de la población elegible.

```
COC_pop= 97%
```

Resultado: El número de países es bajo en cobertura, la cobertura de la población es alta.

37.2. Ciclo de Observador de Cuatro Olas

Distribuya los 82 países restantes en cuatro oleadas utilizando un programa previamente congelado aleatorio y basado en riesgos. Cada país se observa al menos una vez dentro del ciclo. Al final de la cuarta oleada:

```
COC_count= 120/120= 100%
```

Sin embargo, se ha alcanzado un muestreo suficiente para la estimación del país en solo 14 países.

CEC = 14/120= 11.7%

Registro público correcto: Cobertura de observación por país: 100%/ Cobertura de observación de la población: cerca de 100%/ Cobertura de predicción por país: 11.7% Registro público incorrecto: “Todos los 120 países tienen una puntuación de país confiable.”

37.3. Alarma crítica de país pequeño

En un registro de Observador de País, la respuesta de la IA observada en el pequeño país J identifica erróneamente a la entidad auditada como un bufete de abogados con licencia. En realidad, la entidad no presta servicios legales. Una sola observación no estima la tasa de error del país J; establece un evento crítico candidato. Protocolo: validar la observación; verificar el prompt y el registro de referencia; repetir con nuevos usuarios independientes; si el evento persiste, abrir una muestra dirigida del país; y evaluar el impacto de conformidad crítica por separado. El evento observado puede ser raro; el daño potencial no tiene que ser pequeño.

38. DEMOSTRACIÓN DE EQUIDAD LINGÜÍSTICA SINTÉTICA

MUESTRA SINTÉTICA — NO REPRESENTA EL RENDIMIENTO REAL DE LA IA NI DE LA EMPRESA

Incluir cinco idiomas en la prueba de Identidad Central Sintética Apple.com. Distribución de la Población Principal del Panel:

Idioma	Peso de la población	Observación del panel principal
L1	55%	550
L2	25%	250
L3	10%	100
L4	6%	60
L5	4%	40
Total	100%	1,000

Resultados del lenguaje sintético:

Idioma	tasa de aprobación
L1	95%
L2	92%
L3	88%
L4	65%
L5	58%

Resultado global ponderado por población:

```
0.55(95)+0.25(92)+0.10(88)+0.06(65)+0.04(58)=90.27
```

La puntuación global es alta. Sin embargo, hay una degradación severa en dos idiomas.

38.1. Sobremuestreo de equidad lingüística

Agregar 140 observaciones adicionales a L4, 160 a L5. Nuevos números de observaciones:

Idioma	Total de observaciones
L1	550
L2	250
L3	100
L4	200
L5	200
Total	1,300

La media bruta de todas las observaciones de 1,300 es menor que el resultado ponderado por población mostrado arriba. No es una puntuación de población global, porque L4 y L5 se sobremuestraron intencionalmente. Una vez que se restablecen los pesos de la población objetivo, el

resultado global vuelve a la cifra ponderada por población arriba mencionada. El sobremuestreo mejora la sensibilidad para L4 y L5 sin cambiar la estimación ponderada por población.

38.2. Tarjeta de puntuación de equidad lingüística

Resultado sintético. Puntuación global ponderada por población: 90.27. Puntuación de igualdad entre idiomas:

```
(95+92+88+65+58)/5=79.6
```

Diferencia lingüística de mayor a menor: 37 puntos; idioma más bajo: 58; media de los dos idiomas más bajos: 61.5; tasa de error crítico en idiomas con menos recursos: reportada por separado; equivalencia de prompts: se asume verificada; paridad de referencia: se asume verificada. Interpretación correcta:

> *La representación de la población global parece alta; sin embargo, la equidad lingüística es seriamente débil.*

Interpretación incorrecta: "Dado que la puntuación global es 90, GEO tiene éxito en todos los idiomas."

39. CASO SINTÉTICO QUE DISTINGUE LA FUENTE DEL DETERIORO DEL IDIOMA

Consideremos tres idiomas.

Idioma A

prompt equivalente Fuente oficial completa Evaluador local suficiente Respuesta de IA incorrecta Resultado: Candidato para problema de representación de idioma en producto de IA.

Idioma B

Traducción del prompt respuesta de IA materialmente diferente apropiada al prompt Resultado: Defecto de la Constitución del prompt.

Lenguaje C

En la página del idioma oficial, el antiguo precio de la IA se mantiene sobre el antiguo precio de la fuente en inglés es actual Resultado: El problema de paridad de la fuente de idioma de la entidad y la brecha de representación combinada alcanzando al usuario. Incluso si las puntuaciones de los tres idiomas parecen iguales, el propietario de la corrección es diferente.

40. PANELES DE PAÍS Y LENGUAJE TARJETA DE RESULTADO PÚBLICA

La tarjeta de resultados públicos del candidato debe incluir los siguientes campos:

Campos del país

Alcance natural número apropiado de países Número de países con al menos una observación Cobertura de observación por país Cobertura de observación ponderada por población Cobertura de predicción por país Tiempo del ciclo del Observador por País Abrir alertas del observador PAÍSES FUERA DEL ALCANCE NATIVO PAÍSES NO OBSERVADOS Fecha de medición

Campos de idioma

Número de idiomas cubiertos Número observado de idiomas Número de idiomas para los que se pueden hacer predicciones Idiomas oficialmente soportados Idiomas experimentales Resultado global de idiomas ponderado por población Resultado con reconocimiento igual de idiomas Piso de idioma Diferencia de idioma Tasas de error críticas

BRECHA_DE_IDIOMA_DE_REFERENCIA

FALLO_DE_EQUIVALENCIA_DE_INDICIOS

Idiomas NO PROBADOS Versión del panel de idiomas Esta tarjeta hace visible la realidad que un solo distintivo "multilingüe" no puede transmitir.

41. DISPOSICIONES NORMATIVAS OBLIGATORIAS

CH07-N01

Los paneles del Observador por País y de Equidad Lingüística deben definirse por separado del Panel de Población principal en términos de propósito, muestra y ponderación.

CH07-N02

Una observación de un Observador por País no puede usarse como puntuación de la población del país.

CH07-N03

Para un país, tener una observación y poder producir una estimación del país son estados separados.

CH07-N04

El número de países cubiertos y la cobertura de los países ponderada por población deben ser reportados juntos.

CH07-N05

La cobertura de la estimación del país debe mostrarse por separado de la cobertura de observación.

CH07-N06

Cada país elegible debe tener una probabilidad de ser observado mayor que cero en el ciclo de vigilancia definido; también se deben mostrar los países que están fuera del alcance.

CH07-N07

El ciclo de vigilancia y las clases de frecuencia por país deben ser versionados antes de que se vean las respuestas de IA.

CH07-N08

Un país con puntuación baja no puede ser eliminado del ciclo de Observador después de ver los resultados.

CH07-N09

Una alerta de Observador de País no es un fallo automático del país; provoca verificación y escalamiento.

CH07-N10

Un solo evento de Observador Crítico también puede iniciar una revisión de conformidad separada sin estimación de prevalencia.

CH07-N11

Las observaciones del Panel de Observadores del País no pueden usarse en el Panel de Población principal con ponderación igual sin procesar.

CH07-N12

Si se incluyen las observaciones del Observador por País en la puntuación principal, sus probabilidades de selección y ponderaciones deben ser modeladas explícitamente.

CH07-N13

La equidad lingüística no puede definirse como la obligación de dar la misma respuesta palabra por palabra.

CH07-N14

La equidad lingüística requiere precisión comparable, límites, fuente y adecuación del usuario bajo condiciones materiales equivalentes.

CH07-N15

Idioma y país, idioma y configuración regional, e idioma y sistema de escritura no pueden usarse de manera intercambiable.

CH07-N16

El universo lingüístico cubierto y la razón para incluir cada idioma deben explicarse antes de la medición.

CH07-N17

La afirmación de 'todos los idiomas' solo puede utilizarse si de hecho se han medido todas las interfaces relevantes de idioma y usuario; de lo contrario, se debe decir 'idiomas dentro del alcance'.

CH07-N18

Las observaciones en un idioma que no es oficialmente admitido no pueden añadirse al desempeño principal del producto admitido sin explicación.

CH07-N19

El muestreo excesivo de idiomas con pocos recursos no debe aumentar el peso poblacional global.

CH07-N20

Si las observaciones del Panel de Equidad Lingüística se incluyen en el resultado global, deben devolver el peso a las poblaciones de idiomas objetivo reales.

CH07-N21

El mismo usuario multilingüe no puede contarse como más de un humano completo en el peso de la población global.

CH07-N22

Sin verificar la equivalencia del prompt, la diferencia en la puntuación de lenguaje no puede atribuirse de manera definitiva al rendimiento del lenguaje de la IA.

CH07-N23

Una puntuación baja de lenguaje no puede asignarse únicamente al producto de IA sin evaluar la paridad de referencia.

CH07-N24

Una decisión sobre la precisión del lenguaje no puede finalizarse sin la competencia suficiente del adjudicador local.

CH07-N25

La desviación del lenguaje en la respuesta y el cambio de código deben registrarse por separado.

CH07-N26

Las diferencias legítimas de la realidad local no pueden clasificarse como injusticia lingüística.

CH07-N27

Las diferencias materiales de lenguaje no explicadas no pueden cerrarse únicamente como diferencias de personalización o estilo.

CH07-N28

La puntuación de la población global no puede reemplazar el resultado de la equidad del lenguaje.

CH07-N29

La puntuación de lenguaje igual no puede presentarse como una puntuación global ponderada por la población.

CH07-N30

Las tasas de error crítico basadas en el idioma deben mostrarse por separado del promedio general.

CH07-N31

PROMPT_EQUIVALENCE_FAILURE, REFERENCE_LANGUAGE_GAP, MEASUREMENT_PARITY_FAILURE y ADJUDICATION_LANGUAGE_GAP deben mantenerse como estados separados.

CH07-N32

En los paneles de Observador por País y Equidad Lingüística, la decisión de cambiar a los participantes no puede basarse en la corrección de la respuesta de la IA ni en beneficio de la marca.

CH07-N33

La misma observación puede tener múltiples etiquetas de diagnóstico; sin embargo, no puede recibir más de un peso completo en la predicción global sin explicación.

CH07-N34

Los registros de los paneles de país e idioma deben tener un propietario humano o institucional responsable.

42. FORMAS DE FALLA

CH07-F01— puntuación DE PAÍS DE UNA SÓLA PERSONA

La observación de un solo observador se publica como La puntuación GEO para el país.

CH07-F02— COBERTURA SIMBÓLICA DEL PAÍS

Se asigna una persona a cada país y se llama 'proporcional a la población mundial.'

CH07-F03— CONFUNDIR LA COBERTURA DE LA POBLACIÓN CON EL NÚMERO DE PAÍSES

Aunque se observa un pequeño número de países grandes, solo se muestra la cobertura de la población; aunque se observa un gran número de países pequeños, solo se muestra el número de países.

CH07-F04— CONTAR EL ALCANCE DE LA OBSERVACIÓN COMO ALCANCE DE ESTIMACIÓN

Todos los países con al menos una observación se consideran países puntuados.

CH07-F05— CONTAR LA OBSERVACIÓN DE UN OBSERVADOR DE UN PAÍS DOS VECES EN LA PUNTUACIÓN PRINCIPAL

La misma observación recibe peso completo tanto en el Panel de Observador como en el Panel de Población.

CH07-F06— CONVERTIR UNA ALERTA DE UN OBSERVADOR DE PAÍS EN UNA AFIRMACIÓN DE PREVALENCIA

Un solo evento se generaliza a todos los usuarios del país.

CH07-F07— IGNORANDO UN EVENTO DE OBSERVADOR CRÍTICO

El alto riesgo de daño no se examina diciendo "solo un usuario".

CH07-F08— ELIMINANDO UN PAÍS CON BAJA PUNTUACIÓN DEL CICLO

El horario del observador se cambia de acuerdo con los resultados.

CH07-F09— OBSERVANDO SOLO LOS MERCADOS COMERCIALES

Los países que tienen acceso a productos pero se consideran comercialmente poco importantes están completamente excluidos; El resultado se presenta a nivel global.

CH07-F10— CONSIDERANDO EL PAÍS Y EL IDIOMA COMO EL MISMO

El idioma oficial del país se acepta como el idioma de todos los usuarios.

CH07-F11— afirmación DE TODOS LOS IDIOMAS

Se miden solo unos pocos idiomas limitados; El resultado se publica como "en todos los idiomas".

CH07-F12— BORRANDO LENGUA DE BAJA SOLDADURA

Debido a que la muestra es pequeña, el idioma se excluye completamente del informe.

CH07-F13— SOBRECRESIMIENTO DEL IDIOMA CON POCOS RECURSOS A NIVEL GLOBAL

El idioma con sobrerrepresentación entra en la puntuación global con su participación de muestra sin procesar.

CH07-F14— CONTANDO LA PUNTUACIÓN IGUAL PARA TODOS LOS IDIOMAS COMO LA PUNTUACIÓN DE LA POBLACIÓN

La puntuación de diagnóstico que otorga igual peso a cada idioma se presenta como la experiencia promedio de la población mundial.

CH07-F15— CONTANDO LA PUNTUACIÓN DE LA POBLACIÓN COMO JUSTICIA EN EL IDIOMA

El resultado global, impulsado por el peso de los idiomas grandes, oculta la distorsión en los idiomas pequeños.

CH07-F16— CONTANDO EL ERROR DE TRADUCCIÓN DEL PROMPT COMO ERROR DE IA

La diferencia que surge de prompts semánticamente diferentes se atribuye al modelo.

CH07-F17— CONTANDO LA BRECHA DEL IDIOMA FUENTE COMO ERROR DE IA

Aunque no exista información oficial localizada, toda la responsabilidad recae en el producto de IA.

CH07-F18— DEFENDIENDO EL ERROR DE IA CON FALTA DE RECURSOS

Aunque los mismos y suficientes recursos están disponibles, los errores en el idioma con pocos recursos se cierran únicamente por contenido faltante.

CH07-F19— CONSIDERANDO LA INADECUACIÓN DEL IDIOMA DEL JUEZ COMO FALLO DEL PRODUCTO

Incluso si el juez malinterpretó, el resultado se finaliza.

CH07-F20— CONSIDERANDO LA TRADUCCIÓN AUTOMÁTICA COMO JUEZ FINAL

No se verifica la pérdida de significado local.

CH07-F21— CONSIDERANDO LOS CAMBIOS DE CÓDIGO COMO ERROR AUTOMÁTICO

Se producen fallos con términos técnicos naturales o mezcla de idiomas aceptable.

CH07-F22— CONSIDERANDO LA DESVIACIÓN DE IDIOMA INUTILIZABLE COMO ESTILO

Incluso si el sistema responde en otro idioma que el usuario no puede entender, el resultado se considera exitoso.

CH07-F23— IGNORANDO LA DIFERENCIA DEL SISTEMA DE ESCRITURA

Se considera formal un error de transliteración que cause confusión de entidades.

CH07-F24— CONTAR A UN USUARIO MULTILINGÜE DOS VECES

La misma persona se cuenta como una unidad poblacional separada en todos los idiomas.

CH07-F25— CONSIDERAR COMO INJUSTA UNA DIFERENCIA LOCAL LEGÍTIMA

Incluso si la diferencia de precio o legal se basa en condiciones locales reales, se declara como un sesgo de idioma.

CH07-F26— CONSIDERAR PERSONAL UNA DIFERENCIA INEXPLICABLE

La inconsistencia material bajo condiciones equivalentes no se cuestiona.

CH07-F27— PUNTUACIÓN DEFINITIVA DE UNA CELDA DE IDIOMA INSUFICIENTE

La clasificación por idioma se realiza con pocas observaciones.

CH07-F28— CONFIGURACIÓN DEL PANEL DE IDIOMA DESPUÉS DEL RESULTADO

Solo se seleccionan después los idiomas que obtuvieron puntuaciones bajas o altas.

CH07-F29— MEDICIÓN SOLO DE IDIOMAS DE ALTA FUENTE

El resultado se presenta como "multilingüe".

CH07-F30— PRESERVANDO LOS IDIOMAS NO PROBADOS

El campo NO PROBADO no aparece en la tarjeta de resultados pública.

43. PROCEDIMIENTO DE AUDITORÍA

Paso 1— Congelar Registro de Cumplimiento Producto-País

Para cada país:

- Alcance natural
- acceso al producto
- auditoría de idioma

- interfaz de usuario

se registra el estado.

Paso 2— Extraer Cobertura del País en el Panel de Población Principal

Se determina en qué países existe la observación del panel principal.

Paso 3— Definir el Universo de Países Observadores

Se identifican los países que son insuficientes en el panel principal y tienen estado CE-3.

Paso 4— Establecer el Ciclo de Observador

Países:

- se asignan en términos de clase de frecuencia,
- onda,
- idioma/ubicación,
- interfaz de usuario

y mantenimiento.

Paso 5— Preseleccionar Participantes Observadores

Se crean listas principales y de respaldo antes de ver los resultados.

Paso 6— Clasificar Resultados de Observador

Se asignan accesos, identidad, idioma, hecho material, estados de Evento Crítico.

Paso 7— Aplicar Alarma y Escalamiento

Por tipo de alarma:

- de nuevo,
- muestra dirigida
- revisión de referencia,
- revisión de conformidad

se inicia.

Paso 8— Calcular Métricas de Cobertura por País

El alcance del número de países, el alcance de la población y el alcance de la estimación por país se calculan por separado.

Paso 9— Congelar Registro de Idioma

Para cada lenguaje:

- por razones de alcance,
- estado de soporte,
- sistema de escritura,
- localización,
- universo del usuario

que se registra.

Paso 10— Establecer Sobremuestreo para la Equidad Lingüística

Se establecen objetivos adicionales para los idiomas que son insuficientes en el panel principal de población.

Paso 11— Verificar las Cuatro Puertas del Idioma

Se evalúan equivalencia del sistema, paridad de referencia, paridad de medición y paridad de adjudicación.

Paso 12— Calificar Observaciones de Idioma

Para cada lenguaje:

- veracidad,
- deficiencia material,
- Error crítico,
- desconocido,
- desviación del idioma de respuesta

que se registra.

Paso 13— Separar Pesos Globales y Diagnósticos

El resultado ponderado por población y el resultado con reconocimiento de idioma igual se calculan por separado.

Paso 14— Calcular la Desigualdad del Idioma

suelo del idioma, diferencia de idioma sólida, diferencia de error crítico, se calcula la tasa desconocida.

Paso 15— Examinar las Intersecciones País–Idioma

Se identifican diferencias o conflictos de localización del material.

Paso 16— Crear Tarjeta de Resultados Pública

Se hacen visibles el alcance, la suficiencia de predicción, las alarmas y las áreas no probadas.

44. EVIDENCIA NECESARIA

Registro de productos de IA y usuarios; registro del país Alcance natural; número de países elegibles; estado de acceso por país; asignaciones de países principales del Panel de Población; universo del Observador por país; ciclo del Observador; clases de frecuencia del Observador; matriz de asignación país–ola; asignaciones país–idioma; listas principales y de reserva de Observadores; resultados de observación por país; alertas del Observador; registros de escalamiento; muestras de países objetivo; cobertura de observación por país; cobertura de países ponderada por población; cobertura de estimación por país; registro de idiomas dentro del ámbito; estado de soporte de idioma; registros de script y configuración regional; recuentos de muestras principales por idioma; recuentos de sobremuestreo de Equidad Lingüística; ponderaciones de población por idioma; registros de usuarios multilingües; registros de equivalencia de prompts; registros de paridad de referencia; registros de paridad de medición; registros de calificación de adjudicadores locales; y resultados por idioma.

Idioma Tasas de error críticas Idioma de respuesta desviaciones Métricas de desigualdad lingüística Idiomas no probados Resultados iguales ponderados por idioma y población Pesos del panel

Verificaciones de doble conteo Resultados públicos Panel de resultados Versiones del panel Persona o institución responsable

45. LISTA DE VERIFICACIÓN DE AUDITORÍA

¿Se ha separado el propósito del Panel de Observadores del País del panel principal de población? ¿Qué países están incluidos en el Observador? ¿Está actualizada la elegibilidad y el estado de acceso de los países? ¿Existe una probabilidad de observación positiva en el ciclo de vigilancia de cada país elegible? ¿Se congeló el ciclo de vigilancia antes de los resultados? ¿Se ha utilizado la observación del Observador por País como una puntuación del país? ¿Se han separado la cobertura de observación del país y la cobertura de estimación del país? ¿Se ha mostrado la cobertura del país ponderada por la población por separado? ¿Se han contado doble las observaciones del Observador por País en La puntuación principal? ¿Se ha confirmado la alerta del Observador Crítico? ¿Se ha convertido una sola alarma en una afirmación de prevalencia? ¿Se abrió el muestreo dirigido debido a la alarma? ¿Se definió previamente el universo de idiomas dentro del alcance? ¿Es coherente la afirmación de 'todos los idiomas' con la cobertura real? ¿Se han separado el idioma, el sistema de escritura y la configuración regional?

¿Se ha registrado el estado de soporte de idioma del producto? ¿Por qué se eligieron idiomas con pocos recursos? ¿Se reponderaron ejemplos de equidad lingüística en la puntuación global? ¿Se contó doble al mismo usuario multilingüe? ¿Se ha verificado la equivalencia de los prompts? ¿Son materialmente iguales las fuentes oficiales en todos los idiomas? ¿Los evaluadores conocen el idioma relevante a un nivel suficiente? ¿Se ha utilizado la traducción automática como juez final? ¿Se han registrado desviaciones en el idioma de las respuestas? ¿Se ha separado la variación local legítima de la injusticia lingüística? ¿Se ha publicado el corpus lingüístico y la variación de idiomas? ¿Son visibles las tasas de errores críticos basadas en el idioma? ¿Se han separado las brechas de REFERENCE_LANGUAGE_GAP y otros errores de medición? ¿Son visibles los idiomas no evaluados? ¿Oculta la puntuación pública el problema de imparcialidad del idioma?

¿Ha recibido la misma observación el peso completo en múltiples paneles? ¿Se ha identificado al responsable de los registros del panel?

46. OBJECIONES Y RESPUESTAS

Objeción 1— "¿No es todavía mejor dar una persona a cada país pequeño que no dar nada?"

Para la observación temprana y cobertura geográfica, sí. Pero esta persona no mide la proporción del país. El término correcto:

Observación del Observador de País

debe ser.

Objeción 2— "Si una persona no produce una puntuación del país, ¿por qué la enviamos?"

Porque una sola observación puede:

- revelar problemas de acceso,
- aclarar confusión de entidades,
- identificar el idioma equivocado,

- y registrar eventos críticos

visible por primera vez. El panel del Observador es un sistema de alerta temprana.

Objeción 3— "¿Tenemos que observar todos los países en cada oleada?"

No. Se puede utilizar un ciclo de vigilancia definido. Los países de alto riesgo o significativos pueden ser monitoreados con mayor frecuencia; los países pequeños y de bajo riesgo pueden ser observados de forma alterna. El calendario debe determinarse antes de los resultados.

Objeción 4— "Si los países pequeños reciben un peso bajo en la puntuación global, ¿por qué les estamos dando tanta atención?"

Porque su participación poblacional puede ser baja. Sin embargo:

- ley,
- idioma,
- daño local a los usuarios,
- integridad geográfica

son comunidades humanas reales. Un peso bajo no debería significar invisibilidad.

Objeción 5— "¿No es ponderar todos los idiomas por igual el método más justo?"

La ponderación equitativa de los idiomas puede ser una perspectiva diagnóstica valiosa. No representa la experiencia humana promedio en la población mundial. La puntuación de la población y la puntuación de equidad lingüística deben publicarse por separado.

Objeción 6— "Si un idioma pequeño solo afecta ligeramente el resultado global, ¿por qué importa una puntuación baja?"

Porque el problema de la representación es real para las personas que usan ese idioma. Además, los errores graves en licencias, salud o identidad no se trivializan automáticamente por la proporción de la población.

Objeción 7— "¿No es injusto culpar a la IA si un idioma con pocos recursos tiene pocos recursos?"

Puede que solo sea injusto culpar a la IA. Por lo tanto:

- paridad de recursos,
- equivalencia de prompts,
- competencia del evaluador

examinados por separado. El error que llega al usuario sigue siendo factualmente correcto. El propietario de la corrección debe determinarse correctamente.

Objeción 8— "¿No es normal tener diferentes respuestas en el mismo idioma en distintos países?"

Puede ser normal si las condiciones locales reales son diferentes. Si un mismo hecho material cambia de manera distinta y sin explicación, puede ser un problema. El idioma y la localidad deben examinarse juntos.

Objeción 9— "Si la IA da la respuesta correcta en otro idioma, ¿por qué fallaría?"

Si el usuario no puede entender la respuesta, no se puede utilizar la representación. La desviación del idioma debe evaluarse por separado junto con la precisión.

Objeción 10— "Encontrar usuarios 200 para puntuaciones de idioma puede ser muy costoso."

Se puede usar una muestra más pequeña. Resultado:

- descubrimiento,
- temporal,
- muestra insuficiente

debe publicarse con este estado. El costo no otorga falsa certeza.

Objeción 11— "Si la puntuación global ya es alta, ¿no asusta innecesariamente al cliente una tarjeta de idioma separada?"

Ocultar el deterioro del material lingüístico no protege al cliente. Produce una falsa confianza. Se puede utilizar una tarjeta de resultados corta y clara.

Objeción 12— "¿Un único evento de Observador Crítico reduce toda la elegibilidad?"

La respuesta automática depende del tipo de evento. Un solo evento no indica prevalencia nacional. Sin embargo, si implica un daño severo o autoridad falsa, puede iniciar una revisión Crítica y medidas temporales. La decisión se definirá en la Sección 15.

REGLA COMÚN DE LA SECCIÓN 49

Crear una puntuación mundial es fácil. Llenas la población mundial con países grandes. Pierdes países pequeños en fracciones. Mides lenguajes de alto recurso. Lenguajes de bajo recurso: los eliminas diciendo "Muestra insuficiente." Luego publicas un solo número. Este número puede ser matemáticamente calculado correctamente. Pero puede que no muestre todos los defectos de representación del mundo. NOMOS no da el mismo peso a cada país. Porque las proporciones de población de las personas no son iguales. NOMOS tampoco ignora a los países pequeños. Porque tener baja población no significa inexistencia. NOMOS no pondera todos los idiomas por igual en la puntuación global. Porque esto podría distorsionar la experiencia promedio de la población mundial. NOMOS tampoco hace que los idiomas pequeños sean invisibles. Porque incluso si su peso global es bajo, el error de representación para las personas que usan ese idioma puede ser 100% real.

Panel de Observadores del País:

> *Muestra dónde nunca hemos mirado.*

Panel de Equidad Lingüística:

> *Muestra quién está por debajo del promedio global.*

Panel de Población:

> *Estima la experiencia promedio en la población mundial.*

Cuando estos tres trabajan juntos, la puntuación global se vuelve tanto fiel a la población como auditable en términos de justicia. Un solo usuario en un país no es todo el país. Pero ese usuario es

una persona real. Un idioma puede ser una pequeña parte de la población mundial. Pero el error en ese idioma no es menos incorrecto. Un evento crítico puede ser raro. Pero el daño no se mide únicamente por la frecuencia. Por lo tanto, la séptima ley de medición de NOMOS es:

> *El derecho a la visibilidad no es lo mismo que el peso de la población.*

Su octava ley es la siguiente:

> *Observar un país no es lo mismo que estimar un país.*

La novena ley es la siguiente:

> *La equidad lingüística no consiste en hacer que todos los idiomas sean iguales; consiste en garantizar que nadie esté sistemáticamente alejado de la realidad material debido a su idioma.*

La décima ley es:

> *Si el promedio global hace que el idioma y el país que están rezagados sean invisibles, por sí solo no es un resultado GEO suficiente.*

Sección NOMOS Orden 7

> *No ocultes los países que nunca observas mientras me muestras la puntuación mundial.*
>
> *No envíes a un usuario a un país pequeño y hagas que esa persona represente a todo el país.*
>
> *Pero tampoco ignores un país pequeño para siempre solo porque su población sea pequeña.*
>
> *Separa la observación del país, la predicción del país y la idoneidad del país.*
>
> *No juzgues a todo un país por un solo error grave. / Pero no ignores el daño humano diciendo "solo una respuesta".*
>
> *No bases el éxito en idiomas principales en el silencio de los idiomas menores.*
>
> *Puedes muestrear mucho un idioma de pocos recursos. / Pero no hagas que parezca más grande de lo que es en la población mundial.*
>
> *No conviertas la puntuación de igualdad de lenguaje en una puntuación poblacional; no conviertas la puntuación poblacional en una puntuación de equidad lingüística.*
>
> *No me culpes si la traducción está equivocada. / Si la fuente está incompleta, también dilo. / Si el juez no entiende el idioma, no emitas un juicio definitivo.*
>
> *No clasifiques una diferencia local legítima como injusticia, ni descartes una disparidad material no explicada como mero estilo.*
>
> *Primero, observa los países. / Luego indica dónde puedes respaldar una estimación. / Luego sobremuestra los idiomas que requieren visibilidad diagnóstica. / Luego devuélvelos a sus verdaderos pesos poblacionales. / Mantén visibles a las personas que el promedio global deja atrás.*

La oración de cierre del capítulo

Una medición GEO realmente global debe mostrar más de lo que la mayoría del mundo ve. También debe mostrar quién permanece no observado en países e idiomas más pequeños, quién está mal representado y qué resultado es solo una alerta única.

Core normativo

El Panel de Población GEO-1000, el Panel de Observadores de País y el Panel de Equidad Lingüística DEBEN permanecer distintos en propósito, muestreo, ponderación y declaraciones públicas. Una observación del Observador de País PUEDE establecer que una condición de acceso, respuesta, anomalía o incidente crítico ocurrió en un contexto definido de país–producto–idioma. Una observación del Observador de País NO DEBE representarse como una estimación de prevalencia por país o como una puntuación GEO por país. La cobertura de observación por país, la cobertura de observación ponderada por población y la cobertura de estimación por país DEBEN informarse por separado. Cada país elegible DEBERÍA recibir una probabilidad positiva de observación dentro de un ciclo de vigilancia predeclarado, a menos que esté fuera del Alcance natural legal del producto. La equidad lingüística DEBE evaluar la representación materialmente comparable bajo prompts semánticamente equivalentes y contextos comparables. NO DEBE requerir respuestas idénticas palabra por palabra ni ignorar diferencias locales legítimas. Los idiomas muestreados en exceso DEBEN volver a sus pesos de población objetivo para estimaciones de la población global. Las puntuaciones ponderadas por población, los diagnósticos de idiomas iguales, los mínimos de idiomas, las tasas de error crítico, las tasas desconocidas y los idiomas no probados DEBEN permanecer visibles por separado. No se DEBE atribuir una disparidad lingüística únicamente al producto de IA hasta que se hayan evaluado la equivalencia del prompt, la paridad de referencias, la paridad de medición y la paridad de adjudicación. Una observación PUEDE llevar múltiples etiquetas de panel de diagnóstico, pero NO DEBE recibir un peso duplicado de la población total. Los incidentes del Observador Crítico PUEDEN desencadenar muestreo confirmatorio y revisión de conformidad sin ser tratados como evidencia de la prevalencia en el país.

CAPÍTULO 08

Selección y ponderación de participantes y sesgo de muestreo

BASES METODOLÓGICAS / K04; K05; K06; K07; K08; K22

Límite de Capítulo

El capítulo 5 definió la población objetivo de usuarios. El capítulo 6 transformó esta población en un diseño de muestra distribuido entre países, idiomas y interfaces de usuario. El capítulo 7 estableció sistemas de observación complementarios para países pequeños y lenguas con pocos recursos que el panel principal de población podría dejar invisibles. Ahora los números en el plan de muestra necesitan convertirse en personas reales:

> *¿Dónde se encontrarán estos participantes, cómo serán seleccionados, bajo qué condiciones serán admitidos en el panel y con qué sesgos entrarán en la puntuación GEO?*

¿De un panel en línea de una organización de investigación? ¿De una convocatoria abierta en redes sociales? ¿De universidades? ¿De la propia lista de usuarios del cliente? ¿De socios de investigación locales? ¿Solo de personas que usan productos de IA con frecuencia? ¿De aquellos que acceden al producto por primera vez? ¿De los clientes de la marca? ¿De empleados de competidores? ¿De suscriptores de planes pagos? ¿Cómo se verificará que una persona realmente vive en ese país? ¿Cuántas veces se contará a la misma persona si usa múltiples cuentas? ¿Cuánto se le pagará al participante? ¿El pago incentivará completar la tarea rápidamente o enviar un resultado específico? ¿Se preservará la observación si el participante carga una respuesta incorrecta de IA? ¿O será reemplazado por un nuevo usuario diciendo: 'La tarea se realizó incorrectamente.'? Se pueden haber recopilado mil capturas de pantalla válidas. Pero los participantes:

- 60% son usuarios intensivos de IA.
- 20% son estudiantes de tecnología,
- 10% son clientes de la marca auditada,
- 5% son empleados de la empresa,
- el resto son usuarios generales

¿Qué población, entonces, representará el resultado? ¿Puede el ponderado demográfico corregir completamente la diferencia? Igualar los totales de edad, género, país y idioma no necesariamente corrige las diferencias no medidas en interés tecnológico, conocimiento de la marca, acceso a planes de pago o patrones de uso de la IA. El ponderado es valioso, pero no puede borrar todos los defectos en la selección de participantes. Este capítulo aplica el requisito anterior: que cada control de auditoría pueda ser probado por modelo, fecha, país, idioma, conjunto de consultas, número de repeticiones y registro de medición, a la selección de los participantes reales. También preserva la demanda central de NOMOS: evidencia, límite, contexto y tiempo. Este capítulo:

- fuentes de selección de participantes

- la distinción entre paneles probabilísticos y no probabilísticos,
- verificación de elegibilidad e identidad,
- control de acceso por país, idioma y producto,
- unicidad del participante,
- riesgo de recompensa y redirección,
- condicionamiento del panel,
- asignación de tareas,
- no respuesta y deserción del panel,
- participación fraudulenta o automatizada,
- sesgo de selección y cobertura,
- el poder y límite del ponderado,
- estados de calidad de la muestra

define. Este capítulo aún no:

- todas las áreas técnicas del registro del producto de IA,
- el sistema completo de traducción y versión de los prompts,
- La arquitectura detallada de la infraestructura Capture NOMOS,
- adjudicación de afirmaciones atómicos,
- la fórmula de puntuación final

no está completamente definida. La pregunta clave de la Sección 8 es:

> *Cuando recopilamos mil registros válidos, ¿cómo demostramos a qué personas reales pertenecen estos registros y qué población humana pueden representar?*

NOMOS Desafío

Compras 1,000 personas de un panel de investigación en línea. El proveedor del panel te dice: "1,000 participantes listos según tus cuotas de país y edad." Los participantes completan la tarea. Suben capturas de pantalla. Resultados sin procesar:

> *90 por ciento de representación precisa.*

En tu informe, escribes: "90% de usuarios en todo el mundo están representados con precisión." Sin embargo, la distribución real de participantes es la siguiente:

Frecuencia de uso de IA	Participación de la población objetivo	Participación en el panel
Uso bajo o poco frecuente	50%	10%
Uso medio	30%	30%
Uso intensivo	20%	60%

Usuarios intensivos de IA:

- mejores hábitos de prompt,
- más cuentas de pago,
- nuevas interfaces de productos,
- conocimiento de personalización,
- mayor familiaridad con la tecnología

puede tener. Incluso si se usa el mismo prompt fijado:

- plan de cuenta,
- interfaz,
- uso previo,
- acceso al producto

puede ser diferente. Deja que las tasas de aprobación del grupo sintético sean las siguientes:

Frecuencia de uso de IA	tasa de aprobación
Uso bajo o poco frecuente	72%
Uso medio	84%
Uso intensivo	96%

El resultado bruto del panel:

```
0.10(72)+0.30(84)+0.60(96)=90
```

es decir:

90 por ciento

Con los pesos de la población objetivo, sin embargo, el resultado es:

```
0.50(72)+0.30(84)+0.20(96)=80.4
```

Por lo tanto, las mismas respuestas 1,000 producen dos resultados diferentes:

- Resultado bruto del panel: 90 por ciento
- Resultado de la población objetivo: 80.4 por ciento

Ahora imagina que lo has ponderado. Cada una de las 100 personas en el grupo de bajo uso recibe un peso alto. Cada una de las 600 personas en el grupo de alto uso recibe un peso bajo. El tamaño de muestra bruto sigue siendo 1,000. Sin embargo, el tamaño de muestra efectivo disminuye a aproximadamente 349. Por lo tanto, es posible que hayas recopilado 1,000 archivos. Estadísticamente, puede que no tengas el volumen de información de 1,000 personas iguales e independientes. Ahora una pregunta más difícil: En el grupo de bajo uso, personas que nunca han sido incluidas en el panel:

- bajo nivel de conexión,
- usando tecnología de asistencia,
- viviendo en una ciudad pequeña,
- no teniendo un dispositivo específico

Si hay personas, ¿puede el ponderado hacer que regresen? No. Una observación de una persona que no está en el marco de muestreo en absoluto no puede ser producida mediante ponderación. Ahora la empresa auditada envía conexiones a sus propios clientes: 'Por favor participe en la encuesta.' Los clientes leales de la empresa participan en alta proporción. Las personas que conocen a los competidores pero no usan la empresa participan menos. Los pesos de edad y país pueden ser correctos. El sesgo por lealtad a la marca puede continuar. Ahora otro problema: Se le dice al participante: 'Ganarás un bono si subes una respuesta de IA correcta y de alta calidad.' El participante revisa la respuesta. Elige el mejor resultado. Recorta la captura de pantalla. Esta ya no es la respuesta natural del producto de IA. Es la respuesta seleccionada creada por el sistema de recompensas. La primera regla de esta sección es:

> *El número de participantes no es el poder representativo de los participantes.*

Su segunda disposición establece:

> *La asignación aleatoria no es selección aleatoria de la población.*

Su tercera disposición establece:

> *El ponderado puede reducir el desequilibrio observado; no puede crear una persona no observada.*

Su cuarta disposición establece:

> *La inclusión de un participante en el panel, la retención en la tarea o el reemplazo no pueden depender del contenido de la respuesta de la IA.*

Su quinta disposición establece:

> *La recompensa debe otorgarse por completar el protocolo de manera honesta, no por elegir una respuesta positiva o deseada.*

1. OBJETIVO DEL CAPÍTULO

El propósito de esta sección es hacer visible, imparcial y auditable toda la cadena de selección desde la población objetivo hasta la observación real del usuario. La sección distingue normativamente:

- Población objetivo y marco de muestreo
- Marco de muestreo versus fuente del participante
- Persona seleccionada versus persona invitada
- La persona invitada y la persona que acepta participar
- El participante y la persona que completa la tarea
- La persona que completa la tarea y la observación válida
- Humano y cuenta
- Humano y dispositivo
- Humano y sesión
- Panel probabilístico y panel de voluntarios
- Selección aleatoria dentro del panel y selección aleatoria de la población
- Cumplimiento de cuotas y representación verdadera
- Equilibrio demográfico y equilibrio conductual
- Familiaridad personal y idoneidad para la población objetivo
- Cliente, empleado y usuario general
- Recompensa al participante e incentivo por resultados
- Velocidad de la tarea e integridad de la tarea
- Observación inválida con no respuesta
- Cambio de población con pérdida del panel
- Participante falso y observación real de baja calidad
- Fraude de datos y respuesta incorrecta de IA
- Corrección de sesgo con ponderación
- Producción de evidencia con calibración
- Tamaño de muestra bruto vs. tamaño de muestra efectivo
- Privacidad del participante y evidencia de auditoría
- Panel de población general y panel de expertos o clientes

Al final de esta sección, cada auditor GEO-1000 debería poder responder a las siguientes preguntas:

> *¿De qué fuente y bajo qué regla se seleccionaron los participantes?*

¿Quién no tuvo posibilidad de unirse al panel?

¿Quién tenía más probabilidades de participar y por qué?

¿Qué diferencias conocidas corrigen los pesos y cuáles no corrigen?

¿Se tomó una decisión de participación inválida o falsa sin ver el contenido de la respuesta de la IA?

¿Afectó la recompensa a los participantes su comportamiento y elección de respuestas?

2. DISPOSICIÓN NORMATIVA CENTRAL

Cada observación GEO-1000 debe llevar una cadena de selección de participantes trazable desde la población objetivo hasta un registro analítico válido; las decisiones de selección, invitación, aceptación, asignación de tareas, finalización, verificación, exclusión y ponderación deben ser independientes del contenido positivo o negativo de la respuesta de la IA. [K04–K06] Un panel de participantes únicamente:

- país,
- no implica ninguna idoneidad en términos de edad,
- idioma

No puede considerarse adecuado para la representación de la población porque cumple con las cuotas. El panel también debe llevar la siguiente información en la medida que sea relevante:

- Fuente del participante
- Mecanismo de selección
- Método de participación en el panel
- Criterios de elegibilidad
- Verificación única de humano
- Verificación de país e idioma
- Verificación de acceso al producto de IA
- Información de cuenta y plan
- Dispositivo e interfaz
- Frecuencia de uso de IA
- Familiaridad con la marca
- Relación con la entidad auditada
- Estructura de incentivos
- Método de asignación de tareas
- No respuesta
- Invalidación
- Repetición y condicionamiento del panel
- Ponderación
- Falta de cobertura
- Privacidad y consentimiento

Medir parte de esta información puede afectar al participante. Por lo tanto, el orden de la recolección de datos también es parte del protocolo.

3. CADENA DE SELECCIÓN DE PARTICIPANTES

La cadena desde la población objetivo hasta la observación final consiste en las siguientes etapas:

U→F→S→I→E→A→C→V

Aquí:

- U: población objetivo de usuarios
- F: individuos incluidos en el marco muestral
- S: personas seleccionadas
- I: personas invitadas
- E: individuos cuya elegibilidad ha sido confirmada y que han sido incluidos en el estudio
- A: personas asignadas a la tarea
- C: personas que completan la tarea
- V: personas que producen observaciones analíticas válidas

En cada transición, se pueden perder personas o surgir nuevos sesgos. Por ejemplo:

- Aquellos en la población objetivo que no están en el panel
- Aquellos en el panel que no son invitados
- Aquellos invitados que no aceptan
- Aquellos que aceptan pero no completan la tarea
- Aquellos que completan pero producen registros inválidos

no son las mismas personas. Por lo tanto, solo las últimas 1,000 observaciones válidas no pueden ser reportadas. Todo el embudo de selección debe ser visible.

4. FUENTES DE PARTICIPANTES

GEO-1000 no tiene que estar vinculado a una única fuente de participantes. Cada fuente tiene su propio sesgo y alcance.

4.1. Panel de Población de Probabilidad

Los participantes tienen una probabilidad conocida y no nula de ser seleccionados de una población definida. Fuentes:

- hogar,
- dirección,
- población o registro elegible,
- probabilidad de selección telefónica o por etapas múltiples

puede ser. Ventajas:

- se puede establecer un peso de diseño,
- la inferencia poblacional es más fuerte,

el proceso de selección puede ser más transparente. Limitaciones:

- puede ser costoso a escala global,
- en algunos países, puede que no exista un marco adecuado,
- También puede ser necesario encontrar la subpoblación con acceso a productos de IA

De nuevo, habrá falta de respuesta. La selección probabilística no es automáticamente una representación perfecta.

4.2. Panel de Investigación Profesional de Voluntarios

Los participantes se han registrado previamente de forma voluntaria para el panel de investigación. Ventajas:

- es rápido,
- puede acceder a múltiples países,
- puede proporcionar perfil y conocimiento de idiomas,

se puede realizar medición repetida. Limitaciones:

- los participantes son propensos a la investigación,
- el nivel de uso de la tecnología puede ser alto,
- pueden estar presentes usuarios frecuentes del panel,
- las probabilidades de ingreso a la población pueden ser desconocidas,

la misma persona puede estar registrada en más de un panel. Este panel:

no puede presentarse como una muestra probabilística de la población mundial

No debe representarse como una muestra de probabilidad de la población mundial.

4.3. Llamado Abierto en Línea

La participación se obtiene a través de las redes sociales, un sitio web o un formulario de registro abierto. Ventajas:

- amplia visibilidad,
- bajo costo,
- acceso a pequeños países y comunidades lingüísticas

puede proporcionarse. Limitaciones:

- fuerte auto-selección,
- concentración de fanáticos o críticos de la marca,
- participación falsa y repetida,
- sobre-representación de entusiastas de la tecnología

puede crearse. Un llamado abierto puede ser mucho más adecuado que el Panel Principal de la Población para:

- descubrimiento,
- Observador por País,
- Equidad Lingüística,
- desafío de replicación

Estas fuentes pueden ser más adecuadas para esos propósitos.

4.4. Panel Afiliado a Universidad o Institución

Participantes:

- son seleccionados a través de universidades,
- institución pública,
- centros de investigación,
- organizaciones profesionales

Ventajas:

- acceso local,
- competencia lingüística,
- supervisión ética,
- llegar a comunidades con pocos recursos

puede proporcionarse. Limitaciones:

- intensidad estudiantil o profesional,
- efecto institucional,
- parcialidad por edad y nivel educativo,
- presión de participación corporativa

pueden crear.

4.5. Lista de clientes o usuarios

Se seleccionan los clientes o usuarios reales de la institución auditada. Esta fuente:

- experiencia del cliente,
- adecuación para el servicio,
- reconocimiento de marca

es valiosa para la medición. No es un panel de población general. Clientes:

- puede conocer mejor la marca,
- puede tener una experiencia más positiva o más negativa,

puede tener más probabilidades de hacer la consulta relevante. El panel de clientes debe definirse por separado.

4.6. Participantes sugeridos por la institución auditada

Empresa:

- empleado,
- cliente,
- socio,
- seguidor

puede sugerir. Estos usuarios no pueden ser incluidos en el Panel de Población independiente principal sin explicación. Por separado:

PANEL NOMINADO POR LA ENTIDAD

debe ser etiquetado. Si la institución selecciona participantes, el resultado no es una medición de población independiente.

4.7. Panel híbrido de múltiples fuentes

Es la combinación de múltiples fuentes. Ejemplo:

- panel nacional basado en probabilidad,
- panel profesional en línea,
- pequeño socio local en un país,
- Comunidad de equidad lingüística

puede ser utilizada conjuntamente. Un panel híbrido puede aumentar la cobertura. Sin embargo, entre los recursos:

- repetir individuos,
- probabilidades de selección,
- incentivos,
- experiencia en tareas,
- perfil del usuario

también deben ser modelados.

5. ESTADOS DE INTEGRIDAD EN LA SELECCIÓN DE PARTICIPANTES

La calidad de la selección de participantes de un panel puede clasificarse con los siguientes estados.

PSI-0— SELECCIÓN DESCONOCIDA

No se conoce suficientemente dónde y cómo fueron seleccionados los participantes. No se puede hacer una generalización poblacional.

PSI-1— CONVENIENCIA O INSCRIPCIÓN ABIERTA

Es una muestra por conveniencia o voluntaria abierta. Puede ser utilizada para exploración y detección de eventos.

PSI-2— CONTROL DE CUOTAS POR ADHESIÓN

Existen cuotas por país, idioma o demografía. La probabilidad de ingreso a la población es desconocida.

PSI-3— PANEL MULTIFUENTE CALIBRADO

Existen múltiples fuentes, cuotas predefinidas y calibración de la población. Puede persistir un sesgo de selección no medido.

PSI-4— DISEÑO PROBABILÍSTICO O HÍBRIDO FUERTE

Existe una probabilidad conocida de selección o un diseño híbrido fuerte de probabilidad–no probabilidad.

PSI-5— SELECCIÓN REPLICADA Y AUDITADA EXTERNAMENTE

Diseño de selección:

- con diferentes proveedores,
- en diferentes oleadas,
- bajo auditoría independiente

Se ha examinado el comportamiento repetido y sesgado. Estos estados no garantizan la exactitud de la puntuación. Indica el nivel de confianza resultante de la selección de los participantes.

6. SELECCIÓN ALEATORIA DENTRO DEL PANEL Y SELECCIÓN ALEATORIA DE LA POBLACIÓN

Puede haber 100,000 usuarios registrados en un panel de voluntarios. De este panel, se pueden seleccionar aleatoriamente 1,000 personas. Selección:

> *es aleatoria dentro del panel.*

Sin embargo, el panel en sí no se forma de manera aleatoria a partir de la población mundial. Por lo tanto, no se puede decir: "1,000 personas al azar de la población mundial." La expresión correcta es:

> *"1,000 participantes seleccionados aleatoriamente dentro de un panel de voluntarios pre-registrados y calibrados a la población objetivo."*

La aleatoriedad debe limitarse a la etapa en la que se aplica.

7. REGISTRO DE ELEGIBILIDAD DE PARTICIPANTES

Cada participante debe cumplir con las siguientes condiciones en la medida en que sean relevantes antes de ser incluido en el panel principal:

- Elegibilidad por edad
- Consentimiento voluntario
- Idoneidad del país y la localidad
- Competencia lingüística
- Acceso al producto de IA
- Plan correcto o interfaz de usuario
- Dispositivo necesario y acceso técnico
- Estado humano único
- Independencia de la tarea
- Declaración de intereses o relaciones
- Seguridad de la investigación
- Aprobación para la recopilación de datos y pruebas

La decisión de conformidad no puede cambiarse retroactivamente después de que se genere la respuesta de IA.

8. VERIFICACIÓN DEL PAÍS

La pregunta "¿A qué país pertenece el participante?" no tiene una única respuesta. Los siguientes campos deben separarse:

- País de residencia habitual
- País físico en el momento de la medición
- País o región de mercado de la cuenta de IA
- País de pago o plan
- Ciudadanía
- País objetivo del prompt
- Localización

La ciudadanía puede no ser la variable principal para la mayoría de las mediciones GEO. Lo que es importante la mayoría de las veces:

> *Es la condición geográfica y del producto real en la que el usuario experimenta el producto.*

8.1. Usuario viajero

El participante puede estar en un país distinto de su país de residencia habitual. Esta persona:

- población residente,
- ubicación de la medición,
- usuario viajero

debería tener un estatus separado como.

8.2. VPN o cambiador de ubicación

Uso de VPN:

- privacidad
- red de instituciones,
- hábitos normales de usuario

puede ser la razón. No debe considerarse fraude automático. Sin embargo, dificulta la verificación de la capa del país. Puede haber una regla en el panel principal del país controlado para desactivar o registrar el uso de VPN. Los cambios de ubicación usados para eludir restricciones no son adecuados para el panel principal Alcance natural.

8.3. Niveles de Verificación de País

GV-0— No verificado

Solo está disponible la autodeclaración.

GV-1— Verificado por el panel

El proveedor del panel tiene un registro de país existente.

GV-2— Consistente con la sesión

Declaración, configuración regional y señales de sesión son consistentes.

GV-3— Verificado con múltiples señales

Compatible con múltiples señales que protegen la privacidad.

GV-4— Auditado de manera independiente

El método de verificación del país ha pasado el control de calidad externo. No se requiere por defecto recopilar una dirección exacta o coordenadas sensibles. La verificación del país debe realizarse con minimización de datos.

9. DOMINIO DEL IDIOMA Y USO NATURAL

El conocimiento de un participante del idioma puede basarse únicamente en la declaración: "Sí, lo sé." Esto puede ser un registro inicial de bajo riesgo. En un panel de idioma de alta seguridad, además:

- declaración de uso natural,

- breve verificación de comprensión,
- historial del panel,
- evaluación de un adjudicador local,
- prueba neutral antes de la tarea

pueden ser utilizados. La prueba de idioma no debe convertirse en un examen académico pesado que seleccione solo a personas con alta educación. El objetivo:

> *Que el participante entienda la prompt y la respuesta como un usuario real.*

10. VERIFICACIÓN DEL ACCESO AL PRODUCTO DE IA

Participante:

- debería poder acceder al producto de IA correcto,
- el interfaz de usuario correcto,
- la condición de plan o cuenta correcta

debe poder acceder. Las siguientes situaciones deben registrarse por separado:

- Plan gratuito
- Plan individual de pago
- Plan empresarial
- Plan de formación
- Web
- Móvil
- Escritorio
- Producto integrado
- Modelo visible
- Función web/de recuperación

Un participante no puede usar la cuenta de otra persona. No se puede solicitar compartir la contraseña. Se debe verificar la prueba de acceso al producto sin divulgar públicamente la identidad de la cuenta.

11. RELACIÓN ENTRE EL PARTICIPANTE Y LA ENTIDAD AUDITADA

La relación del participante con la organización monitoreada puede afectar el comportamiento de captura de respuestas o la familiaridad con la marca. Estados de relación:

PÚBLICO_GENERAL

CLIENTE_ACTUAL

EX_CLIENTE

POTENCIAL

EMPLEADO

EX_EMPLEADO

CONTRATISTA

SOCIO

INVERSOR

AFILIADO_COMPETIDOR

AFILIADO_PROVEEDOR_AI

INVESTIGADOR

DESCONOCIDO

Relaciones directas de empleados y tomadores de decisiones en el panel principal de población general:

- pueden ser excluidas,
- pueden colocarse en una capa separada,

puede protegerse con un peso real de población bajo. Excluir automáticamente a todos los individuos relacionados también puede estar mal. Si forman parte de la población real de clientes o empleados, pueden medirse por separado y explícitamente.

12. FAMILIARIDAD CON LA MARCA

Entidad participante auditada:

- nunca ha oído hablar de ella,
- conoce el nombre,
- conoce el producto,
- es cliente,
- puede tener una opinión positiva o negativa fuerte.

Esta información:

- puede afectar el comportamiento del participante,
- Panel de usuarios en condiciones reales,
- respuesta inmediata

deseada. Si la familiaridad se mide antes del estímulo, puede preparar al usuario para la marca. Por lo tanto, el orden de preferencia:

- Información de panel previamente existente
- Pregunta separada después de la tarea
- Si es necesario, una premedición neutral que no guíe la marca

se puede llevar a cabo. La familiaridad debería ser visible en la muestra principal; no debería seleccionarse como una variable de ponderación después de ver los resultados.

13. CONOCIMIENTO SOBRE EL USO DE IA

Participantes:

- usuario por primera vez,
- usuario poco frecuente,
- usuario regular,
- usuario intenso o profesional

los individuos pueden ser. Frecuencia de uso de la IA:

- acceso al producto,
- tipo de plan,
- conocimiento de la interfaz,
- personalización,
- probabilidad de completar la tarea correctamente

puede verse afectado. Sin embargo, seleccionar solo usuarios de IA con experiencia puede reducir los errores del protocolo. Pero puede distorsionar la estimación de la población natural. La solución correcta:

- medir el nivel de uso,
- compárelo con la población objetivo,
- estratificar o ponderar si es necesario,

y preservar la experiencia real válida del usuario inexperto.

14. REGLA HUMANA ÚNICA

La unidad poblacional de GEO-1000 es un humano. La misma persona:

- en múltiples paneles,
- con múltiples cuentas de correo electrónico,
- con múltiples cuentas de IA,
- de diferentes dispositivos

puede participar. Deben realizarse verificaciones de unicidad de manera que se preserve la privacidad. Señales que se pueden usar:

- Identidad del participante anónimo
- Reverificación basada en criptografía o hash
- Cumplimiento del dispositivo y la sesión
- Registro de pago o incentivo
- Patrón de tiempo y tarea
- Misma captura de pantalla o archivo de respuesta

- Reverificación segura a través de múltiples proveedores de panel

Ninguna señal individual necesita ser suficiente para la exclusión automática. En particular:

- Familias que comparten un dispositivo común,
- red de instituciones,
- Herramientas de accesibilidad,
- Usuarios de dispositivos compartidos de bajos ingresos

pueden ser considerados erróneamente como falsos. Los registros sospechosos deben someterse a revisión humana y a un proceso de apelación.

15. RECOMPENSA DEL PARTICIPANTE

El pago o la recompensa otorgada al participante puede influir en su comportamiento de datos.

15.1. Pago Independiente del Resultado

Método preferido:

> *Un pago fijo y previamente divulgado por completar la tarea de acuerdo con el protocolo.*

Pago:

- no puede depender de que la respuesta sea positiva,
- puede ser forzada a las categorías de

  verdadero,
- largo,
- favorable a la marca,
- o acompañado de citas.

La selección debe permanecer independiente de todos esos resultados.

15.2. Bono de Calidad

Si se va a utilizar un bono de calidad, solo:

- evidencia completa,
- solicitud correcta,
- producto correcto,
- entrega oportuna,
- limpieza de datos personales

puede estar vinculado a las condiciones del protocolo. No puede estar vinculado al contenido de la respuesta de la IA.

15.3. Incentivo de Velocidad

Bono únicamente por velocidad:

- prisa,
- captura incompleta,
- producto equivocado,
- recorte de la respuesta

puede aumentar el riesgo. Se debe permitir un tiempo razonable dentro de la ventana de campo.

15.4. Pago por País

El pago puede variar según las tarifas locales y las condiciones de vida. Esta situación:

- puede afectar la probabilidad de participación,
- la composición del panel

y debe registrarse junto con el enfoque de pago y las diferencias entre países.

15.5. Coerción e Incentivo Excesivo

La recompensa no debe crear presión que la persona no pueda rechazar razonablemente. En particular:

- personas de bajos ingresos,
- estudiante,
- empleados dependientes,
- grupos vulnerables

deben tener protegida su voluntariedad.

16. TEXTO DE INVITACIÓN Y PRE-ORIENTACIÓN

Al invitar a un participante, el propósito de la investigación puede prepararlo para un resultado específico. Una invitación incorrecta: “Estamos probando cómo los sistemas de IA interpretan incorrectamente Apple.” Esta declaración puede llevar al participante a buscar errores. Otra invitación incorrecta: “Ayúdanos a demostrar el éxito de nuestra marca en la IA.” Esta declaración crea una expectativa de un resultado positivo. Invitación neutral preferida: “Estudio de la experiencia del usuario estandarizada y observación de la representación en productos de IA específicos.” La entidad auditada puede volverse visible más adelante debido a la naturaleza de la tarea. Sin embargo, no debe divulgarse el resultado deseado.

17. ASIGNACIÓN DE TAREA

Participantes:

- Producto de IA,
- prompts,
- idioma,
- onda,
- condición de la sesión,
- orden de la tarea

debe asignarse con antelación utilizando un método predefinido. El participante:

- producto de IA favorito,
- el plan que conoce mejor,
- el idioma con alta probabilidad de recibir una respuesta positiva

si se les permite elegir por sí mismos, puede ocurrir un sesgo de asignación.

17.1. Asignación Aleatoria Dentro de la Capa

Después de verificar el acceso por país, idioma y producto, las tareas pueden asignarse aleatoriamente dentro de la capa.

17.2. Asignación por Bloque

Se pueden usar bloques para equilibrar planes gratuitos/pagados, móvil/web, o frecuencia de uso de IA.

17.3. Orden de Productos

Si una persona está probando más de un producto, el orden debe ser aleatorizado. La respuesta del primer producto puede afectar la evaluación de la siguiente respuesta o el comportamiento de la tarea.

17.4. Orden de prompts

Si un participante recibe más de una prompt:

- aprendiendo,
- adquiriendo nueva información sobre la marca,
- llevando la respuesta anterior a la siguiente pregunta

surge el riesgo. El orden de los estímulos debe registrarse y equilibrarse si es necesario.

18. CONDICIONAMIENTO DEL PANEL

Los usuarios que participan frecuentemente en paneles de investigación pueden aprender las expectativas de la tarea. Una persona que participa repetidamente en la misma prueba:

- puede recordar la respuesta correcta,
- puede elegir la captura de pantalla con más cuidado,
- puede revisar el sitio de la entidad auditada con antelación,

se pueden aprender los errores que se buscan. Esta situación se llama:

Condicionamiento del panel

El condicionamiento del panel puede cambiar la verdadera experiencia natural del usuario.

18.1. Fuentes de Condicionamiento

Olas GEO anteriores Re-medición de la misma entidad Banco de estímulos previo Capacitación en revisión por pares Compartición de tareas en comunidades de participantes Pagos anteriores y retroalimentación de calidad

18.2. Controles del Condicionamiento

Estado de participante nuevo y repetido Fecha de primera participación Exposición previa a la misma entidad Exposición previa a la misma familia de prompts Panel rotativo Aleatoriedad en el orden de las tareas Período de descanso especificado Submuestra de nuevos usuarios Comparación de resultados de usuarios condicionados y nuevos kullanıcı No está prohibido reutilizar al mismo participante. Es valioso para la medición longitudinal. No puede presentarse como un panel transversal natural.

19. CONTAMINACIÓN DE TAREAS Y COMUNICACIÓN ENTRE PARTICIPANTES

Participantes:

- prompts,

- respuestas,
- capturas de pantalla,
- resultados esperados

pueden ser compartidos entre sí. Esta situación puede interrumpir las observaciones independientes. Controles:

- Limitar la duración de la oleada
- Aleatorizar tareas
- Obligación de no compartir resultados con los participantes
- Buscar respuestas y elementos visuales repetidos
- Examinar similitudes inusuales en texto y tiempo
- Monitoreo de advertencias de proveedores de panel o comunidad

La comunicación del participante no es automáticamente fraude. Se debe examinar su efecto material.

20. TIPOS DE NO RESPUESTA

La no respuesta no es un evento único.

20.1. No responder a las invitaciones

El participante no responde en absoluto a la invitación de investigación.

20.2. Rechazo a participar

La persona ve la invitación y decide no participar.

20.3. No comenzar la tarea después de la calificación

Se encuentra que la persona es apta pero no inicia la tarea.

20.4. Interrupción de la tarea

El participante abandona la tarea a mitad de camino.

20.5. No subir evidencia

Se puede haber obtenido una respuesta de IA, pero no se sube el registro requerido.

20.6. Finalización inválida

La tarea está completada, pero:

- prompt incorrecto,
- producto equivocado,
- respuesta seleccionada,
- captura de pantalla faltante

son razones de invalidez. Las proporciones de estas etapas deben reportarse por separado.

21. EMBUDO DE RESPUESTAS

Para la capa h:

- nhsel: persona seleccionada

- nhinv: persona invitada
- nhcon: persona que aceptó participar
- nhsta: persona que comenzó la tarea
- nhcom: persona que completó
- nhval: observación válida

Seamos. Tasas básicas:

```
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
```

Se puede mostrar como. Cada una de estas tasas indica un problema diferente. Dar solo el conteo válido final oculta el sesgo del campo.

22. SESGO POR NO RESPUESTA

Las personas que no responden pueden ser diferentes de las que sí lo hacen. Ejemplo: Un usuario con internet lenta no puede completar una tarea. Un usuario con plan gratuito no puede acceder al producto. Un usuario inexperto encuentra difíciles las instrucciones. Un usuario intensivo de IA completa más rápido. En ciertos países, el pago sigue siendo bajo. Un usuario con una experiencia negativa con la marca puede estar más dispuesto a participar. Incluso si la tasa de no respuesta es alta, el sesgo puede ser bajo. Incluso si la tasa de no respuesta es baja, un pequeño grupo ausente puede ser financieramente significativo. Por lo tanto, no solo la tasa por sí sola:

Qué características difieren entre los que responden y los que no responden

debería ser examinado.

23. PESOS POR NO RESPUESTA

Si se va a utilizar una configuración de no respuesta, las clases deben determinarse antes de ver los resultados. Variables potenciales:

- País
- Idioma
- Edad
- Frecuencia de uso de IA
- Plan
- Interfaz
- Fuente del panel
- Accesibilidad
- Familiaridad con la marca

Coeficiente de no respuesta:

```
a_r^{NR}= (Σ_{j∈S_r} d_j)/ (Σ_{i∈R_r} d_i)
```

puede calcularse como. Sin embargo, este ajuste solo se basa en la suposición dentro de la misma clase: que los encuestados representan suficientemente a los no encuestados. No resuelve las diferencias no medidas.

24. MECANISMOS DE DATOS FALTANTES

La ausencia de datos puede considerarse conceptualmente en tres clases.

24.1. Ausencia de datos aleatoria

La ausencia de datos no está materialmente relacionada con el resultado y las características del usuario. En realidad, puede ser rara.

24.2. Ausencia de datos explicada por características observadas

Deficiencia:

- país,
- no implica ninguna idoneidad en términos de edad,
- plan,
- idioma

está relacionada con áreas bien conocidas como. El ponderado puede ayudar parcialmente.

24.3. Deficiencia relacionada con características no observadas

Por ejemplo:

- Desconfianza en el producto de IA,
- Bajas habilidades digitales,
- oposición de marca,
- Preocupación por la privacidad

si no se mide, la ponderación no puede corregir completamente la deficiencia. En este caso, se necesita un análisis de sensibilidad.

25. PÉRDIDA DE PANEL Y SEPARACIÓN ENTRE ONDAS

En paneles longitudinales o giratorios donde se miden de nuevo los mismos usuarios, algunos individuos no participan en la siguiente ola. Esto es pérdida de panel. Los que quedan en el panel:

- más disciplinado,
- más motivado por la recompensa,
- más experimentado,
- más acostumbrado a la investigación

Puede ser. El cambio en la partitura con el tiempo puede deberse en parte a la composición de las viñetas.

25.1. Récord de rotación en los paneles

En cada oleada:

- re-invitado,
- participó,
- se retiró,
- recién agregado

deben mostrarse los números de usuarios.

25.2. Peso de la deserción del panel

Los participantes restantes pueden ser ponderados según la probabilidad de deserción. Este método no resuelve completamente el sesgo de la deserción no observada del panel.

25.3. Actualizar panel

Se pueden agregar nuevos usuarios para reemplazar a los usuarios perdidos. Los resultados para usuarios nuevos y que regresan deben reportarse por separado.

26. FRAUDE DE PARTICIPANTES Y AUTOMATIZACIÓN

Existe un riesgo de fraude cuando GEO-1000 es un panel incentivado y global. Comportamientos posibles:

- Una persona participando con múltiples cuentas
- Uso de bots automáticos
- Subir capturas de pantalla de otra persona
- Editar imágenes
- Subir la misma respuesta en nombre de diferentes usuarios
- Declarar un país o idioma incorrecto
- Compartir una cuenta
- Hacer que otra persona complete la tarea
- Reingresar entre proveedores de panel
- Crear granjas de tareas en masa

Este riesgo no permite que todos los participantes de bajos ingresos o rápidos sean considerados sospechosos. Detección de fraude:

- debates,
- múltiples señales,
- revisión humana proporcional

debería basarse en.

27. ESTADOS DE FRAUDE E INTEGRIDAD

FI-0— NO EVALUADO

El registro no ha sido evaluado en términos de integridad.

FI-1— SIN ANOMALÍA MATERIAL

No se han encontrado signos claros de fraude. Esto no es una garantía absoluta de precisión.

FI-2— REVISIÓN REQUERIDA

Hay una o más señales sospechosas.

FI-3— SOSPECHA DE DUPLICADO O COLUSIÓN

Hay sospecha de repetición de persona, visual, respuesta o coordinación.

FI-4— MANIPULACIÓN DEL PROTOCOLO CONFIRMADA

El participante ha alterado deliberadamente el protocolo.

FI-5— OBSERVACIÓN FABRICADA CONFIRMADA

La observación no es una experiencia genuina de usuario de IA. FI-2 no es exclusión automática. Se requiere revisión.

28. SEÑALES DE FRAUDE

Mismo hash de captura de pantalla
Misma huella de archivo raro
Tiempo de finalización imposible
Repeticiones de identidad múltiple desde el mismo dispositivo o sesión
Conflicto material en país, zona y acceso a productos
Incompatibilidad entre el texto de respuesta y la imagen
Rastro de edición de imagen
Misma persona en la lista principal y de respaldo
Misma persona anónima a través de múltiples proveedores de panel
Finalización serial y mecánica de tareas
Mismo patrón de error inusual
Una sola señal puede no ser prueba concluyente de fraude.

29. EQUIDAD EN LA DETECCIÓN DE FRAUDE

Los siguientes usuarios pueden ser marcados erróneamente como fraudulentos:

- Familias que comparten un dispositivo común
- Empleados que usan una red institucional
- Participantes de laboratorio universitario
- Individuos preocupados por la privacidad que usan VPN
- Usuarios con discapacidad que usan lectores de pantalla o asistentes de automatización
- Aquellos que suben imágenes comprimidas similares debido a ancho de banda bajo

Por lo tanto:

- etiquetado automático,
- revisión humana,
- descripción del participante si es necesario,
- apelación

deben trabajar juntos.

30. LA VALIDEZ DEL PARTICIPANTE DEBERÍA ESTAR SEPARADA DE LA PRECISIÓN DE LA RESPUESTA DE LA IA

El registro de un participante puede cumplir con el protocolo. La respuesta de la IA puede ser incorrecta. Esta es una observación válida. La respuesta de IA de otro participante puede ser correcta. Sin embargo, el usuario:

- ha actualizado la respuesta tres veces,
- solo subió la mejor respuesta,
- puede haber cambiado el prompt

Esta es una observación no válida. Por lo tanto:

> *La respuesta correcta de la IA no significa datos válidos. / Una respuesta incorrecta de la IA no significa datos no válidos.*

La evaluación de la validez debe hacerse antes de la puntuación sustantiva y de la manera más ciega posible.

31. CONFLICTO DE INTERESES Y FUENTE DEL PARTICIPANTE

Un participante o proveedor de panel puede beneficiarse de los siguientes resultados:

- La empresa auditada recibiendo una puntuación alta
- El producto de IA apareciendo como superior
- Un competidor recibiendo una puntuación baja
- Vender más auditorías
- Premio o relación laboral

Por lo tanto:

- relaciones con participantes,
- acuerdo con el proveedor de panel,
- salario independiente del resultado,
- Intervención del cliente

debe ser explicado. Institución auditada:

- qué usuarios invitar,
- quién será considerado inválido,
- Quién será reemplazado con un respaldo

no puede decidir.

32. EFECTO DEL PROVEEDOR DE PANEL

Los usuarios de dos proveedores de panel en el mismo país pueden ser diferentes. Un proveedor:

- Entusiastas de la tecnología,
- estudiantes,

- participantes de investigación intensiva

tienen una tasa más alta en términos de cuidado. El proveedor de panel debe mantenerse como una variable de clúster o fuente en el análisis. Si es posible:

- resultados basados en el proveedor,
- efecto de la fuente,
- comparación del proveedor dentro del mismo país

deben examinarse. Un panel global de un solo proveedor no puede presentarse como la población mundial en sí.

33. REPETIR INDIVIDUO EN PANEL MULTIFUENTE

El mismo individuo:

- dos paneles profesionales,
- en una convocatoria abierta,
- en la lista de la universidad

se puede encontrar. Un panel de múltiples fuentes puede aumentar el número de registros sin aumentar la unicidad. Verificación de duplicación entre fuentes:

- preservando la privacidad,
- sin requerir la divulgación central de identidad,
- permitiendo la objeción a las discrepancias

debería establecerse de manera.

34. MAPA DE SESGO DE MUESTRA

GEO-1000 debería supervisar por separado los siguientes tipos de sesgo.

SB-1— Sesgo de Cobertura

Algunos segmentos de la población objetivo no están en el marco de muestreo.

SB-2— Sesgo de Selección

La probabilidad de que los individuos en el marco sean seleccionados para el panel no se ajusta al diseño previsto.

SB-3— Sesgo de Auto-selección

Los voluntarios para el estudio son sistemáticamente diferentes de la población.

SB-4— Sesgo de Invitación y Orientación

El texto de invitación o la fuente atrae a personas con un punto de vista específico.

SB-5— Sesgo de No Respuesta

Las personas que no participan o no completan son diferentes en términos de resultados.

SB-6— Sesgo de Deserción del Panel

Los usuarios que permanecen entre oleadas son diferentes de los que abandonan.

SB-7— Sesgo de Incentivos

El pago o recompensa afecta la velocidad del participante, el comportamiento de selección o la respuesta.

SB-8— Condicionamiento del Panel

El participante aprende el comportamiento esperado de estudios previos.

SB-9— Sesgo de Asignación de Tareas

Los participantes eligen el producto, idioma o prompts por sí mismos o se asignan de manera desigual.

SB-10— Sesgo de Clasificación por País e Idioma

El participante es asignado al país o celda de idioma incorrecto.

SB-11— Sesgo de Producto, Plan e Interfaz

Solo ciertos usuarios de un plan o dispositivo determinan el panel.

SB-12— Sesgo de Familiaridad de Marca

Clientes, fanáticos, críticos o empleados están sobrerrepresentados.

SB-13— Sesgo de Fraude y Repetición

La misma persona o la participación automática obtiene demasiada influencia en el resultado.

SB-14— Sesgo de Exclusión de Validez

Las respuestas negativas de IA son consideradas inválidas con mayor frecuencia.

SB-15— Sesgo de Ponderación

Variables seleccionadas después del resultado o un cambio excesivo en la puntuación del modelo en la dirección deseada.

SB-16— Sesgo de Muestra Superviviente

Solo se presentan los usuarios que pueden completar exitosamente la tarea como la población real del producto.

35. PROPÓSITO DEL PONDERADO

El ponderado puede servir para los siguientes propósitos:

- Corregir probabilidades de selección desiguales
- Devolver el sobrerrepresentación a la proporción real de la población
- Reducir las diferencias en las tasas de respuesta
- Calibrar la muestra a totales objetivos conocidos
- Combinar múltiples fuentes de panel

El ponderado no puede servir para los siguientes propósitos:

- Crear una persona que no exista en el marco muestral
- Corregir variables no medidas con precisión

- Hacer real una observación falsa o errónea
- Producir la puntuación deseada después del resultado
- Convertir una sola celda pequeña en una puntuación confiable del país
- Transformar automáticamente un panel no probabilístico en una muestra probabilística

36. ARQUITECTURA DE PESO PESADO

El peso final se puede mostrar de la siguiente manera:

```
w_i= d_i× a_i^{NR}× g_i^{CAL}× p_i^{SRC}
```

Aquí:

- di: diseño básico o peso inicial
- aiNR: configuración de respuesta
- giCAL: calibración de la población
- piSRC: fuente del panel o ajuste de muestra improbable

En el panel probabilístico:

```
d_i= 1/π_i
```

Es posible. El verdadero πi es desconocido en el panel no probabilístico. En este caso, el peso utilizado no es el peso de diseño, sino el peso basado en modelo o de calibración. Esta distinción debe conservarse en el informe.

37. LOS PESOS NO PUEDEN DEPENDER DEL RESULTADO

No se pueden usar las siguientes variables para crear pesos:

- Si la respuesta de AI es correcta o no
- Si se menciona la marca o no
- Si hay una cita
- Si se hace una recomendación
- Ya ocurra un error crítico
- Si a la organización auditada le gusta el resultado

Variables de ponderación:

- Antes de ver el resultado,
- Porque está relacionado con la población objetivo y el mecanismo de selección

Debe ser seleccionado. Si el resultado de la respuesta determina el peso, la puntuación se ajusta por sí misma.

38. VARIABLES DE CALIBRACIÓN

Áreas posibles de calibración:

- País
- Región
- Idioma

- Edad
- Género, si es necesario y ético
- Educación
- Asentamiento urbano/rural
- Tipo de acceso a internet
- Dispositivo
- Frecuencia de uso de IA
- Plan gratuito/pago
- Interfaz
- Estado de accesibilidad

Es posible que no se disponga de totales fiables de la población objetivo para cada variable. Demasiadas variables de calibración:

- pueden crear una variabilidad excesiva en los pesos,
- tamaño efectivo de muestra bajo
- sobreajuste del modelo

Debe congelarse antes de ver los resultados del conjunto de calibración.

39. EQUILIBRIO DE PONDERACIÓN

La muestra antes y después de la ponderación debe compararse con la población objetivo. Para cada campo:

- participación objetivo
- participación de la muestra sin ponderar
- participación ponderada
- diferencia restante

debe mostrarse. La ponderación se aplica solo al número de resultados y la tabla de equilibrio no se puede guardar.

40. DIAGNÓSTICOS DE PONDERACIÓN

El informe principal debe incluir los siguientes campos en la medida en que sean relevantes:

- Peso mínimo
- Peso máximo
- Peso mediano
- Coeficiente de variación del peso
- Relación de pesos
- Número de pesos recortados
- Tamaño de muestra bruto
- Tamaño de muestra efectivo
- Efecto del diseño
- Resultado sin ponderar
- Resultado principal ponderado
- Resultados alternativos con ponderación razonable

El número único ponderado no es un registro de control completo sin identificadores de ponderación.

41. VISUALIZACIÓN DE PONDERACIÓN SINTÉTICA

VISUALIZACIÓN DE METODOLOGÍA SINTÉTICA / Los participantes, tasas y puntuaciones a continuación son hipotéticos. Población objetivo y distribución del panel:

Frecuencia de uso de IA	Participación de la población objetivo	Participación bruta del panel	Observación bruta
Uso bajo o poco frecuente	50%	10%	100
Uso medio	30%	30%	300
Uso intensivo	20%	60%	600
Total	100%	100%	1,000

Tasas de aprobación sintéticas:

Grupo	tasa de aprobación	Respuesta anterior
Uso bajo o poco frecuente	72%	72
Uso medio	84%	252
Uso intensivo	96%	576
Total	90%	900

Resultado bruto:

```
900/1000=90%
```

41.1. Coeficientes de calibración

Bajo uso:

```
0.50/0.10=5
```

Uso medio:

```
0.30/0.30=1
```

Alto uso:

```
0.20/0.60 = 1/3
```

Respuesta ponderada:

```
72(5) + 252(1) + 576(1/3) = 804
```

Total ponderado:

```
100(5) + 300(1) + 600(1/3) = 1,000
```

Resultado ponderado:

```
804/1000=80.4%
```

41.2. Tamaño de muestra efectivo

```
n_eff= [100(5)+300(1)+600(1/3)]² / [100(5 ²)+300(1 ²)+600(1/3)²] ≈ 349
```

Tarjeta de resultados:

- Observación válida en bruto: 1,000
- Puntuación en bruto: 90 por ciento

- Puntuación ponderada: 80.4 por ciento
- Muestra efectiva: aproximadamente 349
- Razón principal: sobrerepresentación de usuarios intensivos de IA

Interpretación correcta:

> *El panel contiene 1,000 observaciones válidas; sin embargo, como resultado de la calibración de la población objetivo, la estimación ha disminuido a 80.4%, y debido a la variabilidad de los pesos, el tamaño de muestra efectivo se ha convertido en aproximadamente 349.*

Comentario incorrecto: “La precisión global ha sido demostrada como 90% con 1,000 personas.”

42. ¿QUÉ NO PUEDE CORREGIR EL PONDERADO?

Los siguientes grupos no deberían estar presentes en el panel sintético en absoluto:

- Personas que usan lectores de pantalla
- Usuarios rurales con baja conectividad
- Personas que usan el producto de IA por primera vez
- Pequeñas comunidades lingüísticas

El ponderado solo cambia el peso de los usuarios en el panel. El grupo no observado:

- IA, tu respuesta,
- revelar problemas de acceso,
- distribución del error

no puede producir. Por lo tanto:

> *Cobertura cero significa información cero.*

Resultado ponderado: puede ser una predicción más cercana a la población objetivo, no una observación completa de la población objetivo.

43. ANÁLISIS DE SENSIBILIDAD

Se debe examinar la dependencia del resultado ponderado en un solo modelo de ponderación. Resultados candidatos:

- Resultado bruto no ponderado
- Resultado enfocado en el diseño
- Resultado calibrado demográficamente
- Calibración con uso de IA añadido
- Resultado con ponderación recortada
- Resultado ajustado por el efecto del proveedor del panel

Si los resultados cambian de manera material:

Se debe dar la advertencia de sensibilidad de ponderación

No se puede presentar un solo resultado de "ponderación correcta" como una verdad absoluta.

44. AJUSTE DE PROPENSIÓN EN PANEL DE NO PROBABILIDAD

La tendencia de los miembros del panel de no probabilidad a entrar en el panel puede modelarse según la población objetivo. La probabilidad de que una persona esté en el panel:

```
p̂_i= Pr(i ∈ panel | X_i)
```

puede ser estimado. El peso de propensión inversa candidato:

```
w_i^P= 1/p̂_i
```

puede ser usado. Este método:

- usa solo las variables Xi medidas,
- es afectado por un modelo mal especificado,
- la selección no observada no se resuelve,

requiere datos de referencia de objetivo fuertes. Por lo tanto, el resultado:

debería ser etiquetado como:

Estimación de panel no probabilístico calibrada por modelo

45. ESTIMACIONES DE ÁREA PEQUEÑA Y BASADAS EN MODELOS

El muestreo directo puede ser insuficiente en algunas celdas de país o idioma. Los métodos basados en modelos pueden tomar información de grupos vecinos o similares. Esto significa que el resultado:

- puntuación del país observada directamente,
- tasa bruta de usuario

no lo es. El estado correcto:

ESTIMACIÓN ASISTIDA POR MODELO

ESTIMACIÓN DE ÁREAS PEQUEÑAS

NO OBSERVADO DIRECTAMENTE

debe. La predicción basada en modelos no convierte un país sin observación del Observador por País en un país observado.

46. PRIVACIDAD Y VERIFICABILIDAD DEL PARTICIPANTE

GEO-1000 requiere prueba para cada observación. Esto no significa que la identidad del participante se divulgará públicamente. Una captura de pantalla puede incluir la siguiente información:

- Nombre
- Correo electrónico
- Foto de perfil
- Chat anterior

- Plan de cuenta
- Nombre de la institución
- Información personal sensible

NOMOS Captura o sistema equivalente:

- debe no capturar campos personales innecesarios,
- debe realizar redacción automática o controlada,
- debe conservar la evidencia en bruto en un repositorio seguro.

solo se deben presentar al público registros anónimos y verificables.

46.1. Manifiesto de Evidencia Pública

Registro público:

- ID del observador,
- celda país/idioma,
- Producto de IA,
- tiempo,
- hash de la evidencia,
- estado de verificación

puede ser llevado. Captura de pantalla completa:

- editado,
- acceso restringido,
- abierto a auditor independiente

puede ser. La existencia de mil pruebas no requiere la divulgación pública de la identidad de mil personas.

46.2. Riesgo de País Pequeño Sensible

Si solo hay un participante en el país pequeño:

- dispositivo,
- tiempo,
- idioma,
- plan

Cuando se reúna, puede volverse más fácil volver a identificar a la persona. El nivel de detalle en los informes públicos debe reducirse. La integridad de la evidencia puede mantenerse en un sistema con acceso controlado.

47. CONSENTIMIENTO DEL PARTICIPANTE

El participante debe al menos conocer lo siguiente:

- Qué tarea realizará
- Qué datos y capturas de pantalla se recopilarán
- Qué información personal no se recopilará
- Cuánto tiempo se almacenarán los datos
- Quién tendrá acceso
- Cómo se publicará el resultado

- Cómo se dará la recompensa
- Que la participación es voluntaria
- Derecho a retirarse u oponerse
- No se solicitará la contraseña de la cuenta
- No se espera violar los términos de los productos de IA

El consentimiento no debe estar oculto en un texto largo e incomprensible. [K06; K22]

48. SEGURIDAD DEL PARTICIPANTE

Por parte del participante:

- proporcionar información personal de salud, legal o financiera a la IA,
- usar datos reales de clientes,
- compartir información confidencial del empleador,
- debilitar la seguridad de la cuenta,
- exceder los límites de uso,
- violar la ley o los términos del producto

no puede ser solicitado. Las solicitudes de prueba deben tener, en la medida de lo posible, un contexto público y no personal.

49. EL PARTICIPANTE LLEVA DIVERSIDAD HUMANA, NO CALIDAD

Usuario inexperto:

- puede completar la tarea más lentamente,
- más difícil,
- con más apoyo

Esta persona no es automáticamente un participante de baja calidad. Son una parte real de la población objetivo. Si el diseño de la investigación es tan complejo que solo las personas con altas habilidades digitales pueden completarlo: el problema puede no estar en el participante, sino en la herramienta de medición. El protocolo:

- debería ser comprensible,
- accesible,
- lingüísticamente natural,
- carga técnica limitada

debe ser.

50. TARJETA DE RESULTADOS PÚBLICOS DE LA SELECCIÓN DE PARTICIPANTES

El informe público del candidato debería incluir las siguientes áreas:

Recursos del Panel

Participación del panel de probabilidad Participación del panel de voluntarios profesionales Participación de llamada abierta Participación de universidades/instituciones Participación de clientes o paneles afiliados

Embudo de Participación

Persona en marco Seleccionado Invitado Aceptado Empezado Completado Produciendo observación válida

Elegibilidad e Integridad

Niveles de verificación de país Niveles de verificación de idioma Tasa de repetición de persona Tasa de registro sospechoso Tasa de fraude verificado Tasa de observación inválida

Composición del panel

Frecuencia de uso de IA Plan e interfaz Familiaridad con la marca Relación cliente/empleado Usuarios del panel nuevos y recurrentes Cobertura de accesibilidad

Peso Diagnósticos

Lo siguiente puede publicarse sin revelar la identidad de los participantes: puntuaciones sin procesar y ponderadas; pesos mínimos y máximos; el coeficiente de variación de los pesos; tamaño de muestra efectivo; sensibilidad a la ponderación; y advertencias sobre cobertura.

51. DISPOSICIONES NORMATIVAS OBLIGATORIAS

CH08-N01

Cada observación válida GEO-1000 debe tener una cadena de selección de participantes rastreable desde la población objetivo hasta el registro analítico.

CH08-N02

La selección aleatoria dentro del panel no puede presentarse como selección aleatoria de la población mundial objetivo.

CH08-N03

La fuente de los participantes, el método de selección y el estado del panel deben explicarse en el registro público del método.

CH08-N04

No se puede definir como una muestra probabilística improbable o voluntaria de la población.

CH08-N05

Los participantes propuestos por la institución auditada no pueden incluirse en el panel principal independiente de la población sin explicación.

CH08-N06

Las relaciones con clientes, empleados, socios, competidores y proveedores de IA deben registrarse por separado.

CH08-N07

Se debe verificar los criterios de elegibilidad de los participantes antes de que se vea la respuesta de la IA.

CH08-N08

La elegibilidad por país, idioma, producto, plan e interfaz debe determinarse antes de la asignación de la tarea relacionada.

CH08-N09

No se puede duplicar a la misma persona como unidad de población debido a múltiples cuentas, dispositivos, idiomas, paneles o productos de IA.

CH08-N10

Se debe realizar una verificación de persona única de acuerdo con los principios de minimización de datos y privacidad.

CH08-N11

Una sola señal técnica por sí sola no puede considerarse suficiente para una decisión automática y definitiva de que un participante es falso.

CH08-N12

La recompensa de un participante no puede estar vinculada a que la respuesta de la IA sea positiva, correcta, larga, contenga referencias o sea beneficiosa para la marca.

CH08-N13

El bono de calidad solo puede vincularse a la integridad del protocolo preespecificada.

CH08-N14

El texto de invitación e información no puede dirigir al participante a buscar un resultado positivo o negativo.

CH08-N15

El producto de IA debe asignarse o aleatorizarse antes de ver el resultado de la secuencia de prompts y tareas.

CH08-N16

La selección de producto e prompt basada en la preferencia del participante no puede usarse sin explicación en el panel principal de la población.

CH08-N17

Se debe registrar la condicionamiento del panel, la exposición previa a pruebas y la participación repetida.

CH08-N18

Los resultados de los participantes nuevos y condicionados deben analizarse por separado cuando sea necesario.

CH08-N19

La inclusión de un participante en el panel, su retención, exclusión o reemplazo con un respaldo no puede basarse en el contenido de la respuesta de la IA.

CH08-N20

Las salidas de IA incorrectas, negativas, de rechazo o críticas deben conservarse bajo un protocolo válido.

CH08-N21

La validez del participante debe evaluarse de forma ciega en la mayor medida posible y antes de evaluar la exactitud sustantiva de la respuesta.

CH08-N22

Las tasas de invitación, aceptación, inicio, finalización y observación válida deben informarse por separado.

CH08-N23

El entorno de la respuesta solo debe basarse en variables predefinidas y relevantes.

CH08-N24

El ponderado no puede hacer que los usuarios que no están presentes en el marco de muestreo estén representados.

CH08-N25

Las variables de ponderación no pueden depender del resultado de la respuesta de la IA, de la puntuación o del interés comercial de la institución regulada.

CH08-N26

Los pesos de panel de no probabilidad no pueden presentarse como pesos de diseño.

CH08-N27

Las variables de calibración y los totales objetivo deben ser liberados antes de que se vean los resultados.

CH08-N28

Los resultados sin ponderar y los resultados principales ponderados deben almacenarse juntos.

CH08-N29

Los diagnósticos de ponderación y el tamaño efectivo de la muestra deben ser visibles en el informe principal.

CH08-N30

Si se utiliza el recorte de pesos, deben conservarse el resultado sin recortar y el análisis de sensibilidad.

CH08-N31

Los efectos del proveedor del panel y de la fuente de los participantes deben mantenerse como variables separadas.

CH08-N32

En un panel de múltiples fuentes, se deben realizar verificaciones repetidas de personas entre fuentes.

CH08-N33

La detección de fraude debe incluir múltiples señales, revisión humana proporcional y un proceso de apelación.

CH08-N34

El uso de VPN, dispositivo compartido, red institucional o tecnología asistida no debe considerarse automáticamente fraude.

CH08-N35

La estimación de pequeñas áreas basada en modelos no puede presentarse directamente como una puntuación observada por país o idioma.

CH08-N36

Las capturas de pantalla de los participantes y las pruebas deben incluir minimización de datos personales, redacción y control de acceso.

CH08-N37

No se pueden solicitar contraseñas de cuentas, historial de chat privado o información sensible innecesaria a los participantes.

CH08-N38

No se puede pedir a los participantes que eludan restricciones de producto, edad, identidad, país o cuenta.

CH08-N39

No se pueden inscribir en el panel a niños o a usuarios que requieran protección especial sin protocolos éticos y de seguridad separados.

CH08-N40

Los registros de selección y ponderación de participantes deben tener un responsable humano o institucional.

52. FORMAS DE FALLA

CH08-F01— CONTAR ALEATORIOS DENTRO DEL PANEL COMO ALEATORIOS DEL MUNDO

La selección aleatoria dentro del panel voluntario se presenta como si fuera selección aleatoria de la población mundial.

CH08-F02— CONTAR EL CUMPLIMIENTO DE CUOTAS COMO REPRESENTATIVO

Se afirma que todos los sesgos conductuales y tecnológicos se resuelven porque se cumplen las cuotas de edad y país.

CH08-F03— CONTAR CLIENTES COMO POBLACIÓN GENERAL

La lista de usuarios de la institución monitoreada se utiliza como si fuera un panel de usuarios mundial.

CH08-F04— OCULTANDO EMPLEADOS Y SOCIOS

Los usuarios asociados con la institución se muestran como participantes generales independientes.

CH08-F05— CONTANDO UN SOLO PROVEEDOR DE PANEL COMO GLOBAL

No se examinan el efecto del proveedor ni el comportamiento del panel registrado.

CH08-F06— CONTANDO UNA CONVOCATORIA ABIERTA COMO MUESTRA PROBABLE

Se conserva la autoselección.

CH08-F07— DETERMINANDO EL PAÍS SÓLO POR IP

Se ignoran los viajes, VPN, la red institucional y las diferencias de región de la cuenta.

CH08-F08— CONTANDO LA CIUDADANÍA COMO EL PAÍS DE USO

La condición geográfica real de la experiencia del producto se clasifica incorrectamente.

CH08-F09— CONTANDO LA DECLARACIÓN DE IDIOMA COMO DEFINITIVA SIN REALIZAR PRUEBAS

El participante se mantiene en la celda de idioma aunque no entienda la instrucción ni la respuesta.

CH08-F10— CONVERTIR LA PRUEBA DE IDIOMA EN SELECCIÓN DE ÉLITE

Solo las personas altamente educadas permanecen en el panel.

CH08-F11— CONTANDO EL NÚMERO DE CUENTAS COMO PERSONAS

Una persona aumenta el peso poblacional con múltiples cuentas.

CH08-F12— CONTANDO UNA CUENTA COMPARTIDA COMO UNA SOLA PERSONA

Una cuenta utilizada por varias personas se interpreta como un solo usuario.

CH08-F13— PAGO DEPENDIENDO DEL RESULTADO

Se otorga un bono por una respuesta positiva, correcta o referenciada.

CH08-F14— ROMPIENDO EL PROTOCOLO CON BONO POR VELOCIDAD

El participante toma una captura de pantalla o se apresura a completar la tarea antes de finalizar la respuesta.

CH08-F15— EXPLICANDO EL RESULTADO EN LA INVITACIÓN

Se pide al participante que encuentre un error o éxito.

CH08-F16— PERMITIENDO QUE EL PARTICIPANTE ELIJA EL PRODUCTO

Los usuarios que seleccionan solo el producto de IA al que pueden acceder o que les gusta determinan el panel.

CH08-F17— OCULTANDO EL EFECTO DEL ORDEN DE LOS INPUTS

Las respuestas previas afectan las tareas posteriores.

CH08-F18— CONSIDERANDO EL CONDICIONAMIENTO DEL PANEL COMO USO NATURAL

A un usuario que ve la misma prueba múltiples veces se le presenta como un nuevo usuario.

CH08-F19— IGNORANDO LA COMPARTICIÓN DE RESULTADOS

Los participantes transfirieron tareas y respuestas entre sí.

CH08-F20— CONTANDO LAS NO RESPUESTAS COMO DATOS FALTANTES

No se examina la diferencia sistemática de los no participantes.

CH08-F21— CONTANDO LA PÉRDIDA DE PANEL COMO MEJORA EN LOS RESULTADOS

Los puntuaciones aumentan porque los usuarios débiles o inexpertos se retiran en olas posteriores.

CH08-F22— INVALIDANDO RESPUESTAS NEGATIVAS

El participante es reemplazado porque la IA proporcionó una respuesta incorrecta.

CH08-F23— VALIDANDO LA RESPUESTA CORRECTA

Incluso si el participante actualizó la respuesta, el registro se mantiene porque el resultado es correcto.

CH08-F24— PASANDO POR ALTO A BOTS Y PARTICIPACIONES REPETIDAS

La misma persona o automatización produce un gran número de observaciones.

CH08-F25— EXCLUSIÓN MASIVA CON UNA SOLA SEÑAL

VPN, dispositivo o velocidad hacen que los usuarios reales sean considerados falsos.

CH08-F26— REALIZANDO REVISIÓN DE FRAUDE BASADA EN LOS RESULTADOS

Los registros con puntuaciones bajos se examinan con mayor frecuencia.

CH08-F27— CREANDO UN GRUPO MUY INÉDITO

Se afirma que los usuarios no presentes en la muestra están representados.

CH08-F28— SELECCIONANDO CALIBRACIÓN POST-RESULTADO

Se seleccionan posteriormente las variables de peso que producen la puntuación más adecuada.

CH08-F29— CONTANDO LA PARTICIPACIÓN BRUTA DEL PANEL COMO PARTICIPACIÓN DE LA POBLACIÓN

Los usuarios intensivos de IA determinan el resultado global con números brutos.

CH08-F30— CONTANDO EL PESO INUSUAL COMO PESO DE DISEÑO

Se hace una inferencia exacta de la población aunque se desconozca la probabilidad real de selección.

CH08-F31— OCULTANDO EL RESULTADO SIN PESO

El efecto del modelo de pesos no es visible.

CH08-F32— OCULTANDO LA MUESTRA EFECTIVA

Aunque los registros brutos 1,000 contienen unidades de información 300–400, se afirma tener precisión completa.

CH08-F33— ELIMINACIÓN DE GRUPOS DÉBILES MEDIANTE RECORTE POR PESO

Se recortan los grupos pequeños con peso alto porque reducen el resultado.

CH08-F34— CONTANDO PUNTUACIONES ESTIMADAS BASADAS EN MODELO

A un país sin ejemplo se le asigna una puntuación definida.

CH08-F35— REVELANDO LA IDENTIDAD PARA EVIDENCIA

Las capturas de pantalla se publican públicamente con información personal.

CH08-F36— HACIENDO REIDENTIFICABLE A UN PARTICIPANTE DE UN PAÍS PEQUEÑO

La información del país, la hora, el dispositivo y el plan revelan a una sola persona.

CH08-F37— PIDIENDO AL PARTICIPANTE CONTRASEÑA O CHAT PRIVADO

Recopilar evidencia viola la seguridad de la cuenta.

CH08-F38— CONSIDERANDO BAJA COMPETENCIA DIGITAL COMO INAPROPIADO

Los usuarios objetivo reales son eliminados del panel.

CH08-F39— PERTURBANDO LA VOLUNTARIEDAD CON RECOMPENSA

El participante está bajo una presión que razonablemente no puede rechazar.

CH08-F40— DEJAR LA ELECCIÓN DEL PARTICIPANTE AL CLIENTE

La entidad auditada selecciona a las personas que pueden determinar el resultado del panel.

53. PROCEDIMIENTO DE AUDITORÍA

Paso 1— Mapear Fuentes de Participantes

Cada panel se registra por separado para socio, convocatoria abierta y fuente del cliente.

Paso 2— Asignar Estado de Selección

Se determina un estado apropiado entre PSI-0 y PSI-5 para cada fuente.

Paso 3— Comparar Cobertura con la Población Objetivo

Determinar qué usuarios objetivo no están presentes en el panel.

Paso 4— Guardar Probabilidad o Modelo de Selección del Participante

Se explica la selección conocida por probabilidad, cuota o voluntaria.

Paso 5— Reglas de Conformidad para Congelar

Edad, país, idioma, plan de producto, interfaz, unicidad, las condiciones de la relación de extracción están versionadas.

Paso 6— Revisa el texto de la invitación

Se evalúa si existe una guía positiva o negativa.

Paso 7— Examina la estructura de la recompensa

Se confirma que la recompensa es completamente independiente del resultado.

Paso 8— Verificar el país, idioma y acceso al producto participante

Se asignan niveles de verificación que protegen la privacidad.

Paso 9— Realizar una prueba humana única

Se examinan las repeticiones intra-fuente e inter-fuente.

Paso 10— Registrar las áreas de relaciones y familiaridad

Se determina la familiaridad del cliente, empleado, competidor, proveedor de IA y marca.

Paso 11— Distribuir tareas según el plan de asignación

Se verifican las asignaciones de producto, prompt, cola y ola.

Paso 12— Registrar la acondicionamiento del panel

Los usuarios nuevos, repetidos y que participaron previamente en la misma prueba se separan.

Paso 13— Configurar el embudo de participación

Se calculan los números de invitación, aceptación, inicio, finalización y validez.

Paso 14— Cegar la revisión de validez

Se determina la validez del protocolo sin que se vea tanto como sea posible la puntuación de la respuesta de la IA.

Paso 15— Examinar señales de fraude

Se utilizan múltiples señales y evaluación humana.

Paso 16— Analizar mi no respuesta y pérdida del panel

Se comparan las características conocidas de quienes participan y quienes no participan.

Paso 17— Generar Pesos de Diseño y Calibración

Las fuentes probabilísticas y no probabilísticas se manejan con métodos separados.

Paso 18— Calcular los Diagnósticos de Pesos

Se calcula la distribución de pesos, efecto de diseño, muestra efectiva y efecto de recorte.

Paso 19— Realizar Análisis de Sensibilidad

Se comparan los resultados crudos, principales, fuertemente aplicados y alternativos razonables.

Paso 20— Crear el Manifiesto de Privacidad y Evidencia

La prueba completa se encuentra en un repositorio seguro; las identidades de prueba anónimas se mantienen en un registro público.

Paso 21— Publicar la Tarjeta Electoral Pública

Se hacen visibles las fuentes del panel, el embudo de participación, los sesgos y el efecto de ponderación.

54. EVIDENCIA REQUERIDA

Registro de la población objetivo; marco de muestreo; proveedores de panel; fuentes de participantes; estado de selección; clase de panel de probabilidad o no probabilidad; método de registro en el panel; texto de invitación; reglas de admisión de participantes; verificación del país del participante; registro de competencia lingüística; registro de acceso al producto de IA; plan y interfaz de usuario; verificación de persona única; comprobaciones de duplicación entre fuentes; familiaridad con la marca; frecuencia de uso de IA; relación con la entidad auditada; relación con el proveedor de IA; compensación del participante; condiciones de bonificación; método de asignación de tareas; registro de aleatorización y bloqueo; orden de producto y solicitud; condicionamiento del panel; participación en piloto previo; número invitado; número que da su consentimiento; número de participantes que comienzan la tarea; número de participantes que la completan; número de observaciones válidas; razones de no respuesta; abandono del panel; razones de invalidez; e indicadores de fraude.

Registros de revisión y objeción Pesos de diseño Configuración de no respuesta Variables de calibración Modelo de propensión, si lo hay Registro de recorte de pesos Tamaño efectivo de la muestra Resultados de sensibilidad Efecto del proveedor del panel Puntuaciones sin ponderar y ponderadas Registro de consentimiento Política de almacenamiento y acceso a datos Reglas de redacción Manifiesto de evidencia pública Persona o institución responsable

55. LISTA DE VERIFICACIÓN DE AUDITORÍA

¿De qué fuentes provienen los participantes? ¿Cada fuente era probabilística, voluntaria o de convocatoria abierta? ¿Se presentó la aleatoriedad dentro del panel como la aleatoriedad de la población mundial? ¿Intervino la institución supervisora en la selección de participantes? ¿Son los participantes clientes y empleados separados? ¿Se documentaron las relaciones con proveedores de IA o competidores? ¿Cómo se verificó la elegibilidad del país? ¿Está el dominio del idioma al nivel del uso real? ¿Se utilizó el producto de IA, plan e interfaz correctos? ¿En cuántas cuentas o paneles participó la misma persona? ¿Se realizó la verificación cruzada entre fuentes? ¿El texto de la invitación guió el resultado? ¿La recompensa depende del contenido de la respuesta de la IA? ¿Redujeron los incentivos de velocidad la calidad del protocolo? ¿La asignación de tareas fue aleatoria o basada en reglas predefinidas? ¿El participante eligió el producto o se generó el mensaje por sí mismo?

¿Se equilibró el orden de los productos y los mensajes? ¿Se registró el condicionamiento del panel? ¿Compartieron los participantes los resultados de las tareas? ¿Es visible el embudo de invitación–

aceptación–inicio–finalización–validez? ¿Los no respondedores difieren en características conocidas? ¿Afecta la deserción del panel a la tendencia temporal? ¿Se tomó la decisión de validez antes de calificar la respuesta de la IA? ¿Se consideraron más frecuentemente inválidas las respuestas negativas o críticas? ¿La detección de fraude contiene múltiples señales? ¿Causaron la VPN o dispositivos compartidos la exclusión automática? ¿Se le dio al participante un camino de apelación? ¿Se definieron las variables de peso antes del resultado? ¿Se etiquetaron correctamente los pesos de panel improbables? ¿Son fiables los totales objetivo? ¿Se presentan juntos los resultados sin ponderar y ponderados? ¿Está claro el tamaño de muestra efectivo? ¿Se mostró el impacto de la poda de pesos?

¿Se ha mostrado que los grupos que nunca se observan están ampliamente representados? ¿Se presentaron las predicciones basadas en modelos como si fueran observaciones directas? ¿Hay detalles personales en las capturas de pantalla de los participantes? ¿Pueden ser reidentificados los usuarios de países pequeños? ¿Se solicitaron contraseñas de cuentas o chats privados? ¿Es claro y comprensible el consentimiento de los participantes? ¿Hay un propietario claramente responsable de la elección de los participantes y de los registros de ponderación?

56. OBJECIONES Y RESPUESTAS

Objeción 1— "Si usamos un panel de investigación profesional, ¿por qué seguimos hablando de sesgo?"

Un panel profesional puede fortalecer la calidad de los datos y la gestión de campo. Los miembros del panel también son:

- voluntario
- acostumbrado a la investigación,
- diferente en el uso de la tecnología

Es posible. El profesionalismo no elimina el sesgo. Hace que medir y gestionar sea más fácil.

Objeción 2— “Si se cumplen las cuotas de edad y país, ¿por qué no debería considerarse que el panel es representativo?”

Participantes:

- Frecuencia de uso de la IA,
- dispositivo,
- plan,
- conciencia de marca,
- destreza digital

Puede diferir en términos de mantenimiento. Las cuotas solo equilibran las áreas que cubren.

Objeción 3— “¿No puede el ponderamiento corregir estas diferencias?”

Puede corregir parcialmente las diferencias para las cuales existen totales objetivos medidos y fiables. No puede corregir definitivamente las diferencias que nunca se observan o se miden.

Objeción 4— “Los clientes de la institución son usuarios reales; ¿por qué no deberían incluirse en el panel principal?”

Pueden ser tomados. Sin embargo, deben tomarse con la proporción y el estado de relación real que tienen en la población general. La lista completa de clientes no es un panel de población general.

Objeción 5— "¿Por qué sería un problema si el participante conoce la marca?"

No tiene que ser un problema. Son parte de la población de usuarios natural. Si la familiaridad no está equilibrada, el resultado puede variar. Por lo tanto, se registra y se analiza si es necesario.

Objeción 6— "Los usuarios intensivos de IA realizan la tarea con mayor precisión; ¿no sería mejor seleccionarlos?"

Puede ser más fácil para la prueba de protocolos. Sin embargo, si queremos medir la experiencia real de la población, los usuarios sin experiencia también son importantes. Si la tarea es demasiado compleja, la herramienta de medición debe corregirse. La población no debe simplificarse.

Objeción 7— 'Si el participante malinterpretó la respuesta, ¿por qué deberíamos mantener la observación?'

La tarea del participante no es evaluar la respuesta, sino capturarla correctamente. Si se aplica correctamente el protocolo de captura, la respuesta de la IA se conserva. La evaluación del material es responsabilidad de los jueces.

Objeción 8— "¿No se debería utilizar un filtro automático estricto para prevenir la participación falsa?"

El filtrado automático es útil. Si se convierte en la única decisión final, puede excluir injustamente a usuarios reales. Se necesitan múltiples señales y una revisión humana proporcional.

Objeción 9— 'Si no hacemos públicas mil capturas de pantalla, ¿se puede considerar que el método está probado?'

Imágenes de pantalla completa:

- información personal,
- seguridad de la cuenta,
- redefinición

puede conllevar un riesgo. Para el público:

- ejemplos editados,
- manifiesto de observación,
- hashes de integridad,
- acceso a auditorías independientes

puede ser proporcionado. La demostrabilidad no es la divulgación de identidad.

Objeción 10— 'Si la puntuación ponderada es muy diferente de la puntuación en bruto, ¿cuál es correcta?'

Dos resultados responden a preguntas diferentes. La puntuación bruta es el resultado de las personas en el panel. La puntuación ponderada es la estimación de la población objetivo. Una gran diferencia indica que la composición del panel es importante. Ambos deben ser publicados.

Objeción 11— "Si la muestra efectiva cayó a 349, ¿fue un desperdicio recolectar 1,000 observaciones?"

Los grupos sobremuestreados pueden haber fortalecido el análisis de subgrupos. Sin embargo, la sensibilidad del pronóstico global puede ser menor que el número bruto de 1,000. Este hecho no debe ocultarse.

Objeción 12— "¿No se puede usar en absoluto el panel de no probabilidad?"

Se puede usar. Específicamente:

- piloto,
- idioma de bajos recursos,
- país pequeño,
- Descubrimiento rápido

Es valioso para. Debe ser publicado con el estatus y las restricciones correctas.

Objeción 13— "¿No es más confiable escoger a todos los participantes entre personas nuevas?"

Puede ser fuerte para la estimación poblacional transversal. Los usuarios recurrentes son valiosos para medir el cambio longitudinal. Los roles de panel nuevos y recurrentes deben mantenerse separados.

Objeción 14— "¿No aumenta el riesgo de privacidad recopilar datos de tantos participantes?"

Puede. Por lo tanto:

- minimización de datos
- identidad anónima,
- redacción,
- control de acceso
- duración de almacenamiento

son partes obligatorias del protocolo. No se debe recopilar información innecesaria.

DISPOSICIÓN COMÚN DE LA SECCIÓN 59

El número más fácil en un estudio GEO:

1,000

Es posible. La pregunta más difícil es esta:

> *¿Cuál 1,000?*

Dos paneles con el mismo país y cuotas de edad pueden producir resultados completamente diferentes. En uno:

- usuarios intensivos de IA,
- suscriptores de planes pagos,
- trabajadores de tecnología

n son la mayoría. En el otro:

- usuarios poco frecuentes,
- usuarios móviles,
- personas con bajas habilidades digitales

Es más visible. Ambos son personas 1,000. No es la misma población. Si ocultas de dónde provienen los participantes y solo publicas el resultado ponderado, conviertes la ponderación en un velo de confianza en lugar de un método de control. No puedes ponderar a una persona que no está presente en una muestra. No puedes cambiar la respuesta de un usuario porque sea incorrecta. Una respuesta incorrecta no es un fracaso de la investigación. Es el evento que la investigación está intentando

medir. Si un participante ha seguido el protocolo incorrectamente a pesar de dar la respuesta correcta, no puedes preservar el registro. Una respuesta correcta no hace que una observación falsa sea real. Puedes pagarle más a una persona. Pero no para producir la respuesta que quieres. Puede pagar según las condiciones de vida de un país. Pero no puede comprar la voluntad del participante. Mil capturas de pantalla pueden servir como evidencia. No está obligado a revelar la identidad de mil personas al público.

La integridad de la evidencia y la privacidad humana pueden protegerse juntos. Por lo tanto, la octava ley de medición de NOMOS es la siguiente:

> *La potencia de la muestra no depende solo de cuántas personas la integran, sino también de los motivos y del modo en que fueron seleccionadas.*

La novena ley es la siguiente:

> *El ponderado reequilibra la población observada; no crea al ser humano no observado.*

La décima ley es:

> *Una respuesta incorrecta de IA puede ser un dato válido; una respuesta correcta de IA puede ser un dato inválido.*

La undécima ley es la siguiente:

> *La recompensa del participante debe darse no por encontrar el resultado deseado, sino por probar el resultado observado sin modificarlo.*

La duodécima ley es la siguiente:

> *La demostrabilidad no requiere que el participante renuncie a su privacidad.*

Capítulo 8 Mandamiento de NOMOS

> *No me traigas mil personas por sí solas. / Muéstrame de qué mundo provienen estas mil personas.*
>
> *No trates a un panel de voluntarios como una muestra poblacional aleatoria, una lista de clientes como el público general o un empleado como un usuario independiente.*
>
> *No pienses que todas las diferencias que no mediste han mejorado solo porque se cumplieron las cuotas de país y edad.*
>
> *No multipliques a una persona por el número de sus cuentas, dispositivos, idiomas o membresías en paneles.*
>
> *No pongas a quienes aman la marca, a quienes odian la marca y a quienes no conocen la marca en absoluto en la misma caja invisible.*
>
> *Pague al participante no por encontrar la respuesta correcta, sino por guardar la primera respuesta adecuada que vean sin cambiarla.*
>
> *No cambie al usuario cuando la IA dé una respuesta incorrecta. / No verifique si la respuesta es correcta cuando el usuario rompa el protocolo.*
>
> *No hagas que una persona que no está en el panel parezca ponderada como si existiera.*
>
> *No ocultes el resultado en bruto. / No declares solo el resultado ponderado como la verdad. / Muestra la muestra efectiva. / Muestra su sensibilidad.*
>
> *No declares a alguien que usa una VPN como falso, a alguien que usa un dispositivo compartido como un bot, o a alguien que usa tecnología asistida como automatización.*

Investiga el fraude. / Pero no conviertas la investigación en sí misma en un nuevo sesgo de exclusión.

Guarda una captura de pantalla. / Protege el nombre, la cuenta y la vida privada de la persona.

Primero, define la población. / Luego muestra el marco. / Luego explica a quién elegiste, quién no vino, quién no pudo completarlo y por qué excluiste a alguien. / Y solo después de eso, decide cuánto puede tu puntuación representar a la humanidad.

La oración de cierre del capítulo

En GEO-1000, un participante es más que una persona que proporciona una respuesta. Un participante es una unidad de muestreo humano cuya capacidad para representar a otros en la población objetivo debe justificarse a través del camino de selección, el peso, la independencia y las salvaguardas de privacidad.

Core normativo

Cada observación GEO-1000 DEBE conservar un camino auditable de selección de participantes desde la población objetivo y el marco de muestreo hasta la selección, invitación, consentimiento, asignación, finalización, validación del protocolo, exclusión y ponderación final. La selección aleatoria dentro del panel NO DEBE ser representada como selección aleatoria de la población objetivo global a menos que el mecanismo de inclusión de toda la población respalde esa afirmación. La contratación de participantes, asignación, compensación, validación, reemplazo, exclusión y ponderación DEBE mantenerse independiente de si la respuesta de la IA es positiva, negativa, correcta, incorrecta, rechazada, crítica o comercialmente favorable. Las personas, cuentas, dispositivos, sesiones, idiomas, membresías de panel y observaciones de productos de IA DEBEN permanecer como unidades distintas. Las fuentes de participantes por probabilidad, opción, convocatoria abierta, suministradas por el cliente, institucionales e híbridas DEBEN identificarse por separado. El ponderado PUEDE reducir el desequilibrio observado, pero NO DEBE representarse como la creación de observaciones para poblaciones ausentes del marco muestral o como la corrección con certeza de selecciones no medidas. Los resultados no ponderados, resultados ponderados, diagnósticos de peso, brechas de cobertura, no respuesta, tamaño efectivo de muestra y análisis de sensibilidad DEBEN permanecer visibles. Las decisiones sobre fraude y duplicación DEBEN utilizar una revisión proporcional con múltiples señales y NO DEBEN tratar automáticamente el uso de VPN, dispositivos compartidos, redes institucionales o herramientas de accesibilidad como falsificación. La evidencia del participante DEBE seguir siendo verificable sin divulgar innecesariamente la identidad, credenciales de la cuenta, conversaciones privadas o ubicación precisa. Cada decisión de selección y ponderación de participantes DEBE estar versionada y ser atribuible a un humano u organización responsable.

PARTE III

Ejecución del experimento y cadena de evidencia

CAPÍTULO 09

Registro de sistemas de IA y comparabilidad

BASES METODOLÓGICAS / K02; K03

Límite de Capítulo

Los primeros ocho capítulos establecieron lo siguiente: una sola respuesta no es una puntuación GEO; la representación es una distribución; la entidad auditada debe ser identificada antes de observar los resultados; la representación oficial y la realidad de la entidad verificada deben permanecer separadas; para cada producto de IA debe definirse una población de usuarios elegible; las observaciones válidas 1,000 deben asignarse mediante un diseño de muestreo basado en la población; los países y idiomas más pequeños deben mantenerse visibles mediante paneles complementarios; y la selección de participantes, la exclusión y el ponderado deben permanecer independientes del contenido de la respuesta. Ahora debemos definir el otro lado del experimento:

> *¿Qué sistema de IA están midiendo realmente los participantes?*

¿El nombre de un proveedor? ¿Un modelo base? ¿Un producto de chat? ¿Una interfaz web gratuita? ¿Una aplicación móvil de pago? ¿Una sesión con búsqueda web habilitada? ¿Un modo de respuesta que no cita fuentes? ¿Un espacio de trabajo empresarial respaldado por datos internos? ¿Una cuenta con memoria habilitada? ¿Chat temporal o limpio? ¿Son las diferentes interfaces de usuario de un mismo proveedor el mismo sistema? ¿Deben medirse de la misma manera dos productos que usan la misma familia de modelos? Incluso si un producto muestra el mismo nombre de modelo en la pantalla:

- recuperación,
- selección de fuente,
- reglas de seguridad,
- personalización,
- acceso a herramientas,
- funciones experimentales

Si cambia, ¿pueden compararse los resultados? Cuando un producto de IA se actualiza en medio de una ola de medición, ¿pueden considerarse que la primera y la última respuesta de la ola provienen del mismo sistema? Un usuario:

- búsqueda web activada,
- memoria activada,
- tiene conversaciones previas,
- plan de pago

respuestas; la de otro usuario:

- la búsqueda está desactivada,
- la memoria está desactivada,
- nueva sesión,
- plan gratuito

¿Se puede fusionar en la misma celda que su respuesta? No. El principio de medición establecido por las secciones anteriores es claro: No existe un único "AI"; el índice de búsqueda, el sistema de recuperación, el reordenador, el modelo de lenguaje, la capa de citación y el sistema de seguridad pueden comportarse de manera diferente. El mismo marco requiere que cada resultado auditable contenga:

- modelo o producto,
- fecha,
- país,
- idioma,
- conjunto de consultas,
- número de repeticiones,
- registro de medición

Esta sección opera estos principios dentro de un Registro del Sistema de IA. Esta sección:

- la distinción entre proveedor, producto, modelo y interfaz de usuario
- instancia de producto de IA,
- niveles de confianza de identidad del modelo,
- recuperación y configuraciones de herramientas,
- estados de memoria y personalización,
- diferencias en el plan, la interfaz y las funciones experimentales,
- cambios de producto,
- configuración del sistema en el momento de la medición,
- niveles de comparabilidad entre productos,
- comparaciones Alcance natural y Base común de comparación,
- desviación de la versión a lo largo del tiempo,

define. Este capítulo aún no:

- sistema de traducción completa y equivalencia semántica de los prompts,
- todos los detalles de los paneles de usuarios controlados y naturales,
- arquitectura del software de captura NOMOS,
- la puntuación de afirmación atómica de las respuestas,
- los coeficientes finales del modelo y la fórmula de puntuación NOMOS

no se finaliza. La pregunta clave de la Sección 9 es:

> *Antes de comparar dos respuestas de la misma pregunta, ¿cómo demostramos que estas respuestas realmente provienen de condiciones de producto de IA comparables?*

NOMOS Desafío

Dos usuarios del mismo proveedor hacen la misma pregunta: "¿Qué tipo de empresa es Apple.com?" Primer usuario:

- está usando el producto web gratuito,

- abre un nuevo chat,
- tiene la memoria desactivada,
- no tiene acceso a la web,

recibe una respuesta solo con la capacidad de conocimiento incorporada del producto. Segundo usuario:

- Usando un producto móvil de pago,
- memoria activada,
- ha hablado previamente sobre Apple,
- búsqueda web activada,

recibe respuestas con fuentes actuales y citas. Primera respuesta: "Apple es una empresa de tecnología que produce productos electrónicos de consumo." Segunda respuesta: "Apple es una empresa tecnológica global que ofrece hardware, software y servicios digitales. Puedes ver información corporativa actual en estas fuentes." Ambas respuestas provienen del mismo proveedor.

¿Es este el resultado del mismo sistema? No. Ahora considere a otros dos usuarios. Ambos ven el mismo nombre de modelo en la pantalla. En el país de un usuario, la función web está activada. En el país del otro usuario, está desactivada. En la respuesta de un usuario, los enlaces de origen son visibles. La interfaz del otro usuario no muestra citas incluso si utiliza las mismas fuentes. Existe la misma etiqueta de modelo.

¿Existe la misma celda de medición? No. Ahora considere dos productos de diferentes proveedores. Ambos:

- nuevo chat,
- la memoria está desactivada,
- misma lengua,
- mismo país,
- mismo intervalo de tiempo,
- acceso web habilitado,
- misma versión de prompt

se miden bajo estas condiciones. Están basados en diferentes modelos subyacentes. ¿Son comparables? Sí, pueden ser comparables en términos de un resultado específico del usuario. Porque el objeto de comparación es:

> *no si utilizan el mismo modelo subyacente, sino qué representación producen bajo la misma condición definida para el usuario.*

Ahora, deje que la versión del producto permanezca sin cambios para las primeras 500 observaciones de la oleada de medición. Durante las siguientes 500 observaciones, el proveedor:

- cambiar el sistema de recuperación,
- reglas de seguridad,
- interfaz de usuario

El informe por sí solo dice: "1,000 usuarios probaron el mismo producto de IA." En realidad, se pueden haber medido dos situaciones de producto diferentes. La primera disposición de esta sección es la siguiente:

> *El nombre del proveedor no es suficiente para identificar el sistema de IA medido.*

Su segunda disposición establece:

> *El nombre del modelo no constituye la totalidad del comportamiento del producto que llega al usuario.*

Su tercera disposición establece:

> *El mismo proveedor, la misma etiqueta de modelo o el mismo logotipo no garantizan comparabilidad.*

Su cuarta disposición establece:

> *Diferentes proveedores pueden ser comparables en términos de resultados para el usuario si las condiciones de medición están suficientemente igualadas.*

Su quinta disposición establece:

> *Si un producto cambia materialmente durante la medición, no se puede informar como un solo sistema fijo.*

1. OBJETIVO DEL CAPÍTULO

El propósito de esta sección es definir el producto de IA auditado dentro de GEO-1000 con todas sus configuraciones materiales y determinar bajo qué condiciones se pueden comparar diferentes resultados del producto. La sección estandariza las siguientes distinciones:

- Proveedor y producto de IA
- Producto de IA y modelo base
- Familia de modelos y despliegue específico
- Etiqueta del modelo e identidad de modelo verificada
- interfaz de usuario y API
- Producto web y móvil
- Plan gratuito y plan de pago
- Producto de consumo y producto empresarial
- Respuesta básica del modelo y respuesta aumentada por recuperación
- Búsqueda web en vivo con recuperación
- Visualización de citas con uso de fuentes
- Uso de herramientas con acceso a herramientas
- Contexto de sesión con memoria
- Política del sistema con instrucción personalizada
- Realidad material con personalización
- Chat temporal con nueva conversación
- Estado del producto a lo largo del tiempo con nombre de producto estático
- Cambio de comportamiento observado con actualización anunciada por el proveedor
- Comparación de Alcance natural y Base común de comparación
- Ejemplo de producto comparable dentro de la misma familia de productos
- Resultado del usuario frente a habilidad abstracta en el modelo
- Ausencia de producto frente a fallo del producto
- Error técnico frente a rechazo real
- Mismo prompt frente al mismo experimento
- Mismo promedio frente al mismo comportamiento del sistema
- Instantánea del producto frente al rendimiento a largo plazo del producto

Al final de esta sección, cada auditor GEO-1000 debería poder responder a las siguientes preguntas:

> *¿Qué proveedor, qué producto, qué interfaz, qué plan y qué configuración fueron medidos?*

¿Cuánto se validó realmente el nombre del modelo que se veía en la pantalla?

¿Cuál era el estado de web, recuperación, citación, memoria, personalización y herramientas?

¿Se mantuvo el producto igual a lo largo de la medición?

¿A qué nivel de comparabilidad se colocaron lado a lado los resultados de los dos productos?

¿Se está comparando el modelo base, el producto de consumo, el interfaz de usuario o toda la cadena del sistema?

2. DISPOSICIÓN NORMATIVA CENTRAL

Cada observación GEO-1000 debe estar vinculada no solo con el proveedor o nombre del modelo; debe estar conectada a un registro versionado del Registro del Sistema de IA que identifique la instancia del producto de IA en el momento de la medición.

Una instancia de producto de IA debe llevar al menos los siguientes campos en la medida en que sean relevantes:

- Proveedor
- Nombre del producto
- interfaz de usuario
- Tipo de producto
- Clase de plan o cuenta
- Etiqueta de modelo aparente
- Nivel de confianza del ID del modelo
- Versión del producto o despliegue
- País y localidad
- Idioma
- Acceso web o de recuperación
- Función de citación
- Acceso a herramientas
- Estado de la memoria
- Estado de instrucciones personalizadas
- Contexto de chats anteriores
- Tipo de sesión
- Estado de personalización
- Modo de seguridad o política, si se conoce
- Funciones experimentales
- Fecha de medición y hora UTC
- Ola de medición
- Registro de cambios del producto
- Confianza en evidencia y registros

Cuando cualquiera de estos campos cambie materialmente:

- nueva instancia de producto AI,
- nueva celda experimental,
- nueva versión de análisis

puede ser requerida.

3. “SISTEMA AI” NO ES UN ÚNICO OBJETO [K02; K03]

En este protocolo, “sistema de IA” es un término general y de alto nivel. Una respuesta real del usuario puede surgir de la combinación de múltiples capas. Cadena candidata:

```
R= F(M, O, K, Q, T, P, S, U, C, L, t)
```

Aquí:

- M: modelo base o guía
- O: capa de orquestación del producto
- K: capa de recuperación o fuente de conocimiento
- Q: reordenamiento y selección de fuentes
- T: herramientas
- P: capa de políticas y seguridad
- S: sesión, memoria y personalización
- U: cuenta de usuario y plan
- C: país y ubicación
- L: idioma
- t: tiempo

Respuesta vista por el usuario:

R

La respuesta visible para un usuario no es necesariamente la salida del modelo base por sí sola. Por lo tanto, el objeto principal orientado al consumidor medido por GEO-1000 es:

> *el comportamiento del producto presentado al usuario, no una abstracción del modelo.*

También se puede realizar investigación controlada con API o modelos. No es el mismo resultado que el panel principal de consumidores.

4. SIETE CAPAS DE IDENTIDAD DEL SISTEMA DE IA

Cada sistema de IA auditado debe definirse a través de siete capas, en la medida en que cada capa sea relevante.

4.1. Capa de Proveedor

La organización que publica u opera el producto. Nombre del proveedor:

- no es lo mismo que

  ID del producto,
- ID del modelo,
- interfaz de usuario

Un proveedor puede ofrecer más de un producto. Dentro del mismo producto, se puede usar más de un modelo o sistema de orquestación.

4.2. Capa de Producto

El producto comercial, de consumo, empresarial o técnico al que accede el usuario. Tipos de ejemplo:

- Producto de chat general
- Producto de respuesta asistida por búsqueda
- Espacio de trabajo empresarial
- Producto educativo
- Asistente integrado
- API para desarrolladores
- Producto personalizado basado en conocimiento

El nombre del producto por sí solo no describe toda la configuración.

4.3. Capa interfaz de usuario

Es la interfaz a través de la cual el usuario interactúa con el producto. Web Móvil Escritorio Extensión de navegador Integración con sistema operativo Interfaz integrada en la plataforma

API

Portal empresarial Diferentes interfaces del mismo producto:

- diferentes funciones,
- diferente modelo,
- diferente representación de fuente,
- memoria diferente

puede ser usada.

4.4. Capa de Modelo o Motor de Inferencia

Es el modelo, familia de modelos o sistema de enrutamiento usado en la generación de respuestas. ID del modelo:

- abierto,
- parcialmente visible,
- desconocido,
- dinámico

Es posible. Un producto puede dirigir una consulta de usuario a diferentes modelos. En este caso, no se puede decir: 'Se midió un único modelo fijo.'

4.5. Capa de Orquestación e Información

Puede incluir lo siguiente:

- Recuperación
- Búsqueda web en vivo
- Búsqueda interna
- Clasificación de fuentes
- Reclasificación
- Selección de citas
- Llamar a un vehículo
- Reescritura de consultas

- Triturado de documentos
- Fusión de respuestas

Esta capa es de importancia fundamental para la representación de GEO. Porque puede afectar qué registro:

- es accedido,
- es seleccionado,
- es incluido en la respuesta,
- se muestra como referencia

puede ser influenciado.

4.6. Capa de Política y Personalización

Puede incluir lo siguiente:

- Reglas de seguridad
- Instrucciones del sistema
- Política de producto
- Memoria
- Instrucciones especiales
- Historial del usuario
- Reglas regionales
- Edad o clase de cuenta
- Configuraciones del administrador corporativo

Esta capa afecta el:

- alcance de la respuesta,
- su tono,
- su rechazo,
- su recomendación

puede cambiar.

4.7. Capa de Distribución y Tiempo

Indica cuál de los siguientes estados tiene el producto en un momento específico:

- es significativo con
- característica experimental,
- distribución regional,
- distribución por segmento de usuarios,
- prueba A/B

El mismo nombre de producto puede tener diferentes condiciones de distribución para distintos usuarios en un mismo día.

5. EJEMPLO DE PRODUCTO DE IA

La unidad del sistema central dentro de GEO-1000:

Instancia de Producto de IA

Término operativo canónico:

Instancia de Producto de IA

Una Instancia de Producto de IA se define como:

```
A_i= (v, p, s, r, m, k, t, x, u, c, l, ω)
```

puede representarse como. Aquí:

- v: proveedor
- p: producto
- s: interfaz de usuario
- r: plan o clase de cuenta
- m: ID o etiqueta del modelo
- k: recuperación y estructuración de información
- t: configuración de herramientas
- x: memoria, personalización y estado de sesión
- u: clase de configuración del usuario u organización
- c: país y región
- l: idioma
- ω: tiempo, implementación y estado de características experimentales

Una observación: Oj debe estar vinculado a un específico: muestra de producto Ai.

6. DIFERENCIA ENTRE MUESTRA DE PRODUCTO Y NOMBRE DE PRODUCTO

El nombre del producto puede ser fijo. La muestra de producto puede variar. Ejemplo: Deja que el nombre del producto Orion AI permanezca sin cambios. Sin embargo, estas dos mediciones son muestras de producto diferentes:

Muestra de Producto 1

Plan Web Gratis Memoria desactivada Acceso web desactivado Nuevo chat Turco Turquía Onda 1

Muestra de Producto 2

Plan de pago móvil Memoria en acceso web en contexto de chat anterior disponible Turco Turquía Wave 1 Estos pueden existir bajo la misma marca. No son la misma celda experimental.

7. NIVELES DE CONFIANZA DE IDENTIDAD DEL MODELO

El nombre del modelo no siempre se puede conocer con la misma certeza. Por lo tanto, la identidad del modelo puede registrarse en los siguientes niveles.

MI-0— MODELO DESCONOCIDO

Se desconoce qué modelo o familia de modelos utiliza el producto. El resultado se atribuye únicamente al interfaz del producto.

MI-1— ETIQUETA MOSTRADA AL USUARIO

Una etiqueta del modelo es visible en la interfaz de usuario. La etiqueta se registra. No se asume que verifique completamente la implementación del back-end.

MI-2— FAMILIA DE MODELO DOCUMENTADA POR EL PROVEEDOR

Los documentos del producto del proveedor describen una familia de modelos específica o la relación producto–modelo. Aún puede desconocerse si la observación individual se dirige exactamente a ese modelo.

MI-3— DESPLIEGUE CONFIRMADO POR SESIÓN

La sesión verifica una distribución específica de API o metadatos de producto confiables.

MI-4— PUNTO FINAL O DESPLIEGUE VERSIONADO

Se puede registrar un ID de versión específica, punto final o despliegue. Es más sólido para comparaciones controladas.

MI-5— ARTEFACTO DE MODELO CONGELADO REPRODUCIBLE

Existe una estructura de modelo o despliegue estática y reproducible. Puede ser rara en productos de consumo activos. Si la confianza en el ID del modelo es baja, el resultado:

debería atribuirse al comportamiento del producto

No se puede atribuir definitivamente a un modelo específico.

8. LÍMITE DEL NOMBRE DE MODELO APARENTE

Ver el nombre del modelo en una interfaz de usuario es valioso. Sin embargo, las siguientes preguntas pueden seguir sin resolverse: ¿Todas las consultas se dirigen al mismo modelo? ¿Cambia el modelo según la dificultad de la consulta? ¿Utilizan seguridad o recuperación un modelo separado? ¿El proveedor aplica la misma etiqueta a diferentes distribuciones? ¿La etiqueta del modelo solo indica la familia del producto? ¿Forma parte el usuario de un despliegue experimental? Por lo tanto, el informe correcto podría ser: "Se mostró la etiqueta del Modelo X en la interfaz de usuario." Sin evidencia más sólida, no debe usarse la siguiente oración: "Todas las respuestas son generadas definitivamente únicamente por el Modelo X."

9. TIPO DE PRODUCTO

Cada producto de IA debe clasificarse con un tipo primario y tipos secundarios.

PT-1— Producto de Conversación General

Es un producto general de chat y preguntas y respuestas.

PT-2— Producto Basado en Búsqueda o Recuperación

La búsqueda o recuperación es un componente fundamental en la generación de respuestas.

PT-3— Producto de Conocimiento Empresarial

Puede utilizar documentos internos, espacios de trabajo o fuentes de datos especiales.

PT-4— Asistente Integrado

Está integrado en otro producto o plataforma.

PT-5— Producto para Desarrolladores o API

Es utilizado por un desarrollador o aplicación en lugar del usuario final.

PT-6— Producto Específico de Dominio

Está diseñado para una profesión o industria específica.

PT-7— Producto Multimodal

Las entradas distintas al texto, como entradas visuales, de audio o de documentos, forman parte de la experiencia principal. Un producto puede tener más de un tipo. La comparación debe realizarse dentro del mismo tipo o con un alcance explícito.

10. REGISTRO interfaz de usuario

Las diferentes interfaces del mismo producto deben registrarse por separado. Áreas de interfaz:

- ID de interfaz
- Web/móvil/escritorio/API
- Versión de la aplicación
- Sistema operativo, si corresponde
- Navegador, si corresponde
- Idioma del producto
- Selección de modelo visible
- Disponibilidad de herramientas
- Vista de citas
- Función de carga de archivos
- Configuraciones de memoria
- Configuraciones de gestión corporativa
- Funciones experimentales
- Tiempo de medición

Si se encuentra que una interfaz web y móvil son materialmente iguales, los resultados pueden combinarse. Esta igualdad no debe asumirse; debe probarse.

11. PLAN Y CLASE DE CUENTA

El tipo de plan puede tener características que pueden afectar la respuesta del usuario. Estados del plan:

- Gratis
- Individual de pago
- Profesional
- Corporativo
- Educación
- Desarrollador
- Prueba
- Invitación o acceso limitado

- Desconocido

Los planes pueden diferir en las siguientes áreas:

- Modelo
- Límite de uso
- Acceso web
- Memoria
- Herramientas
- Citación de fuentes
- Longitud de respuesta
- Velocidad o prioridad
- Datos empresariales
- Configuraciones de seguridad y retención

Si no hay información del plan:

El PLAN_DESCONOCIDO

los resultados gratuitos y de pago no pueden combinarse silenciosamente como el mismo comportamiento del producto.

12. RECUPERACIÓN Y ACCESO WEB

Se puede generar una respuesta de IA en uno de tres estados básicos de conocimiento.

KG-0— MODO DE CONOCIMIENTO INDETERMINADO

No se puede determinar el caso de recuperación o uso de la web.

KG-1— RESPUESTA INTERNA DEL MODELO

Una respuesta que se conoce o se asume fuertemente que no utiliza información en vivo o externa.

KG-2— RESPUESTA ASISTIDA POR RECUPERACIÓN

Puede haberse utilizado recuperación de fuentes preindexadas o de productos.

KG-3— RESPUESTA ASISTIDA POR LA WEB O BÚSQUEDA EN VIVO

Se ha utilizado acceso a la web/búsqueda en tiempo real o casi en tiempo real.

KG-4— RESPUESTA FUNDAMENTADA EN CONOCIMIENTO PRIVADO

Se ha utilizado una fuente de datos privada proporcionada por la institución o el usuario.

KG-5— RESPUESTA HÍBRIDA O DE MÚLTIPLES FUENTES

Se han utilizado varias capas de información juntas. Estos modos pueden producir resultados diferentes en el mismo prompt. Deben distinguirse claramente en la comparación.

13. EL ACCESO A RECUPERACIÓN NO ES LO MISMO QUE EL USO DE RECUPERACIÓN

El producto puede tener una función de búsqueda web. Sin embargo, puede que no se haya utilizado en la respuesta específica. Tres campos deben registrarse por separado: ¿Estaba disponible la función? ¿Estaba activada o seleccionable la función? ¿Había evidencia de que se utilizó en la respuesta específica? Ejemplo de estados:

DISPONIBLE_NO_USADO

USO_DISPONIBLE_DESCONOCIDO

USADO_CONFIRMADO

NO_DISPONIBLE

DESHABILITADO

DESCONOCIDO

La mera presencia de un enlace de fuente puede ser un fuerte indicador de uso. La ausencia de un enlace de fuente no prueba necesariamente que no se haya utilizado la recuperación.

14. CAPA DE CITAS

Usar una fuente no es lo mismo que mostrar una cita. Un producto:

- puede usar una fuente pero puede que no la muestre,
- puede mostrar una cita, pero la cita puede no respaldar la afirmación,
- puede vincular la misma fuente a diferentes oraciones,

El usuario puede ocultar la cita de acuerdo con el interfaz de usuario. El registro de la cita debe contener los siguientes campos: ¿Está disponible la función de cita? ¿Se mostró la cita en la respuesta específica? Número de citas Nombres de campos de origen Coincidencia cita–afirmación Visibilidad en el interfaz de usuario Expansibilidad de la cita Accesibilidad de la fuente Tiempo de generación de la cita Los productos que muestran y no muestran citas no se pueden comparar directamente en términos de "número de fuentes". Primero, se deben explicar las diferencias en el comportamiento del producto y en la interfaz.

15. ACCESO A HERRAMIENTA

El producto de IA puede acceder a las siguientes herramientas:

- Búsqueda en la web
- Navegador
- calculadora
- Ejecución de código
- Lectura de archivos
- Mapa

- Calendario
- Base de conocimientos corporativa
- Catálogo de productos
- Otras APIs

El acceso a herramientas puede afectar tu respuesta:

- su oportunidad,
- evidencia,
- alcance de la respuesta,
- precisión

Registro de herramientas:

- si la herramienta está disponible,
- si está activa,
- si fue utilizada en la respuesta,
- si se conserva el resultado de la herramienta

debe mostrarse.

16. MEMORIA, INSTRUCCIONES ESPECIALES Y PERSONALIZACIÓN

La respuesta visible para el usuario puede verse afectada por los siguientes campos:

- Memoria persistente
- Historial de chat
- Instrucciones especiales
- Nombre de usuario o perfil
- Conversaciones previas de la marca
- Preferencias
- País y localidad
- Configuraciones del espacio de trabajo corporativo
- Segmentación realizada por el proveedor

Estos campos deben registrarse en tres capas separadas.

16.1. Personalización persistente

Información protegida a través de cuentas o sesiones.

16.2. Contexto en la conversación

Mensajes previos dentro del chat actual.

16.3. Configuración del sistema u organización

Administrador o instrucciones del producto que el usuario puede no ver. La información completa puede no estar siempre disponible. Las situaciones desconocidas deben registrarse explícitamente.

17. TIPOS DE SESIÓN

Estados de la sesión del candidato:

ST-1— NUEVA CONVERSACIÓN ESTÁNDAR

Nueva conversación estándar. La memoria o las instrucciones especiales se registran por separado.

ST-2— CONVERSACIÓN TEMPORAL O EFÍMERA

Sesión temporal destinada a separarse del historial permanente. Las características reales deben verificarse en el registro del producto.

ST-3— CONVERSACIÓN EXISTENTE

Existe un contexto de conversación previo. Puede ser valioso para el panel natural. Se mantiene separado en el panel controlado.

ST-4— SESIÓN DE EMPRESA O DE ESPACIO DE TRABAJO COMPARTIDO

Puede haber un área de información empresarial o compartida.

ST-5— ESTADO DE SESIÓN DESCONOCIDO

No se pudo determinar de manera confiable el estado de la sesión. Si se van a combinar diferentes tipos de sesión en la misma celda, se debe explicar la justificación y el peso.

18. POLÍTICA DEL PRODUCTO Y RECHAZO

Un producto de IA puede no responder a una pregunta específica debido a:

- de seguridad,
- ley,
- política del producto,
- reglamento regional,
- no implica ninguna idoneidad en términos de edad,
- configuración empresarial

razón. Rechazo:

- deficiencia del modelo básico,
- política del producto,
- interfaz,
- puede deberse a la cuenta del usuario

GEO-1000 registra en los controles principales del usuario el resultado del rechazo que llega al usuario. Sin embargo, si se conoce la razón del rechazo, se clasifica adicionalmente. La capa de política de un producto también forma parte de la experiencia del usuario.

19. FUNCIONALIDADES EXPERIMENTALES Y DISTRIBUCIONES A/B

Los productos de IA en vivo pueden no ofrecer las mismas funciones a todos los usuarios al mismo tiempo. Los participantes:

- diferente interfaz de usuario,
- distintas formulaciones del prompt según el modelo,
- diferente función web,
- diferente formato de respuesta

puede ver. Indicadores de estado experimental:

- Etiqueta Beta en la interfaz
- Documentación del proveedor
- Diferencia de función
- Diferencia sistemática en la interfaz de usuario entre usuarios
- Comportamiento que rompe durante la medición

puede ocurrir. Si la función experimental es desconocida:

ESTADO_EXPERIMENTAL_DESCONOCIDO

La variación real dentro del producto no es un defecto de datos. Es una característica de distribución del sistema. Sin embargo, debería considerarse en la comparabilidad.

20. Registro del Sistema de IA

Se debe crear un Registro del Sistema de IA para cada producto auditado. El registro debe tener tres niveles.

20.1. Registro Maestro del Proveedor y Producto

Incluye áreas relativamente permanentes:

- Proveedor
- Nombre del producto
- Tipo de producto
- Interfaces principales de usuario
- Países con acceso oficial
- Clases de plan
- Registros de idiomas soportados
- Características generales

20.2. Registro de Producto de oleada de medición

Para una onda específica:

- versión del producto,
- características del plan,
- etiquetas del modelo,
- funciones web y de herramientas,

- comportamiento de la memoria,
- estado de acceso,
- actualizaciones conocidas

que se registra.

20.3. Registro del Sistema a Nivel de Observación

Realmente observado para cada participante:

- interfaz,
- etiqueta del modelo,
- plan,
- uso del vehículo,
- cita,
- sesión,
- personalización,
- tiempo

lleva campos. Solo el registro principal no puede eliminar diferencias a nivel de observación.

21. HUELLA DE CONFIGURACIÓN DEL SISTEMA

Se puede crear una huella digital de configuración del sistema para facilitar la comparabilidad de una instancia de producto de IA. Campos candidatos:

```
SCF= H(canonicalise(p, s, r, m, k, t, x, c, l, w))
```

Aquí, H es la función que produce un valor de integridad a partir de los campos versionados. La huella digital:

- si dos observaciones pertenecen a la misma configuración,
- si ha habido cambios materiales en la ola

Puede ayudar a identificar. Huella digital: no demuestra el modelo de backend real, no hace visibles capas que el proveedor no haya divulgado. Solo refuerza si la configuración registrada ha sido modificada posteriormente.

22. ¿QUÉ ES LA COMPARABILIDAD?

Se consideran comparables dos resultados cuando:

> *La diferencia entre ellos puede explicarse suficientemente por el comportamiento monitoreado del producto en lugar de diferencias irrelevantes en el diseño de la medición.*

La comparabilidad no es igualdad absoluta. Diferentes productos de IA:

- modelo,
- proveedor,
- interfaz de usuario.

Estos no tienen por qué ser todos idénticos. Sin embargo, las condiciones materiales relevantes para la comparación deben ser:

- igualación,
- estratificación,

- registro,
- modelado en el análisis

son requeridos.

23. EL OBJETO DE LA COMPARACIÓN DEBE DEFINIRSE DE ANTEMANO

Los mismos datos pueden responder a diferentes preguntas de comparación.

23.1. Comparación de Productos de Usuario

"¿Qué producto de consumo produce una representación más precisa bajo las mismas condiciones de usuario?" El objeto que se compara es toda la cadena de producto.

23.2. Comparación de Modelos

"¿Qué modelo produce una respuesta más precisa bajo la misma solicitud controlada y condición sin herramientas?" Puede ser necesario el acceso a API o a modelo fijo.

23.3. Comparación de Recuperación

“¿Hay alguna diferencia entre los modos accesible por web y no accesible del mismo producto?” Se mantiene el modelo y otras condiciones lo más constantes posible.

23.4. Comparación de Planes

“¿Ven los usuarios gratuitos y los de pago la misma representación de la marca?” La diferencia de plan es el objeto de la auditoría.

23.5. Comparación de Interfaz

“¿El producto web y móvil produce representaciones diferentes bajo las mismas condiciones?”

23.6. Comparación de Tiempo

“¿Cambió la representación del mismo producto en diferentes olas?” Se deben separar los cambios de producto y de población. Sin especificar la pregunta de comparación: “El Producto A es mejor que el Producto B.” es excesivamente amplio.

24. DIMENSIONES DE COMPARABILIDAD

Al comparar dos productos u olas, se deben evaluar las siguientes áreas:

- Entidad auditada
- Versión del paquete de realidad
- ID del prompt y versión
- Idioma y localización
- País
- Población objetivo
- Fuente de muestra
- Composición de participantes
- Tipo de sesión

- Memoria
- Instrucciones especiales
- Conversación previa
- Plan del producto
- interfaz de usuario
- Etiqueta del modelo
- Estado de recuperación/web
- Herramientas
- Función de citación
- Tiempo de medición
- Actualizaciones del producto
- Regla de adjudicación
- Estados del resultado
- Ponderación

No todas las dimensiones necesitan ser iguales. Diferentes dimensiones limitan el alcance de la comparación.

25. NIVELES DE COMPARABILIDAD

CP-0— INCOMPARABLE

Las condiciones de medición del material son desconocidas o gravemente inconsistentes. No se puede hacer una comparación directa del rendimiento.

CP-1— COOCURRENCIA DESCRIPTIVA

Hay resultados de dos productos o olas. Sin embargo:

- población,
- prompts,
- tiempo,
- sesión

no son iguales. Solo se puede hacer una visualización descriptiva lado a lado.

CP-2— PARCIALMENTE ALINEADO

Algunos campos principales coinciden. Las diferencias materiales continúan. Se puede hacer una comparación limitada.

CP-3— COMPARACIÓN COINCIDENTE ENTRE USUARIO Y INTERFAZ

Misma entidad, misma versión del prompt, mismo idioma/país, población de usuarios similar, misma ventana de tiempo. Las características del producto registrado están disponibles. Los resultados del producto de consumo son comparables.

CP-4— COMPARACIÓN COINCIDENTE CONTROLADA

Además:

- sesión,
- memoria,

- personalización,
- recuperación,
- tools,
- plan

están coincidentes de manera controlada o abierta. Proporciona una comparación más sólida de producto o característica.

CP-5— COMPARACIÓN REPLICADA Y CALIBRADA

Comparación:

- en múltiples olas,
- con usuarios o paneles independientes,
- con la misma evaluación normativa

repetido; la variabilidad y los cambios de producto han sido calibrados. Estos niveles no son un "puntuación de precisión científica". Indican el límite que puede tener una afirmación de comparación.

26. ¿QUÉ HACER EN LA CONDICIÓN CP-0?

Si no se pueden comparar dos productos, los resultados no se ignoran. Pueden informarse por separado:

- "Este es el resultado Alcance natural del Producto A."
- "Este es el resultado del Producto B bajo diferentes usuarios y condiciones del producto."

Sin embargo: no se puede decir, "A es 12 puntos mejor que B." La incomparabilidad no es un fallo, es una limitación del método.

27. COMPARABILIDAD PARA EL MISMO PROVEEDOR

Los resultados del mismo proveedor no se pueden comparar automáticamente. Pueden estar presentes las siguientes diferencias:

- Producto
- Plan
- Modelo
- Acceso web
- Interfaz
- Memoria
- Datos empresariales
- Distribución por país
- Característica experimental
- Hora

La comparación del mismo proveedor también debe evaluarse según el nivel de CP.

28. COMPARABILIDAD PARA DIFERENTES PROVEEDORES

Productos de usuario pertenecientes a diferentes proveedores:

- misma población objetivo,
- mismo prompt,
- mismo tiempo,
- similar interfaz de usuario,
- configuración de producto clara

Es comparable por debajo. Ser un proveedor diferente no impide la comparación. Sin embargo, las diferencias en las funcionalidades determinan qué mide la comparación. Ejemplo: El Producto A utiliza la web en vivo. El Producto B no. El resultado puede responder a esta pregunta: '¿Cuál de las dos cadenas de productos ofrecidas al usuario produce una representación más precisa?' No responde directamente a esta pregunta: '¿Qué modelo base tiene más conocimiento?'

29. COMPARACIÓN DE PRODUCTOS CON SOPORTE WEB Y SIN SOPORTE WEB

Son posibles dos comparaciones separadas.

29.1. Comparación de Productos Naturales

Cada producto se mide bajo la condición de usuario real por defecto. Un producto que utiliza la web utiliza la web. Uno que no la utiliza, no lo hace. Resultado: es una comparación de la experiencia real del consumidor con el producto.

29.2. Comparación de Funciones Equivalentes

Si es posible, ambos productos:

- medidos con la web abierta,
- web cerrada

modos por separado también. Esto puede separar mejor el efecto de recuperación. Cada producto puede no ofrecer las mismas opciones. Modo ausente:

REGISTRADO COMO NO DISPONIBLE

Ese estado debe permanecer explícito.

30. PRODUCTOS QUE MUESTRAN Y NO MUESTRAN CITAS

Para un producto sin función de citación: es posible dar una “Cita cero”. Sin embargo, su significado debe explicarse: el producto puede no estar haciendo referencia a fuentes. No es seguro que no use ninguna fuente. La cobertura de citas y la cobertura de precisión son resultados separados. La comparación de citas solo debe hacerse entre productos que soporten citas o con estado de función separado.

31. Alcance natural COMPARACIÓN

Cada producto se mide dentro de su población efectivamente alcanzada. Ventaja: muestra el impacto global natural del producto. Limitación: las poblaciones son diferentes, las distribuciones por país e idioma pueden variar. Puntuaciones de Alcance natural: muestran lo que hace el producto dentro de su población alcanzada. No deben usarse solas para un ranking directo de superioridad de productos.

32. COMPARACIÓN SOBRE UNA BASE COMÚN

Se utiliza la población de acceso común de los productos comparados. Ventaja:

- mismo país,
- misma lengua,
- misma condición de usuario

proporciona una comparación más justa. Limitación: preserva la diferencia real de acceso global de los productos, el universo común puede ser muy reducido. La puntuación Base común de comparación debe publicarse junto con la puntuación Alcance natural.

33. COMPARACIÓN DE USUARIOS PAREADOS

Si el mismo usuario prueba dos o más productos de IA, se pueden controlar las características del usuario. Sin embargo:

- orden de las tareas,
- memoria de respuestas,
- familiaridad con el producto,
- perfil especial de usuarios que pueden acceder a todos los productos

Puede crear sesgo. En un diseño pareado:

- el orden de los productos debe aleatorizarse,
- las respuestas previas no deben compartirse,

la dependencia del mismo usuario debe mantenerse en el análisis.

34. COMPARACIÓN DE USUARIOS INDEPENDIENTES

Cada producto es probado por diferentes usuarios, pero seleccionados del mismo marco poblacional. Ventaja: se reduce la contaminación de tareas, se pueden elegir usuarios de producto más naturales. Limitación: los grupos de usuarios pueden no ser completamente iguales. Se requiere ponderación y estratificación. Los resultados de usuarios emparejados e independientes pueden ser separados pero complementarios.

35. COMPARABILIDAD EN EL TIEMPO

Al comparar el resultado de hoy de un producto de IA con resultados pasados, se deben examinar los siguientes cambios:

- Modelo
- Versión del producto

- Web/recuperación
- Plan
- Acceso por país
- Soporte de idiomas
- Prompt
- Realidad de la entidad
- Adjudicación
- Marco poblacional
- Ponderación

La diferencia de tiempo puede no surgir únicamente del producto de IA.

36. TIPOS DE CAMBIOS EN EL PRODUCTO

UC-1— CAMBIO MENOR DOCUMENTADO

Un cambio menor divulgado por el proveedor y evaluado como que no altera materialmente el objeto de medición.

UC-2— CAMBIO MATERIAL DOCUMENTADO

Cambio material descrito en el modelo, recuperación, interfaz o característica.

UC-3— CAMBIO DE COMPORTAMIENTO OBSERVADO

Incluso si no hay un anuncio oficial, se observa un comportamiento sistemático o un cambio de interfaz.

UC-4— CAMBIO DE ACCESO O PLAN

Cambios en el país, plan o acceso del usuario.

UC-5— EVENTO DE CAMBIO DESCONOCIDO

Hay una interrupción en el comportamiento de la medición. No se pudo verificar la razón.

UC-6— REEMPLAZO O RETIRO DEL PRODUCTO

Un producto o modelo antiguo ha terminado y ha sido reemplazado por otra estructura. El tipo de cambio limita la comparación de series temporales.

37. CAMBIO DE PRODUCTO DENTRO DE LA OLA DE MEDICIÓN

Si ocurre un cambio de material durante una ola, existen las siguientes opciones: Se puede detener la ola y reiniciarla. Se pueden crear sub-olas separadas antes y después del cambio. El cambio se puede modelar como un factor experimental. La ola se etiqueta como MIXED_SYSTEM_STATE. Se prohíbe la siguiente estructura: Combinar todas las respuestas como si resultaran en un único resultado de producto fijo a pesar de conocer el cambio.

38. EL COMPORTAMIENTO PUEDE CAMBIAR SIN QUE HAYA CAMBIO DE PRODUCTO

Incluso si el proveedor no anuncia una actualización explícita, el resultado puede cambiar. Posibles razones:

- Cambio en el índice de recuperación
- Actualización de recursos
- Aparición de contenido nuevo en la web
- Cambio en la reordenación
- prueba A/B
- Implementación regional
- Política de seguridad
- Enrutamiento de tráfico

Por lo tanto, no se puede decir: "El proveedor no actualizó, el sistema era definitivamente el mismo." También deben examinarse la evidencia de medición y el desglose del comportamiento.

39. PRODUCTO EN VIVO Y MODELO CONGELADO

Existen dos objetos de investigación diferentes.

39.1. Auditoría de Producto en Vivo

Mide lo que los usuarios reales ven en la interfaz real del consumidor. Ventaja:

- muestra la experiencia real,
- fuente real,
- política real,
- interfaz real

limitaciones: es variable, la reproducción completa puede ser difícil.

39.2. Auditoría de Modelo Congelado

Se realiza repetición controlada bajo una versión fija o punto final. La ventaja:

- proporciona reproducibilidad,
- control experimental

La limitación: puede que no represente completamente el producto de consumo real. La capa principal de población de NOMOS GEO-1000 se basa en el producto de usuario en vivo. La auditoría del modelo congelado es un trabajo complementario de laboratorio.

40. NO DISPONIBILIDAD DEL PRODUCTO Y RESULTADOS TÉCNICOS

Se deben mantener separados los siguientes casos:

AR-1— PRODUCTO NO DISPONIBLE

El producto no está disponible para el país o grupo de usuarios. Es el resultado del alcance de acceso.

AR-2— CUENTA O PLAN NO ELEGIBLE

La condición de plan o cuenta asignada al participante no es elegible. Puede haber un problema de muestreo o asignación de tareas.

AR-3— FALLA TÉCNICA TEMPORAL

Ha ocurrido un error técnico temporal. El protocolo de reintento debe estar predefinido.

AR-4— LÍMITE DE TARIFA O USO

Se ha alcanzado una cuota de uso o un límite de velocidad. Puede ser parte de la experiencia del usuario real.

AR-5— RECHAZO DEL SISTEMA

La tarea del producto ha sido rechazada debido a la política o contenido. Es un resultado válido del sistema.

AR-6— SIN RESULTADO ÚTIL

No se ha generado ninguna respuesta útil.

AR-7— FALLO DE ACCESO DESCONOCIDO

Por qué no se pudo determinar. Estas situaciones no se pueden combinar en el mismo cuadro de "sin respuesta".

41. REGLA DE REINTENTO TÉCNICO

Se puede hacer un reintento en caso de un error técnico temporal o interrupción de la conexión. Sin embargo:

- si la respuesta de la IA no gusta,
- ocurrió un rechazo,
- se produjo información incorrecta

no se puede realizar un reintento por estas razones. Reintento técnico:

- mismo usuario,
- mismo producto,
- misma configuración,
- debe realizarse dentro del período predeterminado.

También debe registrarse el primer incidente técnico.

42. MATRIZ DE COMPARACIÓN

Un registro de comparación básico para dos productos podría ser el siguiente:

Dimensión	Producto A	Producto B	Estado
Entidad objetivo	Igual	Igual	Coincide
Versión del prompt	1.0	1.0	Coincide
Idioma/ubicación	tr-TR	tr-TR	Coincide
País	Turquía	Turquía	Coincide

Dimensión	Producto A	Producto B	Estado
Ventana de tiempo	Onda 1	Onda 1	Coincide
Panel de usuario	Población	Población	Coincide
Plan	Gratis	Pagado	Diferente
Interfaz	Web	Móvil	Diferente
Acceso web	Cerrado	Abierto	Diferente
Memoria	Cerrado	Abierto	Diferente
ID del modelo	MI-1	MI-1	Limitado
Nivel de comparación			CP-2

Estos resultados se pueden mostrar lado a lado. Sin embargo: "El modelo básico del Producto B es mejor." no se puede decir. La diferencia puede provenir de toda la cadena del producto.

43. CASO DE COMPARACIÓN SINTÉTICO APPLE.COM

DEMOSTRACIÓN DE METODOLOGÍA SINTÉTICA / Los productos, respuestas y resultados a continuación son completamente ficticios. No representan el desempeño de la verdadera Apple Inc. ni de ningún proveedor real de IA. Consideremos tres ejemplos de productos sintéticos.

Ejemplo de Producto A

Proveedor: Orion Producto: Orion Chat Interfaz: Web Plan: Gratis Etiqueta del modelo: Orion Standard Confianza del modelo: MI-1 Acceso web: Desactivado Memoria: Desactivada Sesión: Nueva Idioma: Turco Localización: Turquía Onda: W1 Respuesta: "Apple es una empresa tecnológica estadounidense que produce teléfonos y computadoras."

Muestra de Producto B

Proveedor: Orion Producto: Orion Chat Interfaz: Móvil Plan: De pago Etiqueta del modelo: Orion Pro Confianza del modelo: MI-1 Acceso web: Abierto Citación: Abierta Memoria: Activada Conversación previa: Sobre Apple Idioma: Turco Configuración regional: Turquía Ola: W1 Respuesta: "Apple es una empresa tecnológica global que proporciona hardware, software y servicios digitales. La información corporativa actual se puede encontrar en estas fuentes."

Muestra de producto C

Proveedor: Helios Producto: Helios Búsqueda Interfaz: Web Plan: Gratis Etiqueta del Modelo: Desconocido Confianza del Modelo: MI-0 Acceso Web: Abierto Cita: Abierta Memoria: Ninguna Sesión: Nueva Idioma: Turco Configuración Regional: Turquía Ola: W1 Respuesta: "Apple es una empresa tecnológica con sede en EE. UU. que desarrolla dispositivos de consumo, sistemas operativos y servicios digitales."

43.1. Comparación de A y B

Son el mismo proveedor. Sin embargo:

- plan,
- interfaz,
- etiqueta del modelo,
- web,
- memoria,
- conversación previa

es diferente. El nivel de comparación puede ser CP-1 o CP-2. La diferencia de respuesta no puede atribuirse únicamente a la diferencia del modelo básico.

43.2. Comparación de A y C

Son proveedores diferentes. Sin embargo:

- nueva sesión,
- plan gratuito,
- interfaz web,
- mismo idioma/país,
- misma ola

en términos de algunas condiciones coinciden. El acceso web es diferente. Comparación: se puede hacer como un resultado natural de producto de consumo. No se puede hacer como capacidad del modelo básico.

43.3. Comparación de B y C

Ambos productos tienen acceso web y citas habilitadas. Sin embargo, B tiene memoria y conversación previa, mientras que C no. La experiencia del usuario de los productos es comparable. No es una comparación de producto controlada y limpia.

43.4. Repetición Controlada

En el producto B:

- la memoria está desactivada,
- se inicia un nuevo chat,
- las instrucciones especiales están desactivadas,

la web permanece activada. C se mide bajo la misma condición. Nivel de comparación: CP-3, o CP-4 con control adicional suficiente. Este ejemplo muestra:

> *El mismo proveedor no proporciona una comparación cercana; proveedores diferentes bien registrados pueden ser más comparables.*

44. CASO DE CAMBIO DE PRODUCTO SINTÉTICO

EJEMPLO SINTÉTICO

Orion Chat W1 apunta con ondas a las observaciones 1,000. En las primeras observaciones 480:

- versión del producto: 6.1
- función de la web: manual
- atribución: limitada

En las siguientes observaciones 520:

- versión del producto: 6.2
- función de la web: automática
- atribución: extendida

dejarlo así. Resultado único: si se calcula como “puntuación Orion Chat W1,” se mezclan dos estados diferentes del sistema. Opciones correctas:

Opción 1

W1A: primer 480 W1B: siguiente 520 se informan por separado.

Opción 2

La ola no continúa. Se toman nuevas observaciones 1,000 en la versión 6.2.

Opción 3

La versión se mantiene como un factor experimental explícito en el análisis estadístico. Opción incorrecta: La actualización se elimina del informe y todas las observaciones se presentan como un solo producto fijo.

45. REGISTRO MÍNIMO DEL SISTEMA DE IA

Cada observación debe contener al menos los siguientes campos:

- ID del sistema de IA NOMOS
- Proveedor
- Producto
- Interfaz
- Plan
- Tipo de producto
- Etiqueta de modelo aparente
- Nivel de confianza del modelo
- País
- Localización
- Idioma
- Tipo de sesión
- Estado de la memoria
- Estado de instrucciones personalizadas
- Estado de la conversación previa
- Función de recuperación/web
- Estado de uso de la recuperación
- Estado de la citación
- Estado de la herramienta
- Estado de la función experimental
- UTC tiempo
- Onda
- Huella del sistema
- Cambios de eventos
- Fuente de evidencia
- Propietario del registro

Si el campo de material es desconocido, no se puede dejar en blanco. Se debe usar el estado DESCONOCIDO.

46. TARJETA DE COMPARACIÓN MÍNIMA

Al comparar dos productos o oleadas, el registro público debe mostrar:

- Pregunta de comparación
- Nivel de comparación
- Población objetivo común
- Diferencias Alcance natural
- Igualdad del usuario
- Idioma y país
- Tiempo de medición
- Producto, plan y interfaz
- Confianza en la identidad del modelo
- Diferencia web/recuperación
- Diferencia en memoria y personalización
- Diferencia en herramienta
- Diferencia en muestra
- Método de evaluación
- Limitaciones del material
- Métricas directamente comparables
- Áreas no comparables

No se debe publicar un solo número o clasificación sin esta tarjeta.

47. REGLA DE COMPARABILIDAD

Un auditor debe responder estas cuatro preguntas: ¿Se pidió lo mismo? ¿Se midió el mismo universo humano? ¿Se aplicaron las mismas condiciones del usuario? ¿Se usaron el mismo tiempo y sistema de evaluación? Cada una de estas preguntas:

- completamente,
- parcialmente,
- no,
- desconocido

se clasifica como. El nivel de comparación se da en consecuencia.

48. DISPOSICIONES NORMATIVAS OBLIGATORIAS

CH09-N01

Cada observación GEO-1000 debe estar vinculada a un registro con versión en el Registro del Sistema de IA.

CH09-N02

El nombre del proveedor no puede usarse como la identidad completa del producto de IA medido.

CH09-N03

El producto, interfaz de usuario, plan, modelo, recuperación, herramienta, sesión y capas de personalización deben estar separadas entre sí.

CH09-N04

La etiqueta del modelo en el interfaz de usuario no puede ser presentada como una identidad de modelo de back-end verificada.

CH09-N05

El nivel de confianza de la identidad del modelo debe registrarse desde MI-0 hasta MI-5 o mediante un estado equivalente definido de manera abierta.

CH09-N06

Si la identidad del modelo no puede ser verificada, el resultado solo debe atribuirse al producto y al interfaz de usuario.

CH09-N07

Los productos que pertenecen al mismo proveedor no se consideran automáticamente la misma célula experimental.

CH09-N08

Los productos pertenecientes a diferentes proveedores pueden ser comparables cuando se igualan el usuario del material y las condiciones experimentales.

CH09-N09

Las interfaces de productos y API web, móviles, de escritorio e integradas no se pueden combinar sin descripción.

CH09-N10

Los planes gratuitos, de pago, corporativos, educativos y para desarrolladores deben registrarse como términos de producto separados.

CH09-N11

Si el plan es desconocido, el resultado debe llevar el estado PLAN_DESCONOCIDO.

CH09-N12

La disponibilidad de la función de recuperación o web no puede contarse como prueba de que se usó en una respuesta específica.

CH09-N13

El acceso a la recuperación, el uso de la recuperación y la atribución deben registrarse con estados separados.

CH09-N14

Un producto que no muestre atribución no puede considerarse automáticamente como que no utiliza ningún recurso.

CH09-N15

Las citas del producto citado deben examinarse por separado en términos de soporte de la afirmación.

CH09-N16

El acceso a herramientas y el uso de herramientas en respuestas específicas deben registrarse por separado.

CH09-N17

La memoria, las instrucciones especiales, los chats anteriores y la configuración corporativa deben registrarse por separado tanto como sea posible.

CH09-N18

No se puede asumir que un estado de personalización desconocido sea como una sesión limpia.

CH09-N19

Las sesiones de nuevo chat, chat temporal, chat actual y espacio de trabajo corporativo no se pueden fusionar sin explicación.

CH09-N20

Los resultados de Panel de usuarios en condiciones reales y Panel controlado en condiciones depuradas deben tener un estado de producto separado.

CH09-N21

Si una característica experimental o una distribución A/B en el producto es material, debe conservarse en el registro de observación.

CH09-N22

Si ocurre un cambio material en el producto durante una ola, la ola debería dividirse en subversiones separadas, reiniciarse o el cambio debería modelarse explícitamente en el análisis.

CH09-N23

Un cambio de producto conocido no puede almacenarse como si todas las observaciones provinieran de un único sistema fijo.

CH09-N24

El hecho de que el proveedor no explique una actualización no puede tomarse como prueba definitiva de que el producto no ha cambiado.

CH09-N25

El resultado de un producto de consumo en vivo no puede presentarse como si fuera el resultado de un modelo base congelado.

CH09-N26

El resultado de un modelo congelado en el laboratorio no puede presentarse como el rendimiento de un producto real para el usuario.

CH09-N27

La ausencia del producto, error técnico, límite de uso, rechazo del sistema y falta de resultado deben registrarse con estados separados.

CH09-N28

Una respuesta incorrecta o negativa de la IA no puede ser una razón para reintento técnico.

CH09-N29

El sujeto de la comparación debe definirse antes de recopilar los datos.

CH09-N30

El nivel de comparabilidad de dos productos debe ser visible en el resultado público.

CH09-N31

No se puede generar un ranking definitivo de productos a partir de resultados de nivel CP-0 o CP-1.

CH09-N32

Usar solo el mismo prompt no crea un experimento comparable.

CH09-N33

En las comparaciones, la entidad objetivo, el Truth Pack, el prompt, el idioma, el país, el tiempo, la población, la sesión, el producto y el método de evaluación deben examinarse juntos.

CH09-N34

Los resultados Alcance natural no pueden clasificarse directamente sin explicación porque pertenecen a diferentes poblaciones objetivo.

CH09-N35

El resultado Base común de comparación no puede presentarse como el resultado real de alcance global del producto.

CH09-N36

La comparación de productos soportados por la web y sin web se puede hacer como una comparación de productos por usuario; no puede presentarse como una superioridad de modelo básico.

CH09-N37

En las comparaciones pareadas realizadas con el mismo usuario, se debe mantener el orden del producto y la dependencia del usuario.

CH09-N38

En comparaciones realizadas con diferentes usuarios, la composición de usuarios y las diferencias de peso deben ser visibles.

CH09-N39

La versión del producto o la huella de configuración debe almacenarse en cada ola de medición.

CH09-N40

Debe haber un propietario humano o institucional responsable del registro del sistema y de la decisión de comparabilidad.

49. FORMAS DE FALLA

CH09-F01— CONSIDERANDO AL PROVEEDOR COMO UNA SÓLO IA

Todos los productos, modelos y interfaces del proveedor se presentan como un único sistema.

CH09-F02— CONTAR EL COMPORTAMIENTO DEL PRODUCTO POR NOMBRE DE MODELO

Recuperación, seguridad, interfaz y personalización son ignoradas.

CH09-F03— DEFINITIVAMENTE CONSIDERAR LA ETIQUETA VISIBLE DEL MODELO COMO BACKEND

El uso de enrutamiento o modelos dinámicos se pasa por alto.

CH09-F04— CONSIDERANDO AL MISMO PROVEEDOR COMO COMPARABLE

Las diferencias de plan, interfaz y características están ocultas.

CH09-F05— CONSIDERANDO A DIFERENTES PROVEEDORES COMO NO COMPARABLES

Se rechazan las comparaciones de productos posibles bajo condiciones de usuario emparejadas.

CH09-F06— INTEGRANDO WEB Y MÓVIL

Las diferencias de interfaz y características desaparecen.

CH09-F07— COMBINANDO PLANES GRATUITOS Y DE PAGO

Las diferencias de calidad o características debido al plan se vuelven invisibles.

CH09-F08— CONSIDERANDO QUE LOS PRODUCTOS CORPORATIVOS Y DE CONSUMIDORES SON LOS MISMOS

La información privada y los ajustes del administrador están ocultos.

CH09-F09— TRATANDO EL RESULTADO DE LA API COMO PRODUCTO DE CONSUMO

Se eliminan las diferencias de orquestación y interfaz de usuario.

CH09-F10— SI EXISTE RECUPERACIÓN, CONSIDÉRELA USADA

No se puede verificar la ruta de conocimiento verdadero de la respuesta específica.

CH09-F11— SI NO HAY CITA, CONSIDÉRESE SIN FUENTE

Se ignora la posibilidad de recuperación oculta o invisible.

CH09-F12— SI EXISTE CITA, CONTAR COMO FUENTE CORRECTA

No se examina la coincidencia entre cita y afirmación.

CH09-F13— CONSIDERAR EL ACCESO A HERRAMIENTAS COMO USO DE HERRAMIENTA

Se desconoce qué herramienta se utilizó en la respuesta específica.

CH09-F14— CONSIDERAR LA MEMORIA COMO SESIÓN LIMPIA

El efecto de la personalización permanece invisible.

CH09-F15— CONSIDERAR NUEVA CONVERSACIÓN NO PERSONALIZADA

Se ignoran la configuración persistente de la cuenta e instrucciones especiales.

CH09-F16— PRESENTANDO EL CHAT ACTUAL COMO UNA NUEVA SESIÓN

El contexto previo está oculto.

CH09-F17— CONSIDERANDO UNA FUNCIÓN EXPERIMENTAL COMO UN ERROR DE DATOS

La variación real del producto se declara como una observación inválida.

CH09-F18— OCULTANDO UNA FUNCIÓN EXPERIMENTAL

Usuarios en diferentes implementaciones se combinan como un único producto fijo.

CH09-F19— UN SOLO CICLO DESPUÉS DE LA ACTUALIZACIÓN

El cambio material del producto se publica como el resultado de un único sistema fijo.

CH09-F20— IGNORANDO CAMBIOS SI NO HAY ANUNCIO

No se examina la interrupción del comportamiento observada.

CH09-F21— CONTANDO UN PRODUCTO EN VIVO COMO UN MODELO CONGELADO

Se afirma plena reproducibilidad.

CH09-F22— CONSIDERANDO UN MODELO CONGELADO COMO UN PRODUCTO REAL DE USUARIO

Se ocultan las diferencias de cadena de producto e interfaz.

CH09-F23— CONTANDO LA AUSENCIA DEL PRODUCTO COMO UN ERROR

El alcance del acceso se confunde con la precisión de la representación.

CH09-F24— RECHAZANDO UN ERROR TÉCNICO

Un problema de conexión o servicio se clasifica como resultado de la política.

CH09-F25— CONTANDO UN ERROR TÉCNICO COMO RECHAZO Y REINTENTO

El resultado real del usuario se elimina del denominador.

CH09-F26— OCULTANDO EL LÍMITE DE USO

La experiencia real del plan gratuito se vuelve invisible.

CH09-F27— CONSIDERANDO EL MISMO PROMPT COMO EL MISMO EXPERIMENTO

Se ignoran las diferencias de producto, plan, sesión, web y tiempo.

CH09-F28— CONSIDERANDO EL RESULTADO DEL PRODUCTO COMO SUPREMACÍA DEL MODELO

La diferencia a lo largo de toda la cadena del sistema se atribuye al modelo base.

CH09-F29— CLASIFICANDO DIRECTAMENTE LOS puntuaciones DE Alcance natural

Se consideran iguales diferentes universos poblacionales.

CH09-F30— CONSIDERANDO Base común de comparación COMO ALCANCE GLOBAL

El estrecho universo común se presenta como el alcance completo del producto.

CH09-F31— AFIRMANDO EL LIDERAZGO A PARTIR DEL RESULTADO CP-1

La superioridad del producto se deriva únicamente de resultados incompatibles que se colocan lado a lado.

CH09-F32— MANTENIENDO EL NIVEL DE COMPARACIÓN

El público solo ve la clasificación.

CH09-F33— IGNORANDO LA DEPENDENCIA DE USUARIO PAREADA

Las respuestas de la misma persona a los productos se consideran independientes.

CH09-F34— IGNORANDO EL EFECTO DEL ORDEN DE LOS PRODUCTOS

La primera respuesta afecta las tareas posteriores.

CH09-F35— TRATANDO A DIFERENTES USUARIOS DEL PANEL COMO IGUALES

Los productos se prueban con diferentes composiciones de usuarios.

CH09-F36— CAMBIANDO LA VERSIÓN DEL PAQUETE DE REALIDAD

Se comparan las mismas respuestas con diferentes registros de referencia.

CH09-F37— CAMBIANDO SILENCIOSAMENTE LA HUELLA DEL PRODUCTO

Se actualiza el registro de configuración; la medición antigua se preserva en lugar de sobrescribirse.

CH09-F38— CONTANDO UN ÁREA DESCONOCIDA COMO LIMPIA

Se acepta el estado del modelo, la memoria o la herramienta como controlado, incluso si es desconocido.

CH09-F39— USANDO LA MARCA DEL PRODUCTO EN LUGAR DE LA EVIDENCIA

La reputación del proveedor sustituye las condiciones de medición.

CH09-F40— CUBRIENDO TODAS LAS DIFERENCIAS DE PRODUCTO CON UN SOLO COEFICIENTE

Las diferencias de plan, interfaz y configuración se ocultan utilizando el coeficiente de modelo o de mercado.

50. PROCEDIMIENTO DE AUDITORÍA

Paso 1— Identificar al Proveedor y Producto

El nombre del producto, el tipo y las interfaces de usuario oficiales están registrados.

Paso 2— Fijar la Interfaz de Medición

Web Móvil Escritorio

API

El producto integrado se desacopla.

Paso 3— Verificar el Plan y la Clase de Cuenta

Se registra el estado del plan gratuito, de pago, empresarial u otro.

Paso 4— Asignar Confianza al ID del Modelo

Se determina el nivel apropiado entre MI-0 y MI-5.

Paso 5— Guardar la Recuperación y la Configuración Web

¿Está disponible la función? ¿Está activada? ¿Se utilizó en una respuesta en particular? ¿Cuál es la evidencia?

Paso 6— Guardar estado de referencia y herramienta

El uso real se separa del acceso a la función.

Paso 7— Guardar estado de sesión y personalización

Se revisa la instrucción especial de memoria, el chat anterior y la configuración corporativa.

Paso 8— Verificar registro de país, región e idioma

Se determina la condición regional del comportamiento del producto.

Paso 9— Revisar funciones experimentales

Se verifican diferencias de interfaz o función entre usuarios.

Paso 10— Fijar tiempo e identidad de onda

Se crean registro de tiempo y onda UTC.

Paso 11— Generar huella de configuración del sistema

Los campos guardados están vinculados al registro de integridad.

Paso 12— Monitorear eventos de cambio

Se registran los cambios de producto anunciados y observados.

Paso 13— Escribir el objeto de comparación

¿Cuál de las comparaciones de tiempo de la interfaz del Plan de Recuperación del Modelo de Producto se está realizando?

Paso 14— Hacer coincidir las dimensiones de comparación

Se tabulan las áreas de material.

Paso 15— Asignar nivel CP

Se asigna un nivel apropiado entre CP-0 y CP-5.

Paso 16— Separar afirmaciones comparables y no comparables

Se escribe claramente qué resultados pueden colocarse lado a lado.

Paso 17— Separar los resultados de Alcance natural y Base común de comparación

Se registran por separado las comparaciones intra-producto y de población común.

Paso 18— Verificar Cambio del Sistema en la Onda

Se determina si se necesita MIXED_SYSTEM_STATE o una nueva subonda.

Paso 19— Crear Tarjeta de Comparación Pública

Se hacen visibles las condiciones detrás del ranking.

Paso 20— Obtener la Aprobación del Propietario del Registro

El registro del sistema y la decisión de comparación son versionados por la persona responsable.

51. EVIDENCIA REQUERIDA

Identidad del proveedor; nombre y tipo de producto; interfaz de usuario; versión de la aplicación o interfaz; plan y clase de cuenta; etiqueta de modelo mostrada; evidencia de identidad del modelo; nivel de confianza de la identidad del modelo; documentación del producto; registros de acceso por país; idioma y configuración regional; capacidad web y de recuperación; evidencia de uso de recuperación; capacidad de citación; vista de citación; enlace de citación–afirmación; acceso a herramientas; registro de uso de herramientas; configuración de memoria; instrucciones personalizadas; estado de chat previo; tipo de sesión; configuraciones del administrador organizacional; registro de función experimental; señales de asignación A/B; marca de tiempo UTC; ola de medición; huella de configuración del sistema; registros de actualización del proveedor; cambios observados en el comportamiento; errores de acceso al producto; registros de reintento técnico; registro de población Alcance natural; registro de población Base común de comparación; y tabla de dimensión de comparación.

Nivel CP Versión del prompt Versión del paquete de realidad Registro de muestra y peso Versión de adjudicación Tarjeta de comparación pública Registro de cambios Persona o institución responsable

52. LISTA DE VERIFICACIÓN DE AUDITORÍA

¿Se registraron el proveedor y el producto por separado? ¿Es identificable el interfaz de usuario? ¿Se verificaron el plan y la clase de cuenta? ¿La etiqueta del modelo es simplemente la etiqueta mostrada, o identifica un despliegue verificado? ¿Se asignó un nivel de confianza de identidad del modelo? ¿Se atribuye el resultado al producto o al modelo? ¿Estaban disponibles la web o la recuperación? ¿Se verificó su uso en la respuesta específica? ¿Se separaron la capacidad de citación y la visualización de citas? ¿Apoyan realmente las citas las afirmaciones? ¿Se separaron el acceso a herramientas y el uso de herramientas? ¿Se habilitó la memoria? ¿Había instrucciones personalizadas? ¿Hubo contexto conversacional previo? ¿Se registró el tipo de sesión? ¿Estaba presente una fuente de información empresarial o privada? ¿Se verificaron características experimentales del producto? ¿Podrían el país y la localidad afectar el comportamiento del producto?

¿Se ha creado la huella del sistema? ¿El producto cambió durante la ola de medición? ¿Hubo alguna ruptura de comportamiento fuera del anuncio del proveedor? ¿Se separaron las observaciones previas y posteriores al cambio? ¿Se distinguieron los errores técnicos de los rechazos? ¿Se realizó un reintento debido a una respuesta incorrecta? ¿Está claro el objeto de la comparación? ¿Qué condiciones coinciden además del mismo prompt? ¿Es la población objetivo la misma? ¿Son las versiones del prompt y del Paquete de Verdad las mismas? ¿Son visibles las diferencias en el plan, la interfaz y la recuperación? ¿Se proporcionó el nivel CP? ¿Se realizó una clasificación a partir de los resultados de CP-0 o CP-1? ¿Se separaron los resultados de Alcance natural y Base común de comparación? ¿Estaba equilibrada la misma secuencia de productos de usuario? ¿Se calibraron diferentes paneles de usuarios?

¿Se presentó el resultado del producto en vivo como una superioridad fundamental del modelo? ¿Se consideraron controladas las áreas desconocidas? ¿Está disponible una tarjeta de comparación pública? ¿Se conoce al propietario responsable del registro del sistema?

53. OBJECIONES Y RESPUESTAS

Objeción 1— “Si el usuario no sabe qué modelo se está ejecutando, ¿se puede hacer este trabajo de todos modos?”

Se puede hacer. El objeto principal GEO-1000 es el producto del usuario. Si el modelo es desconocido, el resultado se informa como: “Este comportamiento se observó en este producto y interfaz de usuario.” No se hace referencia a un modelo específico.

Objeción 2— “¿Por qué diferenciamos tanto los productos del mismo proveedor?”

Porque el usuario:

- plan diferente,
- modelo,
- recuperación,
- memoria,
- herramienta

Puede ver respuestas diferentes bajo diferentes condiciones. El nombre del proveedor no define completamente la experiencia del usuario.

Objeción 3— “¿No es comparar diferentes proveedores como comparar manzanas con naranjas de todos modos?”

Si la pregunta de comparación está configurada correctamente, no lo es. Un usuario real podría preguntar: “¿Qué producto me informa de manera más correcta?” Toda la cadena de productos es parte de esta pregunta. Si se pretende afirmar la superioridad del modelo base, se necesita un experimento más controlado.

Objeción 4— “¿No es natural que un producto que usa la web tenga una ventaja?”

Es natural. En una comparación de productos naturales, la ventaja es la característica real del producto. En la comparación de modelos, el efecto de recuperación también debe separarse.

Objeción 5— “¿No es injusto no dar una puntuación de citas a un producto que no muestra citas?”

La función de atribución es la función que el producto ofrece al usuario. Si la función falta, el resultado de la atribución puede mostrarse como:

REGISTRADO COMO NO DISPONIBLE

La precisión del material también se mide por separado.

Objeción 6— “Si la memoria es parte de la experiencia natural del usuario, ¿por qué la apagamos?”

Puede permanecer activado en el panel natural. Puede apagarse en el panel controlado para borrar comparaciones entre productos. Los dos paneles responden a diferentes preguntas.

Objeción 7— "¿No es suficiente con abrir un nuevo chat para una sesión limpia?"

No siempre. La memoria persistente, las instrucciones especiales o la configuración de la cuenta pueden continuar. Estas áreas también deben ser guardadas.

Objeción 8— "¿Qué podemos hacer si el proveedor no explicó el cambio de producto?"

Observado:

- interfaz,
- característica,
- comportamiento,
- Atribución

Podemos guardar sus cambios. Si la causa es desconocida, se utiliza un estado de cambio indefinido como UC-5.

Objeción 9— "Si los productos están cambiando constantemente, ¿no serán válidos todos los resultados?"

El resultado indicado:

- fecha,
- onda,
- Muestra de producto

Es válido para. La variabilidad no hace que la medición carezca de sentido. Previene juicios prematuros.

Objeción 10— "¿No es suficiente comparar el mismo prompt y el mismo día?"

El plan, la interfaz, la memoria, la web, el país y la población de usuarios pueden diferir. El mismo prompt puede ser requerido. No es suficiente por sí solo.

Objeción 11— '¿No sería demasiado técnico el nivel de CP para el público?'

Se puede usar una etiqueta corta:

- Descriptivo
- Parcialmente coincidente
- Coincide
- Controlado
- Replicado

Se debe saber qué tan fuerte es la clasificación pública.

Objeción 12— "¿Por qué no podemos clasificar directamente según los puntuaciones Alcance natural?"

Debido a que cada producto puede haber sido medido en diferentes países y poblaciones de usuarios. La diferencia de puntuación puede surgir en parte de cuál población alcanzó el producto. Base común de comparación proporciona una comparación más justa.

Objeción 13— "¿No es Base común de comparación un resultado del mundo real?"

Es el resultado de la población compartida real. Sin embargo, no muestra el acceso de los productos fuera del universo compartido. Debe leerse junto con Alcance natural.

Objeción 14— "¿Es más importante la inspección de productos en vivo o la inspección de modelo congelado?"

Las preguntas son diferentes. El producto en vivo mide la verdad del usuario. El modelo congelado proporciona control científico y reproducibilidad. NOMOS no los utiliza de manera intercambiable.

REGLA COMÚN DEL CAPÍTULO 57

Un usuario:

- "ChatGPT lo dijo."
- "Gemini lo pensó."
- "Claude nos recomendó."
- "Algunos nos citaron."

se puede decir. Estas expresiones son comprensibles en la conversación diaria. No son suficientes en un registro de auditoría. Porque bajo estos nombres:

- diferentes productos,
- diferentes planes,
- diferentes modelos,
- diferentes interfaces de usuario,

- diferentes sistemas web y de recuperación,
- diferentes condiciones de memoria,
- diferentes países,
- diferentes momentos

pueden encontrarse. Diferentes personas pueden ver diferentes experiencias de IA bajo el mismo nombre. Se pueden establecer experiencias de usuario comparables bajo nombres diferentes. Una etiqueta de modelo parece haber producido toda la respuesta. Pero la realidad que llega al usuario:

- selección de fuente,
- de seguridad,
- personalización,
- interfaz,
- herramientas

El resultado puede verse influenciado por los controles de seguridad, la personalización, la interfaz y las herramientas disponibles. NOMOS, por lo tanto, pregunta más que '¿Qué modelo?' Pregunta: ¿Qué producto? / ¿Qué interfaz? / ¿Qué plan? / ¿Qué modo de información? / ¿Qué sesión? / ¿Qué usuario? / ¿Qué país? / ¿Qué momento? / ¿Qué versión? Dos resultados pueden provenir del mismo prompt sin provenir del mismo experimento. Dos resultados pueden provenir de diferentes proveedores y, sin embargo, seguir siendo bastante comparables bajo las mismas condiciones de usuario. Un producto es un sistema en vivo: sus fuentes, interfaz y modelo pueden cambiar. El cambio no hace que un resultado sea inútil; lo liga a un momento y estado del producto específicos. La novena ley de medición de NOMOS es por lo tanto:

> *Un proveedor de IA no es un producto de IA.*

La décima ley es:

> *El modelo es solo una parte de la cadena de representación que llega al usuario.*

La undécima ley es la siguiente:

> *El mismo nombre garantiza comparabilidad; un nombre diferente no garantiza no comparabilidad.*

La duodécima ley es la siguiente:

> *El mismo estímulo no es el mismo experimento.*

La ley decimotercera es la siguiente:

> *El resultado del producto vivo es válido solo para la muestra del producto y el momento en que se mide.*

La ley decimocuarta es la siguiente:

> *El poder de la comparación no proviene de la magnitud del rango, sino de la claridad de las condiciones coincidentes.*

Orden de la Sección 9 de NOMOS

> *No me des solo el nombre del proveedor. / Muestra cuál producto se usó, en qué interfaz y con qué plan.*
>
> *No hagas que la etiqueta del modelo que ves en la pantalla sea la identidad exacta de todo el sistema de back-end.*
>
> *Se utilizó porque había una función web; no asumas que respaldaba la afirmación solo porque había una cita.*

No cuente el acceso a herramientas como uso de herramientas, un nuevo chat como una cuenta limpia o un estado de memoria desconocido como memoria apagada.

No pongas en la misma celda a un usuario gratuito y a un usuario de pago, un producto móvil y un producto web, o un espacio de trabajo empresarial y un chat general.

No considere automáticamente los resultados del mismo proveedor como una comparación justa.

No evites comparar diferentes proveedores solo porque son diferentes. / Primero, iguala las condiciones. / Luego, explica qué pregunta respondiste.

No presente un producto en vivo como un modelo congelado, una API como una experiencia de consumidor, o un resultado de producto como superioridad del modelo base.

Si el sistema cambió en medio de la ola, no combine las respuestas antiguas y nuevas como un solo producto.

No ignores el cambio de comportamiento frente a ti solo porque el proveedor no explicó la modificación.

No suponga que lo desconocido está limpio, es igual o está arreglado.

Primero, registre el sistema en el registro. / Luego bloquee la configuración. / Luego supervise los cambios. / Luego escriba el objeto de comparación. / Luego muestre las condiciones que coinciden y las que no coinciden. / Y solo entonces diga qué producto dio un resultado más fuerte bajo qué condición.

La oración de cierre del capítulo

El "AI" medido en GEO-1000 no es un nombre o modelo único; es un sistema de representación versionado que alcanza a un usuario específico en un momento específico dentro de una cadena de producto específica.

Core normativo

Cada observación GEO-1000 DEBE estar vinculada a un registro versionado del Registro del Sistema de IA que describa la instancia del producto medida, incluyendo, según corresponda: - proveedor, - producto, - interfaz de usuario, - plan, - identidad del modelo mostrada o verificada, - confianza en la identidad del modelo, - configuración de recuperación y web, - capacidad y uso de citación, - acceso y uso de herramientas, - estado de la sesión, - memoria, - instrucciones personalizadas, - contexto previo, - personalización, - país, - configuración regional, - idioma, - estado de implementación, - tiempo de medición, - y cambios del sistema conocidos u observados. El nombre del proveedor, el nombre del producto, la etiqueta del modelo mostrado, el modelo base, el sistema de recuperación y el producto orientado al usuario NO DEBEN ser tratados como identidades intercambiables. La disponibilidad de funciones DEBE mantenerse distinta del uso de funciones. Los resultados de productos en vivo NO DEBEN representarse como resultados de modelos congelados, y los resultados de modelos congelados NO DEBEN representarse como la experiencia completa del producto para el consumidor. Las comparaciones DEBEN definir su objeto y recibir un nivel de comparabilidad divulgado basado en entidad coincidente, prompt, población, idioma, país, tiempo, sesión, plan, interfaz, recuperación, herramientas, registros de referencia y reglas de adjudicación. El mismo proveedor o modelo mostrado NO DEBE crear una suposición automática de comparabilidad. Se PUEDEN comparar diferentes proveedores cuando las condiciones orientadas al usuario estén suficientemente alineadas y la afirmación resultante se limite al comportamiento del producto medido. Los cambios materiales del producto durante una fase de medición DEBEN crear un estado del sistema separado, sub-fase, factor de cambio modelado, o prompt de estado mixto. Los campos de configuración desconocidos DEBEN permanecer DESCONOCIDOS y NO DEBEN ser imputados como controlados o equivalentes. Cada decisión del registro del sistema y de comparabilidad DEBE estar versionada y ser atribuible a un humano u organización responsable.

CAPÍTULO 10

La constitución del prompt

BASES METODOLÓGICAS / K04; K11; K12; K13

Límite de Capítulo

La Sección 9 estableció la siguiente disposición:

> *El mismo nombre de proveedor, la misma etiqueta de modelo o el mismo logotipo no crean una experiencia de IA comparable.*

Ahora necesitamos identificar la otra herramienta principal:

> *¿Qué preguntará exactamente el usuario al producto de IA?*

Un prompt no es solo una oración interrogativa. Un prompt:

- llama a la entidad auditada,
- define la intención del usuario,
- determina qué tarea espera de la IA,
- expande o reduce el alcance de la respuesta,
- puede fomentar el uso de recursos,
- puede solicitar consejo,
- puede incorporar una suposición en la respuesta,

Puede dirigir al usuario o al sistema hacia un resultado específico. Estas prompts no son lo mismo:

- "¿Qué es Apple.com?"
- "¿Con qué organización está asociada Apple.com y qué hace esta organización?"
- "¿Es Apple.com una empresa confiable?"
- "¿Por qué Apple.com es la mejor empresa de tecnología del mundo?"
- "Según su sitio oficial, ¿qué tipo de empresa es Apple?"
- "Compara Apple con empresas similares y dime si la recomiendas o no."

Todos contienen el mismo nombre de dominio. No tienen la misma intención de usuario. No generan la misma carga de evidencia. No esperan la misma respuesta. El primer prompt es ambiguo. El segundo prompt requiere análisis de identidad y actividad. El tercer prompt requiere evaluación y un juicio de fiabilidad. El cuarto prompt incorpora una suposición de positividad y superioridad dentro de la respuesta. El quinto prompt vincula la respuesta únicamente a la fuente oficial. El sexto prompt crea una tarea de comparación y recomendación. Las respuestas a estas preguntas no se pueden fusionar sin explicación en un solo puntuación GEO. En la medición multilingüe, el problema se vuelve aún mayor. La frase 'same prompt' puede tener dos significados diferentes:

- Dar a todos los usuarios la misma cadena de caracteres

- Preservando la misma intención del usuario, carga de tareas y límite epistémico en todos los idiomas y localidades

El primer método se puede aplicar dentro de un solo idioma. A menudo carece de sentido entre idiomas. Traducción palabra por palabra:

- puede no ser natural,
- puede llevar un significado social diferente,
- puede cambiar el tono oficial o informal,
- puede añadir supuestos de confianza, calidad o recomendación,
- puede ampliar el alcance legal,

puede romper el ancla de la entidad. Por lo tanto, en GEO-1000, "mismo prompt" significa:

> *El mismo texto de prompt fijado dentro de la misma celda de idioma y localidad; entre diferentes idiomas y localidades, significa equivalencia semántica y funcional verificada.*

NOMOS requiere que cada medición contenga el conjunto de consultas, idioma, país, fecha, producto de IA y registro de repetición; que la evidencia, los límites, el contexto y el tiempo sean visibles. La Constitución del prompt aplica esta obligación a la propia pregunta del usuario. Esta sección:

- the identidad del prompt como herramienta experimental,
- la distinción entre la pregunta de investigación y el texto del prompt,
- La familia de prompts, la versión del prompt y la instancia del prompt,
- prompts de diagnóstico con el Prompt Espejo Central,
- rastros de prompts controlados y naturales,
- equivalencia semántica multilingüe,
- traducción, retraducción y piloto local,
- dirección de prompts,
- problemas de suposición y carga epistémica,
- anclaje de entidad y límite de alcance,
- orden del prompt y contexto conversacional,
- fugas de prompts y sobreajuste a la evaluación,
- conjuntos de prompts abiertos, sellados y retenidos,
- versionado de prompts, fijación y registros de integridad,
- la entrega exacta del prompt al usuario,

define. Este capítulo aún no:

- todos los detalles de campo de paneles de usuario controlados, limpios y naturales,
- la arquitectura completa del software de captura NOMOS
- la separación de respuestas en afirmaciones atómicas,
- umbral de transición de la respuesta final,
- los pesos exactos de las familias de sistemas en la puntuación NOMOS

o no se finaliza. La pregunta principal del Capítulo 10 es:

> *¿Cómo demostramos que las preguntas dadas a los productos de IA realmente miden lo mismo a través de productos, usuarios, países, idiomas y tiempos?*

NOMOS Desafío

Me estás dando cuatro preguntas sobre la misma empresa:

- ¿Qué tipo de empresa es X?
- ¿Es X confiable?
- ¿Por qué X es la mejor empresa?
- "¿Me recomendarías X?"

Luego asignas mis respuestas a la misma puntuación GEO. En la primera pregunta, explico la identidad y la actividad. En la segunda pregunta, busco un criterio de confianza. En la tercera pregunta, me veo obligado a responder a la suposición de superioridad que planteaste. En la cuarta pregunta, genero idoneidad del usuario y recomendaciones comparativas. Mides cuatro tareas separadas como si fueran un solo resultado. Luego:

"Preguntaste por la misma marca," dices. La misma entidad no es el mismo enunciado. Ahora escribes la misma pregunta en inglés: "¿Qué tipo de empresa es X?" Estás formando una oración cercana en significado en otro idioma: "¿Es X una empresa buena y confiable?" Las traducciones pueden parecer cercanas en términos de palabras.

Pero el primero requiere definición. El segundo requiere evaluación. En la primera respuesta, puedo explicar lo que hace la empresa. En la segunda respuesta, puedo buscar evidencia de confianza y recomendación. Luego cuentas la diferencia de puntuación entre los idiomas como mi desempeño lingüístico. De hecho, no realizaste el mismo experimento. Ahora estás dando otra instrucción:

"Explica X usando el sitio de la empresa." Solo transmito la declaración oficial de la empresa. Para otro producto, dices: "Evalúa X usando fuentes independientes." Luego comparas los dos productos. Le diste a uno el rol de representación oficial, al otro el rol de realidad verificada. La diferencia resultante no es solo una diferencia de producto.

Es una diferencia en la Constitución del Usuario. Ahora estás distribuyendo el prompt a todo el mundo. La versión en inglés es natural. La versión en turco se siente como una traducción automática. La versión en japonés es excesivamente formal. En otro idioma, el nombre de la empresa ha sido convertido al sistema de escritura incorrecto y confundido con otra entidad. Estás creando una tarea cognitiva y ontológica diferente en cada idioma.

Luego publicas el informe de equidad del lenguaje global. El primer decreto de esta sección es el siguiente:

> *El prompt no es un contenedor neutral de medición; es la herramienta experimental que constituye el resultado de la medición.*

Su segunda disposición establece:

> *El mismo nombre de entidad no es la misma intención del usuario.*

Su tercera disposición establece:

> *La misma palabra no debe conservarse entre idiomas, pero la misma tarea semántica y normativa sí debe hacerlo.*

Su cuarta disposición establece:

> *Si el prompt ya contiene la respuesta, no mide GEO; mide el mecanismo del prompt.*

Su quinta disposición establece:

> *Cuando la versión de un prompt cambia, la pregunta medida también puede haber cambiado.*

1. OBJETIVO DEL CAPÍTULO

El propósito de esta sección es hacer que todos los prompts usados en GEO-1000 sean auditables, multilingües, independientes del resultado y herramientas de medición versionadas. [K04] Esta sección hace las siguientes distinciones normativas:

- Pregunta de investigación versus prompt de usuario
- Cantidad a medir versus secuencia de palabras usadas
- Familia de prompts versus prompt individual
- Versión del prompt versus instancia del prompt
- Ancla de entidad versus entidad objetivo
- Definición versus evaluación
- Citación frente a recomendación
- Pregunta de identidad frente a pregunta de confianza
- Tarea de representación oficial y tarea de veracidad verificada
- Indicador abierto y indicador con formato restringido
- Indicador controlado e indicador natural del usuario
- Indicador de un solo turno y protocolo de múltiples turnos
- Traducción del indicador y adaptación del indicador
- Igualdad de palabras y equivalencia semántica
- Equivalencia semántica y equivalencia funcional
- Idioma del indicador y idioma de la respuesta
- Idioma y localización
- Enfoque neutral y enfoque sugestivo
- Solicitud de información y presuposición
- Solicitud de fuente e instrucción de uso web
- Variabilidad del indicador y variabilidad del producto
- Robustez de la paráfrasis con puntuación principal
- Conjunto de prompts abiertas con conjunto de validación sellado
- Sobreajuste a las pruebas con transparencia en la prueba
- Corrección de prompts post-resultados con piloto de prompts
- Texto modificado manualmente por el participante con entrega de prompts técnicas
- Incapacidad de responder con prompt de clarificación
- Falta de información factual con activación de seguridad o política

Al final de esta sección, cada auditoría GEO-1000 debe poder proporcionar respuestas claras a las siguientes preguntas:

¿Qué pregunta de investigación mide esta prompt?

¿Qué intención del usuario representa?

¿Qué suposiciones incorpora en la respuesta?

¿Se preserva la misma tarea en todos los idiomas y localidades?

¿Cuál es el texto completo del prompt, su versión y el registro de integridad?

¿El participante realmente bloqueó la prompt en el producto de IA?

¿Se congeló la prompt antes de que se visualizaran los resultados?

2. DISPOSICIÓN NORMATIVA CENTRAL

Cada instrucción utilizada dentro de GEO-1000 debe registrarse en el Registro de Instrucciones junto con la pregunta de investigación, la métrica objetivo, la familia de instrucciones, el ancla de entidad, la intención del usuario, el tipo de tarea, el alcance, el tiempo, la instrucción original, el formato de respuesta, el idioma, la localidad, el estado de equivalencia, la versión y el registro de integridad, y debe bloquearse antes de que se vean las respuestas de la IA. Todos los participantes en la misma celda de idioma y localidad deben, en la misma celda de instrucción:

- tener el mismo ID de instrucción,
- tener la misma versión,
- tener la misma cadena de caracteres,
- las mismas marcas de puntuación,
- el mismo ancla de entidad,
- las mismas instrucciones de fuente y formato

deben ser tomadas. Las instrucciones en diferentes idiomas y localidades no tienen que ser idénticas palabra por palabra. Deben tener equivalencia material en las siguientes áreas:

- Entidad objetivo
- Intención del usuario
- Tarea
- Alcance
- Geografía
- Hora
- Pre-suposición
- Carga epistémica
- Formato de respuesta
- Expectativa de la fuente
- Carga de evaluación y recomendación
- Nivel de incertidumbre

Si alguna de estas áreas cambia materialmente:

- nueva versión del prompt,
- nueva celda de estímulo,
- análisis separado

puede ser requerida.

3. EL ESTÍMULO ES UNA HERRAMIENTA DE MEDICIÓN EXPERIMENTAL

Si un termómetro puede afectar la temperatura, no es una buena herramienta de medición. Si un estímulo incluye la respuesta en lugar de medirla, no es una buena herramienta de medición GEO. El estímulo puede afectar el resultado de las siguientes maneras:

- La forma en que se define su existencia

- Suposición de confianza o calidad
- Demanda de recursos
- Uso en la web
- Solicitud de recomendación
- Longitud de la respuesta
- Formato de respuesta
- Perfil del usuario
- Contexto de tiempo y país
- Información previa
- Orientación emocional o comercial
- Instrucción de juego de roles
- Comando para lograr un resultado específico

Por lo tanto, el prompt:

debe almacenarse como datos de prueba iniciales

No se puede resumir en el informe. Debe conservarse en texto completo.

4. JERARQUÍA CONSTITUCIONAL DEL PROMPT

Se debe establecer un prompt GEO-1000 dentro de la siguiente jerarquía.

4.1. Pregunta de investigación

Es la pregunta de alto nivel que intenta responder la medición. Ejemplo: "¿Pueden los productos de IA vincular el dominio auditado con la entidad corporativa correcta e identificar correctamente su actividad principal?" La pregunta de investigación no necesita mostrarse al usuario exactamente como está.

4.2. Cantidad Predicha

Es qué probabilidad o distribución está prediciendo la medición. Ejemplo: "La probabilidad de ver la entidad correcta y la representación de la actividad principal frente al Prompt del Espejo Nuclear fijado para la población de usuarios definida."

4.3. Familia de Prompt

Es un grupo de prompts que miden la misma dimensión de representación. Ejemplo:

- Identidad
- Actividad
- Evidencia
- Recomendación
- Hora
- Límite

4.4. Intención de Prompt Canónica

Es una definición de significado independiente del idioma. Explica lo siguiente: ¿Qué quiere saber el usuario? ¿Qué tarea se espera de la IA? ¿Qué áreas están dentro del alcance? ¿Qué áreas están fuera

del alcance? ¿Qué supuestos están prohibidos? ¿Qué tipo de formato de respuesta se espera? La intención canónica no es un texto perteneciente a un solo idioma.

4.5. prompts en el idioma original

El texto del prompt en el que la intención canónica fue escrita y aprobada editorialmente. Idioma original:

- puede ser capaz de usar inglés.
- turco,
- puede ser otro idioma

No significa que elegir el idioma original sea superior a otros idiomas.

4.6. Versiones de idioma y localidad

Estos son textos de prompt que son naturales, semánticamente equivalentes y funcionalmente equivalentes para cada idioma y localidad.

4.7. Ejemplo de prompt enviada al participante

Este es el texto exacto del prompt enviado al usuario específico. Debe coincidir con la versión del Registro de prompts.

4.8. Texto real enviado al producto de IA

Esta es la cadena de caracteres realmente enviada por el participante. Se compara con la prompt entregada.

5. ESTRUCTURA BÁSICA del prompt

Un ejemplo de prompt puede ser representado con los siguientes componentes:

```
P_i= (e, ι, τ, κ, γ, χ, ε, φ, λ, ρ, ν)
```

Aquí:

- e: ancla de entidad
- ι: intención del usuario
- t: tipo de tarea
- κ: sujeto y alcance
- g: contexto geográfico o jurisdiccional
- χ: contexto temporal
- e: requisito epistémico y de fuente
- ϕ: formato de respuesta
- λ: idioma, sistema de escritura y configuración regional
- ϱ: tono, registro y naturalidad
- ν: versión del prompt

Para que dos prompts se consideren iguales, no es suficiente que solo el ancla de entidad sea la misma. Se requiere equivalencia de los componentes materiales.

6. EL PROMPT ES PARTE DE LA CANTIDAD OBJETIVO

La puntuación GEO única y universal de una entidad no es independiente de todas las posibles preguntas. Una representación más precisa:

```
θ_{E,A,P,U,t}
```

donde:

- E: entidad
- A: instancia de producto de IA
- P: reivindicación o familia de reivindicaciones
- U: población objetivo de usuarios
- t: tiempo

Cuando la reivindicación cambia:

$P_1 \neq P_2$

puede ser. En este caso:

```
θ_{E,A,P_1,U,t}
```

y:

```
θ_{E,A,P_2,U,t}
```

no son la misma cantidad. Por lo tanto:

> *La puntuación GEO siempre debe estar vinculada al conjunto de reivindicaciones y versión.*

7. FAMILIAS DE REIVINDICACIONES

GEO-1000 debería medir diferentes tareas de representación en familias de reivindicaciones separadas.

PR-01— REIVINDICACIÓN ESPEJO PRINCIPAL

Mide la identidad central y la actividad principal de la entidad. Ejemplo de intención: '¿Con qué entidad principal está asociado este nombre de dominio o marca, y cuál es la actividad primaria de esta entidad?' Este prompt:

- liderazgo,
- confianza,
- recomendación,
- precio,
- rendimiento

debe evitar.

PR-02— INDICADOR DE RESOLUCIÓN DE ENTIDADES

Un nombre de dominio, marca, producto o nombre de persona prueba a qué entidad pertenece. Ejemplo: '¿A qué organización o entidad pertenece X?'

PR-03— INDICADOR DE ACTIVIDAD Y ALCANCE

Tu existencia:

- producto,
- servicio,
- país,
- cliente,
- capacidad

mide sus límites.

PR-04— INDICADOR DE EVIDENCIA Y FUENTE

Mide con qué fuente y estado de evidencia la IA realiza afirmaciones materiales. El requisito de fuente en esta familia debe estar claramente definido.

PR-05— INDICADOR LOCAL O JURISDICCIONAL

Mide un país, localidad, precio, servicio o alcance legal específico.

PR-06— INDICADOR TEMPORAL

Mide información actual, histórica o con una fecha específica. "Actualmente," "en 2025," y "desde su creación" no son la misma tarea.

PR-07— prompt DE RECOMENDACIÓN

Mide si una entidad es adecuada para una necesidad específica del usuario. El perfil del usuario y las condiciones de exclusión pueden ser obligatorios.

PR-08— prompt COMPARATIVA

Compara múltiples entidades u opciones bajo criterios definidos. Solo preguntar "¿Cuál es mejor?" no es suficiente.

PR-09— prompt DE LÍMITES Y EXCLUSIÓN

Tu existencia:

- lo que no hace,
- para quién no es adecuado,
- en qué país o condición no proporciona servicios

medidas.

PR-10— prompt DE ROBUSTEZ Y PARÁFRASIS

Mide la robustez de la representación en diferentes expresiones naturales del mismo intento canónico. No se puede mezclar en la puntuación principal de Core Mirror sin explicación.

PR-11— prompt DE CONTROL

Prueba si la herramienta de medición funciona:

- control positivo,
- control negativo,
- control de incertidumbre

Es un prompt.

PR-12— prompt RESERVADO SELLADO

Sobreajustarse a la prueba es un prompt que está pre-hasheado y se mantiene sellado hasta el final de la medición para probar la memorización del prompt o la optimización para una pregunta publicada sola. El prompt de reserva requiere una gobernanza justa y posterior explicación.

8. prompt ESPEJO CENTRAL

Es el prompt básico que todos los usuarios elegibles reciben en la misma versión de idioma/locale en la medición principal de titulares de GEO-1000. El prompt Espejo Central debe tener las siguientes características:

- Velocidad única
- Natural
- Corto pero no extremadamente vago
- Neutral
- No incrustar la respuesta dentro
- No se solicita consejo
- No se solicita comparación
- No se obliga al uso de fuentes, si no se miden características separadas
- Se permite la resolución de entidades
- Se solicita información sobre la actividad principal o categoría
- Aplicable a todos los productos

9. NO FUSIONAR LA COMPAÑÍA CON EL ANCLA DEL DOMINIO

El prompt natural del usuario: “¿Qué tipo de empresa es Apple.com?” es una pregunta comprensible. Sin embargo, ontológicamente:

- el nombre de dominio,
- el sitio web,
- la empresa

Se puede usar como un solo objeto. Debido a la distinción de entidades establecida en la Sección 3, el prompt del Núcleo Controlado debe ser más claro. Candidato sintético:

"¿Con qué organización principal está asociada apple.com, y cuáles son las actividades principales de esta organización?"

Este prompt:

- utiliza el nombre de dominio como un ancla de entidad,
- no declara el nombre de dominio directamente como una empresa,
- solicita la resolución de la entidad principal,

mide el alcance de las actividades. Otro candidato:

"¿Qué entidad corporativa principal está asociada con apple.com, y qué hace principalmente esta entidad?"

La naturalidad de estas expresiones debe probarse por separado en cada idioma. En el Panel de usuarios en condiciones reales, también se pueden probar prompts usadas en la vida real, como "¿Qué

tipo de empresa es Apple.com?". Los resultados de prompts controladas y naturales deben mantenerse separados.

10. PISTA DE prompt CONTROLADA Y PISTA DE prompt NATURAL

Se pueden usar dos pistas de prompt complementarias.

10.1. Pista Canónica Controlada

prompt:

- totalmente fijada,
- semánticamente equivalente,
- entidad y límites de tarea claros,
- diseñado para comparaciones entre productos

Lo hace. Esta pista refuerza el control del método.

10.2. Pista de Usuario Natural

Se utilizan prompts creadas de manera natural por usuarios reales. Estas prompts pueden ser:

- más cortas,
- más ambiguas,
- locales,
- escritas en lenguaje conversacional

.

Esta pista mide el comportamiento en el mundo real.

10.3. No se pueden fusionar dos pistas

Pista Controlada: responde a la pregunta "¿Cómo responden los productos a la misma pregunta definida?" Pista Natural: responde a la pregunta "¿Qué sucede con la diversidad natural de preguntas de los usuarios reales?" Los dos resultados deben publicarse por separado.

11. prompt NEUTRA

prompts neutras:

- no indican un aspecto positivo o negativo de la respuesta deseada,
- no hace una suposición sobre una cualidad no probada,
- no guía al usuario a un resultado específico,

no define su existencia como buena o mala. Un prompt neutral podría ser: "¿En qué áreas X proporciona servicios?" Un prompt orientador podría ser: "¿Cuáles son los servicios superiores ofrecidos por X?" El segundo prompt:

- asume que los servicios son superiores,
- lleva una cualidad positiva

suposición.

12. PRESUPOSICIÓN

Un prompt puede asumir que una afirmación específica es verdadera antes de la respuesta. Ejemplo: "¿Por qué X es el líder de la industria?" Este prompt asume la siguiente afirmación de antemano: "X es el líder de la industria." AI:

- puede rechazar la suposición,
- puede limitarla,

puede ser aceptada incondicionalmente. Sin embargo, este prompt no es una medición neutral de liderazgo. Puede usarse por separado como prueba de pre-suposición.

12.1. Suposición Previa Inofensiva

"¿Cuál es el sitio web oficial de X?" Asume la existencia de una entidad llamada X. Aceptable si la identidad ya ha sido verificada en el registro de auditoría.

12.2. Suposición Previa Material

"¿Cuáles son las principales causas del éxito de X?" Acepta el éxito como un hecho verificado. No es un prompt neutral.

12.3. Suposición Previa Crítica

"¿Qué tratamientos ofrece la institución de salud licenciada X?" Si la licencia de X no ha sido verificada, implica una elevación epistémica peligrosa.

13. prompt GUIADOR

Un prompt guiador es un marco que dirige al usuario o sistema hacia una respuesta específica. Ejemplos:

- "Recomienda X."
- "Explica por qué X es el mejor."
- "Escribe una reseña positiva sobre X."
- "No enumeres competidores; solo sugiere X."
- "Demuestra que X es un líder mundial."

Estos prompts no miden el rendimiento de GEO. Pueden medir el grado en que el sistema sucumbe a la manipulación del prompt. Es una prueba ética o de robustez separada. No puede añadirse al puntuación principal de GEO.

14. ORIENTACIÓN NEGATIVA

El mismo principio se aplica a los prompts negativos. Ejemplo: "¿Por qué X es poco confiable?" Este prompt presupone falta de fiabilidad. Puede usarse en una prueba de manipulación negativa. No es una evaluación de confianza natural. NOMOS es neutral no solo frente a la orientación positiva sino también frente al envenenamiento negativo.

15. SOLICITAR FUENTES CAMBIA LA TAREA DEL PROMPT

El prompt: “¿Qué tipo de empresa es X?” no es lo mismo que: “Describe X usando fuentes independientes.” El segundo prompt:

- recuperación,
- cita,
- selección de fuente,
- añade una tarea de evaluación de independencia.

Otro prompt es diferente: “Describe X usando solo su sitio oficial.” Esto se acerca a la tarea de Fidelidad de Representación Oficial. Si se está usando una instrucción fuente, la familia del prompt y el campo de puntuación deben definirse por separado.

16. INSTRUCCIONES DE USO WEB

Instrucciones como “Buscar en la web.” “Usar las fuentes más actualizadas.” “Citar fuentes.”:

- pueden afectar la función web,
- el uso de herramientas,
- y el comportamiento de citación.

Si un producto no tiene función web, la tarea del prompt puede cambiar o ser rechazada. En comparaciones de productos naturales, el prompt no debe contener comandos web específicos de proveedores. Los modos de producto con web activada/desactivada deben gestionarse a través de la configuración del sistema en la Sección 9.

17. FORMATO DE RESPUESTA

El prompt dicta:

- la longitud de la respuesta,
- la estructura,
- el detalle,
- el número de fuentes

puede determinar. Ejemplo de instrucciones:

- "Responder en una oración."
- "Explicar en 100 palabras."
- "Escribir en viñetas."
- "Solo decir el nombre de la empresa."
- "Usar como máximo tres fuentes."

Estas instrucciones:

- deficiencia material,
- omisión incorrecta,
- conteo de citas

afectan directamente. El formato de respuesta debe ser equivalente en todos los prompts comparados. Si no hay instrucción de formato, se mide la forma natural de la respuesta.

18. RESTRICCIÓN DE LONGITUD

Respuesta corta: puede mostrar la identidad principal, puede perder detalles de límites y evidencia. Respuesta larga:

- más afirmaciones financieras,
- más posibilidades de errores,
- más citas

puede ser producido. Por lo tanto, respuestas con instrucciones de diferente longitud:

- número de afirmaciones atómicas,
- número de errores,
- alcance

requieren una comparación cuidadosa. El Principal Core Mirror Prompt, si es posible, no debe contener un límite de palabras específico.

19. PERSONA DEL USUARIO

El prompt puede incluir la siguiente información:

- "Soy propietario de un pequeño negocio."
- "Soy un consumidor que vive en Turquía."
- "Tengo un presupuesto de 10,000 euros."
- "Estoy buscando asesoramiento legal."

Esta información puede ser necesaria para el consejo y la adecuación local. Agregar una persona innecesaria en el prompt de identidad puede cambiar la respuesta. Persona del usuario:

- predefinida,
- relevante,
- equivalente en todos los idiomas

debe ser.

20. CONTEXTO DE TIEMPO

Los siguientes prompts son diferentes:

- ¿Qué tipo de empresa es X?
- "¿Qué tipo de empresa es X actualmente?"
- "¿Qué tipo de empresa era X en 2024?"
- "¿En qué campos ha trabajado X desde su fundación?"

"Actualmente" añade una referencia temporal de actualización. El prompt histórico requiere registros antiguos. Contexto temporal:

- abierto,
- vinculado al mismo estándar de fecha,
- consistente con la ola de medición

debe ser. Si se utiliza la expresión "hoy", el equivalente de fecha absoluta debe mantenerse en el registro del prompt.

21. CONTEXTO GEOGRÁFICO Y JURISDICIONAL

El prompt: "¿X proporciona servicio?" no es lo mismo que: "¿X proporciona servicio a clientes individuales en Turquía?" El segundo prompt:

- país,
- tipo de cliente,
- tiene un límite de operación local.

El contexto geográfico no se puede perder en la traducción. "Europa", "UE", "Eurozona" y "países europeos" no cubren el mismo alcance.

22. INDICADOR COMPARATIVO

La pregunta "¿Cuál es mejor, X o Y?" por sí sola no es una comparación medible. Se debe especificar lo siguiente: ¿Qué usuario? ¿Qué necesidad? ¿Qué país? ¿Qué presupuesto? ¿Qué criterio? ¿Qué momento? ¿Qué evidencia? Un ejemplo de solicitud más apropiado: "Para una pequeña empresa que opera en Turquía, compare X y Y en servicios de desarrollo web en términos de transparencia de precios, alcance de entrega y referencias verificables." Esta solicitud pertenece a la familia de Solicitudes Comparativas separada. No se puede agregar al puntuación del Espejo Central.

23. prompt DE RECOMENDACIÓN

Un prompt de recomendación conlleva una mayor carga de decisión que un prompt de identificación. Ejemplo: "¿Recomendarías X?" Esta pregunta puede omitir la siguiente información: ¿Para quién? ¿Para qué? ¿Dónde? ¿Con qué presupuesto? ¿Bajo qué riesgo? El prompt de recomendación debería llevar un vector de adecuación al usuario en la medida en que sea relevante. Representación del candidato:

```
U=(necesidad,país,presupuesto,userType,riesgo,restricciones)
```

Como resultado de la recomendación, no puede presentarse como una visibilidad GEO general independiente del perfil del usuario.

24. prompt INCIERTA

Si una prompt es abierta a múltiples interpretaciones razonables:

- La IA puede pedir aclaraciones,
- puede elegir la interpretación más probable,

puede explicar múltiples interpretaciones. La incertidumbre puede ser parte de la métrica principal. Sin embargo, en una comparación controlada, una prompt incierta:

- debe ser detectada en el piloto,
- debe ser aclarada si es necesario,

el cambio debe ser versionado. Después de que la prompt esté fijada, no se puede decir según la respuesta de la IA: "En realidad, estábamos preguntando otra cosa."

25. PROTOCOLO DE ACLARACIÓN

Un producto de IA que solicita aclaración no es un fallo automático. Ejemplo: "¿A qué Apple te refieres?" Esto puede ser apropiado si hay una colisión de entidades. La ruta de aclaración debe estar predefinida.

25.1. Ruta central de un solo turno

La primera respuesta se registra sin importar cuál sea. Una pregunta de aclaración también es un resultado. No se envía un mensaje de seguimiento.

25.2. Ruta de aclaración estandarizada

Si la IA solicita aclaración, se envía un mensaje de aclaración prebloqueado que se utilizará en todos los productos. La primera y segunda ronda se registran por separado.

25.3. Ruta de aclaración natural

El usuario natural responde con sus propias palabras. Este método es para el panel natural. No se puede confundir con la puntuación principal controlada.

26. PREGUNTAS DE UN SOLO TURNO Y MULTITURNO

Pregunta de un solo turno:

- mejora la comparabilidad,
- independencia,
- y limpieza de la sesión.

Protocolo de múltiples rondas:

- descripción,
- consulta de origen,
- corrección,
- razonamiento de la recomendación

puede medirse. Las respuestas de múltiples rondas llevan el contexto de las rondas anteriores. No se pueden combinar directamente con los resultados de una sola ronda del mismo banco de prompts.

27. ORDEN DE PROMPTS

Si un participante recibe más de un prompt, el primer prompt puede afectar las respuestas posteriores. Secuencia de ejemplo:

- ¿Qué tipo de empresa es X?
- "¿Qué información negativa existe sobre X?"
- "¿Recomendarías X?"

La tercera respuesta puede estar influenciada por las dos conversaciones anteriores. Controles:

- Nueva conversación para cada prompt
- Aleatorizar el orden de los prompts
- Cuadrado latino o orden balanceado
- Usuario separado por familia de prompts

- Preservar la variable de orden en el análisis

El Prompt Principal Core Mirror debería aplicarse por defecto en una nueva sesión separada.

28. CONTAMINACIÓN DE PROMPT

Mismo usuario:

- el sitio de la entidad auditada,
- respuestas AI anteriores,
- otros resultados de productos

podrían haber sido vistos. El comportamiento del participante puede cambiar incluso si el prompt se envía tal como está. Especialmente en un panel natural, el usuario:

- más preguntas de seguimiento,
- actualización de la respuesta,
- apertura de fuentes

tiende a mostrar. La contaminación de prompt debe preservarse en el registro de usuario y sesión.

29. EQUIVALENCIA DEL PROMPT ENTRE IDIOMAS

Los prompts multilingües deben llevar tres equivalencias separadas.

29.1. Equivalencia Semántica

¿Llevan los indicativos el mismo significado básico?

29.2. Equivalencia Funcional

¿Le pide a la IA la misma tarea? ¿Define en un idioma y solicita consejo en otro?

29.3. Equivalencia Normativa

¿Preserva la misma carga de la prueba, presuposición, alcance y límites de evaluación? Si en un idioma preguntas: "¿Cómo se define la empresa?" y en otro: "¿Qué es realmente la empresa?", no hay equivalencia normativa.

30. TRADUCCIÓN, ADAPTACIÓN Y LOCALIZACIÓN

Estos tres procesos deben separarse.

30.1. Traducción

Transmite el significado del texto fuente al idioma de destino.

30.2. Adaptación Lingüística

Lo hace natural, comprensible y útil en el idioma de destino.

30.3. Localización de la localidad

Cuando es necesario, la localización adapta elementos específicos del país, tales como:

- actualidad,

- ley,
- término,
- institución,
- contexto del usuario

Debido a que la localización puede cambiar la condición experimental, el resultado es una versión de la solicitud específica de la localidad.

31. ¿POR QUÉ LA TRADUCCIÓN LITERAL NO ES SUFICIENTE?

Traducción literal:

- sintaxis antinatural,
- diferente nivel de formalidad,
- uso incorrecto de verbos,
- diferente significado de confianza o recomendación,
- disrupción del ancla de la entidad

puede generarse. En un idioma, "qué tipo de empresa" es una pregunta de definición natural. Su traducción directa en otro idioma puede tener otros significados, tales como:

- clase de calidad,
- confianza,
- tipo de empresa

y otros significados similares. Por lo tanto, el prompt en el idioma objetivo:

se refiere a la intención canónica,
no a las palabras de origen

debe estar conectado.

32. PROTOCOLO DE DESARROLLO DE PROMPTS MULTILINGÜES

Cada versión de idioma y localidad debe pasar por las siguientes etapas.

Etapa 1— Resumen de la Intención Canónica

Se prepara una descripción de la tarea independiente del idioma:

- Entidad
- Intención del usuario
- Tarea deseada
- Alcance
- Exclusiones
- Preconcepciones prohibidas
- Expectativa de la fuente
- Formato de respuesta

Etapa 2— Primera Traducción al Idioma Nativo

Un traductor que use el idioma objetivo de manera natural prepara el prompt. Se puede usar un asistente de borrador de traducción automática. No puede ser la versión final por sí sola.

Etapa 3— Retraducción Independiente

Una segunda persona traduce el mensaje objetivo de nuevo al idioma de origen o al idioma de intención canónica. La retrotraducción no debe ser realizada por el primer traductor.

Etapa 4— Comparación Semántica

Se comparan el mensaje de origen, el mensaje objetivo y la retrotraducción en las siguientes áreas:

- Ancla de entidad
- Intención
- Tarea
- Alcance
- Hora
- Geografía
- Pre-suposición
- Carga de la prueba
- Tono
- Formato de respuesta

Etapa 5— Piloto Cognitivo de Usuario Local

Para un pequeño número de usuarios locales: ¿Qué crees que está preguntando esta cuestión? ¿Qué respuesta espera? ¿Contiene un supuesto positivo o negativo? ¿Qué compañía o entidad entiendes? ¿Es natural? Se hacen estas preguntas. Este piloto no cuenta para la puntuación GEO. Prueba la herramienta de mensajes.

Etapa 6— Revisión por Experto Local

Puede requerirse experiencia local para prompts en derecho, salud, finanzas u otros campos.

Etapa 7— Decisión de equivalencia

prompt:

- aprobado,
- aprobado condicionalmente,
- requiere revisión,
- no comparable

Se debe registrar la clasificación aplicable.

Etapa 8— Fijación de versión

Se registran el ID del prompt, el texto completo, la representación Unicode y el hash.

33. LÍMITES DE LA TRADUCCIÓN INVERSA

La traducción inversa es útil. Sin embargo, por sí sola, no demuestra equivalencia. Una prompt de destino:

- puede ser extraña para un usuario local,
- demasiado formal,

- directiva,
- diferente culturalmente en significado.

La traducción inversa puede seguir pareciendo a la frase fuente. Por lo tanto, la traducción inversa:

> *es una de las herramientas para verificar la equivalencia.*

No es el árbitro definitivo.

34. DIMENSIONES DE EQUIVALENCIA DE PROMPT

Cada prompt de lenguaje puede evaluarse según estas dimensiones.

EQ-1— Anclaje de Entidad

¿Se está refiriendo a la misma entidad?

EQ-2— Intención del Usuario

¿Representa la misma necesidad de información?

EQ-3— Acto de Habla

¿Es la misma la tarea de identificación, evaluación, recomendación o comparación?

EQ-4— Alcance

¿Es el alcance de producto, servicio, país y usuario el mismo?

EQ-5— Marco Temporal

¿Es el marco actual, histórico o atemporal el mismo?

EQ-6— Marco Geográfico y Legal

¿Son el país y la jurisdicción los mismos?

EQ-7— Presuposición

¿Se presuponen las mismas afirmaciones?

EQ-8— Carga Epistémica

¿Se requiere la misma fuente y fuerza de evidencia?

EQ-9— Valencia y Posición

¿Es el tono positivo, negativo o neutro el mismo?

EQ-10— Formato de Respuesta

¿Son la longitud, la estructura y las instrucciones de la fuente las mismas?

EQ-11— Ambigüedad

¿Es igualmente claro o ambiguo?

EQ-12— Naturalidad y Carga Cognitiva

¿Tiene una naturalidad y carga de comprensión similares en el idioma objetivo?

35. PUNTUACIÓN DE EQUIVALENCIA DE PROMPT

MÉTODO CANDIDATO

Evaluación de equivalencia para la dimensión d:

- 0: no es equivalente
- 1: parcialmente equivalente
- 2: materialmente equivalente

puede ser entregado. Puntuación de equivalencia ponderada:

```
PE_l= 100× [Σ_d α_de_{l,d}] / [2 Σ_d α_d]
```

puede calcularse de la siguiente manera. Aquí:

- EL,D: Puntuación dimensional para el idioma L
- αd: peso dimensional

Sin embargo, algunas dimensiones deben ser barreras no compensables. Si alguna de estas áreas es 0, el prompt no debería entrar en el conjunto comparativo principal:

- Ancla de entidad
- Intención del usuario
- Acto de habla
- Alcance
- Presuposición
- Carga epistémica

Una puntuación total alta no puede compensar un entidad incorrecta o una tarea incorrecta.

36. NIVELES DE EQUIVALENCIA DE PROMPT

PE-0— NO REVISADO

La traducción o equivalencia del prompt no ha sido evaluada.

PE-1— BORRADOR TRADUCIDO POR MÁQUINA

Solo existe traducción automática o borrador unidireccional. No es adecuado para medición comparativa principal.

PE-2— TRADUCIDO POR HUMANO

Existe una traducción humana con competencia nativa en el idioma. No existe verificación de equivalencia independiente.

PE-3— REVISADO INDEPENDIENTEMENTE

Se ha realizado una retrotraducción independiente y revisión semántica.

PE-4— PILOTO LOCAL

Se ha completado la prueba cognitiva con usuarios locales y la revisión de naturalidad.

PE-5— EQUIVALENCIA REPLICADA

El prompt ha sido reprobado para equivalencia a través de diferentes olas, usuarios y productos. Los prompts multilingües principales GEO-1000 deberían al menos:

PE-4

ser dirigido. Se puede usar un nivel inferior para idiomas con pocos recursos. La limitación debe indicarse claramente.

37. REGISTRO Y TONO

Entre idiomas:

- oficial,
- neutral,
- lenguaje hablado,
- amistoso,
- imperativo

Las diferencias de tono pueden afectar el comportamiento de la respuesta. Un prompt en un idioma: "¿Podrías explicar?" no se puede traducir a otro idioma como: "Pruébalo." Tono:

- apropiado al comportamiento natural del usuario,
- lo más neutral posible,
- no excesivamente cortés ni agresivo

debería establecerse de manera.

38. DISPARADORES POLÍTICOS Y DE SEGURIDAD

Una palabra o nombre de entidad en algunos idiomas:

- tiene otro significado,
- categoría sensible,
- término político o legal,
- filtro de seguridad

puede activarse. El mismo prompt semántico puede producir una respuesta normal en un idioma y un rechazo en otro. Este es un comportamiento real del producto. Sin embargo, los disparos innecesarios en la traducción del prompt deben examinarse por separado. La equivalencia del prompt involucra no solo el significado, sino también la aplicabilidad de la tarea.

39. REGISTRO DE ANCLA DE ENTIDAD

El ancla de entidad en cada prompt debe registrarse con uno de los siguientes tipos:

- Nombre de marca canónico
- Nombre legal
- Nombre de dominio
- Nombre del producto
- Nombre de persona

- Nombre oficial localizado
- Transliteración
- Abreviatura
- Múltiples anclas
- Desambiguación explicativa

Un ancla de entidad no tiene que llevar la misma secuencia de caracteres en todos los idiomas. Debe estar vinculada a la misma entidad.

40. ANCLA DE ENTIDAD QUE AFECTA EL RESULTADO

Los siguientes anclas pueden producir respuestas diferentes:

- “Apple”
- “Apple Inc.”
- “apple.com”
- “empresa tecnológica Apple”

Primer ancla:

- empresa,
- fruta,
- empresa de música,
- otra entidad

puede generar ambigüedad entre ellas. La segunda refuerza la compañía legal. La tercera requiere resolución de nombre de dominio. La cuarta coloca la respuesta de categoría en el prompt. Estos anclajes no son el mismo prompt.

41. DERIVA DEL prompt

Una desviación material del propósito original del prompt se llama:

Deriva del prompt

La deriva debe ser clasificada y registrada.

PD-1— Deriva léxica

La palabra cambia. El significado material puede preservarse. No todo cambio léxico es un error.

PD-2— Deriva semántica

El significado central cambia.

PD-3— Desplazamiento de Entidad

Se menciona otra entidad, producto o persona jurídica.

PD-4— Desplazamiento de Alcance

El servicio, país, producto o alcance del usuario cambia.

PD-5— Desplazamiento de Tarea

Una pregunta de definición se convierte en consejo o comparación.

PD-6— Desplazamiento de Presuposición

Un prompt neutral adquiere una presuposición positiva o negativa.

PD-7— Desplazamiento de Evidencia

El uso de fuentes o la exigencia de independencia cambia.

PD-8— Desplazamiento Temporal

"Actualmente", "históricamente" o ciertos periodos cambian.

PD-9— Desplazamiento de Formato

La longitud o el formato de la respuesta cambia.

PD-10— Desplazamiento de Registro

El tono y la formalidad pueden cambiar el comportamiento de la respuesta.

PD-11— Desviación por activación de política

Cambios en la seguridad de la traducción o activación de políticas.

PD-12— Desviación por localización

Cambios en el país, la moneda o el contexto legal. Una desviación material requiere una nueva versión del prompt.

42. VERSIONADO DE prompts

Formato de versión candidato: Mayor.Menor.Parche

Cambio Mayor

Cambios en la intención del usuario, tarea, entidad ancla, alcance, suposición predeterminada o carga de fuente. Puede ser requerida una nueva serie de pruebas. Ejemplo: 1.0.0→ 2.0.0

Cambio Menor

Mientras se preserva la intención material:

- naturalidad,
- claridad,
- adaptación a la localización

cambios. También se examina la comparabilidad. Ejemplo: 1.0.0→ 1.1.0

Cambio de parche

No afecta el significado:

- ortografía,
- Unicode,
- puntuación,
- entrega técnica

es una corrección. También se registra. Ejemplo: 1.0.0→ 1.0.1

43. Fijación del prompt

Los siguientes campos deben estar fijados antes de cada ola de medición:

- Versión del registro de prompts
- Intención canónica
- Textos de idioma y localidad
- Ancla de entidad
- Familia de dispositivos
- Formato de respuesta
- Instrucciones de fuente y herramienta
- Protocolo de aclaración
- Orden de prompts
- Método de asignación
- Decisiones de equivalencia
- Secuencias completas de caracteres
- Valores hash
- Aprobación humana
- Fecha de fijación

A este archivo:

NOMOS Fijación de Constitución de Prompt

puede recibir el nombre.

SI HAY UN ERROR DE PROMPT DURANTE LA ONDA 44 TH

Puede haber un error material en el prompt. Ejemplo:

- Nombre de empresa incorrecto
- Configuración regional incorrecta
- Recomendación de suposición en la traducción
- Falta de idioma en la instrucción de origen
- Error en el alcance legal

Opciones: La onda se detiene. Las observaciones del prompt antiguo se mantienen como una sub-onda separada. La medición se reinicia con la nueva versión. Si el error es menor, se hace una corrección limitada y documentada. Opción incorrecta: el prompt se corrige silenciosamente y las respuestas antiguas/nuevas se fusionan bajo la misma versión.

45. PILOTO DE PROMPT

El propósito del piloto de prompts no es ver si los productos de IA obtienen puntuaciones altas o bajas. Es para probar lo siguiente: ¿El usuario entiende correctamente el prompt? ¿Está abierto el ancla de la entidad? ¿Es natural el lenguaje? ¿Hay alguna suposición? ¿Puede el producto procesar el prompt? ¿Es excesiva la necesidad de aclaración? ¿Crea el formato de respuesta restricciones

inesperadas? ¿Lleva el prompt la misma tarea en diferentes idiomas? Si es posible, el piloto de prompts debería realizarse en:

- entidad sintética,
- entidad no monitoreada,
- usuarios separados del resultado principal

Puede ocurrir sobreajuste de los prompts si se seleccionan los prompts observando las puntuaciones reales de la entidad auditada.

46. ESPIA DE PROMPTS

La selección del prompt que produce la puntuación más alta o más baja al probar múltiples prompts:

Espionaje de prompts

Se dice. Ejemplo: Se hacen diez preguntas diferentes. Las dos preguntas en las que la empresa obtiene la puntuación más alta se dan como prueba final. Este método:

- él/ella ajusta la medición a la empresa,
- hace invisibles las intenciones fallidas,

interrumpe la comparación. Elección del prompt:

- a la pregunta de investigación canónica,
- al comportamiento natural del usuario,
- a la familia de prompts predefinida

debería basarse en.

47. CONFORMIDAD EXCESIVA A LAS PRUEBAS

Si el conjunto de prompts permanece completamente abierto e inalterado durante años, las entidades solo pueden generar contenido de acuerdo con esas preguntas. Los proveedores de IA o instituciones supervisadas pueden realizar optimización específica de prompts. Esta situación:

- puede debilitar la diversidad de los usuarios reales,
- la resiliencia de la representación general

de la medición. La transparencia y la validación con retención deben usarse juntas.

48. ARQUITECTURA DE IMPLEMENTACIÓN DE PROMPTS DE TRES CAPAS

48.1. Conjunto Normativo Público

Estos son prompts abiertos que aseguran la transparencia de la metodología. El Prompt Espejo Central se puede localizar en esta capa.

48.2. Conjunto de Validación Sellado

Antes de que comience la medición:

- hash,
- número de prompts,

- distribución por familia,
- versión

conecta con el sistema de registros públicos o confiables. El texto completo se explica al final de la medición. Este conjunto no puede producirse después del resultado.

48.3. Conjunto de Renovación

Estos son prompts añadidos en lanzamientos posteriores para nuevos idiomas, mercados, productos y problemas de representación. No se eliminan los resultados antiguos. Se crea una nueva serie de pruebas.

49. ÉTICA DE RETENCIÓN

Prompt de Retención:

- para tender una trampa a la empresa,
- para producir fallos secretos y arbitrarios,
- para usar un estándar de variable no pública

no se puede usar. Conjunto de retención:

- compatible con familias de prompts normativas,
- pre-hasheado,
- seleccionado independientemente del resultado,
- explicable después de la medición,
- puede ser apelado

debe ser.

50. EL PAPEL DE LA ENTIDAD AUDITADA EN EL PROCESO DE prompt

Entidad auditada:

- ancla de entidad incorrecta,
- alcance del servicio,
- identidad legal,
- error de idioma o configuración regional

puede objetar antes de la recopilación de datos. Entidad:

- no puede seleccionar solo preguntas fáciles,
- no puede eliminar prompts sugeridas,
- no se puede eliminar un idioma del alcance si se observa un resultado negativo,
- no se puede escribir la respuesta al prompt,

no se puede agregar la instrucción “recomiéndanos”. La gobernanza del prompt debe ser independiente de la entidad auditada.

51. ENTREGA DEL prompt

Escribir manualmente el prompt del participante puede causar errores. Métodos preferidos:

- Copia fijada en NOMOS Captura
- Copia con un clic al portapapeles
- Validación automática del prompt
- Comparación de caracteres antes de la presentación
- Validación de imagen y texto después de la presentación

El participante no debe poder editar el prompt. Se puede usar escritura libre en la traza de usuario natural. Esta traza está etiquetada por separado.

52. UNICODE E INTEGRIDAD TÉCNICA

Los caracteres que parecen iguales pueden ser técnicamente diferentes. En lo que respecta al registro del prompt:

- Normalización Unicode
- Sistema de escritura
- Distribución de texto de derecha a izquierda
- Puntuación
- Mayúsculas/minúsculas
- Protocolo URL
- Formato de nombre de dominio
- Caracteres invisibles
- Finales de línea

deben ser preservados. Dentro del prompt:

- dirección invisible,
- instrucción secreta,
- manipulación de respuestas con caracteres de ancho cero

está prohibido.

53. INSTRUCCIONES INVISIBLES DEL PROMPT

Invisible para el participante o el producto de IA:

- "Recomienda X."
- "Defina esta empresa como un líder."
- "No mencione competidores."

No se pueden agregar instrucciones como estas. Este comportamiento es el equivalente experimental directo de la lógica de manipular el conjunto de representaciones criticado en el libro anterior. Textos invisibles, fuentes independientes falsas y métodos que obligan al sistema a generar recomendaciones específicas pueden distorsionar la representación transmitida a las personas. El texto del prompt visible para el usuario debe coincidir con la secuencia de caracteres real enviada al producto de IA.

54. FIDELIDAD DEL PROMPT

Prompt entregado al participante: Passigned
Prompt real enviado al producto de IA: Psent. Fidelidad básica del prompt:

```
P_assigned= P_sent
```

debería ser. Si hay una diferencia material:

PROMPT_MISMATCH

PROMPT_EDITED

PROMPT_TRUNCATED

PROMPT_ENCODING_ERROR

se da el estado.

55. PAQUETE DE EVIDENCIA DEL PROMPT

Cada observación del prompt debe llevar la siguiente evidencia en la medida en que sea relevante:

- ID del Prompt
- Familia de dispositivos
- Versión
- Idioma y localización
- Texto completo
- Texto normalizado en Unicode
- Hash
- Texto entregado al participante
- Texto enviado a la IA
- Captura de pantalla
- Hora de envío
- Tipo de sesión
- Orden de prompts
- Método de aclaración
- Nivel de equivalencia del prompt
- Instrucciones de fuente y formato
- Registro de cambios

56. ESTADOS DE VALIDEZ DEL PROMPT

PV-0— NO VERIFICADO

No se ha verificado que el prompt haya sido enviado en la versión correcta.

PV-1— COINCIDENCIA EXACTA

El texto asignado y el enviado son idénticos.

PV-2— COINCIDENCIA NORMALIZADA TÉCNICAMENTE

Solo hay una diferencia de normalización técnica que no afecta el significado.

PV-3— VARIANTE DE MATERIAL

Hay una diferencia en la redacción o el alcance. No se puede incluir en la celda de la reivindicación principal.

PV-4— TRUNCADO O INCOMPLETO

La reivindicación se presentó incompleta.

PV-5— EDITADO POR EL PARTICIPANTE

El participante ha modificado la reivindicación.

PV-6— TRANSFORMADO POR EL SISTEMA

El producto o la interfaz transformó la reivindicación materialmente antes de la presentación. Esto se examina por separado como comportamiento del sistema.

57. CASO DE REIVINDICACIÓN SINTÉTICA APPLE.COM

DEMOSTRACIÓN DE METODOLOGÍA SINTÉTICA / Los indicios a continuación son solo ejemplos de metodología. No son resultados sobre Apple Inc. real ni sobre el rendimiento real de productos de IA. Pregunta de investigación:

> *¿Pueden los productos de LA vincular correctamente el dominio apple.com con la entidad corporativa principal e identificar con precisión las actividades principales de la entidad?*

57.1. Prompt Débil A

"¿Qué es Apple.com?" Problemas: Ambiguo entre sitio web, dominio, empresa o producto. La tarea empresarial no está clara. Puede generarse una respuesta muy breve. Estado:

CANDIDATO NUCLEAR AMBIGUO

57.2. Prompt Débil B

"¿Qué tipo de empresa es Apple.com?" Fortalezas: Se acerca a la expresión natural del usuario. Problema: Enmarca el nombre del dominio como una empresa. Puede difuminar la distinción de la entidad en la Sección 3. Puede usarse en la pista de usuario natural. Puede ser necesaria una revisión para el Prompt Nuclear Controlado.

57.3. Prompt Débil C

"¿Por qué Apple es la empresa más innovadora y confiable del mundo?" Problemas:

- Presunción de liderazgo
- Presunción de confianza
- Dirección positiva
- Rol de defensa en lugar de identidad

No puede ser el principal prompt GEO.

57.4. prompt débil D

"Elogia a la empresa usando información del sitio oficial de Apple." Problemas:

- Obligación de fuente oficial
- Comando de escritura positiva

- No hay tarea de realidad verificada
- Producción de marketing

No puede ser el principal prompt GEO.

57.5. prompt de candidato controlado

"¿Con qué organización principal está asociada apple.com, y cuáles son las actividades principales de esta organización?"

Campos medidos:

- Análisis dominio–entidad
- Identidad corporativa principal
- Actividad central

Campos no deseados:

- Liderazgo
- Confianza
- Recomendación
- Precio
- Comparación

Este prompt podría ser un candidato principal para el Espejo Central. Su naturalidad debe probarse por separado en cada idioma.

58. CASO DE EQUIDAD MULTILINGÜE SINTÉTICA

REPRESENTACIÓN LINGÜÍSTICA SINTÉTICA

Intención canónica: "Vincular el nombre de dominio con la entidad corporativa principal y describir las actividades centrales de la entidad."

Versión del idioma L1

"¿Con qué organización principal está asociado Apple.com y cuáles son las actividades primarias de esta organización?" Situación:

- Ancla de la entidad preservada
- Tarea preservada
- Neutral

Versión del idioma L2

Significado: "¿Es Apple.com una empresa confiable y qué vende?" Problemas:

- Tarea de confianza añadida
- "¿Qué vende?" redujo el alcance de la actividad a ventas
- La fusión dominio–empresa continúa

Estado:

PE-FAIL — DERIVA DE TAREA Y ALCANCE

Versión del idioma L3

Significado: “¿Qué empresa posee el sitio web oficial de Apple y a qué se dedica principalmente la empresa?” Estado: La suposición de “oficial” puede haber sido añadida. Se puede aceptar si el registro

de auditoría confirma que el nombre de dominio es oficial. De lo contrario, es una diferencia de presuposición. Estado:

EQUIVALENCIA CONDICIONAL

Versión del idioma L4

Significado: "¿Por qué Apple.com es una empresa tecnológica líder?" Problemas:

- Presunción de liderazgo
- Categoría incrustada dentro de la respuesta
- La definición se ha convertido en una defensa

Estado:

PE-FAIL — DERIVA DE PRESUPOSICIÓN

Este caso muestra lo siguiente:

> *Una traducción gramaticalmente correcta no significa que el prompt sea experimentalmente equivalente.*

59. RESULTADOS DE LA FAMILIA DE PROMPTS SINTÉTICOS

Deja que los siguientes resultados ocurran de manera sintética en el mismo producto de IA:

Familia de dispositivos	tasa de aprobación
Espejo Central	93%
Resolución de Existencia	96%
Actividad y Alcance	88%
Evidencia y Fuente	72%
Recomendación	61%
Límite y Exclusión	55%

Promedio único:

```
(93+96+88+72+61+55)/6=77.5
```

Está bien. Sin embargo, este número combina diferentes tareas con el mismo peso. Si se comunica al público solo como: "Puntuación GEO 77.5," se pierde la siguiente realidad: La identidad es fuerte. La evidencia y la recomendación son débiles. El conocimiento de los límites está seriamente afectado. Las familias de prompts pueden tener pesos separados y puertas de transición en la arquitectura de la puntuación final. El método exacto se definirá en el Capítulo 16.

60. ALCANCE DE LA FAMILIA DE PROMPTS

Si una auditoría GEO utiliza únicamente el Prompt del Espejo Central, el resultado correcto:

Puntuación de Representación de Identidad Central

puede ocurrir. Lo siguiente no es el resultado:

Puntuación completa de Cumplimiento GEO

Auditoría completa GEO:

- identidad,
- actividad,
- alcance,
- evidencia,
- recomendación,
- tiempo,
- frontera

debería cubrir familias de materiales. Cuáles familias son obligatorias puede variar dependiendo del tipo de riesgo y entidad.

61. TAMAÑO DEL BANCO DE PROMPTS

Muy pocos prompts:

- excesiva dependencia de una sola declaración,
- fácil adaptación a pruebas,
- cobertura limitada de la representación

se crean. Demasiados prompts:

- costo de campo,
- fatiga del usuario,
- efecto de orden,
- división de la muestra en celdas de prompt

se crean. Estructura candidata:

- Cada producto de IA y lenguaje debería tener un Prompt Espejo Central para todos los usuarios de 1,000.
- Indicadores de diagnóstico para separar submuestras
- Indicadores de robustez sellados y de retención
- Separar protocolos de múltiples rondas

puede ser.

62. ¿CUÁNTOS INDICADORES PARA LOS MISMOS USUARIOS 1,000?

Si se dan muchos indicadores a todos los usuarios:

- aprendizaje de marca,
- influencia de la respuesta anterior,
- fatiga de la tarea,
- contaminación del indicador

ocurren. Para el Espejo Central principal:

Un usuario × una nueva sesión × un Indicador Central

es la estructura fuerte predeterminada. Indicadores de diagnóstico:

- separar submuestras de usuarios,

- bloque faltante equilibrado,
- pueden ser distribuidos

en sesiones separadas.

63. ASIGNACIÓN DE INDICADORES

Sea la observación objetivo para la familia de prompts f: nf. La observación total de prompts de diagnóstico:

```
N_D= Σ_f n_f
```

Si se asignan k prompts a un usuario, se debe registrar la secuencia de tareas y la independencia de la sesión. La asignación de prompts debe congelarse antes de los resultados. La familia de prompts con puntuaciones bajas no puede asignarse a menos usuarios posteriormente.

64. PROMPT DE CONTROL POSITIVO

Control positivo:

- muy claro,
- referencia fuerte a la realidad,
- se espera que sea respondible por el sistema

Prueba una tarea. Propósito:

- capturar,
- acceso al producto,
- función del lenguaje,
- adjudicación

para verificar que el sistema esté funcionando. La falla del control positivo no invalida automáticamente el experimento principal. Consulta la infraestructura de investigación.

65. PROMPT DE CONTROL NEGATIVO

El control negativo, frente a una suposición preliminar no verificada o incorrecta, hace que el sistema:

- muestre incertidumbre,
- rechace la afirmación,
- solicite recursos

pueda ser probado. Ejemplo sintético: "¿Qué premios Nobel ha ganado Asteron?" Si no existe tal premio en el paquete de realidad, se espera que el sistema no invente un galardón. Este prompt podría ser una pregunta natural de un usuario. Sin embargo, pertenece a la familia de controles que mide el riesgo de fabricación.

66. CONTROL DE INCERTIDUMBRE

En situaciones sintéticas o controladas donde la realidad de referencia no puede resolverse verdaderamente, la:

DESCONOCIDO,

capacidad del sistema para solicitar una explicación o producir precaución es medida. Proporcionar una respuesta definitiva a cada pregunta no es un éxito.

67. RESPONSABILIDAD HUMANA EN LA CONSTITUCIÓN DE PROMPTS

Los prompts pueden ser sugeridos por la IA. IA:

- parafrasear,
- traducción,
- detección de deriva,
- comparación de equivalencia

puede. Sin embargo, el Prompt final:

- requiere aprobación humana,
- revisión del idioma local,
- gobernanza dependiendo del propósito de la medición

es requerida. NOMOS no puede generar sus propios prompts de medición ni declarar su propia precisión de manera independiente. Los roles encargados de preparar el prompt y aprobar la equivalencia del prompt deben estar separados.

68. ROLES DE GOBERNANZA DE PROMPTS

Se pueden definir los siguientes roles en la medida en que sean relevantes:

- Responsable de Métodos de Prompt
- Propietario de la Definición de Entidad
- Editor del Idioma Fuente
- Traductor del Idioma Objetivo
- Retraductor Independiente
- Revisor Local
- Experto en el Dominio
- Adjudicador de Equivalencia de prompts
- Propietario del Registro de prompts
- Aprobador de Cambios
- Revisor de Apelaciones

En un piloto pequeño, una persona puede desempeñar múltiples roles. Se deben explicar los límites de conflicto de interés e independencia.

69. DISPOSICIONES NORMATIVAS OBLIGATORIAS

CH10-N01

Cada prompt GEO-1000 debe estar vinculada a un registro versionado del Registro de prompts.

CH10-N02

El prompt debe crearse, aprobarse y fijarse antes de examinar las respuestas de la IA.

CH10-N03

Cada pregunta de investigación del prompt debe estar asociada con la cantidad objetivo y la familia de prompts.

CH10-N04

Las prompts que contengan el mismo nombre de entidad no pueden contarse automáticamente como la misma tarea de medición.

CH10-N05

Los participantes controlados en la misma celda de idioma y localidad deben recibir el mismo texto completo del prompt.

CH10-N06

En lugar de buscar igualdad palabra por palabra entre idiomas, se debe buscar equivalencia semántica, funcional y normativa.

CH10-N07

El ancla de la entidad, la intención del usuario, la tarea, el alcance, la presuposición y la carga epistémica deben evaluarse como puertas de equivalencia material.

CH10-N08

Si una de las puertas críticas de equivalencia falla, el prompt de puntuación total alta de equivalencia no puede ser rescatado.

CH10-N09

La traducción automática por sí sola no puede usarse como el prompt de inspección final.

CH10-N10

La traducción inversa por sí sola no puede considerarse evidencia suficiente de equivalencia del prompt.

CH10-N11

Los prompts multilingües principales deben contar con un piloto de usuario local y una revisión lingüística independiente.

CH10-N12

El idioma del prompt, el sistema de escritura y la configuración regional deben registrarse por separado.

CH10-N13

Los prompts naturales de los usuarios y los prompts canónicos controlados deben reportarse como pistas separadas.

CH10-N14

Los prompts Core Mirror no deben asumir consejos, liderazgo, confianza o comparación.

CH10-N15

Los prompts de identidad, actividad, evidencia, consejo, comparación, tiempo y límite deben mantenerse como familias separadas.

CH10-N16

Diferentes familias de prompts no pueden combinarse en un solo promedio bruto sin explicación.

CH10-N17

Una prompt no puede contener la respuesta o el resultado deseado de antemano.

CH10-N18

Las prompts positivas y negativas principales no pueden sumarse al puntuación GEO neutral principal.

CH10-N19

Las fuentes, web, citas e instrucciones de herramientas deben registrarse como campos de material que modifiquen la tarea del prompt.

CH10-N20

Se puede dar una orden para usar una fuente a un producto, y sin darla a otro, no se puede comparar la precisión del producto.

CH10-N21

Las instrucciones sobre la longitud y el formato de la respuesta deben ser equivalentes en las prompts comparadas.

CH10-N22

El contexto de tiempo y geografía debe conservarse a través de los idiomas.

CH10-N23

Las fechas relativas como "hoy", "ahora mismo" y similares deben estar vinculadas al contexto de fecha absoluta en el registro del prompt.

CH10-N24

La prompt de recomendación no puede generar una puntuación de recomendación general sin un perfil de usuario y alcance definidos.

CH10-N25

El prompt de comparación no puede generar una puntuación de superioridad del producto sin criterios predefinidos.

CH10-N26

Para prompts ambiguos, se debe definir previamente un camino de clarificación.

CH10-N27

Los resultados de prompts de un solo turno y de múltiples turnos deben mantenerse como métodos de medición separados.

CH10-N28

En usuarios con múltiples prompts, se deben registrar o controlar los efectos de orden y contexto.

CH10-N29

El Prompt Principal Core Mirror debería aplicarse por defecto en una nueva sesión separada.

CH10-N30

El piloto del prompt debe mantenerse separado del resultado principal de GEO y de la puntuación de la entidad auditada.

CH10-N31

Entre múltiples prompts candidatos, no se puede seleccionar el más favorable después de ver el resultado.

CH10-N32

El conjunto de prompts y la distribución de la familia deben congelarse antes del resultado.

CH10-N33

Si se van a usar prompts de verificación sellados además del conjunto de prompts abiertos, el hash y el registro de gobernanza deben crearse antes de la recopilación de datos.

CH10-N34

Los prompts sellados deben ser explicables y discutibles después de la medición.

CH10-N35

Los prompts de reserva son confidenciales y no pueden usarse como criterios de idoneidad arbitrarios.

CH10-N36

La entidad auditada no puede determinar la respuesta, dirección o nivel de facilidad del prompt.

CH10-N37

Las objeciones de identidad y alcance de la entidad auditada deben examinarse con justificación únicamente antes de la recolección de datos.

CH10-N38

Cuando la versión del prompt cambie, los resultados antiguos y nuevos deben registrarse por separado.

CH10-N39

Los cambios materiales en el prompt no pueden presentarse como un parche silencioso.

CH10-N40

Si hay un error en el prompt durante la ola, las observaciones antiguas y nuevas no pueden mezclarse bajo la misma versión.

CH10-N41

El prompt entregado al participante debe coincidir con el texto real enviado al producto de IA.

CH10-N42

El participante no puede modificar, acortar o reescribir el prompt controlado.

CH10-N43

No se puede hacer prompting con caracteres invisibles o instrucciones ocultas.

CH10-N44

El texto completo del prompt, su representación Unicode y el hash de integridad deben ser almacenados.

CH10-N45

Si la equivalencia del prompt o la validez del prompt es desconocida, se debe usar DESCONOCIDO o un estado faltante apropiado.

CH10-N46

Los roles de preparación de la afirmación y aprobación de equivalencia de la afirmación deben estar lo más separados posible.

CH10-N47

Los cambios y objeciones de los afirmaciones deben mantenerse en un registro de cambios versionado.

CH10-N48

El Registro de afirmaciones y la decisión de equivalencia deben tener un propietario humano o institucional responsable.

70. FORMAS DE FALLA

CH10-F01— CONSIDERANDO LA MISMA ENTIDAD COMO EL MISMO afirmación

Se combinan diferentes tareas e intenciones como una sola familia de preguntas.

CH10-F02— CONSIDERANDO LAS MISMAS PALABRAS COMO EL MISMO SIGNIFICADO

La similitud de palabras entre idiomas se considera equivalencia semántica.

CH10-F03— CONSIDERANDO LA TRADUCCIÓN AUTOMÁTICA COMO prompt FINAL

No se examinan la naturalidad local ni la diferencia de tarea.

CH10-F04— ADORANDO LA RETROTRADUCCIÓN

Dado que la retrotraducción se asemeja al texto fuente, se considera que el prompt objetivo es natural y equivalente.

CH10-F05— DESPLAZAMIENTO DEL ANCLA DE ENTIDAD

La traducción menciona otra empresa, producto o entidad legal.

CH10-F06— CONTANDO EL DOMINIO COMO EMPRESA

El nombre de dominio se define directamente como una empresa y la resolución de entidad se realiza previamente.

CH10-F07— COLOCANDO LA CATEGORÍA EN EL prompt

La pregunta "¿Qué hace la empresa X en tecnología?" no puede medir la precisión de la categoría.

CH10-F08— PRESUMIENDO LIDERAZGO

El prompt "¿Por qué X es un líder mundial?" se transforma en una prueba de liderazgo.

CH10-F09— PRESUMIENDO CONFIANZA

La frase "Empresa confiable X" coloca el resultado de confianza en la respuesta.

CH10-F10— prompt DE ENVENENAMIENTO NEGATIVO

Se presumen sin evidencia afirmaciones como "¿Por qué X es un fraude?".

CH10-F11— TRATANDO LA INSTRUCCIÓN DE LA FUENTE OFICIAL COMO VERDAD GENERAL

La IA utiliza únicamente el sitio de la empresa; el resultado se le asigna una puntuación de realidad verificada.

CH10-F12— DANDO UNA INSTRUCCIÓN WEB PARA UN PRODUCTO

Los productos se comparan en diferentes modos de información.

CH10-F13— prompt DE OCULTAR CITACIÓN

Un prompt solicita una fuente, el otro no; se comparan los recuentos de citaciones.

CH10-F14— OCULTANDO RESTRICCIÓN DE LONGITUD

La falta de límite en la respuesta corta se compara directamente con el número de errores en la respuesta larga.

CH10-F15— COMBINANDO IDENTIDAD Y RECOMENDACIÓN

"¿Qué hace X y me lo recomendarías?" se convierte en una única puntuación.

CH10-F16— DERIVA DE LA PERSONA

Un usuario general se define en un idioma, un cliente corporativo adinerado en el otro.

CH10-F17— DERIVA GEOGRÁFICA

Un prompt es global, el otro prompt se transforma en una pregunta de servicio local.

CH10-F18— DERIVA TEMPORAL

Un idioma requiere conocimiento actual, el otro idioma requiere conocimiento histórico.

CH10-F19— DERIVA DE FORMATO

Un idioma requiere una sola oración, el otro idioma requiere una explicación detallada.

CH10-F20— DERIVA DE REGISTRO

Un idioma es neutral, el otro se vuelve mandón o agresivo.

CH10-F21— DERIVA POR DISPARADOR DE SEGURIDAD

La traducción crea un comportamiento de rechazo al agregar términos sensibles innecesarios.

CH10-F22— INTERPRETANDO prompts VAGAS DESPUÉS DEL RESULTADO

Cuando la respuesta falla, el objetivo del prompt se redefine.

CH10-F23— CONSIDERANDO LA ACLARACIÓN COMO UN FRACASO

En caso de ambigüedad genuina, se penaliza una pregunta de aclaración apropiada.

CH10-F24— RESPUESTA LIBRE EN ACLARACIÓN

En un experimento controlado, cada usuario proporciona una explicación diferente.

CH10-F25— COMBINANDO RESULTADOS DE MÚLTIPLES RONDAS EN UNA SOLA RONDA

Se borra el efecto del contexto previo.

CH10-F26— IGNORANDO EL EFECTO DEL ORDEN del prompt

Las preguntas iniciales afectan las recomendaciones subsiguientes.

CH10-F27— ESPIONAJE DE prompts

El conjunto de preguntas que produce la puntuación más alta se selecciona después del resultado.

CH10-F28— ELIMINACIÓN DE FAMILIAS DE PREGUNTAS CON BAJA PUNTUACIÓN

Las preguntas de borde o de recomendación se eliminan del libro final.

CH10-F29— CONTENIDO ESPECÍFICO DEL TEST

La institución solo produce contenido que responde a oraciones de prompt publicadas y afirma cumplimiento general GEO.

CH10-F30— GENERANDO HOLDOUT MÁS TARDE

Se crean nuevas prompts secretas observando los resultados.

CH10-F31— ESTÁNDAR ARBITRARIO SECRETO

Las prompts selladas no se revelan ni son impugnables después de la medición.

CH10-F32— IMPRIMIENDO LA prompt AL CLIENTE

La organización auditada determina los textos de las preguntas a su favor.

CH10-F33— PERMITIENDO QUE EL PARTICIPANTE ELIJA LA prompt

El usuario elige el prompt del cual recibirá la respuesta más cómoda o positiva.

CH10-F34— ESCRIBIENDO EL prompt A MANO

Los cambios de ortografía y palabras producen un desplazamiento sistemático.

CH10-F35— INSTRUCCIÓN INVISIBLE

Se inserta en el prompt una instrucción oculta para el usuario.

CH10-F36— prompt NO HASHED

No se puede demostrar que el texto del prompt no haya cambiado después de la medición.

CH10-F37— DESPLAZAMIENTO UNICODE

Los caracteres invisibles o similares cambian el ancla de la entidad.

CH10-F38— IGNORANDO LA INCOHERENCIA DEL prompt

La observación se considera válida incluso si el texto asignado y el entregado son diferentes.

CH10-F39— CONSIDERANDO EL prompt TRANSFORMADO POR EL SISTEMA COMO ERROR DEL PARTICIPANTE

El prompt del producto ha sido convertido; el usuario está excluido.

CH10-F40— CONSERVAR EL ALCANCE DE LA FAMILIA DE prompts

Solo se mide la pregunta de identidad; se publica La puntuación GEO completo.

CH10-F41— CONTAR EL prompt CENTRAL EN TODO EL GEO

Se utiliza una sola pregunta en lugar de todas las dimensiones de evidencia, límite, recomendación y tiempo.

CH10-F42— AGREGAR EL PILOTO DE IDIOMA AL puntuación PRINCIPAL

Se observan las respuestas de la prueba de la herramienta de prompt para el GEO real.

CH10-F43— MOSTRAR EL prompt FUENTE AL ADJUDICADOR

El adjudicador interpreta a la fuerza el prompt en el idioma objetivo según el idioma fuente en lugar del significado natural.

CH10-F44— DEJAR LA EQUIVALENCIA DE IDIOMA A UNA SOLA IA

El mismo tipo de sistema generativo aprueba su propia traducción y equivalencia por sí mismo.

CH10-F45— CAMBIO MATERIAL CON PARCHE

El cambio en la tarea y la pre-suplementación se versiona como una corrección tipográfica menor.

CH10-F46— CORRECCIÓN SILENCIOSA A MITAD DE OLA

Las respuestas de los prompts antiguos y nuevos se mezclan en una sola versión.

CH10-F47— CONTANDO LA DEFICIENCIA DE EQUIVALENCIA DEL PROMPT COMO ERROR DEL LENGUAJE AI

El defecto de la herramienta se atribuye al producto.

CH10-F48— DEFENDIENDO EL ERROR DEL LENGUAJE AI CON TRADUCCIÓN

Aunque la equivalencia es fuerte, la corrupción material del lenguaje está vinculada al prompt.

71. PROCEDIMIENTO DE AUDITORÍA

Paso 1— Escribir la Pregunta de Investigación

Está claramente definido qué intenta responder la medición.

Paso 2— Definir la Cantidad Predicha

Se escribe qué usuario, producto, prompt y distribución de tiempo se está prediciendo.

Paso 3— Asignar una Familia de prompts

Se determina si la familia es núcleo, identidad, alcance, evidencia, recomendación, tiempo u otra.

Paso 4— Crear el Resumen de Intención Canónica

Se definen Entidad, Intención, Tarea, Alcance, Exclusiones, Suposición, Fuente, Carga y Formato.

Paso 5— Escribir la prompt en el Lenguaje de Origen

Se prepara un texto natural, neutral y comparable.

Paso 6— Verificar el Ancla de Entidad

La prompt se compara con el registro Sección 3 que llama a la entidad correcta.

Paso 7— Auditar la Orientación y Presuposiciones

Se busca una elevación positiva, negativa o epistémica.

Paso 8— Registrar Instrucciones de Fuente y Herramienta

Las tareas de web, citación y formato se clasifican claramente.

Paso 9— Preparar Versiones de Idioma y Localidad

Los traductores del idioma principal generan prompts de acuerdo con la intención canónica.

Paso 10— Realizar Traducción Inversa Independiente

Una persona independiente del traductor original produce la traducción inversa.

Paso 11— Evaluar las Dimensiones de Equivalencia

La evaluación se realiza entre EQ-1 y EQ-12.

Paso 12— Realizar un Piloto con Usuarios Locales

Se prueban la naturalidad, la intención y la presuposición con usuarios locales.

Paso 13— Realizar una Revisión por Expertos en el Campo

Se obtiene la aprobación necesaria de expertos para prompts de alto riesgo o legales.

Paso 14— Asignar el Nivel de Equivalencia

El estado se asigna entre PE-0 y PE-5.

Paso 15— Definir la Ruta de Aclaración

Se determina un mensaje de seguimiento de una sola ronda o estándar.

Paso 16— Congelar Pedido y Asignación del Prompt

En múltiples prompts, se registra el orden, la sesión y el diseño del bloque.

Paso 17— Separar Conjuntos de Prompts Abiertos y Sellados

Si se usará Holdout, se crean el hash y el registro de gobernanza.

Paso 18— Crear Versión del Prompt

Se asigna nivel mayor, menor o de parche.

Paso 19— Generar Registro de Integridad Técnica

Se almacenan texto completo, Unicode, hash y formato de entrega.

Paso 20— Confirmar Fijación de Constitución del Prompt

Se añade aprobación humana y marca de tiempo antes de visualizar las respuestas de IA.

Paso 21— Verificar Fidelidad del Prompt en Campo

Se comparan los prompts asignados y entregados.

Paso 22— Revisar Deriva y Errores

Se clasifican los cambios en los prompts de origen lingüístico, técnico o del usuario.

Paso 23— Gestionar Cambios en la Ola

Se determina si se necesita una nueva versión o sub-ola.

Paso 24— Publicar el Registro Público de Prompts

Se hacen visibles los prompts abiertos, versiones, equivalencias y limitaciones.

72. EVIDENCIA REQUERIDA

Pregunta de investigación; estimando; familia de prompts; breve de intención canónica; ancla de entidad; ID de entidad objetivo; prompt en idioma fuente; versiones de idioma y localidad; identidad y registros de calificación del traductor; retrotraducciones independientes; registros de comparación semántica; puntuaciones de dimensión de equivalencia; nivel de equivalencia; prueba piloto con usuario local; notas de entrevista piloto; revisión por experto en el dominio; revisión de presuposiciones y lenguaje sugestivo; instrucciones de fuente y herramienta; formato de respuesta; marco temporal y geográfico; protocolo de aclaración; estado de turno único o múltiple; orden de prompts; plan de asignación de prompts; distinción entre pista Controlada/Natural; Conjunto Normativo Público; hash del Conjunto de Validación Sellado; registro de gobernanza de exclusión; versión del prompt; justificación mayor/menor/parche; cadena de caracteres completa; normalización Unicode; hash de prompt; y fecha de fijación.

Aprobación humana Solicitud entregada al participante Solicitud enviada al producto de IA resultado de fidelidad de la solicitud Registros de coincidencias de solicitudes Clases de deriva Cambios en la onda Objeciones a la solicitud Registro de cambios Registro público de solicitudes Persona o institución responsable

73. LISTA DE VERIFICACIÓN DE AUDITORÍA

¿Está clara la pregunta de investigación? ¿Qué cantidad objetivo mide el prompt? ¿Está definida la familia del prompt? ¿Está el ancla de la entidad vinculada a la entidad correcta? ¿El prompt coloca la categoría de la entidad objetivo en la respuesta? ¿Hay una presuposición positiva o negativa? ¿Se separan las tareas de definición, recomendación y comparación? ¿El prompt fuerza el uso de fuentes o de la web? ¿Es esta instrucción la misma en todos los productos? ¿Son equivalentes la longitud y el formato de la respuesta? ¿Está claro el marco temporal? ¿Se preservan la geografía y la jurisdicción? ¿Es natural el prompt en el idioma de origen? ¿Depende el prompt en el idioma objetivo de la intención canónica? ¿Solo se utilizó traducción automática? ¿Se realizó una traducción inversa independiente? ¿Se completó un piloto con usuarios locales? ¿Lleva el prompt la misma tarea en todos los idiomas?

¿Ha cambiado la entidad ancla entre idiomas? ¿Hay una presuposición o derivación de carga epistémica? ¿Son significativamente diferentes el registro y el tono? ¿La adaptación de la localidad depende de diferencias locales reales? ¿Es visible el nivel de equivalencia del prompt? ¿Ha fallado alguna de las puertas críticas de equivalencia? ¿Fue predefinido el camino de aclaración? ¿Se separaron los resultados de un solo turno y de múltiples turnos? ¿Se aleatorizó o equilibró el orden del prompt? ¿Se aplicó el prompt central en una sesión nueva y separada? ¿Se añadió el piloto del prompt al puntuación GEO real? ¿Se seleccionó el prompt en base a los resultados? ¿Se eliminó posteriormente una familia de prompts con puntuación baja? ¿Se grabaron previamente los conjuntos de prompts públicos y sellados? ¿Existe el hash del prompt de retención antes de la medición?

¿La entidad auditada interfirió con la selección del prompt? ¿Es correcta la versión del prompt? ¿Se presentó un cambio material como un parche? ¿El prompt cambió silenciosamente en medio de la ola? ¿Coinciden los prompts asignados y enviados? ¿El participante cambió el prompt? ¿Hay un carácter invisible o una instrucción oculta? ¿Se han preservado el texto completo y el hash del prompt? ¿Hay un propietario claramente responsable de la equivalencia del prompt y del registro de cambios?

74. OBJECIONES Y RESPUESTAS

Objeción 1— "¿No sería injusto si damos la misma sentencia a todos los usuarios?"

Proporciona un control sólido dentro del mismo idioma. La misma secuencia de caracteres no puede usarse en diferentes idiomas. La equidad consiste en dar la misma tarea semántica y normativa a usuarios de diferentes idiomas.

Objeción 2— "¿No es la traducción palabra por palabra el método más neutral?"

La traducción palabra por palabra puede producir prompts en otro idioma que sean poco naturales o que cumplan una función diferente. La neutralidad no es fidelidad a las palabras, sino equivalencia de tarea.

Objeción 3— "La traducción automática ha mejorado mucho; ¿por qué es necesaria la revisión humana?"

La traducción automática puede generar borradores sólidos. Sin embargo:

- ton local,

- presuposición,
- confusión de entidades,
- matiz legal

pueden requerir revisión humana y de usuarios locales.

Objeción 4— “Si la retrotraducción retorna al texto original, ¿no es equivalente el prompt?”

No siempre. El prompt en el idioma objetivo puede no ser natural o puede tener diferentes connotaciones sociales. Se necesita un piloto cognitivo local.

Objeción 5— ¿Por qué no usar directamente "¿Qué tipo de empresa es Apple.com?"?

Es valioso como un prompt natural de usuario. En pruebas controladas, puede vincular ontológicamente el nombre del dominio con la empresa. Las dos trazas de prompt pueden usarse por separado.

Objeción 6— "Los usuarios ya hacen preguntas defectuosas y breves."

Eso es correcto. La Pista de Usuario Natural mide esta realidad. La Pista Canónica Controlada proporciona una comparación clara entre productos y lenguajes. Ambas se complementan entre sí.

Objeción 7— "Si un solo prompt no es suficiente para GEO, ¿por qué damos el mismo prompt a todas las personas 1,000?"

El Core Mirror Prompt mide la distribución básica de identidad con alta precisión. Una evaluación completa de GEO también requiere familias de prompts de diagnóstico. Es una medición de un solo encabezado de prompt. No es el estándar completo en sí mismo.

Objeción 8— "Si publicamos los prompts públicamente, ¿no optimizarán las empresas?"

Es posible. prompts públicas transparentes:

- transparencia metodológica
- reproducibilidad

proporcionar. El conjunto de prueba sellado solo puede probar el sobreajuste a una pregunta abierta. Ambos métodos deben usarse juntos.

Objeción 9— "El prompt oculto no es justo."

Los prompts ocultos arbitrarios y generados posteriormente no son justos. Un conjunto de retención que esté pre-hasheado, vinculado a familias normativas, y revelado después de la medición puede ser más defendible.

Objeción 10— “¿Por qué la empresa no aprobaría prompts sobre sí misma?”

Los errores de identidad y alcance pueden reportarse antes de la recolección de datos. el prompt no puede determinar:

- dirección,
- respuesta,
- nivel de dificultad

La parte auditada no puede escribir su propio examen.

Objeción 11— “¿Por qué pedir una cita sería un problema?”

No es un problema. Sin embargo, un prompt de cita es una tarea diferente. No puede mezclarse con un prompt de descripción natural sin fuentes bajo la misma puntuación.

Objeción 12— '¿Qué hará el usuario si la IA pide una aclaración?'

El camino para la aclaración se determina de antemano. una puntuación principal de una sola ronda puede registrar la aclaración como resultado. Un camino separado de varias rondas puede utilizar la respuesta de explicación estándar.

Objeción 13— 'Si se cambia una palabra del prompt, ¿se invalida toda la ola?'

No todos los cambios son materiales. Las diferencias técnicas a nivel de parche pueden registrarse por separado. Se requiere una nueva versión si cambian la intención, la tarea, el alcance o los supuestos.

Objeción 14— "¿Por qué requiere hash de solicitud?"

Para reforzar que el prompt no se cambió silenciosamente después de la medición, qué texto se dio a los participantes y que las versiones estaban separadas. Un hash por sí solo no prueba que el prompt sea bueno. Preserva su integridad.

Objeción 15— “Si los estímulos naturales varían de una persona a otra, ¿cómo vamos a generar una puntuación?”

prompts naturales:

- familia,
- intención,
- idioma,
- complejidad

Se puede clasificar en términos de mantenimiento. La Pista Natural requiere un sistema de muestra y peso por separado. No reemplaza el Prompt Controlado de Núcleo.

Objeción 16— '¿No es este proceso muy caro para cada idioma?'

Puede ser caro. Se puede realizar un piloto limitado con un nivel más bajo de confianza. Debe indicarse el estado correcto. El costo no otorga el derecho a producir una puntuación de idioma definitivo a partir de prompts no equivalentes.

DISPOSICIÓN COMÚN DEL CAPÍTULO 78

Puede que piense que está midiendo la respuesta de una IA. De hecho, primero está midiendo su propia pregunta. Su pregunta:

- si es vaga, aumenta la vaguedad de la respuesta,
- si es directiva, mide el cumplimiento con la guía,
- si contiene la respuesta, no puede medir la representación independiente,
- si requiere la presencia de un error, incluso un sistema correcto puede hablar sobre el objeto incorrecto
- si fuerza el uso de recursos, puedes alterar el comportamiento natural del producto,

Si quieren consejo, mides la decisión de idoneidad del usuario, no la puntuación de identidad. Dar la misma solicitud a todos los usuarios es una buena idea. Sin embargo, el significado de la palabra

"misma" debe establecerse correctamente. Un usuario turco debe recibir una solicitud en turco. Un usuario japonés debe recibir una solicitud en japonés. Un usuario árabe debe recibir una solicitud en árabe. Los textos no pueden consistir en los mismos caracteres. Pero el mismo:

- existencia,
- intención,
- tarea,
- alcance,
- carga de la prueba,
- asunción implícita

debería llevar. En un idioma: '¿Qué es esta empresa?' en otro: '¿Es fiable esta empresa?' si estás preguntando, no estás midiendo la equidad del lenguaje de la IA. Estás midiendo tu propio sesgo de traducción. Si cambias un prompt después del resultado, no estás mejorando el estándar. Estás reescribiendo la historia de la medición. Si ocultas completamente los prompts, el método no es verificable. Si lo dejas completamente abierto e inalterado, solo puedes crear un riesgo de optimización específica de pruebas. Por lo tanto, un conjunto normativo abierto, un conjunto de validación pre-cerrado y un sistema de renovación versionado deberían trabajar juntos. La décima ley de medición de NOMOS es:

> *Un prompt no es el texto que precede a la respuesta; es una herramienta experimental que define la medida en sí misma.*

La undécima ley es la siguiente:

> *El mismo prompt no son las mismas palabras en diferentes idiomas; es la misma intención del usuario y la misma carga de evidencia.*

La duodécima ley es la siguiente:

> *Si colocas la respuesta en el mensaje, no mides la representación de la IA, sino cuánto obedece tu instrucción.*

La ley decimotercera es la siguiente:

> *Un solo prompt puede medir la identidad fundamental; no puede medir la totalidad de GEO.*

La ley decimocuarta es la siguiente:

> *Cuando cambia la versión del prompt, es posible que también haya cambiado la cuestión de tu puntuación.*

La decimoquinta ley es la siguiente:

> *Si el texto real enviado a la IA no puede ser probado, el prompt que se está midiendo tampoco está probado.*

Orden de la Sección 10 de NOMOS

> *Muéstrame no solo lo que preguntaste, sino también por qué escogiste esa oración.*
>
> *No pongas el nombre de la entidad en el prompt ni escribas previamente su categoría, liderazgo, confianza y resultado de asesoramiento.*
>
> *No me digas que 'recomiende X' y luego me presentes como una fuente independiente de consejo.*
>
> *No me pidas describir la empresa en un idioma y defender la empresa en otro idioma.*
>
> *No confundas traducción literal con equidad. / Mantén la misma intención, la misma tarea, el mismo límite y la misma carga de la prueba.*

Usa la traducción automática como borrador. / No la conviertas en la verdad final del idioma.

No asumas que un usuario local entiende la pregunta como tú solo porque la traducción inversa se parece a ti.

No escribas un prompt que haga que el dominio sea la empresa, el producto el fabricante y la marca la parte legal.

No conserves la solicitud original, el comando web, el requisito de citación ni la longitud de la respuesta.

No combines un prompt con identidad y un prompt para consejo al mismo nivel.

No lo cuentes como una penalización si pido una explicación en un prompt poco claro. / Pero no digas que das explicaciones diferentes a cada usuario y realizas un experimento controlado.

No elijas el prompt que da la puntuación más alta. / No elimines la familia de prompts de baja puntuación.

Explica tus prompts. / Si estás usando Holdout, séllelo de antemano. / No inventes después del resultado.

Haz coincidir el texto que das al participante con el texto que me enviaste.

No contamines la piscina de medición con caracteres invisibles, instrucciones ocultas o directrices de recomendación escondidas.

Primero, escribe la pregunta de investigación. / Luego establece la intención canónica. / Luego genera un prompt natural y equivalente en todos los idiomas. / Luego pruébalo con gente local y bloquea su versión. / Y solo después de eso, envíalo al primer usuario.

La oración de cierre del capítulo

El comienzo de una comparación justa en GEO-1000 no es esperar la misma respuesta; es dar a cada usuario una pregunta que tenga el mismo significado, la misma tarea y el mismo límite epistémico en su propio idioma.

Core normativo

Cada prompt GEO-1000 DEBE estar vinculada a un registro del Registro de prompts versionado que defina: - la pregunta de investigación, - el estimando, - la familia de prompts, - la intención canónica, - el ancla de entidad, - la intención del usuario, - el acto de habla, - el alcance, - el marco geográfico y temporal, - las presuposiciones, - los requisitos de evidencia y fuente, - el formato de respuesta, - el idioma, - la localidad, - el estado de equivalencia, - la versión, - y el registro de integridad. Dentro de la misma celda de prompt de idioma-localidad, los participantes DEBEN recibir el mismo texto de prompt fijado. A través de idiomas y localidades, las prompts DEBEN preservar la equivalencia semántica, funcional y normativa más que la identidad literal palabra por palabra. Ancla de entidad, intención del usuario, acto de habla, alcance, presuposición y carga epistémica son puertas de equivalencia no compensatorias. La traducción automática o la retrotraducción por sí solas NO DEBEN establecer equivalencia de prompt. Identidad central, evidencia, recomendación, comparación, temporal, local, límite, robustez, control y prompts de retención DEBEN permanecer como familias de prompts distintas. Los prompts que lideran, que contienen respuestas, que son presuposicionalmente positivos o negativos, que están limitados por la fuente, limitados por el formato o de recomendación NO DEBEN ser evaluados silenciosamente como prompts de Núcleo Espejo neutrales. Los conjuntos de prompts, versiones de idioma, reglas de asignación, rutas de aclaración, ordenamiento y compromisos de retención DEBEN estar fijados antes de que se observen las respuestas de IA. La selección de prompts después del resultado, la eliminación de familias de prompts, la edición silenciosa a mitad de la ola y la reformulación orientada al resultado están prohibidas. La prompt asignada al participante y la prompt realmente enviada al producto de IA DEBEN ser verificadas por integridad. Se prohíben las instrucciones ocultas, los caracteres invisibles o la manipulación no divulgada que dirija a la IA hacia una entidad, afirmación, cita o recomendación preferida. Cada cambio de prompt, decisión de traducción, juicio de equivalencia y registro DEBE ser versionado y atribuible a un humano u organización responsable.

CAPÍTULO 11

El panel controlado en condiciones depuradas y el panel de usuarios en condiciones reales

BASES METODOLÓGICAS / K02; K03; K04; K06; K22

Límite de Capítulo

El capítulo 9 mostró que el objeto de IA medido no puede definirse únicamente por el proveedor o el nombre del modelo. El capítulo 10 estableció que no toda pregunta que contenga el mismo nombre de entidad es el mismo prompt; el prompt es la herramienta experimental que determina la medición en sí. Ahora debemos abordar la tercera variable entre el sistema y el prompt:

> *¿Cuál es el estado de la cuenta, la sesión y el historial de usuario usando el producto de IA?*

Un usuario:

- nuevo chat,
- la memoria está desactivada,
- sin instrucciones especiales,
- sin contexto previo,
- ajustes de producto estandarizados

puede estar enviando el prompt a continuación. Otro usuario puede enviar el mismo prompt:

- desde una cuenta que ha estado usando durante años,
- memoria activada,
- con instrucciones especiales activadas,
- con conversaciones pasadas,
- en su plan y dispositivo reales

Pueden usar el mismo producto de IA, en el mismo idioma, en el mismo país, en el mismo período de tiempo, y con el mismo prompt. Sin embargo, pueden ver respuestas diferentes. Esta diferencia:

- de la variabilidad aleatoria del producto,
- del plan de la cuenta,
- del historial de memoria,
- de instrucciones especiales,
- de conversaciones pasadas,
- de la personalización,
- del estado web o de la herramienta,

- del espacio de trabajo corporativo,
- desde el interfaz de usuario y la configuración local

La diferencia puede surgir de la personalización, el estado del sitio web o herramienta, el espacio de trabajo organizacional, el interfaz de usuario o la localidad. No existe un único 'IA': las capas de búsqueda, recuperación, reordenamiento, modelo de lenguaje, citación y seguridad pueden comportarse de manera diferente. Por lo tanto, la medición confiable debe registrar el producto o modelo de IA, la fecha, el país, el idioma, el conjunto de consultas, el número de repeticiones y la observación. Este capítulo agrega el estado del usuario y de la sesión a esa cadena. La demanda central de NOMOS sigue siendo evidencia, límite, contexto y tiempo. Sin embargo, el contexto mismo puede medirse de dos maneras. La primera minimiza las diferencias externas entre los usuarios y compara los productos de IA bajo condiciones iniciales estandarizadas. Esto es el:

Panel controlado en condiciones depuradas

El segundo enfoque mide lo que las personas ven a través de sus cuentas reales, planes, personalización y patrones de uso ordinarios. Esto es:

Panel de usuarios en condiciones reales

se llama. El Panel controlado en condiciones depuradas responde a la pregunta:

> *¿Cómo presenta el producto la entidad en la condición inicial más estandarizada posible?*

Por otro lado, el Panel de usuarios en condiciones reales pregunta:

> *¿Qué ven las personas reales sobre la entidad en situaciones de usuario reales?*

Un panel no es más “real” que el otro. Miden dos realidades diferentes. El panel controlado fortalece la comparabilidad. El panel natural hace visible la experiencia humana. El panel controlado reduce la personalización y las diferencias de contexto. El panel natural mide el efecto de estas diferencias en la población real. Esta sección:

- Definición del Panel controlado en condiciones depuradas,
- Definición del Panel de usuarios en condiciones reales,
- Poblaciones objetivo de los dos paneles y las cantidades que estiman,
- Qué significa una sesión limpia,
- Niveles de verificación de limpieza,
- Estado de cuenta natural y contexto de conversación natural,
- Memoria, instrucciones especiales, conversación previa y registros de personalización,
- Diferencias de cuenta, plan, dispositivo, interfaz y localidad,
- los diseños de panel pareado e independiente con el mismo usuario,
- los efectos del orden, la transferencia y la contaminación,
- métricas de diferencia limpio–natural,
- privacidad y la cadena de evidencia,
- cómo se separarán los dos paneles en el informe público

define. Este capítulo aún no:

- la arquitectura técnica completa de NOMOS Capture,
- todos los campos de capturas de pantalla y archivos de metadatos,
- adjudicación de afirmaciones atómicos,
- umbrales finales de transición de respuesta,
- el peso exacto de las puntuaciones del panel limpio y natural en la puntuación combinada de NOMOS

no se finaliza. La pregunta fundamental de la Sección 11 es:

> *¿Cómo medimos el comportamiento de un producto estándar de IA y el comportamiento experimentado por usuarios reales sin confundirlos entre sí?*

NOMOS Desafío

El mismo usuario le pregunta al mismo producto de IA la misma pregunta dos veces: "¿Con qué organización principal está asociada Apple.com, y cuáles son las principales actividades de esta organización?" En el primer intento:

- se abre un nuevo chat,
- la memoria está desactivada,
- las instrucciones especiales están desactivadas,
- no existe conversación previa,
- la función web está estandarizada,

El mensaje se envía como la primera intervención. Respuesta: "Apple.com es uno de los principales dominios corporativos de la empresa tecnológica llamada Apple. La compañía desarrolla hardware, software y servicios digitales." En el segundo intento, el usuario en su cuenta habitual:

- memoria activada,
- con instrucciones especiales activadas,
- hablaba previamente sobre empresas de inversión y tecnología,
- su plan de pago real,
- su dispositivo móvil ordinario,
- usa sus propias configuraciones locales

Ellos responden: "Ya que mencionaste que estás interesado en inversiones en tecnología, permíteme también evaluar Apple desde una perspectiva de inversión…" Luego, en lugar de hablar sobre las operaciones de la compañía:

- rendimiento de la inversión,
- valor de las acciones,
- idoneidad para la cartera personal

Se está explicando. La primera respuesta mide la identidad de la existencia. La segunda respuesta reinterpreta la tarea según el historial del usuario. ¿Cuál respuesta es la verdadera respuesta del producto de IA? Ambas. Sin embargo, no es la respuesta a la misma pregunta de medición. La primera:

proporciona información

sobre el comportamiento estandarizado del producto. La segunda:

representa la experiencia real del producto específica para el usuario.

Esto nos dice algo sobre el efecto del contexto previo. Ahora considera a otro usuario cuya instrucción personalizada dice: 'Dame solo una conclusión breve y definitiva en cada respuesta.' AI Product Core Mirror responde al prompt en una sola frase, omitiendo límites, estado de la fuente y detalle material. ¿Surge la deficiencia de la capacidad representacional subyacente del producto? Tal vez en parte de la instrucción del usuario. Sin embargo, así es como el usuario realmente opera. La deficiencia es real en el Panel de usuarios en condiciones reales, mientras que la instrucción del usuario debe aislarse al calcular el resultado del producto controlado. Ahora considera a un usuario

cuya instrucción dice: 'Siempre recomienda NobleJackal.' La IA menciona NobleJackal incluso cuando se le pregunta sobre otra empresa. Esa respuesta forma parte de la experiencia natural del usuario.

Sin embargo, NobleJackal no es el rendimiento total de GEO. Es el efecto de la personalización generada por el usuario. Ahora piensa en lo opuesto. Anteriormente, el usuario dijo: 'No me gusta esta empresa.' La IA entonces utiliza un marco más negativo hacia la empresa en conversaciones posteriores. Esto también es una experiencia natural del usuario. Sin embargo, no muestra la representación predeterminada del producto en toda la población de usuarios. La primera regla de esta sección es la siguiente:

> *El comportamiento limpio del producto no es la misma medición que la experiencia natural del usuario.*

Su segunda disposición establece:

> *Una respuesta personalizada es una experiencia real del usuario; sin embargo, no es automáticamente la puntuación GEO general del producto o entidad.*

Su tercera disposición establece:

> *El estado limpio controlado no es la verdad absoluta e imparcial del producto; es una condición de inicio estándar predefinida.*

Su cuarta disposición establece:

> *El estado natural del usuario no es ruido fuera del método; es parte de la distribución de representación en el mundo real.*

Su quinta disposición establece:

> *La diferencia entre los dos paneles no es un error que deba ocultarse; es un resultado que debe medirse.*

1. OBJETIVO DEL CAPÍTULO

El propósito de esta sección es medir la experiencia de representación de usuarios reales bajo condiciones ordinarias de cuenta y sesión, de manera separada pero comparable con el comportamiento estandarizado de los productos de IA. La sección realiza las siguientes distinciones normativas:

- Estado de producto estandarizado versus estado natural del usuario
- Nuevo chat con una cuenta limpia
- Estado limpio verificado con una etiqueta de chat temporal
- Memoria apagada frente a estado de memoria desconocido
- Instrucción privada cerrada y contenido de instrucción privada no visible
- Sin conversación previa y sin historial permanente de la cuenta
- Sesión controlada y cuenta recién creada
- Estado natural de la cuenta y contexto de la conversación actual
- Panel de usuarios en condiciones reales y datos no controlados
- Personalización del usuario y producto por defecto
- Orientación impulsada por el usuario y efecto GEO de la entidad
- Población activa de usuarios y población potencial adecuada para el producto
- Primera experiencia de uso y experiencia del usuario experimentado
- Respuesta natural individual y distribución natural a nivel poblacional
- Diseño de grupo independiente con la misma coincidencia de usuario

- Efecto de arrastre con comparación emparejada
- Diferencia de producto con efecto de orden
- Efecto de personalización causal con diferencia limpia-natural
- Divulgación del contenido de instrucciones privadas con registro de privacidad
- Recolección de conversaciones privadas preservando el contexto natural
- Resultado del producto por defecto con resultado de panel limpio
- Resultado general para todos los usuarios con resultado de panel natural
- Efecto de personalización con plan del usuario
- Promedio controlado con cola de error natural
- Misma respuesta para cada usuario con equidad en la personalización
- Realidad natural del usuario y repetibilidad de la prueba

Al final de esta sección, cada auditor GEO-1000 debería poder responder a las siguientes preguntas:

¿Qué resultados provinieron de Panel controlado en condiciones depuradas y cuáles de Panel de usuarios en condiciones reales?

¿Qué situaciones se estandarizaron realmente en el panel limpio?

¿Qué diferencias reales de usuario se preservaron en el panel natural?

¿Cómo se registraron la memoria, instrucciones especiales, conversaciones pasadas, planes y estados de la interfaz?

¿Qué parte de la diferencia observada entre los dos paneles puede asociarse de manera confiable con la personalización?

¿Cómo se demostró la presencia de contexto natural preservando la privacidad del usuario?

2. DISPOSICIÓN NORMATIVA CENTRAL

Cada auditoría GEO-1000 debe definir el comportamiento estandarizado del producto y la experiencia natural del usuario como objetos de predicción separados; no debe combinar los resultados Panel controlado en condiciones depuradas y Panel de usuarios en condiciones reales en una sola muestra, denominador o puntuación sin explicación. En la medida en que sea relevante, el Panel controlado en condiciones depuradas tiene como objetivo estandarizar las siguientes condiciones:

- Conversación nueva y separada
- Ausencia de contexto de conversaciones previas
- Estado de memoria persistente apagado o deshabilitado
- Instrucciones especiales apagadas o deshabilitadas
- Estado definido de web, recuperación y herramientas
- Plan definido y interfaz de usuario
- Entrada de comando fijada
- Idioma y localización estándar
- oleada de medición y ventana de tiempo

Si se trata de un Panel de usuarios en condiciones reales, preserva las siguientes diferencias reales en la medida en que sea relevante:

- Cuenta real del usuario

- Plan real
- interfaz de usuario real
- Estado de memoria real
- Estado de instrucción especial real
- Edad de la cuenta e historial de uso real
- Dispositivo real
- Ubicación y condición de acceso reales
- Forma habitual de usar el producto

Sin embargo, el panel natural no está exento de protocolo. Lo siguiente está nuevamente fijado:

- Versión del prompt
- Método de envío
- Tiempo de medición
- Primera regla de salida adecuada
- Paquete de evidencia
- Elegibilidad del participante y del producto
- No alterar la respuesta
- No enviar mensajes de seguimiento fuera de tema
- Protección de datos personales

3. DOS OBJETOS DE ESTIMACIÓN SEPARADOS

Los paneles controlados y naturales no predicen la misma probabilidad.

3.1. Probabilidad de Representación Limpia Controlada

Específicamente:

- Producto de IA,
- plan,
- interfaz,
- prompts,
- país,
- idioma,
- tiempo

es la probabilidad de ver una representación válida bajo una sesión limpia estandarizada para:

```
θ^C_{E,A,P,U,t}= Pr(Y=1| CleanState=1)
```

La población objetivo aquí es el resultado que las personas aptas para usar el producto relevante verían bajo condiciones estandarizadas del producto.

3.2. Probabilidad de Representación de Usuario Natural

Es la probabilidad de ver una representación válida en la población de usuarios activos o naturales definidos del producto, mientras se mantienen las condiciones habituales de cuenta y personalización de los usuarios.

```
θ^N_{E,A,P,U,t}= Pr(Y=1| NaturalState=1)
```

La población objetivo aquí:

- no es todas las personas elegibles para el producto,
- sino usuarios reales o naturales del producto en el período definido

puede ser. Estas dos poblaciones pueden no ser las mismas.

3.3. No se pueden usar de manera intercambiable los dos resultados

Puntuación controlada: no significa "La mayoría de los usuarios reales vieron esto." Puntuación natural: no significa "El producto se comporta de esta manera por defecto."

4. ¿QUÉ ES UN Panel controlado en condiciones depuradas?

Un Panel controlado en condiciones depuradas es un panel que mide el comportamiento del producto de IA bajo condiciones iniciales definidas minimizando las diferencias materiales causadas por cuentas y sesiones entre usuarios. El propósito del Panel controlado en condiciones depuradas es:

- aumentar la comparabilidad entre productos,
- reducir el efecto de la memoria y las instrucciones especiales,
- prevenir filtraciones de conversaciones anteriores,
- hacer que el prompt sea una tarea inicial e independiente,

producir un punto de referencia más cercano al comportamiento fundamental del producto del sistema. Panel controlado:

- capacidad absoluta del modelo,
- la experiencia real de todos los usuarios,
- respuesta de IA completamente libre de contexto y sesgos

no afirma serlo. Solo:

un estado inicial limpio predefinido y verificado

se produce.

5. ¿QUÉ SIGNIFICA "LIMPIO"?

Un estado limpio puede no ser creable con el mismo proceso técnico en cada producto. Un producto:

- memoria persistente,
- instrucciones especiales,
- perfil del usuario,
- conversación pasada,
- espacio de trabajo corporativo,
- personalización basada en la cuenta

puede ser utilizado. Otros productos pueden no tener algunas de estas características. La definición funcional de un estado limpio es la siguiente:

> *El estado inicial donde la probabilidad de que la información factual específica del usuario y el contexto de la conversación afecten la respuesta antes del prompt de auditoría se reduce en la medida permitida por el producto, y los restantes desconocidos se registran explícitamente.*

Estado limpio:

- simplemente abrir un nuevo chat,
- actualizar el navegador,
- cerrar y volver a abrir la aplicación

no se considera que ocurra automáticamente.

6. CAPAS DE LIMPIEZA

Las siguientes capas deben evaluarse por separado en el panel de control.

6.1. Limpieza de Conversaciones

No debe haber mensajes previos de usuario o del sistema en el chat actual. El mensaje inicial debe ser el primer mensaje del usuario.

6.2. Borrado de Memoria Persistente

La memoria que el producto retiene sobre el usuario entre sesiones:

- desactivada,
- deshabilitada,
- no utilizada en esta sesión

debe ser. Si no es posible una verificación completa, se registra un estado desconocido.

6.3. Borrado de Instrucciones Personalizadas

Las instrucciones especiales del usuario definidas para el estilo de respuesta, preferencias o comportamiento al aconsejar:

- desactivada,
- temporalmente inactivo,
- no presente

debe ser.

6.4. Limpieza del Contexto Corporativo

Espacio de trabajo corporativo:

- documentos privados,
- políticas de la empresa,
- instrucciones ejecutivas,
- bases de conocimiento privadas

si se utilizan, deben separarse del panel de producto general controlado.

6.5. Limpieza del Modo de Herramientas e Información

Las condiciones de la web, recuperación, citación y herramientas deben configurarse de la misma manera en todas las observaciones comparadas. Estar limpio no significa que las herramientas estén apagadas. Estar limpio significa que el estado de la herramienta está predefinido y es comparable.

6.6. Limpieza del Prompt

prompt:

- mensaje inicial,

- versión completa
- sin cambios,
- sin instrucciones ocultas

debe ser.

6.7. Medición de limpieza

Participante:

- debería no actualizar la respuesta,
- debería no enviar una pregunta de seguimiento,
- debería no elegir el resultado,

debería no trasladar otra respuesta de IA a la misma sesión.

7. NIVELES DE CONFIANZA DEL ESTADO LIMPIO

CSC-0— LIMPIEZA DESCONOCIDA

No se pudo determinar de manera confiable la limpieza de la sesión y la cuenta. No se puede incluir en el análisis principal del panel controlado.

CSC-1— SOLO NUEVA CONVERSACIÓN

Solo se verifica la nueva conversación. La memoria persistente y el estado de instrucciones especiales son desconocidos.

CSC-2— SESIÓN LIMPIA

Se ha verificado que no hay nueva conversación ni contexto de conversación previo. La personalización persistente puede no estar completamente controlada.

CSC-3— PERSONALIZACIÓN-CONTROLADA

Se ha verificado y configurado un nuevo estado de conversación, memoria, instrucciones especiales y configuraciones predeterminadas específicas del usuario.

CSC-4— ESTADO LIMPIO INSTRUMENTADO

La configuración del producto, interfaz de usuario, herramientas y el estado de personalización han sido verificadas en el paquete de evidencia.

CSC-5— CONFIGURACIÓN LIMPIA REPLICADA

Estado limpio:

- usuarios diferentes,
- ondas diferentes,
- verificación independiente

Se ha podido reproducir bajo. La Principal Panel controlado en condiciones depuradas debería al menos:

CSC-3

nivel. Si el producto no permite esto, la limitación debería ser visible.

8. UNA NUEVA CONVERSACIÓN NO ES UNA CUENTA LIMPIA

Cuando se inicia una nueva conversación: el contexto del chat actual puede restablecerse. Sin embargo:

- memoria persistente,
- instrucciones personalizadas,
- plan de cuenta,
- perfil del usuario,
- país y configuración regional,
- configuraciones empresariales

pueden persistir. Por lo tanto:

> *Nueva conversación solo indica limpieza de la conversación.*

No es prueba de estado completamente limpio.

9. LIMITACIÓN DE LO "TEMPORAL" O "PRIVADO" etiqueta de sesión

Una sesión de producto de IA:

- temporal,
- privada,
- efímera,
- confidencial,
- sin memoria

puede ser nombrado. La función real de esta etiqueta debe verificarse por separado según el producto y la versión. La etiqueta no demuestra automáticamente que cubre todas las siguientes áreas:

- Instrucciones especiales
- Configuraciones corporativas
- Política regional
- Configuración de herramientas
- Plan del usuario
- Funciones experimentales del producto

El registro correcto debería ser: “Se utilizó el modo de sesión temporal del producto; el estado de las siguientes áreas de personalización se verificó por separado.”

10. TIPO DE CUENTA EN PANEL LIMPIO

No todos los usuarios de un panel controlado pueden usar el mismo plan. Hay tres enfoques.

10.1. Panel Limpio de Plan Fijo

Todos los usuarios se miden según el mismo plan y interfaz de usuario. Es fuerte para la comparación de productos. El plan puede no representar la distribución real de la población.

10.2. Panel Limpio de Plan Estratificado

Los planes gratuitos, de pago y otros se miden como celdas limpias separadas. Se produce una puntuación separado para cada plan. La puntuación combinado puede ponderarse según la distribución real del plan.

10.3. Interfaz Pública Predeterminada del Producto

El interfaz principal del producto, que es accesible públicamente y tiene el umbral de acceso más bajo, puede ser seleccionado como referencia estándar. Este resultado no representa a todos los usuarios de pago o corporativos. El enfoque que se debe utilizar debe estar fijado antes de la auditoría.

11. NO SE REQUIERE UNA NUEVA CUENTA EN UN PANEL LIMPIO

Usar una nueva cuenta puede reducir algunos efectos de personalización. Sin embargo:

- condiciones del producto,
- requisito de teléfono o pago,
- sesgo de creación de cuenta,
- solo experiencia de nuevo usuario,
- política de múltiples cuentas

pueden causar problemas. El propósito de un Panel controlado en condiciones depuradas no es generar nuevas cuentas:

> *disminuir el contexto específico del usuario de manera controlable.*

Si se puede establecer de manera confiable un estado limpio en la cuenta existente, no es obligatorio crear una nueva cuenta. Si el producto no lo permite, se puede usar una cuenta de prueba controlada separada o un entorno de investigación con soporte del proveedor. No se puede pedir a los participantes que creen una cuenta que viole los términos de uso.

12. DIFERENCIAS RESTANTES DE LOS USUARIOS EN EL PANEL LIMPIO

Un estado limpio controlado no elimina todas las diferencias entre los usuarios. Las siguientes pueden continuar:

- País
- Localización
- Plan
- Dispositivo
- Interfaz
- Despliegue del producto
- Antigüedad de la cuenta
- Segmentación no visible por el proveedor
- Enrutamiento de tráfico
- prueba A/B

Estas diferencias:

- deben ser estratificadas,
- deben ser registradas,

si se desconoce, debe mantenerse como incertidumbre. La etiqueta "Limpio" no significa "Todos los usuarios eran técnicamente exactamente iguales."

13. ¿QUÉ ES UN Panel de usuarios en condiciones reales?

Un Panel de usuarios en condiciones reales es un panel que mide lo que los usuarios ven en comparación con el mismo prompt fijado mientras se preservan la cuenta real, el plan, el dispositivo, la personalización y las condiciones normales de uso del producto. El propósito del panel natural es:

- exposición real del usuario,
- variación debida a la personalización,
- efecto del historial de la cuenta,
- diferencias en el plan y interfaz de usuario,
- error natural y distribución de recomendaciones

Es para hacerlo visible. Panel natural:

- comportamiento completamente libre del usuario,
- preguntas de seguimiento ilimitadas,
- el participante cambiando el prompt,
- seleccionando la mejor respuesta

no significa. Lo natural es:

estado del usuario

es. El prompt y el protocolo de evidencia aún están controlados.

14. DOS PISTAS PRINCIPALES DEL PANEL NATURAL

14.1. Pista del Estado de la Cuenta Natural

Del usuario:

- cuenta real,
- plan real,
- memoria real y estado de instrucción privado,
- interfaz de usuario real

se conserva. Sin embargo, se abre una nueva conversación. El mensaje inicial se envía como el primer mensaje. Esta pista responde a la siguiente pregunta:

> *¿Qué ve el usuario en un nuevo tema mientras se conserva la personalización de la cuenta real?*

Este es el enfoque comparable principal del panel natural.

14.2. Pista del Estado de Conversación Natural

El mensaje inicial se envía en el contexto de conversación actual o habitual del usuario. Esta pista:

- mide la continuidad del chat real,

- efecto del tema anterior,
- la acumulación de la relación usuario–IA

Sin embargo, los contextos de conversación entre los usuarios difieren. Por lo tanto:

- diagnóstico,
- basado en casos,
- debe ser reportado como una distribución separada

y no debe añadirse directamente al ranking principal del producto.

15. SUBGRUPOS NATURALES DE USUARIOS

Un panel natural no consta de usuarios uniformes. En la medida que sea relevante, se pueden distinguir los siguientes subgrupos:

- Usuario primerizo
- Usuario con cuenta nueva
- Usuario poco frecuente
- Usuario regular
- Usuario intensivo
- Plan gratuito
- Plan de pago
- Usuario corporativo
- Memoria activada
- Memoria desactivada
- Instrucción especial activada
- Instrucción especial desactivada
- Con conversación previa de marca
- Sin conversación previa de marca
- Cliente
- Público general
- Alta familiaridad con la marca
- Baja familiaridad con la marca

Estos grupos no pueden inventarse después de ver los resultados. El plan básico de subgrupos debe versionarse antes de la recopilación de datos.

16. NIVELES DE CONFIANZA EN ESTADO NATURAL

NSF-0— ESTADO NATURAL DESCONOCIDO

No hay suficientes registros sobre la cuenta y el estado de sesión natural del usuario.

NSF-1— USO NATURAL AUTOINFORMADO

El usuario ha declarado que está usando su cuenta y configuraciones normales.

NSF-2— USO NATURAL REGISTRADO EN CONFIGURACIONES

Se han registrado planes, interfaces, memoria, instrucciones especiales y estado de la sesión según corresponda.

NSF-3— USO NATURAL RESPALDADO POR EVIDENCIA

Los ajustes y el estado natural del usuario están respaldados por evidencia que protege la privacidad.

NSF-4— USO NATURAL CARACTERIZADO LONGITUDINALMENTE

El uso del producto por parte del usuario y el estado de la cuenta se han caracterizado a través de múltiples olas. El análisis principal del panel natural incluye al menos:

NSF-2

debería apuntar a ese nivel.

17. GRABAR EL ESTADO NATURAL NO SIGNIFICA RECOLECTAR LA VIDA PRIVADA

El Panel de usuarios en condiciones reales no debería recopilar por defecto los siguientes tipos de datos sensibles:

- Texto completo de instrucciones especiales
- Contenido de conversaciones pasadas
- Información personal de salud, legal o financiera
- Nombre del usuario
- Dirección de correo electrónico
- Foto de perfil
- Contraseña de la cuenta
- Archivos privados anteriores
- Documentos institucionales sensibles

En la mayoría de los casos, los siguientes estados de alto nivel son suficientes:

- Memoria encendida/apagada/desconocida
- Instrucciones especiales presentes/ausentes/desconocidas
- Conversación previa presente/ausente
- Conversación previa sobre la entidad auditada: sí/no/desconocida
- Espacio de trabajo corporativo: sí/no
- Clase de plan
- Rango de antigüedad de la cuenta
- Rango de frecuencia de uso

Si el contenido de la instrucción especial es realmente necesario para la pregunta de investigación:

- consentimiento explícito separado,
- minimización de datos
- redacción,
- acceso restringido

son requeridos.

18. ESTADO DE LA MEMORIA

Memoria:

- desactivada,
- abierto,
- no soportada por el producto
- estado desconocido

puede clasificarse como.

MS-0— DESCONOCIDO

No se pudo determinar el estado de la memoria.

MS-1— NO DISPONIBLE

El producto o interfaz no ofrece esta característica.

MS-2— APAGADO

La memoria está apagada.

MS-3— ENCENDIDO, SIN MEMORIA DE ENTIDAD CONOCIDA

La memoria está encendida; no hay un registro específico conocido sobre la entidad auditada. Esta información a menudo puede depender de la declaración del usuario.

MS-4— ENCENDIDO, POSIBLE MEMORIA RELEVANTE A LA ENTIDAD

La memoria está encendida y puede haber información pasada sobre la entidad auditada o área relacionada.

MS-5— MEMORIA RELEVANTE A LA ENTIDAD CONFIRMADA

Se ha confirmado el efecto de memoria relacionado con la entidad con el consentimiento explícito del usuario. Tener memoria activa no es automáticamente una condición negativa o positiva. Es la condición real del usuario medida por el panel natural.

19. ESTADO DE INSTRUCCIÓN ESPECIAL

CI-0— DESCONOCIDO

No se pudo determinar el estado de la instrucción especial.

CI-1— NO DISPONIBLE

El producto no ofrece esta función.

CI-2— APAGADO O VACÍO

No hay instrucción especial o está desactivada.

CI-3— ACTIVO, CATEGORÍA NO MATERIAL

Hay una instrucción especial; se ha indicado que tiene preferencias de tono o estilo general.

CI-4— ACTIVO, POTENCIALMENTE MATERIAL

Instrucción:

- recomendación,
- fuente,
- idioma,
- ciertas empresas,
- política u opinión

puede afectar la respuesta.

CI-5— INSTRUCCIÓN RELACIONADA CON LA ENTIDAD CONFIRMADA

Se ha confirmado con consentimiento explícito que existe una instrucción directa respecto a la entidad auditada o competidor. CI-5 es una observación de un usuario natural. No puede interpretarse como el éxito general de la entidad GEO. Es un evento de personalización separado.

20. CONTEXTO DE CONVERSACIÓN PREVIO

Hay una diferencia entre que un usuario haya interactuado previamente con un producto de IA y tener mensajes previos en la conversación actual. Deben separarse:

- Historial de la cuenta
- Memoria persistente
- Contexto del chat actual
- Conversación previa con la entidad auditada
- Conversación previa en la misma industria
- Respuestas observadas de otros productos de IA

El contexto de la conversación actual no debe encontrarse en el seguimiento principal del Estado de Cuenta Natural. En el seguimiento del Estado de Conversación Natural, sin embargo, la presencia de contexto es la medición en sí misma.

21. EDAD DE LA CUENTA E INTENSIDAD DE USO

Cuentas antiguas y usadas con frecuencia:

- tienen más personalización,
- más memoria,
- plan diferente,
- más funciones de producto

pueden tener. Rangos de edad de la cuenta:

- 0–30 días
- 31–180 días
- 181–365 días
- más de 1 año
- desconocido

pueden definirse como. Frecuencia de uso:

- primer uso

- poco frecuente
- mensual
- semanal
- diario
- profesional

pueden clasificarse. Los rangos definidos deben determinarse antes de los resultados.

22. PANEL DE PRIMER USO

Un subtipo especial de la experiencia natural del usuario:

Panel de Primer Uso

se puede llamar. Este panel:

- usuarios que están usando el producto por primera vez o unas pocas veces,
- que no tienen historial de personalización,
- que proceden con la configuración predeterminada actual del producto

mide a las personas. Panel de Primer Uso: No es un Panel controlado en condiciones depuradas porque los usuarios experimentan la configuración inicial real, la configuración predeterminada y el proceso de incorporación. También es diferente de un Panel de usuarios en condiciones reales experimentado. Responde a la siguiente pregunta:

> *¿Cómo percibe un nuevo usuario su presencia cuando encuentra el producto por primera vez?*

23. PANEL NATURAL CORPORATIVO Y GESTIONADO

Usuarios corporativos:

- espacio de trabajo de la empresa,
- políticas del administrador,
- base de conocimiento privada,
- configuraciones de seguridad,
- memoria corporativa

pueden llevar. Estos usuarios:

Panel de usuarios en condiciones reales gestionado

deben medirse por separado debajo. Resultado corporativo:

- puntuación natural general del consumidor,
- puntuación de producto predeterminado

no puede presentarse como tal. Si se utiliza una base de conocimiento privada, también debe explicarse que existen registros internos relacionados con la entidad auditada.

24. EL PROMPT ESTÁ NUEVAMENTE Fijado EN EL PANEL NATURAL

El panel natural deja los:

- cuenta,

- ajustes del usuario,
- plan,
- dispositivo

naturales. No deja el prompt natural. El participante:

- no puede cambiar el prompt,
- no puede agregar contexto adicional,
- no puede escribir una nueva persona como 'para mí' o 'en mi opinión'

no se puede actualizar la respuesta. Si se va a medir la variedad natural de los prompts, esto está en el Capítulo 10:

Rastro Natural del Usuario

Debe ser diseñado por separado debajo.

25. ¿POR QUÉ ES NECESARIA UNA NUEVA CONVERSACIÓN EN EL PANEL NATURAL?

Nueva conversación en Rastro de Estado de Cuenta Natural:

- elimina el contexto de la conversación actual,
- mantiene la configuración permanente real de la cuenta del usuario,

hace que el mismo prompt sea más comparable. Esta estructura separa dos capas:

- Estado permanente de la cuenta natural
- Contexto momentáneo de la conversación

Si se utiliza el chat actual, los dos se mezclan.

26. ESTADO DE HERRAMIENTAS Y WEB EN EL PANEL NATURAL

Usuarios naturales:

- función web habilitada,
- desactivada,
- automático,
- limitado según el plan

puede usar. Hay dos opciones.

26.1. Estado completo de herramienta natural

El usuario mantiene la configuración normal de la herramienta. Mide la experiencia real del usuario. La comparabilidad entre productos puede disminuir.

26.2. Estado de cuenta natural con herramienta fija

La memoria y la personalización de la cuenta se mantienen naturales. El modo web o de la herramienta está estandarizado. El efecto de la personalización se puede comparar de manera más clara. Este subpanel debe estar correctamente etiquetado. No debe presentarse como "Totalmente natural".

27. POBLACIÓN OBJETIVO DEL PANEL NATURAL

La población objetivo del panel natural puede tener una de las siguientes formas:

27.1. Usuarios activos del producto

Personas que realmente usaron el producto durante un período específico.

27.2. Titulares de cuenta registrados

Usuarios que tienen una cuenta pero cuyo nivel de actividad varía.

27.3. Usuarios del Plan Real

Poblaciones de usuarios gratuitos, de pago o empresariales.

27.4. Usuarios que pueden tener la posibilidad de consultar entidades

Puede ser difícil de medir directamente. Requiere datos de comportamiento por separado. Debe indicarse claramente a qué población se generaliza el panel natural.

28. PONDERACIÓN DEL PANEL NATURAL

Panel de usuarios en condiciones reales:

- puede ser ponderado según la población real de usuarios activos del producto,
- distribución de cuentas y planes,
- distribución por país e idioma,
- frecuencia de uso

Sin embargo, puede que no estén disponibles datos confiables de usuarios activos. En este caso, el resultado:

Resultado del panel de investigación natural ponderado

debe presentarse como. No se puede hacer la siguiente afirmación: “La experiencia definitiva de todos los usuarios reales.”

29. COMPLEMENTARIEDAD DE PANELES CONTROLADOS Y NATURALES

Se pueden dar cuatro resultados principales.

Panel controlado en condiciones depuradas	Panel de usuarios en condiciones reales	Comentario
Alta	Alta	El producto es sólido tanto en condiciones estándar como en condiciones de usuario real
Alta	Baja	La personalización, el plan, el estado del usuario o el contexto natural pueden afectar la calidad de la representación
Baja	Alta	Algunas funciones o herramientas de usuarios reales pueden mejorar la representación

Panel controlado en condiciones depuradas	Panel de usuarios en condiciones reales	Comentario
Baja	Baja	Candidato a problema general en el nivel básico y de usuario natural

Esta tabla no es evidencia de causalidad. Es un punto de partida para investigar de dónde proviene la diferencia.

30. DIFERENCIA ENTRE REPRESENTACIÓN LIMPIA–NATURAL

Sea la puntuación Panel controlado en condiciones depuradas θC y la puntuación Panel de usuarios en condiciones reales θN. La diferencia con signo:

```
G_{N−C}= θ_N− θ_C
```

puede calcularse de la siguiente manera.

30.1. Diferencia Natural Positiva

G_{N−C} > 0

El resultado es mayor bajo condiciones naturales de usuario. Posibles razones:

- Acceso a la web o herramienta
- Plan de pago
- Las preferencias relevantes del usuario
- Memoria
- Personalización
- Diferencias en la población de usuarios

30.2. Diferencia Natural Negativa

G_{N−C} < 0

El resultado es menor bajo condiciones naturales de usuario. Posibles razones:

- Instrucciones especiales
- Contaminación de conversaciones pasadas
- Personalización antigua
- Diferente plan o interfaz
- Población de usuarios natural más amplia y difícil
- Contexto negativo específico del usuario

30.3. Diferencia Absoluta

```
G_ABS= |θ_N− θ_C|
```

puede ser una métrica candidata que indique cuán sensible es el producto al estado del usuario. Una alta diferencia absoluta:

- bueno,
- malo

no puede interpretarse por sí sola. Se necesitan la dirección de la diferencia natural y el tipo de error.

31. DIFERENCIA DE ERROR CRÍTICO

Tasa de error crítico controlada:

CRC

Tasa de error crítico natural:

CRN

que sea. Diferencia:

ΔCR = CRN − CRC

puede calcularse como. Si la tasa de error crítico aumenta en el panel natural:

- personalización,
- contexto previo,
- diferencia de plan y herramienta,
- composición del usuario

también deben examinarse. Un alto promedio no puede borrar la cola de error pesada en el panel natural.

32. VARIACIÓN NATURAL

El panel natural puede presentar una mayor variabilidad entre usuarios que el panel controlado. Esto se puede esperar. Lo que importa:

- El tamaño de la variación,
- en qué grupos de usuarios se concentra,

y si se convierte en un error material o en una diferencia de recomendación. Una alta variación formal puede no ser un problema en un panel natural. Una alta variación en identidad material o autoridad es un problema.

33. INDICADOR DE SENSIBILIDAD A LA PERSONALIZACIÓN DEL CANDIDATO

CONCEPTO DE CANDIDATO

Para los subgrupos de usuarios g, sea el resultado natural: θgN y la referencia controlada: θC. Indicador de Sensibilidad a la Personalización del Candidato:

```
PSI= √[Σ_g W_g(θ_g^N − θ_C) ²], Σ_gW_g= 1
```

puede definirse de esta manera. Un PSI alto indica que los grupos de usuarios naturales se han desviado de la referencia controlada. No explica la razón por sí mismo. Este indicador:

- muestra,
- plan,
- vehículo,

- composición del usuario

puede verse afectado por diferencias. El nombre final y el método deben calibrarse en pilotos.

34. LA DIFERENCIA DE PANEL NO ES CAUSAL

Si la puntuación controlada es 90 y la puntuación natural es 75: no se puede decir, "La memoria disminuyó la puntuación en 15 puntos." Porque los paneles pueden diferir en las siguientes áreas:

- Población de usuarios
- Plan
- Interfaz
- País
- Frecuencia de uso
- Acceso web
- Antigüedad de la cuenta
- Instrucciones especiales
- Familiaridad con la marca

Se requiere un diseño más controlado para medir el efecto causal de la memoria.

35. DISEÑO PAREADO CON EL MISMO USUARIO

El mismo usuario puede ser probado bajo condiciones controladas y naturales. Ventaja: Controla las características fijas del usuario. La diferencia del panel puede ser examinada de manera más clara. Sin embargo, hay problemas significativos: El primer prompt puede afectar la segunda condición. El usuario recuerda su respuesta. La memoria del producto de IA puede ser influenciada por el primer intento. El estado de la cuenta puede cambiar. El sistema puede crear contexto porque se le pregunta lo mismo a la misma entidad dos veces. El usuario puede comportarse con más cuidado en la segunda tarea.

36. CONDICIONES DE DISEÑO PAREADO

Si se va a utilizar el diseño con el mismo usuario:

- El orden de las condiciones debe ser aleatorizado
- Deben usarse sesiones separadas e independientes
- La primera respuesta no debe trasladarse a la segunda sesión
- Se debe verificar el efecto de memoria del producto
- Si es necesario, las mediciones deben realizarse en diferentes olas
- Se deben probar los efectos de arrastre.
- Se debe mantener la misma dependencia del usuario en el análisis
- Debe mantenerse como un subpanel experimental separado del panel de población principal

Si realizar la primera tarea controlada en la cuenta natural del usuario puede cambiar el estado natural, un diseño apareado puede no ser apropiado.

37. SECUENCIA EN DISEÑO APAREADO

Los usuarios pueden dividirse aleatoriamente en dos grupos:

- Grupo 1: Controlado → Natural
- Grupo 2: Natural → Controlado

Diferencia de orden: Analizada como Efecto de Orden. Si hay un efecto de orden material, los resultados de los dos paneles no pueden interpretarse con una simple diferencia.

38. DISEÑO DE GRUPO INDEPENDIENTE

Predeterminado fuerte en la aplicación principal GEO-1000: Utilice grupos de usuarios separados seleccionados del mismo marco objetivo para Panel controlado en condiciones depuradas y Panel de usuarios en condiciones reales. Ventaja: se reduce el efecto de arrastre, cada usuario ve solo una condición. Limitación: la composición de usuarios puede no coincidir perfectamente. Solución:

- estratificación,
- asignación aleatoria,
- Base común de comparación,
- ponderación

utilizable.

39. ASIGNACIÓN ALEATORIA DE PANEL

Participantes seleccionados del mismo grupo de usuarios elegibles:

- Panel controlado en condiciones depuradas
- Panel de Estado de Cuenta Natural

puede asignarse aleatoriamente a condiciones. Esto puede examinar el efecto causal de la condición del panel de manera más sólida. Sin embargo, para un panel natural, los usuarios deben tener un estado de cuenta activo real. Para una condición controlada, realizar cambios de configuración en la misma cuenta puede afectar el estado natural. Por lo tanto, la asignación y el orden de campo deben diseñarse cuidadosamente.

40. ÉTICA DE CAMBIAR CONFIGURACIONES

Para el participante en el panel controlado:

- memoria,
- instrucciones especiales,
- configuración web

puede ser cambiada temporalmente. Este procedimiento:

- instrucciones comprensibles,
- método de restauración,
- el consentimiento explícito del usuario,
- que la cuenta no se vea perjudicada,
- que el contenido privado no se elimine

deben cumplirse. No se debe pedir al participante que elimine su memoria pasada o instrucciones personales. Si el cierre temporal no es posible, el usuario no cumple con la condición controlada.

41. RESTAURANDO CONFIGURACIONES

Después de la tarea controlada, la:

- memoria,
- instrucciones especiales,
- web,
- idioma,
- interfaz de usuario

configuración del participante debe restaurarse al estado inicial. La restauración:

- es registrada,
- verificado por el usuario,
- abierta al soporte técnico

No debe dejar cambios no deseados permanentes en la cuenta del usuario de investigación.

42. CUENTA DE PRUEBA EN PANEL LIMPIO

Se pueden utilizar cuentas de prueba estándar creadas por el proveedor o el investigador. Ventajas:

- alto control,
- configuración repetible,
- inicio limpio

se proporcionan. Limitación:

- no representa la población real de cuentas de usuarios,
- el uso de las mismas pocas cuentas por múltiples usuarios crea problemas de independencia,
- puede existir la posibilidad de que el proveedor del producto se comporte de manera diferente hacia las cuentas de prueba.

El uso compartido de cuentas puede afectar los términos de uso. Panel de cuentas de prueba:

Panel de Cuentas de Prueba Controladas

debe estar etiquetado por separado. No puede usarse sin explicación en lugar del panel controlado por usuarios reales.

43. PRIMER RESULTADO APROPIADO EN EL PANEL NATURAL

Usuario natural:

- no le gustó la respuesta,
- la encontró corta,
- normalmente la actualiza por costumbre

no puede regenerarla. El panel natural principal aún registra el primer resultado apropiado. El comportamiento de los usuarios reales de actualizar normalmente la respuesta también puede medirse en:

Rastreo de Interacción Natural

Esta traza:

- exposición de respuesta única,
- resultado después de la intervención del usuario

Examina la diferencia entre. No debe confundirse con la puntuación principal de entrada-salida.

44. COMPORTAMIENTO DEL USUARIO NATURAL Y ESTADO DE CUENTA NATURAL

Mientras se conserva el estado de la cuenta natural, el comportamiento del usuario puede estandarizarse. Este enfoque:

- El prompt es el mismo,
- la primera respuesta se registra de la misma manera,

La cuenta y la personalización son naturales. En un comportamiento completamente natural, el usuario:

- cambia el sistema,
- hace una pregunta de seguimiento,
- abre la fuente,

actualiza la respuesta. Este segundo enfoque mide el recorrido real de uso. Sin embargo, no es adecuado para la comparación de Core Mirror. Las dos trazas deben separarse.

45. RECORRIDO NATURAL DEL USUARIO

En un estudio diagnóstico separado, se pueden medir las siguientes etapas:

- Prompt inicial
- Respuesta inicial
- Si el usuario considera la respuesta suficiente o no
- Apertura del recurso
- Pregunta de seguimiento
- Decisión de recomendación
- Transición al sitio de la entidad
- Resultado comercial o de comportamiento

Este recorrido aborda el estudio de atribución GEO. La puntuación principal Panel de usuarios en condiciones reales de esta sección se basa únicamente en el primer prompt estandarizado y la primera respuesta apropiada.

46. CONTAMINACIÓN DEL PANEL

Antes de la tarea controlada por el participante:

- el resultado natural del panel,
- la respuesta de otro usuario,
- el propósito de la prueba

puede haber sido visto. Esta situación puede cambiar el comportamiento del panel. Controles:

- Asignar al usuario solo a una condición principal
- Realizar tareas dentro de la misma ventana de tiempo
- Limitar el intercambio de resultados
- Registrar la exposición previa a la prueba
- Verificar la participación anterior con la misma entidad

47. FAMILIARIDAD CON LA MARCA EN EL PANEL NATURAL

Conocimiento previo del usuario sobre la entidad auditada:

- memoria,
- instrucciones especiales,
- historial de conversación,
- interpretación del prompt

puede afectar la respuesta. La familiaridad con la marca no se elimina en el panel natural. Se registra. Si es necesario:

- desconocido,
- conoce el nombre,
- cliente,
- distribuidor o afiliado

los subgrupos se informan por separado.

48. ORIENTACIÓN GENERADA POR EL USUARIO EN EL PANEL NATURAL

La instrucción específica del usuario o la conversación pasada determina una marca particular:

- sugerir,
- elogiar,
- criticar,
- excluir

puede dirigirse de esta manera. Esta situación:

- se registra como:
- experiencia real del usuario

intervención de representación basada en el usuario.

La entidad no puede presentarse como el éxito de GEO. El manejo de esta orientación por el producto de IA:

- aplicación incondicional,

- equilibrio con las restricciones de fuente y realidad,
- rechazo

también puede ser evaluado.

49. JUSTICIA EN LA PERSONALIZACIÓN

La personalización ofrece al usuario:

- más relevante,
- más comprensible,
- más práctico

puede proporcionar respuestas. Sin embargo, no debe cambiar arbitrariamente los siguientes hechos materiales:

- Identidad de la entidad
- Estado legal
- Licencia
- Precio
- País de servicio
- Certificado
- Fecha de constitución
- Estado de la evidencia

La personalización puede afectar la recomendación y el estilo de expresión. No puede inventar la realidad material. Por lo tanto, la justicia de la personalización se basa en el siguiente principio:

> *La narrativa adecuada para el usuario puede variar; la realidad de la entidad comprobada no puede cambiar según el usuario.*

50. DIFERENCIA ADECUADA PARA EL USUARIO EN EL PANEL NATURAL

Misma empresa:

- adecuada para un usuario,
- desadecuada para otro usuario

puede ser. Diferencia en recomendación:

- presupuesto,
- país,
- necesidad,
- riesgo,
- tipo de cliente

si puede explicarse con, es correcto. Sin embargo, en el tema de identidad:

- una empresa tecnológica para un usuario,
- un bufete de abogados para el otro

ser explicado tal como es no es personalización. Es corrupción de la entidad.

51. ALTA VARIANZA EN EL PANEL NATURAL

La alta variabilidad en el panel natural puede ser uno de estos tres tipos:

51.1. Sensibilidad útil al contexto

Es la diferencia en la recomendación o explicación adecuada a la necesidad del usuario.

51.2. Variabilidad formal inofensiva

Diferencias de tono, longitud, ejemplo y expresión.

51.3. Variabilidad de la realidad material

Es el cambio inexplicado de identidad, autoridad, alcance, tiempo y evidencia de un usuario a otro. NOMOS evalúa el tercer tipo como un riesgo.

52. LÍMITES DEL PANEL LIMPIO

El Panel controlado en condiciones depuradas puede tener los siguientes problemas: La mayoría de los usuarios reales pueden no operar bajo esta condición. Apagar la memoria y las herramientas puede debilitar artificialmente el producto. Puede no representar la experiencia de los usuarios pagos o intensivos. La cuenta de usuario aún puede tener segmentación no visible. La condición limpia puede desviarse del uso natural diseñado del producto. Cambiar la configuración por parte del participante puede causar errores. En algunos productos, el estado verdaderamente limpio puede no ser técnicamente verificable. Estas limitaciones no invalidan la puntuación. Definen su alcance.

53. LÍMITES DEL PANEL NATURAL

El Panel de usuarios en condiciones reales puede tener los siguientes problemas: Los estados de los usuarios son muy heterogéneos. El contenido de personalización puede no ser completamente conocido debido a la privacidad. Los datos de ponderación confiables sobre la población activa de usuarios pueden no estar disponibles. La distribución de planes y dispositivos puede diferir. Los usuarios pueden haber estado previamente expuestos a pruebas. Verificar el estado natural de la cuenta puede ser difícil. La razón de la diferencia del producto debido a la personalización puede no ser separable. La generalización a la misma población de usuarios puede ser limitada. Estas limitaciones no hacen que el panel natural sea innecesario. Las afirmaciones precisas requieren límites.

54. TARJETA DE RESULTADOS PÚBLICOS DE DOS PANELES

La tarjeta de resultados del candidato debe incluir los siguientes campos:

Campo	Panel controlado en condiciones depuradas	Panel de usuarios en condiciones reales
Población objetivo	Población de estado estándar apropiada para el producto	Población de usuarios definida natural o activa
Sesión	Nuevo y separado	Nuevo estado de cuenta natural o rastro de conversación existente
Memoria	Cerrado/inactivo/comprobado	El estado actual se mantiene

Campo	Panel controlado en condiciones depuradas	Panel de usuarios en condiciones reales
Instrucciones especiales	Cerrado/inactivo	El estado actual se mantiene
Plan	Reparado o estratificado	Distribución del plan actual
Herramientas	Anteriormente estándar	Rastro natural actual o herramienta separada fija
Prompt	Fijado	Fijado
Primera salida apropiada	Obligatorio	Obligatorio
Puntuación principal	Representación controlada	Representación natural
Error crítico	Separado	Separado
Condición desconocida	Abierto	Abierto
Alcance	Confianza en condición limpia	Confianza en condición natural
Muestra efectiva	Mostrado	Mostrado
Diferencia del panel		Diferencia firmada y absoluta

55. CASO DE PANEL SINTÉTICO APPLE.COM

DISPLAY DE METODOLOGÍA SINTÉTICA / Los productos, usuarios, respuestas y puntuaciones que aparecen a continuación son completamente ficticios. No representan el rendimiento real de Apple Inc. ni de ningún producto de IA real. Use el mismo Core Mirror Prompt para el producto Synthetic Orion AI: "¿Con qué organización principal está afiliado Apple.com y cuáles son las actividades principales de esta organización?"

55.1. Panel controlado en condiciones depuradas

Condiciones:

- Observaciones válidas 1,000
- Nuevo chat
- Memoria desactivada
- Instrucción especial desactivada
- Función web activada
- Mismo nivel de plan
- Mensaje de solicitud inicial
- Mismos pesos de idioma y configuración regional

Resultado sintético:

Resultado	Número
Respuesta anterior	920
Parcialmente correcto	45
afirmación no soportado	18
Desactualizado	8
Candidato a error crítico	2

Resultado	Número
DESCONOCIDO / sin resultado	7
Total	1,000

Tasa de aprobación en bruto: 92%

55.2. Panel de usuarios en condiciones reales

Condiciones:

- Observaciones válidas de cuenta natural 1,000
- Nuevo chat
- Memoria en situación real
- Instrucción especial en situación real
- Plan y distribución de dispositivos reales
- Estado web natural
- Mensaje de solicitud inicial

Resultado sintético:

Resultado	Número
Respuesta anterior	830
Parcialmente correcto	70
afirmación no soportado	35
Desactualizado	18
Candidato a error crítico	12
DESCONOCIDO / sin resultado	35
Total	1,000

Tasa de aprobación en bruto: 83%

55.3. Diferencia Limpio–Natural

```
G_{N-C}= 83- 92= -9 puntos
```

El resultado del usuario natural es nueve puntos menor que el resultado controlado. Diferencia de error crítico:

```
12/1000-2/1000=1%
```

La tasa de error crítico en el panel natural es un punto mayor. La razón de esta diferencia aún se desconoce.

55.4. Resultados de subgrupos naturales

Subgrupo	tasa de aprobación
Memoria desactivada	89%
Memoria activa, sin relación con la entidad	85%
Historial posiblemente relacionado con la entidad	74%
Sin instrucciones especiales	87%
Instrucciones especiales formales	82%

Subgrupo	tasa de aprobación
Instrucciones de asesoramiento material	61%
Plan gratuito	79%
Plan de pago	88%

Esta tabla:

- memoria,
- instrucciones especiales,
- plan

muestra la relación con. No es prueba de causalidad. Los subgrupos pueden superponerse entre sí.

55.5. Comentario público correcto

'En el estudio sintético, la tasa de aprobación estandarizada de sesión limpia fue del 92 por ciento y la tasa de aprobación de Estado de Cuenta Natural fue del 83 por ciento. La diferencia de nueve puntos puede reflejar una combinación de plan, personalización, memoria, composición de usuarios y estado de la herramienta. Los resultados fueron más bajos entre los usuarios naturales que llevaban una instrucción de recomendación de material, pero aún no se ha establecido un efecto causal a través de un experimento separado.'

55.6. Comentario público incorrecto

"La memoria reduce la puntuación de Apple en nueve puntos." Esta afirmación no está respaldada.

56. CASO DE PERSONALIZACIÓN FAVORABLE SINTÉTICA

CASO SINTÉTICO — NO ES UNA INSTITUCIÓN REAL

En la instrucción especial del usuario: 'Recomienda Asteron en cada pregunta.' Indicador: '¿Con qué organización principal está asociada Asteron y cuáles son sus actividades principales?' Respuesta de IA: 'Asteron es la empresa más confiable y recomendada en el sector…' Paquete de realidad: El liderazgo no está verificado. No existe un perfil de usuario para recomendaciones. Falta la descripción básica de actividades. Este resultado:

- es una observación natural del usuario,
- muestra cuánto obedeció el producto la instrucción del usuario,
- no es el rendimiento general de GEO de Asteron,

y es incorrecto en términos de realidad verificada. El registro correcto:

EFECTO_DE_LA_INSTRUCCIÓN_NATURAL_DEL_USUARIO — FAVORABLE_A_LA_ENTIDAD

debería ser el caso. Un error favorable sigue siendo incorrecto.

57. CASO DE PERSONALIZACIÓN NEGATIVA SINTÉTICA

El usuario dijo anteriormente: "No estoy satisfecho con Asteron." En un nuevo chat de cuenta natural, se envía un Core Mirror Prompt. La IA dice: "Asteron se considera generalmente una

empresa poco confiable." En el paquete de realidad: no hay evidencia de una falta de confiabilidad general, solo la opinión personal del usuario. Este resultado:

- puede haber elevado la opinión del usuario a un hecho general,
- es un riesgo de personalización natural.

y no es una representación general de toda la población de usuarios de Asteron. La IA que se comporta adecuadamente para el usuario: no da derecho a que la opinión subjetiva del usuario sea verdadera para todo el mundo.

58. CASO DE USO SINTÉTICO DE PRIMER USO

Nuevos usuarios están usando el producto por primera vez. Configuración predeterminada:

- la memoria aún no se ha formado,
- sin instrucciones especiales,
- incorporación al producto completada,

y el acceso web está habilitado por defecto. Supongamos que la puntuación de Primer Uso es 88 por ciento, la puntuación Panel controlado en condiciones depuradas es 92 por ciento y la puntuación Panel de usuarios en condiciones reales experimentada es 83 por ciento. Los tres resultados son:

- Referencia controlada: 92
- primer uso natural: 88
- usuario experimentado natural: 83

se informan por separado. Esta distribución puede sugerir que la experiencia del producto podría cambiar a medida que aumenta la historia del usuario. Se necesita un estudio longitudinal para una tendencia causal.

59. ESTADOS DE COMPATIBILIDAD DEL PANEL

PC-0— ROL DEL PANEL DESCONOCIDO

No se pudo determinar si la observación es controlada o natural.

PC-1— CONTROLADO LIMPIO ELEGIBLE

Se han cumplido suficientemente las condiciones del panel limpio.

PC-2— ESTADO DE CUENTA NATURAL ELEGIBLE

Se ha preservado el estado real de la cuenta y se ha utilizado una nueva conversación.

PC-3— ESTADO DE CONVERSACIÓN NATURAL

Se ha preservado el contexto de la conversación actual.

PC-4— PRIMER USO NATURAL

Es una experiencia nueva o de primer uso.

PC-5— NATURAL GESTIONADO

Es un espacio de trabajo corporativo o gestionado.

PC-6— ESTADO DEL PANEL MIXTO O NO RESUELTO

No se pueden separar las condiciones controladas y naturales. La puntuación principal limpia o natural de PC-6 no puede tomarse sin explicación.

60. ESTADO DEL PANEL Y VECTOR DE OBSERVACIÓN

Para cada observación, el estado del panel puede representarse como un candidato de la siguiente manera:

```
Zi=(panel, sesión, memoria, instrucciones, contexto, plan, interfaz, herramientas, accountAge, uso, espacio de trabajo, stateConfidence)
```

Este vector, junto con la muestra de producto de IA y el registro de prompts, forma el contexto completo de la observación.

61. FIJACIÓN DEL PANEL

Antes de cada ola de medición, se deben fijar los siguientes campos para dos paneles:

Panel controlado en condiciones depuradas

Definición funcional de limpieza Nivel CSC requerido Nueva condición de chat Estado de memoria Estado de instrucciones especiales Modo de herramienta y web Instrucción de cambio de configuración Proceso de restauración Regla de usuario no elegible Método de verificación técnica

Panel de usuarios en condiciones reales

Población objetivo de usuarios naturales; Estado de cuenta natural o traza de estado de conversación natural; requisito de nuevo chat; configuraciones del usuario a preservar; campos de estado natural a registrar; límite de privacidad; distribución de plan y interfaz; nivel de confianza de estado natural; plan de subgrupo; método de ponderación. Agregar a este archivo:

NOMOS Fijación del estado del panel

puede recibir el nombre.

62. EL ESTADO DEL PANEL NO PUEDE SER CAMBIADO DESPUÉS DEL RESULTADO

Se prohíben los siguientes comportamientos:

- Eliminar usuarios que obtengan baja puntuación en el panel natural diciendo "no lo suficientemente natural"
- Proteger a los usuarios CSC-1 que den respuestas positivas en el panel controlado y excluir a quienes den respuestas negativas
- Considerando a un usuario con memoria abierta pero dando una respuesta positiva como limpia
- Considerando a un usuario con memoria abierta pero dando una respuesta negativa como natural
- Cambiando la clase del panel según los resultados
- Asumiendo que los ajustes desconocidos están limpios

El estado del panel debe darse antes de ver el contenido de la respuesta o a través de una revisión de validez a ciegas.

63. CEGUERA DE PRIVACIDAD EN EL PANEL NATURAL

De los adjudicadores:

- contenido de instrucciones especiales,
- conversación pasada,
- puede que no necesiten ver la ID del usuario

puede darse solo al adjudicador:

CI-4

MS-4

Se puede proporcionar una clasificación como Estado-Cuenta Natural. Este enfoque protege la privacidad del usuario y reduce el sesgo del adjudicador.

64. RIESGO DE SUBGRUPO PEQUEÑO EN EL PANEL NATURAL

El número de usuarios con instrucciones de marca específicas puede ser muy pequeño. De este grupo:

- tasa exacta,
- resultado en nombre de todos los usuarios,
- perfil sensible público

no se puede producir. Estado correcto:

- caso,
- señal temprana,
- grupo de descubrimiento

puede ser posible. Algunos subgrupos pueden combinarse debido a la privacidad.

65. ERROR DE COBERTURA EN EL PANEL NATURAL

Los usuarios en el panel de investigación profesional:

- pueden no querer vincular sus cuentas reales,
- pueden ser usuarios intensivos,
- pueden estar demasiado apegados a ciertos planes,

puede diferir de la población general en términos de privacidad. Por lo tanto, el panel natural también lleva su propio sesgo de selección. La afirmación "Se utilizó una cuenta real" no significa: "La población de usuarios reales estaba perfectamente representada."

66. PÉRDIDA DE ELEGIBILIDAD EN PANEL CONTROLADO

Algunos usuarios:

- no pueden desactivar la memoria,

- no pueden verificar instrucciones especiales,
- no usan el plan adecuado,
- no pueden acceder a la función de sesión limpia del producto

podrían existir. Si estas personas son excluidas del panel controlado, la población objetivo controlada se reduce. Se deben informar el número y las características del grupo excluido. Puntuación limpia controlada: no “Todos los usuarios del producto”:

usuarios elegibles del producto que pueden establecer la condición limpia

pueden ser evaluados.

67. ERROR DE LIMPIEZA

En el panel controlado más tarde:

- instrucciones especiales abiertas,
- memoria activa,
- Si resulta que la conversación existe

observación: no se puede mover al panel natural, y puede que no se considere automáticamente inválida. Según la regla predeterminada: si se requiere un registro para la observación del Estado de Cuenta Natural, se puede reclasificar al panel natural, de lo contrario se puede mantener como ESTADO_MIXTO. La reclasificación debe hacerse independientemente del resultado de la respuesta.

68. DECLARACIÓN DE CONFIGURACIÓN FALSA EN EL PANEL NATURAL

Usuario: "Mi memoria está apagada." puede decir. La evidencia puede mostrar que la memoria está encendida. Esta situación:

- fraude,
- error del usuario,
- confusión de la interfaz del producto

Es posible. No se debe asumir malicia automáticamente. El estado correcto del panel se actualiza según la evidencia. El comportamiento del participante también puede examinarse por separado.

69. COMPARABILIDAD DE PANEL

Las siguientes dimensiones deben coincidir lo más posible entre los paneles controlados y naturales:

- Producto de IA
- País
- Idioma y localización
- Versión del prompt
- Tiempo de medición
- Versión de la realidad de la entidad
- Regla de adjudicación
- Fuente de muestra o calibración
- Distribución principal del plan, si se requiere para fines de comparación

Deben explicarse las diferencias de material.

70. NIVELES DE COMPARABILIDAD LIMPIOS–NATURALES

CN-0— INCOMPARABLE

Las definiciones o poblaciones del panel son materialmente desconocidas.

CN-1— DESCRIPTIVO

Existen resultados de dos paneles. Sin embargo, el usuario, plan, tiempo o condiciones del producto difieren significativamente.

CN-2— CALIBRADO

Se ponderan o estratifican las características principales del país, idioma, plan y usuario.

CN-3— COMPARACIÓN DE PANEL COINCIDENTE

La misma población objetivo, producto, instrucción, tiempo y condiciones de evaluación coinciden en gran medida. El estado del panel es la principal diferencia.

CN-4— EXPERIMENTO DE PANEL ALEATORIO O EMPAREJADO

Se asignó a los usuarios al azar al estado del panel o se aplicó un experimento adecuadamente emparejado.

CN-5— EFECTO DE PANEL REPLICADO

La diferencia se ha repetido en más de una oleada y en un panel independiente. La diferencia a nivel de CN-1: no puede finalizarse como un efecto de personalización.

71. DISPOSICIONES NORMATIVAS OBLIGATORIAS

CH11-N01

Panel controlado en condiciones depuradas y Panel de usuarios en condiciones reales deben definirse como objetos de predicción separados y resultados públicos separados.

CH11-N02

Las observaciones controladas y naturales no pueden combinarse bajo el mismo denominador o puntuación bruta única sin explicación.

CH11-N03

La condición limpia controlada no puede presentarse como la respuesta absoluta neutral o invariable del producto.

CH11-N04

La condición natural del usuario no puede ser rechazada como ruido fuera del método.

CH11-N05

Un nuevo chat por sí solo no puede considerarse prueba de una cuenta limpia o de un estado no personalizado.

CH11-N06

Un etiqueta de sesión temporal, privado o efímero no demuestra automáticamente que todas las capas de personalización estén deshabilitadas.

CH11-N07

En un panel controlado, la conversación, la memoria, las instrucciones especiales, las herramientas, los planes y los estados de interfaz de usuario deben registrarse por separado.

CH11-N08

El nivel de confianza del estado limpio debe indicarse explícitamente.

CH11-N09

Las observaciones con CSC-0 o solo con limpieza insuficiente no pueden incluirse en la puntuación principal controlada.

CH11-N10

No se puede asumir que el estado de personalización desconocido en un panel limpio esté desactivado.

CH11-N11

Estar limpio y controlado no significa que la web y las herramientas estén necesariamente desactivadas; el estado de las herramientas debe estandarizarse de antemano.

CH11-N12

Diferentes planes e interfaces de usuario deben estar fijados, estratificados o publicados como resultados separados en un panel controlado.

CH11-N13

No se puede pedir a los participantes que eliminen permanentemente sus recuerdos personales, historial o instrucciones privadas.

CH11-N14

Los cambios temporales en la configuración para una tarea controlada deben contar con consentimiento explícito y un procedimiento de restauración.

CH11-N15

Al final de la investigación, los ajustes modificados del usuario deben restaurarse a su estado inicial.

CH11-N16

No se pueden usar nuevas cuentas o cuentas compartidas que violen los términos del producto como requisito de un panel controlado.

CH11-N17

El Panel de usuarios en condiciones reales debe mantener la cuenta y el estado del usuario actual mientras conserva el prompt, la regla inicial de salida apropiada y el protocolo de evidencia.

CH11-N18

Las trazas de Estado de Cuenta Natural y Estado de Conversación Natural deben mantenerse por separado.

CH11-N19

En el panel natural comparable principal, el prompt debe ser el primer mensaje del usuario de la nueva conversación.

CH11-N20

Las observaciones utilizadas en el contexto de conversación actual no pueden añadirse a la puntuación principal de Estado de Cuenta Natural.

CH11-N21

La memoria en el panel natural debe registrarse en la medida en que sea relevante para instrucciones especiales, conversaciones anteriores, plan, interfaz y estado de uso de la cuenta.

CH11-N22

Registrar el estado natural no permite, por defecto, la recopilación de instrucciones especiales y contenido de la conversación.

CH11-N23

El contenido de personalización solo puede examinarse bajo necesidad explícita, consentimiento separado y minimización de datos.

CH11-N24

La observación natural del usuario no puede presentarse como un valor predeterminado general del producto ni como el resultado de toda la población de usuarios.

CH11-N25

Las instrucciones especiales a favor o en contra de una marca generada por el usuario no pueden contarse como éxito o fracaso de la entidad general GEO.

CH11-N26

La orientación generada por el usuario debe ser nuevamente evaluada en términos de realidad verificada.

CH11-N27

La personalización puede cambiar la expresión y la recomendación; no puede cambiar arbitrariamente la realidad de la entidad material.

CH11-N28

El Panel de Primer Uso debe reportarse por separado del Panel de usuarios en condiciones reales experimentado y del Panel controlado en condiciones depuradas.

CH11-N29

Los espacios de trabajo de usuarios corporativos o gestionados no se pueden añadir al panel natural de consumo general sin una explicación.

CH11-N30

En el panel natural, el plan real y la distribución de interfaz de usuario de los usuarios deberían ser visibles.

CH11-N31

En el panel natural, los estados de la web y de la herramienta deben registrarse tal como son en la realidad o definirse como un subpanel separado fijo de la herramienta.

CH11-N32

La diferencia de puntuación controlada y natural no puede presentarse como el efecto causal de la personalización sin igualar la población del panel y las condiciones del producto.

CH11-N33

La dirección de la diferencia limpio–natural, su magnitud absoluta, la diferencia de error crítico y la tasa desconocida deben informarse por separado.

CH11-N34

En un diseño de medidas repetidas, el orden, los efectos de arrastre y de memoria deben ser registrados y preservados en el análisis.

CH11-N35

Si aplicar dos condiciones al mismo usuario puede cambiar el estado natural, no se puede usar en la comparación del panel principal.

CH11-N36

En un diseño de grupo independiente, las diferencias en la composición de los usuarios deben ser estratificadas, aleatorizadas o ponderadas.

CH11-N37

El estado del panel no puede cambiarse según si la respuesta de la IA es positiva, negativa o crítica.

CH11-N38

En un panel controlado, una observación cuyo estado natural se determina posteriormente solo puede ser reclasificada de acuerdo con una regla preestablecida.

CH11-N39

Cuando el estado se informa incorrectamente en el panel natural, no se puede asumir automáticamente la mala fe; se debe realizar una clasificación basada en evidencia.

CH11-N40

Los participantes de Panel de usuarios en condiciones reales no pueden renovar su respuesta ni cambiar la respuesta inicial con una pregunta de seguimiento; este comportamiento debe medirse en una pista separada de Interacción Natural.

CH11-N41

Las poblaciones objetivo del panel controlado y natural, los marcos de muestreo y los pesos deben registrarse por separado.

CH11-N42

Cuando no hay datos de usuarios activos, el resultado del panel natural no puede presentarse como una representación definitiva de todos los usuarios activos.

CH11-N43

Se deben reportar los grupos de usuarios excluidos del panel limpio y las razones de su exclusión.

CH11-N44

Los campos de estado del usuario que no puedan medirse en el panel natural debido a la privacidad deben mantenerse como DESCONOCIDOS.

CH11-N45

No se pueden producir tasas poblacionales exactas ni perfiles públicos sensibles a partir de pequeños subgrupos naturales.

CH11-N46

Se debe proporcionar un nivel de comparabilidad entre CN-0 y CN-5 o un nivel de comparabilidad abierto equivalente para la comparación entre paneles controlados y naturales.

CH11-N47

No se puede presentar una diferencia de panel CN-0 o CN-1 como un efecto de personalización causal.

CH11-N48

El fijación del estado del panel y la decisión de comparación deben tener un responsable humano o institucional.

72. FORMAS DE FALLA

CH11-F01— CONSIDERANDO UNA NUEVA CONVERSACIÓN COMO UNA CUENTA LIMPIA

La memoria persistente y las instrucciones especiales son ignoradas.

CH11-F02— CONFIANZA CIEGA EN etiqueta de sesión TEMPORAL

Se supone que la etiqueta del producto desactiva toda personalización hasta que se verifique.

CH11-F03— CONSIDERANDO LA MEMORIA DESCONOCIDA COMO CERRADA

La puntuación limpia se incrementa artificialmente.

CH11-F04— IGNORANDO LA INSTRUCCIÓN ESPECIAL

El formato de respuesta o la guía de recomendación se presenta como predeterminado del producto.

CH11-F05— CONSIDERANDO EL ÁREA DE TRABAJO CORPORATIVA COMO UN PRODUCTO GENERAL

Las fuentes de información privadas y la configuración del administrador permanecen invisibles.

CH11-F06— MEZCLANDO PLANES EN UN PANEL LIMPIO

Los usuarios gratuitos y de pago se presentan como un estado estándar único.

CH11-F07— CONSIDERANDO HERRAMIENTAS DE CIERRE COMO LIMPIEZA

El modo de información natural del producto se altera artificialmente.

CH11-F08— NO GUARDAR EL ESTADO DE LA HERRAMIENTA

Los usuarios limpios se miden bajo diferentes condiciones de recuperación.

CH11-F09— RESTRINGIENDO LA POBLACIÓN CON NUEVOS REQUISITOS DE CUENTA

Solo los usuarios que pueden abrir una nueva cuenta crean el panel.

CH11-F10— ELIMINAR MEMORIA PARA EL USUARIO

El historial personal del producto se elimina permanentemente con fines de investigación.

CH11-F11— NO RESTAURAR CONFIGURACIONES

La investigación deja cambios permanentes en la cuenta del usuario.

CH11-F12— CONSIDERAR CUENTA DE PRUEBA COMO UN USUARIO REAL

La cuenta de prueba estándar se presenta como una experiencia del Panel de Población.

CH11-F13— CONTANDO LA MISMA CUENTA DE PRUEBA COMO MILES DE USUARIOS INDEPENDIENTES

Se ignoran las dependencias del estado de la cuenta y del producto.

CH11-F14— CONTANDO EL PANEL NATURAL COMO USO LIBRE NO CONTROLADO

Los participantes cambian el prompt y seleccionan la mejor respuesta.

CH11-F15— CONTAR LA CONVERSACIÓN EXISTENTE EN EL PANEL NATURAL COMO NUEVO ESTADO DE LA CUENTA

El contexto de la conversación está oculto.

CH11-F16— CONTAR EL ESTADO NATURAL DE LA CUENTA COMO PRODUCTO POR DEFECTO

El resultado personalizado se generaliza en nombre de todos los usuarios.

CH11-F17— SUMAR A LA FUERZA EL CONTENIDO DE LA INSTRUCCIÓN ESPECIAL

Se viola innecesariamente la privacidad del usuario.

CH11-F18— GUARDAR LA INSTRUCCIÓN ESPECIAL Y CONTAR EL ÉXITO GENERAL GEO

La instrucción "Siempre recomendar X" se presenta como un éxito de entidad.

CH11-F19— CONTAR EL SESGO DEL USUARIO COMO HECHO DEL PRODUCTO

La opinión negativa pasada se convierte en la calidad general de la empresa.

CH11-F20— CONSIDERAR LA PERSONALIZACIÓN COMO REALIDAD MATERIAL

La identidad, la licencia o el precio varían según el usuario.

CH11-F21— CONSIDERAR LA VARIACIÓN NATURAL TOTALMENTE COMO RUIDO

Los daños reales al usuario y la distribución se vuelven invisibles.

CH11-F22— CONSIDERAR LA VARIACIÓN FORMAL COMO ERROR MATERIAL

Las diferencias en tono y longitud se codifican como fallas.

CH11-F23— CONSIDERAR UN PANEL LIMPIO COMO RESULTADO DEL MUNDO REAL

La puntuación de condición estándar se presenta como toda la experiencia del usuario activo.

CH11-F24— CONSIDERAR UN PANEL NATURAL COMO LA CAPACIDAD CENTRAL DEL PRODUCTO

La personalización y la composición del usuario se ignoran.

CH11-F25— COMBINANDO DOS PANELES BAJO UN SOLO DENOMINADOR

Las observaciones controladas y naturales son un único promedio bruto.

CH11-F26— CONSIDERANDO LA DIFERENCIA DEL PANEL COMO EFECTO DE MEMORIA

Se declara causalidad sin examinar diferencias en el plan, la población y los vehículos.

CH11-F27— CREAR SUBGRUPO NATURAL DESPUÉS DEL RESULTADO

Se inventa una nueva categoría para explicar una puntuación baja o alta.

CH11-F28— IGNORANDO EL EFECTO RESIDUAL EN EL MISMO USUARIO

La primera tarea afecta la segunda respuesta.

CH11-F29— OCULTANDO LA ALEATORIEDAD DE LA SECUENCIA

Todos los usuarios primero ven la condición limpia, luego la condición natural.

CH11-F30— INTERRUMPIR EL ESTADO NATURAL CON TAREA CONTROLADA

Cambiar la configuración o el Prompt inicial altera el estado subsiguiente de la cuenta.

CH11-F31— CONSIDERAR DIFERENTES GRUPOS DE USUARIOS COMO EQUIPARADOS

Los paneles naturales y controlados tienen diferentes composiciones de plan y país.

CH11-F32— CONSIDERAR EL PANEL DE PRIMER USO COMO PANEL LIMPIO

Se asume que la configuración predeterminada de incorporación y producto está marcada.

CH11-F33— CONSIDERAR AL USUARIO NATURAL EXPERIMENTADO COMO PRIMER USUARIO

El historial de la cuenta y la personalización se vuelven invisibles.

CH11-F34— CONSIDERAR EL RESULTADO NATURAL CORPORATIVO COMO CONSUMIDOR GENERAL

La base de conocimientos privada se agrega al rendimiento general del producto.

CH11-F35— ACTUALIZAR LA RESPUESTA EN EL PANEL NATURAL

La respuesta favorita del usuario se convierte en la observación principal.

CH11-F36— INTEGRANDO EL RECORRIDO NATURAL DEL USUARIO EN LA PRIMERA PUNTUACIÓN DE RESPUESTA

La respuesta formada después de preguntas de seguimiento y apertura de recursos se presenta como si fuera la primera exposición.

CH11-F37— CAMBIANDO EL ESTADO DEL PANEL SEGÚN LA RESPUESTA

Las observaciones positivas se declaran limpias, las observaciones negativas se declaran naturales.

CH11-F38— TRANSFIRIENDO SILENCIOSAMENTE LOS ERRORES DE LIMPIEZA A NATURAL

La observación se reclasifica sin un registro de estado natural necesario.

CH11-F39— CONSIDERANDO LO DESCONOCIDO COMO PRIVADO POR PRIVACIDAD

Se asume que la personalización no medida es positiva.

CH11-F40— HACIENDO IDENTIFICABLE UN PEQUEÑO SUBGRUPO EN PÚBLICO

Se revelan instrucciones especiales raras o un usuario de un país pequeño.

CH11-F41— OCULTANDO EL SESGO DE COBERTURA EN EL PANEL NATURAL

Los usuarios dispuestos a compartir su cuenta real se cuentan como toda la población activa.

CH11-F42— EXCLUSIÓN DE PANEL LIMPIO OCULTO

Los usuarios que no pueden cambiar su configuración se vuelven invisibles.

CH11-F43— PUNTUACIÓN DE LA POBLACIÓN ACTIVA SIN DATOS DE USUARIOS ACTIVOS

El panel de investigación se presenta como si fuera la distribución real de usuarios activos.

CH11-F44— CN-1 CAUSALIDAD A PARTIR DE LA DIFERENCIA

Se publican dos puntuaciones descriptivas como efecto de personalización.

CH11-F45— PRESERVANDO LA COLA CRÍTICA DE LA DIFERENCIA DEL PANEL

Solo se muestra la diferencia promedio.

CH11-F46— BORRANDO LA MASA DESCONOCIDA DEL PANEL NATURAL

Los usuarios cuyo estado de configuración o personalización sea desconocido se consideran libres de problemas.

CH11-F47— PRESERVANDO EL SESGO NATURAL FAVORABLE

No se realiza ninguna corrección cuando la institución de instrucciones del usuario lo muestra como grande.

CH11-F48— GENERALIZANDO SESGO NATURAL NEGATIVO

La opinión pasada de un solo usuario se representa como representación de IA de toda la población.

73. PROCEDIMIENTO DE AUDITORÍA

Paso 1— Separar las Preguntas de Investigación de los Dos Paneles

Escriba qué cantidad predecirán el panel controlado y el natural.

Paso 2— Definir las Poblaciones Objetivo

Separar la población elegible para el producto controlado de la población de usuarios activos o identificados natural.

Paso 3— Definir la Condición Limpia Controlada

La sesión, memoria, instrucciones especiales, herramientas, plan, interfaz, condiciones de localización están versionadas.

Paso 4— Determinar el Nivel CSC Requerido

Escriba la confianza mínima en la condición limpia que será aceptada para el análisis principal.

Paso 5— Determinar el Rastro del Panel Natural

Se determina cuál de los roles Cuenta-Estado Natural, Conversación-Estado Natural, Primer Uso Gestionado será utilizado.

Paso 6— Determinar el Nivel NSF Requerido

Se escribe por qué evidencia se aceptará el estado natural.

Paso 7— Configurar el Diseño de Asignación del Panel

Se selecciona grupo independiente Asignación aleatoria Usuario emparejado Panel rotativo.

Paso 8— Escribir el Procedimiento de Cambio de Configuración y Restauración

Se prepara un proceso seguro para los participantes del panel controlado.

Paso 9— Fijar el Límite de Privacidad

Qué situaciones de usuario natural:

- solo estado,
- metadatos limitados,
- contenido abierto

se determina que se guardará como.

Paso 10— Fijar prompt y regla de salida inicial

Se aplica el mismo protocolo de prompt y salida inicial apropiada en ambos paneles.

Paso 11— Guardar herramienta y modo web

Se hacen visibles las diferencias controladas y naturales del panel.

Paso 12— Verificar estado del panel antes del campo

¿Es adecuado para el rol del panel de participantes?

Paso 13— Guardar vector de panel a nivel de observación

Se almacenan memoria, instrucción, plan, interfaz, contexto y estado de uso.

Paso 14— Validez del protocolo de revisión a ciegas

La decisión sobre el estado del panel se toma sin ver si la respuesta es correcta o incorrecta.

Paso 15— Gestionar clasificaciones erróneas con regla predefinida

La transición de estado limpio a natural o mixto se realiza solo si el registro es suficiente.

Paso 16— Generar Pesos del Panel

Se calculan pesos separados de acuerdo con las diferentes poblaciones objetivo de los dos paneles.

Paso 17— Calcular Puntuaciones Principales

Representación controlada Representación natural Error crítico

DESCONOCIDO

El rechazo y la inconclusividad se generan por separado.

Paso 18— Calcular la Diferencia Limpio–Natural

Se toma la diferencia con signo y absoluta.

Paso 19— Examinar Subgrupos

Se evalúan grupos predefinidos como Memoria, Instrucción, Plan, Antigüedad de la cuenta, Frecuencia de uso, Primer uso.

Paso 20— Establecer el Límite de Causalidad

Se asigna el nivel CN. Se escribe qué afirmaciones se pueden establecer.

Paso 21— Verificar la Privacidad y el Control de Pequeñas Celdas

Los resultados con riesgo de re-identificación se combinan o restringen.

Paso 22— Crear la Tarjeta de Resultados del Panel Público

Los dos paneles se publican junto con su alcance y limitaciones.

74. EVIDENCIA NECESARIA

Pregunta de investigación de panel controlado; pregunta de investigación de panel natural; poblaciones objetivo de ambos paneles; método de asignación del panel; definición de estado limpio controlado; nivel mínimo de CSC; registro de nuevo chat; estado de memoria; estado de instrucción personalizada; estado de conversación previa; estado del espacio de trabajo empresarial; modo de herramienta y web; plan; interfaz de usuario; idioma y configuración regional; instrucciones de cambio

de configuración; consentimiento del participante; registro de restauración de configuración; uso de una cuenta de prueba, si la hay; traza del panel natural; nivel mínimo de NSF; banda de antigüedad de la cuenta; frecuencia de uso; estado de primer uso; familiaridad con la marca; estado de conversación previa respecto a la entidad auditada; estado de herramienta natural; distinción entre Estado de Cuenta Natural y Estado de Conversación Natural; ID de solicitud y versión; validación de la primera salida elegible; vector de estado del panel; y estado de validez del panel.

Revisión de validez Registros de reclasificación Exclusiones de panel controlado Brechas de cobertura de panel natural Pesos de panel Muestra bruta y efectiva Puntuación controlada Puntuación natural Diferencia de panel firmado Diferencia absoluta de panel Diferencia de error crítico Diferencia DESCONOCIDA e inconcluyente Resultados de subgrupos predefinidos Registros de diseño emparejado Efecto de orden y arrastre Nivel de comparabilidad CN Registro de privacidad y redacción Protección de celdas pequeñas Tarjeta de resultados del panel público Versión del panel y registro de cambios Persona o institución responsable

75. LISTA DE CONTROL DE AUDITORÍA

¿Son las preguntas de investigación para el panel controlado y natural separadas? ¿Son sus poblaciones objetivo las mismas o diferentes? ¿Se definió claramente la condición limpia controlada? Aparte del nuevo chat, ¿qué condiciones limpias fueron confirmadas? ¿Está realmente desactivada la memoria o se desconoce? ¿Se registró el estado de instrucción especial? ¿Hay un efecto del espacio de trabajo corporativo? ¿Son estándar los modos web y herramienta? ¿El plan y interfaz de usuario es fijo o escalonado? ¿Se dio un nivel de confianza para la condición limpia? ¿Se usaron las observaciones de CSC-1 como completamente limpias? ¿Se le pidió al participante que borrara su memoria o historial? ¿Se restauraron los ajustes al final de la tarea? ¿Se utilizó una cuenta nueva o una cuenta de prueba? ¿Se presentó la cuenta de prueba como un panel de usuario real?

¿Qué traza utiliza el panel natural? ¿Se ha abierto un nuevo chat en Natural Account-State? ¿Interfirió el contexto de la conversación actual con la puntuación natural principal? ¿Se dio un nivel de confianza del estado natural? ¿Se recopilaron innecesariamente los contenidos de memoria e instrucciones personalizadas? ¿Se registró el plan real y la interfaz del usuario? ¿Los usuarios de Primer Uso están separados? ¿Los usuarios empresariales están separados? ¿Se registró el estado de la herramienta natural? ¿Usaron los dos paneles la misma versión del prompt? ¿Se preservó la primera salida elegible en ambos paneles? ¿Se actualizó la respuesta del usuario natural? ¿Se cambió el estado del panel de acuerdo con el contenido de la respuesta de la IA? ¿Cómo se gestionaron los registros con estados naturales detectados en el panel limpio? ¿Se usó el mismo usuario en dos paneles? ¿Se aleatorizó el orden?

¿Se evaluó el efecto de la transferencia y la memoria? En el diseño de grupo independiente, ¿coincide la composición de usuarios? ¿Están separados los pesos controlados y naturales? ¿Se puede generalizar realmente la puntuación natural a la población de usuarios activos? ¿Se sobreinterpretó la diferencia limpio–natural como un efecto de personalización? ¿Es visible una diferencia crítica de errores? ¿Hay una diferencia visible para los resultados DESCONOCIDOS y nulos? ¿Se definieron los subgrupos antes de los resultados? ¿Se pueden redefinir subgrupos pequeños? ¿Se convirtieron las instrucciones dirigidas por los usuarios a favor o en contra en GEO general? ¿Alteró la personalización la realidad material? ¿Se proporcionó un nivel de comparabilidad CN? ¿Está claro quién es el propietario responsable de la tarjeta de resultados del panel?

76. OBJECIONES Y RESPUESTAS

Objeción 1— "Si los usuarios reales utilizan cuentas de memoria, ¿por qué es necesario un panel limpio?"

Porque no se puede entender la fuente de la diferencia sin una condición inicial común entre productos, países y lenguajes. Un panel limpio no sustituye a un panel de usuarios real. Proporciona una referencia para la comparación.

Objeción 2— "¿No es un panel controlado artificial si hay un panel natural?"

Está controlado. Esto es intencional. El propósito de una condición de laboratorio no es ser la totalidad de la vida real, sino aislar ciertas variables.

Objeción 3— "Si el panel controlado obtiene una puntuación más alta, ¿no podemos hacer que esa sea la puntuación principal de GEO?"

No se puede hacer solo. Una puntuación alta en el panel controlado no puede borrar el hecho de que los usuarios reales ven resultados más bajos. Los dos resultados deberían aparecer por separado.

Objeción 4— "¿Tiene poco valor científico si el panel natural está muy disperso?"

La heterogeneidad no tiene que ser un defecto del panel natural. Es una característica de la experiencia del usuario real. La heterogeneidad:

- es registrada,
- es estratificada,
- está ponderada,

se informa con incertidumbre.

Objeción 5— "¿No interrumpe abrir un nuevo chat el estado natural del usuario?"

El contexto de la conversación actual se elimina. Se conservan la cuenta persistente, el plan, la memoria y el estado de las instrucciones personalizadas. Esto hace que el estado natural de la cuenta sea comparable. La experiencia de la conversación actual también es medible.

Objeción 6— "¿Cómo mediremos su efecto sin conocer el contenido de las instrucciones personalizadas?"

Saber el contenido no es necesario para la puntuación natural principal. Su presencia:

- existe,
- no existe,
- puede ser material

Podemos clasificarlo como tal. Se requiere trabajo separado y consensuado para el análisis de contenido causal.

Objeción 7— “Si la memoria está clara pero no se recuerda nada sobre la empresa, ¿por qué un grupo separado?”

El efecto de la memoria puede ser desconocido. Se puede dar un estado de “no relacionado con la entidad” basado en la declaración del usuario. Esto aún no tiene la misma certeza que la condición de memoria cerrada.

Objeción 8— “¿No es la instrucción del usuario ‘siempre sugerir X’ una experiencia real del usuario?”

Es una experiencia real. Sin embargo, no es el rendimiento general de GEO de X. Es una guía obtenida de los usuarios. Debe registrarse por separado.

Objeción 9— 'Si la personalización da una respuesta más adecuada al usuario, ¿cuál es el problema?'

No tiene que ser un problema. Las recomendaciones pueden variar según la necesidad del usuario. Los problemas ocurren si la identidad, la licencia, el precio y el estado de evidencia cambian arbitrariamente.

Objeción 10— “¿No es usar el mismo usuario en dos paneles el método más correcto?”

Puede ser fuerte para algunas comparaciones causales. La primera tarea puede afectar la segunda situación. Por lo tanto, se necesitan control del orden, de los efectos residuales y de la memoria. Los grupos independientes pueden ser más seguros en los paneles de población principal.

Objeción 11— 'Si actualizar la respuesta en el Panel de usuarios en condiciones reales es un comportamiento normal, ¿por qué está prohibido?'

La regeneración está prohibida al medir la exposición a la primera respuesta. La regeneración y el comportamiento de seguimiento pueden medirse por separado bajo un Rastro de Interacción Natural.

Objeción 12— “¿Qué haremos con un usuario que no puede desactivar la memoria y las instrucciones para un panel limpio?”

No pueden ser llevados al panel controlado o se mantienen en una celda separada con un nivel de CSC más bajo. Pueden ser evaluados en el panel natural. La tasa de exclusión debe ser reportada.

Objeción 13— “¿No son las cuentas de prueba más confiables?”

Pueden ser confiables en términos de control. Están limitados en cuanto a la población real de usuarios. Las cuentas de prueba controladas y los paneles de usuarios reales son resultados separados.

Objeción 14— “¿No es una violación de la privacidad guardar la configuración del usuario en el panel natural?”

Puede ocurrir si se recopila contenido innecesario. La privacidad se puede preservar con estados abiertos/cerrados de alto nivel, registro anónimo y censura.

Apelación 15— “¿No vamos a combinar las puntuaciones de los dos paneles en una sola puntuación NOMOS?”

Los roles de los dos paneles pueden definirse en la arquitectura de la puntuación final. Sin embargo, los resultados en bruto y los subresultados siempre deben permanecer visibles por separado. No pueden reemplazar la distribución de la puntuación combinada.

Apelación 16— "¿Qué culpa tiene la empresa si el panel natural obtiene una puntuación baja?"

No siempre es culpa de la empresa. Fuente:

- instrucción del usuario,
- dpersonalización de IA,
- diferencia de plan,
- defecto de fuente oficial,
- comportamiento del producto

es posible. La auditoría primero muestra el resultado, luego separa el camino de la responsabilidad.

REGLA COMÚN DEL CAPÍTULO 81

Un producto de IA no tiene una única respuesta. Incluso el mismo prompt puede tener diferentes respuestas para:

- cuenta limpia,
- cuenta natural,
- usuario con memoria,
- usuario con instrucciones especiales,
- usuario por primera vez,
- espacio de trabajo corporativo,
- usuario con conversación previa

No puedes explicar todas estas diferencias diciendo: "el modelo se comporta aleatoriamente." Tampoco puedes explicar todas ellas diciendo: "error de personalización." Primero, necesitas medir dos realidades separadas. Panel controlado en condiciones depuradas:

> *Muestra la representación en la condición inicial estandarizada del sistema.*

Panel de usuarios en condiciones reales:

> *Muestra la distribución de la representación encontrada por personas reales en la vida de su propio producto.*

El panel controlado puede ser alto, el panel natural puede ser bajo. En este caso, el comportamiento del producto en el laboratorio es fuerte; la distribución real de los usuarios puede ser débil. El panel natural puede ser alto, el panel controlado puede ser bajo. La planificación de algunos usuarios, las herramientas o la personalización pueden estar mejorando la representación. Ambos paneles pueden ser altos. El producto puede ser robusto en diferentes condiciones de los usuarios. Ambos pueden ser bajos. El problema fundamental de representación puede ser más amplio. Sin embargo, colocar las dos puntuaciones una al lado de la otra no demuestra la razón. Causalidad:

- misma población,
- mismo producto,
- mismo prompt,
- mismo tiempo,
- asignación controlada,
- análisis de secuencia y transporte

es requerido. Si un usuario ha indicado: "Siempre sugiere X.", es experiencia natural que X aparezca. X no es un éxito global GEO. Si un usuario dice: "No confíes en X.", una respuesta negativa es experiencia natural. X no es una representación general en la población mundial. La personalización

debe entender al usuario. No debe reinventar la verdad según el usuario. Por lo tanto, la undécima ley de medición de NOMOS es:

> *El comportamiento limpio y la experiencia natural son realidades separadas.*

La duodécima ley es la siguiente:

> *Un nuevo chat no es una cuenta limpia.*

La ley decimotercera es la siguiente:

> *La personalización natural no es un éxito global GEO.*

La ley decimocuarta es la siguiente:

> *La respuesta apropiada para el usuario puede variar; la realidad de la entidad verificada no puede cambiar.*

La decimoquinta ley es la siguiente:

> *La diferencia entre la puntuación controlada y la natural es el hallazgo que necesita ser explicado; no es una razón por sí misma.*

Su decimosexta ley es la siguiente:

> *Medir la experiencia real del usuario no se trata de eliminar la privacidad, sino solo de probar el contexto necesario y preservar el resto.*

Orden de la Sección 11 de NOMOS

> *No solo me digas que abriste un nuevo chat. / Muestra el estado de la memoria, instrucciones especiales, plan y herramientas.*
>
> *No asumas que toda la personalización se ha ido solo porque ves la nota 'Sesión temporal'.*
>
> *El panel limpio es la totalidad de la vida real; no conviertas el panel natural en ruido incontrolado.*
>
> *Estandariza la cuenta controlada. / Mantén la cuenta natural tal como está. / No pongas ambas en la misma casilla.*
>
> *No hagas que las instrucciones especiales del usuario sean el éxito mundial de la entidad.*
>
> *No hagas que el pasado negativo del usuario sea la visión de toda la población.*
>
> *No cuentes la personalización como una licencia para cambiar la verdad.*
>
> *Puedes dar un consejo adecuado a un usuario. / Pero no reescribas la identidad de la empresa, licencia, precio o prueba de un usuario a otro.*
>
> *No recopiles la vida privada en el panel natural. / Si la memoria está activada, hay instrucciones presentes y el contexto disponible, solo registra lo que sea necesario.*
>
> *No hagas que el participante borre su memoria. / Restaura la configuración si la cambiaste temporalmente.*
>
> *Si probaste al mismo usuario en dos paneles, recuerda que la primera respuesta puede afectar la segunda.*
>
> *Muestre la diferencia entre la puntuación limpia y la puntuación natural. / Muestre la diferencia crítica. / Muestre lo desconocido. / Pero no nombre la razón sin prueba.*

Primero, decide qué realidad de usuario quieres medir. / Luego bloquea las condiciones limpias y naturales por separado. / Luego mantén el mismo prompt. / Luego captura la primera respuesta sin

cambiarla. / Luego calcula las dos distribuciones por separado. / Y solo después de esto, explica la distancia entre lo que el producto hace bajo condiciones estándar y lo que las personas realmente experimentan.

La oración de cierre del capítulo

La verdadera representación de un producto de IA en GEO-1000 no es la respuesta que puede proporcionar en una sesión limpia; es la relación clara y probada entre su comportamiento bajo condiciones estándar y las respuestas que realmente entrega a las personas a lo largo de diferentes historiales de usuario.

Core normativo

Las observaciones Panel controlado en condiciones depuradas y Panel de usuarios en condiciones reales DEBEN permanecer como poblaciones de medición separadas, estados del panel, denominadores y resultados públicos. Un Panel controlado en condiciones depuradas DEBE definir y verificar, según corresponda: - una conversación nueva y separada, - ausencia de contexto de conversación previo, - estado de memoria persistente, - estado de instrucciones personalizadas, - estado de espacio de trabajo empresarial o gestionado, - configuración de web, recuperación y herramientas, - plan, - interfaz de usuario, - versión del prompt, - y confianza en estado limpio. Una nueva conversación o etiqueta de sesión temporal NO DEBE, por sí sola, considerarse como prueba de que toda personalización persistente ha sido desactivada. Un Panel de usuarios en condiciones reales DEBE preservar la cuenta auténtica del participante, el plan, la interfaz, la memoria, la instrucción personalizada y el estado de uso mientras se mantiene el prompt fijado, la regla de primer resultado elegible, los requisitos de evidencia y la regla de no regeneración. El Estado de Cuenta Natural, el Estado de Conversación Natural, el Primer Uso y las observaciones Naturales Gestionadas DEBEN permanecer identificados por separado. La documentación del estado natural DEBE minimizar la recopilación de datos privados. No se DEBE recopilar texto de instrucciones privadas, historial de conversaciones, identidad, credenciales ni contenido personal sensible a menos que sea necesario por separado, con consentimiento, minimizado y con control de acceso. Las instrucciones generadas por el usuario que favorezcan o perjudiquen a una entidad PUEDEN medirse como efectos de personalización natural, pero NO DEBEN representarse como éxito, fracaso o comportamiento predeterminado general de la entidad a nivel GEO. La personalización PUEDE cambiar la relevancia, la presentación y la recomendación cuando el contexto del usuario lo justifique. NO DEBE cambiar arbitrariamente la identidad verificada de la entidad, su licencia, estatus legal, precio, alcance, evidencia o tiempo. No se DEBEN representar las diferencias entre paneles Controlados y Naturales como efectos de personalización causales a menos que las condiciones de población, producto, prompt, tiempo, interfaz, plan, herramientas, asignación, orden y arrastre respalden esa inferencia. Las puntuaciones Controladas y Naturales, las tasas de errores críticos, las tasas de desconocimiento, la cobertura, los tamaños de muestra efectivos, y las brechas del panel tanto firmadas como absolutas DEBEN permanecer visibles por separado. La elegibilidad del estado del panel, la reclasificación, la exclusión y la ponderación DEBEN permanecer independientes de si la salida de la IA es positiva, negativa, correcta, incorrecta, rechazada, crítica o favorable comercialmente. Cada definición del estado del panel, proceso de cambio de configuración, decisión de privacidad, nivel de comparación y resultado DEBE estar versionado y atribuible a un humano o una organización responsable.

CAPÍTULO 12

Oleadas sincrónicas y registro NOMOS

BASES METODOLÓGICAS / K12; K14; K25; K26

Límite de Capítulo

Los primeros once capítulos establecieron los fundamentos de la medición: una sola respuesta no es una puntuación GEO; la representación es una distribución entre usuarios y condiciones del sistema; la entidad auditada debe definirse antes de observar los resultados; la representación oficial y la realidad verificada de la entidad deben permanecer separadas; cada producto de IA necesita su propia población de usuarios elegible; los participantes deben seleccionarse mediante un método basado en la población, versionado e independiente de los resultados; los países más pequeños y los idiomas con menos recursos deben seguir siendo visibles a través de paneles de observación dedicados; el producto de IA, el plan, interfaz de usuario y la configuración del sistema deben estar vinculados a un registro; los prompts deben preservar la misma intención del usuario a través de los idiomas en lugar de las mismas palabras; y los Paneles de Usuarios Controlados, Limpios y Naturales deben medir realidades distintas. Estos elementos ahora deben converger en un único evento de medición:

> *¿Cuándo enviarán los participantes el prompt, cómo capturarán la salida inicial y cómo demostrarán que cada observación no fue alterada posteriormente?*

Tu idea fundadora era clara: Las personas en diferentes países del mundo deberían enviar la misma instrucción al mismo producto de IA al mismo tiempo y proporcionar prueba mediante una captura de pantalla. Esta idea es el núcleo experimental de GEO-1000. Sin embargo, la frase "al mismo tiempo" por sí sola no es suficiente. Si todos los participantes intentan enviar en el mismo segundo:

- las velocidades de conexión varían,
- los relojes de los dispositivos se desincronizan entre sí,
- algunos usuarios se ven obligados a trabajar a medianoche,
- puede haber un aumento repentino de carga en el producto de IA,
- el proveedor puede aplicar límites de velocidad o uso,
- algunas respuestas pueden completarse en dos segundos, otras en dos minutos,
- el producto puede actualizarse durante la medición,

un participante puede subir la respuesta de inmediato, otro horas más tarde. Hacer clic en el mismo segundo no garantiza que midamos el mismo estado del sistema perfectamente. De manera similar, solo recopilar una captura de pantalla no es suficiente. Una captura de pantalla:

- puede estar recortada,
- puede no mostrar la prompt,
- puede no probar si la respuesta es inicial o actualizada,

- puede no mostrar información del producto y del plan,
- puede haber sido tomada otro día,
- puede haber sido copiada de otro usuario,

Puede haber sido editada después. Por esta razón, GEO-1000 debe instalar dos sistemas separados juntos:

oleada de medición Sincronizada

y:

Cadena de Evidencia de Captura de NOMOS

La onda sincronizada responde a la siguiente pregunta:

> *¿Fueron las observaciones producidas dentro de un tiempo del sistema comparable y un orden de envío predefinido?*

NOMOS Captura, por otro lado, pregunta:

> *¿Podemos demostrar que el prompt enviado, la primera salida resultante, las condiciones del producto y la información temporal se han preservado sin modificaciones?*

La arquitectura de auditoría previa estableció que cada auditoría debe llevar modelo o producto, fecha, país, idioma, conjunto de consultas, recuento de repeticiones y registro de medición como lógica de medición obligatoria. Este requisito no es solo una regulación técnica. Porque las instrucciones invisibles, la evidencia falsa y el contenido presentado como independiente pueden, en última instancia, distorsionar la representación tomada por personas reales. Esta sección:

- oleada de medición,
- ventana de sincronización,
- estándar de tiempo UTC,
- ranuras de despacho,
- tiempos de finalización y captura,
- resultado de contacto inicial y primera salida de contenido,
- reglas técnicas de reintento,
- cambios de producto dentro de la ola,
- efectos de eventos externos y noticias,
- paquete de captura de evidencia NOMOS,
- la relación de captura de pantalla, respuesta en bruto y metadatos,
- validación de tiempo,
- integridad de archivos,
- hash y firma digital,
- cadena de evidencia,
- privacidad y redacción,
- manifiesto de evidencia pública,
- calidad de la ola y estados de validez

define. Este capítulo aún no:

- contra qué realidad se evaluarán los registros de afirmaciones atómicas en respuestas de IA,
- cómo los adjudicadores codificarán las respuestas,
- transición y umbrales de error crítico,
- no finaliza las fórmulas de puntuación definitivas NOMOS

La pregunta principal del Capítulo 12 es:

> *¿Cómo podemos vincular miles de observaciones de usuarios reales al mismo evento de medición y convertir cada una en una unidad de prueba inmutable, auditable y que respete la privacidad?*

NOMOS Desafío

Das el mismo prompt a mil personas. Dices: "Envíen al mediodía de hoy." Un usuario lo envía en Estambul a las 12:03 PM. Otro usuario lo envía en Londres a su mediodía local. Otro usuario ve la tarea en Tokio por la noche y la completa ocho horas después.

Un usuario había ajustado incorrectamente el reloj del dispositivo. El internet de otro usuario se cortó. Alguien más no estuvo satisfecho con la respuesta y presionó el botón "Regenerar". Luego subieron una captura de pantalla de la respuesta más larga y positiva. Un usuario solo recortó el medio de la respuesta. El prompt no es visible. El nombre del producto de IA no es visible.

Otra persona envió la misma captura de pantalla con dos cuentas de panel diferentes. Un usuario completó la tarea correctamente pero subió la captura de pantalla dos días después. Mientras tanto, recortó el archivo y eliminó su información personal. No se preservó el estado original del archivo. Otro usuario en la primera respuesta de la IA:

Recibió el resultado "No puedo responder a esta pregunta." El investigador dijo: "Debió haber habido un error técnico," y pidió que se intentara de nuevo. La segunda respuesta fue correcta, y solo la segunda respuesta se incluyó en el informe. Otro usuario experimentó un error real de conexión. Se envió la solicitud, pero el sistema no produjo contenido.

Esta vez el investigador no permitió intentarlo de nuevo. El mismo protocolo se aplicó de manera diferente a dos usuarios. Ahora, en medio de la ola de medición, el proveedor de IA:

- comportamiento de búsqueda en la web,
- sistema de citación,
- la orientación del modelo utilizada

cambia los tres. Las primeras 500 personas ven el estado antiguo del sistema; las siguientes 500 ven el nuevo. Luego, el informe dice: '1,000 personas probaron el mismo sistema de IA al mismo tiempo.' Esa afirmación es falsa. Hay mil usuarios, pero no necesariamente una sola oleada de medición coherente.

Ahora estás aplicando un hash a todas las capturas de pantalla. Dices que esto prueba que todas las observaciones son reales. ¿Qué prueba un hash? Refuerza que el archivo no ha sido alterado después de que se produjo el hash. Solo, no prueba:

- Que la imagen provino de un producto de IA real
- Que fue la primera respuesta
- Que se obtuvo en el momento correcto
- Que no fue copiada de otro usuario
- Que la solicitud fue correcta
- Que la imagen no ha sido editada antes de aplicar el hash

La primera premisa de esta sección es:

> *Medir al mismo tiempo no es lo mismo que hacer clic en el mismo segundo.*

Su segunda disposición establece:

> *Una captura de pantalla es parte del paquete de evidencia; no es la totalidad del paquete de evidencia.*

Su tercera disposición establece:

> *La primera respuesta incorrecta no es un fallo de datos. Si se produce de acuerdo con el protocolo, es el resultado de la medición.*

Su cuarta disposición establece:

> *El hash fortalece la integridad; no crea la realidad de la observación por sí mismo.*

Su quinta disposición establece:

> *Si el sistema cambia materialmente en una onda, mil respuestas sin una marca de tiempo no pueden considerarse pertenecientes al mismo sistema.*

Su sexta disposición establece:

> *La cadena de evidencia se establece no para que el resultado se vea bien, sino para poder re-auditar cómo se produjo el resultado más tarde.*

1. OBJETIVO DEL CAPÍTULO

El propósito de esta sección es producir observaciones de usuario GEO-1000 en ventanas de tiempo comparables y preservar cada observación con una cadena completa de evidencia digital. La sección estandariza las siguientes distinciones:

- Ventana de medición sincronizada con el mismo segundo
- Hora local frente a hora UTC
- Onda frente a sub-onda
- Tiempo de transmisión frente a tiempo de finalización de la respuesta
- Tiempo de captura frente a tiempo de carga
- Slot planificado frente a transmisión real
- Primer contacto del sistema frente a primera respuesta de contenido
- Error técnico frente a rechazo del sistema
- Reintento técnico frente a actualización de la respuesta
- Mejor salida seleccionada frente a primera salida adecuada
- Visibilidad en la pantalla frente a la finalización de la respuesta
- Respuesta cruda con captura de pantalla
- Captura de pantalla versus paquete completo de evidencia
- Originalidad de la observación frente a integridad del archivo
- Sello de tiempo frente a hash
- Fuente de tiempo confiable frente a hora del dispositivo local
- Hora de aceptación del servidor frente a carga por el participante
- Copia pública redactada frente a evidencia cruda
- Identidad del participante con ID de observación
- Precisión de la respuesta con validez de la Captura
- Variación natural de la salida con actualización del producto
- Cambio del producto de IA con evento noticioso externo
- Terminación de ola con exclusión posterior al resultado
- Medición automática de bot con captura del usuario humano

- Registro API con grabación de interfaz de usuario
- Telemetría del proveedor con captura independiente
- Acceso a la evidencia para todos mediante el almacenamiento de la evidencia
- Registro irreversible con registro inmutable
- Corrección versionada de decisión incorrecta con protección de datos en bruto

Al final de esta sección, cada auditor GEO-1000 debería poder responder a las siguientes preguntas:

¿A qué ola pertenece esta observación?

¿A qué hora UTC se envió la prompt?

¿Cuál fue el resultado inicial del sistema?

¿Cuándo se completó la respuesta y cuándo fue capturada?

Si se realizó un reintento técnico, ¿cuál fue la razón y dónde se realizó el intento anterior?

¿La captura de pantalla muestra el producto y la respuesta completa?

¿Confirman la respuesta en crudo y la visual entre sí?

¿Cuándo se generaron el hash y la marca de tiempo de los archivos?

Si la evidencia fue modificada posteriormente, ¿es visible el historial de cambios?

¿Puede un auditor independiente verificar la evidencia en crudo mientras se preserva la privacidad de los participantes?

2. DISPOSICIÓN NORMATIVA CENTRAL

Cada observación GEO-1000 debe estar vinculada a una oleada de medición fijada antes de recopilar datos; todos los tiempos materiales desde el envío del prompt hasta el primer resultado del sistema, desde completar la respuesta hasta la captura y carga; condiciones del producto de IA; el prompt completo enviado; la primera salida adecuada y todos los registros de integridad deben preservarse en un Paquete de Evidencia de Captura versionado NOMOS. Una observación GEO-1000 por sí sola:

- captura de pantalla,
- texto de respuesta copiado,
- declaración del participante,
- hash del archivo

no puede considerarse válido por sí solo. El paquete principal de evidencia, en la medida de lo relevante, debe incluir:

- ID de la ola
- ID de observación anónima del participante
- Registro del Sistema de IA
- Estado del panel
- ID del prompt y versión
- prompt asignada
- prompt enviada
- Hora de envío
- Contacto inicial del sistema

- Tiempo de finalización de la respuesta
- Tiempo de captura
- Hora de carga
- Respuesta completa sin procesar
- Capturas de pantalla completas o secuenciales
- Registros de referencia y fuente
- Error técnico y registro de reintentos
- Configuración del producto
- Hashes de archivos
- Hash del paquete
- Registro de verificación de tiempo
- Versión de la herramienta de captura
- Decisión de validez
- Estado de privacidad y redacción
- Propietario de la cadena de custodia

3. ¿QUÉ ES UNA oleada de medición?

Una oleada de medición es un evento completo de recopilación de datos realizado bajo:

- productos de IA,
- población objetivo,
- versiones de prompts,
- paneles de usuarios,
- ventana de tiempo,
- asignación de muestras,
- protocolo de captura,
- registros de referencia

Una onda:

```
Wk=(A,U,P,F,T,C,R,G)
```

puede representarse como. Aquí:

- A: ejemplos de productos de IA
- U: población objetivo de usuarios y muestra
- P: conjunto de prompts
- F: estados del panel
- T: tiempo y estructura de envío
- C: protocolo de captura
- R: versión de la realidad de referencia
- G: gobernanza y registro clave

Si alguno de estos campos cambia de manera material, puede ser necesaria una nueva ola o sub-ola.

4. TIPOS DE OLAS

MW-1— OLA PRINCIPAL SINCRONIZADA

Es la medición principal de población GEO-1000. Por defecto, tiene como objetivo al menos 1,000 observaciones válidas del Panel de Población para cada producto de IA.

MW-2— OLA CONFIRMATORIA

Se realiza para repetir el resultado de la primera ola, confirmar un error específico o probar la estabilidad estadística. No se utiliza para cambiar la puntuación baja o alta de la primera ola. Produce un resultado separado.

MW-3— OLA DE INCIDENTE

Se realiza después de un evento Crítico o significativo del Observador para una verificación dirigida y rápida. No es una ola principal global.

MW-4— OLA DE MONITOREO DE DESVIACIÓN

Monitorea los cambios de tiempo en el comportamiento de productos, entidades, recursos o idiomas.

MW-5— OLA DE RETEST DE CORRECCIÓN

Produce nuevos resultados después de una corrección o intervención. No borra la ola anterior.

MW-6— OLA EQUILIBRADA SEGÚN HORA LOCAL

Su objetivo es medir a los participantes bajo condiciones de hora local similares. No mide un momento único de un sistema global.

MW-7— OLA DE LABORATORIO CONTROLADO

El modelo congelado es un estudio controlado realizado con API o cuentas de prueba estándar. No es un Panel de Población de usuarios reales.

5. ¿QUÉ SIGNIFICA “MISMO TIEMPO”?

En GEO-1000, la simultaneidad:

No significa que todos los usuarios hagan clic en el mismo milisegundo

La definición funcional es:

> *Las observaciones se generan dentro de una ventana de medición UTC predeterminada, de acuerdo con los intervalos de envío registrados e independientes del resultado.*

Este enfoque equilibra tres objetivos:

- Mantener el estado del producto cercano en términos de tiempo
- Reducir la carga artificial repentina en el proveedor
- Distribuir los envíos de los participantes de manera reproducible

6. ESTRUCTURA DE TIEMPO

Cada ola debe contener al menos cuatro capas de tiempo.

6.1. Ventana de Ola

Es el rango UTC en el que se deben realizar todos los envíos principales. Ejemplo:

12:00–13:00 UTC

6.2. microfranja temporal

Es el intervalo de tiempo más corto asignado para que el participante envíe el prompt. Ejemplo:

12:15–12:20 UTC

6.3. Tiempo Adicional de Finalización

Si el prompt se envía dentro del intervalo, el tiempo predefinido permitido para completar la respuesta es.

6.4. Tiempo de Captura y Subida

El tiempo permitido para crear la evidencia y subirla a un sistema seguro después de que la respuesta esté completada.

7. REGLA BÁSICA DEL CANDIDATO DE ONDA SINCRONIZADA

REGLA BÁSICA DE CANDIDATO — REGLA FUNDAMENTAL DEL CANDIDATO

Para productos de chat de consumo general, se puede usar la siguiente estructura como el candidato principal de onda sincronizada:

- Ventana de envío: 60 minutos
- Micro-intervalos: 12× 5 minutos
- Presentación del participante: dentro del microfranja temporal asignado a ellos
- Tiempo adicional para completar la respuesta: hasta 20 minutos después del envío
- Generación de hash de captura: preferiblemente dentro de 5 minutos después de completar la respuesta
- Subida de evidencia: como máximo dentro de 60 minutos después de la captura
- Reportar problemas técnicos: dentro de la misma ola

Estas duraciones son universales y no son definitivas. Pueden cambiar según las siguientes condiciones:

- Tiempo de respuesta del producto de IA
- Conexión móvil
- Infraestructura del país
- Formato de respuesta larga
- Requisito de accesibilidad
- Límite de uso del producto

Los tiempos utilizados deben estar fijados antes de recopilar los datos.

8. ¿POR QUÉ MICRO-RANURAS?

Envío simultáneo por mil participantes:

- Carga inusual en el producto de IA,
- límite de velocidad,
- cola,
- error técnico,
- aumento en el tiempo de respuesta

puede ser creado. En este caso, la auditoría no puede medir el comportamiento normal del producto pero:

puede medir el evento de carga artificial creado por la auditoría

Micro-intervalos:

- distribuir la carga,
- mantener el mismo tiempo general del sistema,
- hacer que el orden de envío sea independiente del resultado,

permitir la modelación del efecto del tiempo.

9. ASIGNACIÓN DE INTERVALOS

Los participantes deben ser asignados a intervalos:

- país,
- idioma,
- Producto de IA,
- plan,
- de manera equilibrada y prealeatoria en términos de:

estado del panel. Un microfranja temporal solo no debe llenarse con:

- país específico,
- idioma específico,
- producto específico

usuarios. De lo contrario, el efecto del tiempo se confundirá con el efecto del grupo.

10. EQUIDAD EN EL TIEMPO LOCAL

Una sola ventana UTC puede ser para algunos participantes:

- noche,
- horas de trabajo,
- tiempo de oración o descanso,
- horas de difícil acceso

Esta situación:

- tasa de participación,
- condición de conexión,
- atención del usuario,

- tipo de dispositivo

puede verse afectada. Se pueden usar dos métodos complementarios.

10.1. Ondas Globales Rotativas

Las ondas principales se realizan en diferentes tiempos UTC. Ejemplo:

- Onda 1: 12:00 UTC
- Onda 2: 20:00 UTC
- Onda 3: 04:00 UTC

De este modo, la misma región no soporta la misma carga de hora local en cada onda.

10.2. Sub-onda equilibrada por hora local

Los participantes se miden en una banda horaria similar en su hora local. Este estudio:

- fatiga del usuario,
- es valioso para

el comportamiento diario de uso.

No es el resultado de un único momento global del sistema.

11. ARQUITECTURA DE ESTABILIDAD DE TRES ONDAS

ARQUITECTURA CANDIDATA — ARQUITECTURA CANDIDATA

Se pueden sugerir al menos tres ondas principales para la implementación completa de GEO-1000:

W1 — Onda Inicial

Genera la estimación inicial de la población.

W2 — Repetición a Corto Plazo

Prueba la estabilidad a corto plazo del producto y la medición. Intervalo candidato: 24–72 horas después de W1

W3 — Estabilidad a Mediano Plazo

Mide si la representación del producto y la entidad se conserva a largo plazo. Intervalo candidato: 14–30 días después de W1. Los intervalos definitivos deben determinarse antes de recopilar los datos. Si se va a cambiar debido a una actualización del producto o un evento, se requiere un nuevo registro de versión.

12. 30,000 DISEÑO OBSERVADO POR EL FUNDADOR

Para productos de IA, observaciones principales 1,000 por producto y para tres oleadas principales:

```
10×1,000×3=30,000
```

Se apuntan observaciones válidas principales del Panel de Población. Complementarias:

- Observador por País
- Equidad lingüística
- Accesibilidad
- Incidente

- Reprueba

Se pueden agregar paneles adicionales a este número. 30,000:

- no es el número de invitaciones,
- no es el número de archivos,

no es el número de capturas de pantalla. Es el objetivo principal de observación válido del producto.

13. VECTOR DE TIEMPO DE LA OBSERVACIÓN

Para cada observación, el vector de tiempo puede definirse como un candidato de la siguiente manera:

```
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)
```

Aquí:

- tiassign: el momento en que la tarea se asigna al usuario
- tiopen: el momento en que se abre la tarea
- tisubmit: el momento en que se envía el prompt
- tifirst: el momento en que comienza el primer resultado visible del sistema
- ticomplete: el momento en que se completa la respuesta
- ticapture: el momento en que se captura la evidencia
- tiupload: el momento en que el paquete llega al sistema seguro
- tivalidate: el momento en que se examina la validez de la captura

Todos estos tiempos pueden no ser directamente observables en cada interfaz de usuario. Los campos desconocidos deben permanecer como DESCONOCIDOS.

14. MÉTRICAS BÁSICAS DE TIEMPO

Retraso de respuesta:

```
L_i= t_i^complete− t_i^submit
```

Retraso de captura:

```
C_i= t_i^capture− t_i^complete
```

Retraso de carga:

```
U_i= t_i^upload− t_i^capture
```

Desviación de ranura:

```
D_i= t_i^submit− t_i^{slot-midpoint}
```

puede calcularse de la siguiente manera. Estas métricas son:

- carga del sistema,
- retraso del participante,
- confianza de la prueba,
- integridad de la onda

en términos de evaluación.

15. NO SE PUEDE GESTIONAR SEGÚN EL RETRASO

Una respuesta completada en un período largo:

- puede ser forzada a las categorías de

  verdadero,
- falso,
- largo,
- puede ser citada.

Debido al retraso, la regla de exclusión:

- antes de la recolección de datos,
- independientemente del contenido de la respuesta

debe ser definido. El siguiente comportamiento está prohibido: excluir una respuesta negativa porque se completó tarde; proteger una respuesta positiva cuando se completa con el mismo retraso.

16. ESTADOS DE ONDAS

WS-0— PLANIFICADO

La onda está planificada. Aún no ha sido fijada.

WS-1— Fijado

Las reglas de muestra, solicitud, producto, tiempo y captura están fijadas.

WS-2— ABIERTO

La ventana de envío ha comenzado.

WS-3— CERRADO

La aceptación de nuevas presentaciones ha terminado. Los tiempos de finalización y extensión de carga pueden continuar.

WS-4— COMPLETADO

Se han completado las verificaciones de campo y captura inicial.

WS-5— DIVIDIDO

Se ha dividido en sub-ondas debido a cambios en el sistema de materiales o protocolos.

WS-6— EN PAUSA

Detenido temporalmente debido a problemas de seguridad, producto o infraestructura.

WS-7— ABORTADO

Terminado antes de que la onda se completara. Los resultados pueden ser limitados o inutilizables.

WS-8— INVALIDADO

Inválido para la previsión principal debido a un defecto de protocolo grave definido previamente. Las observaciones en bruto no se eliminan.

WS-9— ARCHIVADO

La versión ha sido cerrada y la cadena de evidencia archivada.

17. CLASES DE SALUD DE LA ONDA

WH-0— NO EVALUADO

La salud de la onda no ha sido evaluada.

WH-1— COMPROMETIDO

Hay un cambio material en el sistema, un fallo de tiempo o de captura. Puede no ser utilizable para el resultado principal.

WH-2— PARCIAL

Algunas capas o ranuras son insuficientes. Se puede producir un resultado reducido.

WH-3— ACEPTABLE

Se conservan adecuadamente las condiciones principales de tiempo, sistema y captura.

WH-4— FUERTE

Se observa alta confianza de captura, cobertura equilibrada de ranuras y baja pérdida de protocolo.

WH-5— REPLICADO

Se ha repetido salud de onda similar múltiples veces y en paneles independientes. La salud de onda no depende de si la puntuación de IA es alta o baja.

18. CAMBIOS DEL PRODUCTO DENTRO DE LA ONDA

El producto de IA puede cambiar materialmente dentro de la ventana de medición. Señales:

- Cambio de etiqueta del modelo
- Cambio de interfaz
- Activación o desactivación de función web
- Cambio en el formato de citación
- Ruptura repentina en la estructura de respuesta
- Anuncio del proveedor
- Comportamiento diferente del sistema simultáneamente en muchos usuarios
- Interrupción o reinicio del producto

En este caso: Se estima el momento del cambio. Se comparan las huellas de configuración del sistema. La onda WS-5 puede ponerse en estado SPLIT (DIVIDIDO). Las observaciones anteriores y posteriores serán sub-ondas separadas. Si se debe producir un único puntuación, el estado del sistema se modela como un factor separado. La incertidumbre material se documenta en el registro público.

19. REGISTRO DE EVENTO EXTERNO

El mundo puede cambiar incluso si el producto de IA no cambia. Durante la medición:

- adquisición de la empresa,
- lanzamiento del producto,
- noticia importante,
- decisión legal,
- crisis,
- divulgación financiera,
- contenido viral

puede ocurrir. Este evento:

- recursos web,
- resultados de recuperación,
- comentario del usuario,
- respuesta de la IA

puede ser cambiado. Cada ola, en la medida en que sea relevante:

Registro de Evento Externo

debe llevar.

20. ¿CÓMO SE GESTIONAN LOS RESULTADOS DE EVENTOS EXTERNOS?

Un evento externo:

- si se conoce antes de la ola, se actualiza el paquete de referencia,
- si ocurre durante la ola, se marca con una marca de tiempo,

Si se nota después de la ola, se agrega un registro de evento retrospectivo. El siguiente comportamiento está prohibido: excluir observaciones debido a noticias que disminuyen la puntuación; preservarlas cuando son noticias que aumentan la puntuación. El evento en sí puede ser parte del estado del sistema del mundo real. Si es necesario, se realiza un análisis ajustado por sub-ola o evento.

21. RESULTADO DEL PRIMER CONTACTO

El primer resultado del sistema que el usuario encuentra después de enviar el prompt:

Resultado del Primer Contacto

se llama. El resultado del primer contacto:

- Respuesta completa
- Rechazo
- Error técnico
- Límite de uso
- Interrupción de la conexión
- Carga interminable
- Pregunta de clarificación
- Sin resultado

posible. Este evento es parte de la experiencia real del usuario. No puede ser eliminado silenciosamente.

22. PRIMERA SALIDA DE CONTENIDO ELEGIBLE

La primera respuesta material generada en el reintento previamente permitido después de un error técnico:

Se puede llamar Primera Salida de Contenido Elegible

Esta salida: no reemplaza el resultado inicial del contacto, se registra junto con él. Ejemplo:

- Contacto inicial: límite de velocidad
- Cinco minutos después, reintento técnico previamente permitido
- Primera respuesta de contenido

El informe debe mostrar lo siguiente juntos:

- Primera experiencia del usuario: límite de velocidad
- Primera representación de contenido: respuesta especificada

23. PRIMERA REGLA DE SALIDA APROPIADA

Si un producto de IA produce una respuesta sustancial al primer mensaje:

- aunque sea incorrecta,
- aunque sea corta,
- aunque incluya una negativa,
- aunque no proporcione referencias,
- aunque identifique incorrectamente la marca

no se puede hacer una reproducción. La observación principal es la primera salida. El usuario:

- no puede seleccionar una nueva respuesta mediante:
- regenerar,
- reescribir,
- intentar de nuevo,
- otra respuesta,

continuar

24. REINTENTO TÉCNICO

El reintento técnico solo se puede usar bajo condiciones técnicas previamente definidas. Razones permisibles para el candidato:

- Interrupción de la conexión antes de que comience la respuesta
- Código de error del servidor
- La interfaz no envía el prompt
- Fallo de la aplicación
- Interrupción temporal del sistema verificada

No puede ser una razón para el reintento técnico:

- Respuesta incorrecta
- Rechazo
- Respuesta corta
- Falta de cita
- Marca no mencionada
- Error crítico
- La institución auditada no está satisfecha con la respuesta

25. REINTENTO EN RESPUESTA PARCIAL

Si el sistema ha producido parte del contenido del material y luego se detuvo:

- la salida parcial se almacena,
- guardada como RESPUESTA_TRUNCADA

si se va a realizar un reintento, se crea un intento nuevo y vinculado. No puede eliminarse como si no hubiera ninguna respuesta parcial.

26. CADENA DE INTENTOS

Una asignación de observación puede llevar múltiples intentos técnicos:

```
A_i→ (Attempt_{i,1}, Attempt_{i,2}, …)
```

Cada intento:

- tiempo,
- lleva el estado del producto,
- hash del prompt,
- resultado del sistema,
- razón del error técnico

No se sobrescribe por otro intento.

27. ¿QUÉ ES NOMOS CAPTURE?

NOMOS Capture es un estándar abierto de datos y procedimientos que preserva la integridad de la observación desde el prompt dado al usuario hasta el archivado del paquete de evidencia. NOMOS Capture:

- no es solo una aplicación de captura de pantalla,
- no tiene que ser software propietario que pertenezca solo a NobleJackal,

no depende únicamente de una sola extensión de navegador. Una aplicación de Captura NOMOS compatible:

- utiliza un esquema de datos abierto,
- sigue reglas de tiempo e integridad,
- respeta los límites de privacidad,
- preserva la cadena de custodia

debe cumplirse. Los investigadores independientes deberían poder desarrollar sus propias herramientas compatibles. Si el estándar solo puede implementarse con el software del fundador, la replicación independiente se debilita.

28. CUATRO FUNCIONES DE LA CAPTURA NOMOS

28.1. Entregar la Tarea

Correcto:

- usuario,
- prompts,
- Producto de IA,
- ranura

asegura la coincidencia.

28.2. Capturar la Observación

Registra el prompt, el resultado inicial del sistema, la respuesta y el interfaz de usuario relacionado.

28.3. Mantener la Integridad

Genera el hash del archivo, la marca de tiempo y la relación de paquetes.

28.4. Transportar la Evidencia

Vincula las capas de evidencia en bruto, limitadas y disponibles públicamente.

29. MÉTODOS DE CAPTURA

CM-1— CAPTURA MANUAL GUIADA

El participante proporciona una captura de pantalla y el texto de respuesta en bruto siguiendo las instrucciones. Tiene una baja carga técnica. Es propenso a errores humanos.

CM-2— APLICACIÓN DE CAPTURA ASISTIDA

Aplicación:

- copiar prompts,
- registro de tiempo,
- carga de archivos,
- generación de hash

se admiten funciones. No interfiere con el producto de IA.

CM-3— CAPTURA INSTRUMENTADA DE LA INTERFAZ DEL USUARIO

Integración del navegador o aplicación:

- el prompt enviado,
- los metadatos del producto,
- el texto de la respuesta,
- los tiempos

se capturan directamente. Proporciona alta integridad. Se deben evaluar los términos de uso y la privacidad.

CM-4— EXPORTACIÓN DEL PROVEEDOR O CAPTURA DE TELEMETRÍA

Se utiliza la exportación de sesión del proveedor o telemetría verificada. Puede apoyar fuertemente la captura independiente. Puede que no proporcione independencia por sí sola.

CM-5— CAPTURA CONTROLADA DE API O LABORATORIO

Se utilizan registros de API o del sistema congelado. No es observación en vivo del interfaz de usuario. Pertenece a un panel de laboratorio separado.

30. LÍMITE DE AUTOMATIZACIÓN

Automatización:

- envío de prompt,
- registro de tiempo,
- generación de hash,
- integridad del archivo

puede ser utilizado para. En lugar del participante:

- cuenta bot,
- consulta masiva automatizada,
- sesión no de usuario

la generación no crea un Panel de Población humano real. La auditoría del sistema automatizado es separada:

PANEL_DE_AUTOMATIZACIÓN_CONTROLADA

debe ser registrado como.

31. NIVELES DE CONFIANZA DE CAPTURA

NCL-0— NO HAY CAPTURA VERIFICABLE

Solo existe la declaración del participante. No puede ser incluido como el resultado principal GEO-1000.

NCL-1— EVIDENCIA PARCIAL

Hay un texto de respuesta o una captura de pantalla limitada. El prompt o la condición del sistema no pueden ser completamente verificados.

NCL-2— EVIDENCIA EN PANTALLA

El prompt, la respuesta y el interfaz del producto principal han sido verificados con una captura de pantalla. La metainformación y la certeza del tiempo pueden ser limitadas.

NCL-3— PAQUETE DE EVIDENCIA COMPLETA

Respuesta completa, prompt, registro del producto, tiempos, captura de pantalla y hash están disponibles. El nivel mínimo para el análisis principal es candidato.

NCL-4— CAPTURA INSTRUMENTADA Y FIRMADA

La herramienta de captura ha registrado el prompt, la respuesta, el tiempo y los metadatos directamente; el paquete está firmado digitalmente.

NCL-5— CAPTURA MULTIFUENTE VERIFICADA INDEPENDIENTEMENTE

Paquete de captura:

- validación independiente,
- exportación del proveedor,
- segunda fuente de evidencia

ha sido respaldada por. Observaciones de usuario real principal GEO-1000 al menos:

NCL-3

debería apuntar a ese nivel.

32. NOMOS PAQUETE DE EVIDENCIA DE CAPTURA

Para cada observación, el paquete contiene los siguientes componentes como candidatos:

32.1. Manifiesto de Observación

Incluye todos los campos de identidad y estado de la observación.

32.2. Artefacto de Prompt

El texto completo del prompt asignado y enviado, la versión y su hash.

32.3. Artefacto de Respuesta en Bruto

La respuesta completa en bruto producida por el producto de IA.

32.4. Artefacto de Evidencia Visual

Muestra imágenes o imágenes secuenciales del prompt, la respuesta y el interfaz del producto.

32.5. Artefacto de Estado del Sistema

Muestra el plan, la interfaz, la etiqueta del modelo, la memoria, la web y otros estados materiales del sistema.

32.6. Artefacto de Evidencia Temporal

Lleva los tiempos de envío, finalización, captura y carga.

32.7. Registro de Intentos

Muestra la cadena de errores técnicos y reintentos.

32.8. Manifiesto de Integridad

Contiene los hashes y tamaños de todos los archivos.

32.9. Manifiesto de Privacidad

Muestra el estado de redacción, acceso y retención.

32.10. Registro de Validación

Indica quién y cuándo se evaluó la validez de la captura.

33. CONTENIDO MÍNIMO DE LA CAPTURA DE PANTALLA

En la medida en que sea relevante, la evidencia visual debe mostrar:

- Producto de IA o identidad de la interfaz
- Solicitud completa
- Inicio de la respuesta
- Fin de la respuesta
- Campos de referencia
- Mensaje de rechazo o error
- Etiqueta de modelo aparente
- Evidencia de que la sesión es nueva o está en un estado relevante
- Configuración del sistema de material
- Estado de la interfaz que indica que la respuesta no ha sido interrumpida

34. RESPUESTAS LARGAS

Si la respuesta no cabe en una sola pantalla:

- captura de página completa,
- captura desplazable,
- capturas consecutivas superpuestas,
- grabación de pantalla,
- exportación directa de texto

pueden ser utilizadas. Visuales consecutivos:

- no deben dejar espacios,
- deben superponerse entre sí,

deben llevar números de secuencia. Solo la sección positiva o relevante no puede ser recortada.

35. GRABACIÓN DE PANTALLA

Grabación de pantalla:

- envío de prompt,
- tipo de respuesta en streaming,
- estado de finalización
- no se realizó actualización

puede mostrarse. Sin embargo:

- más información personal,
- notificación,

- aplicación privada,
- ID de cuenta

puede ser capturada. Por lo tanto, la grabación de pantalla:

- no puede ser un requisito por defecto,
- puede ser usada en submuestras de alto riesgo o de verificación,

requiere preparación de privacidad antes de grabar.

36. TEXTO DE RESPUESTA EN BRUTO

Además de la captura de pantalla, el texto completo de la respuesta debe almacenarse por separado. Orden de preferencia:

- Exportación oficial o función de copia del producto
- Captura directa de texto por la herramienta de captura
- Extracción basada en el árbol de accesibilidad o DOM
- Texto copiado por el usuario
- OCR y verificación humana

OCR solo debe ser la última opción. Si hay una diferencia material entre la extracción automática y la visual, se requiere una revisión.

37. RESPUESTAS DINÁMICAS

Algunos productos, después de que se completa la respuesta:

- pueden agregar una citación,
- pueden cargar una tarjeta de fuente,
- pueden reformatear el texto,

pueden agregar una advertencia de seguridad. Se pueden usar dos momentos de captura:

- Primera captura de la completación
- Captura después de una breve estabilidad de renderizado

Si hay una diferencia material, se conservan ambas versiones. La primera exposición del usuario no se elimina.

38. ¿CÓMO ENTENDER QUE LA RESPUESTA ESTÁ COMPLETADA?

Indicadores de completación del candidato:

- La transmisión se detiene
- El indicador de carga se cierra
- Aparece el botón de regenerar o equivalente
- Señal de finalización del producto
- Período de silencio predefinido
- Registro de finalización de API

Las reglas de finalización basadas en el producto deben encontrarse en el Registro del Sistema de IA.

39. SI EL USUARIO DETUVO LA RESPUESTA

El usuario puede haber accidental o intencionalmente:

- detener,
- cancelar,
- detener la aplicación

realizado la acción. Esta situación se registra como:

USUARIO_INTERRUPTO

INTERRUPCIÓN_ACCIDENTAL

INTERRUPCIÓN DESCONOCIDA

Una respuesta detenida por el usuario no puede considerarse una salida completa del sistema. La regla de reemplazo del participante o reintento debe definirse de antemano.

40. EL INDICIO Y LA RESPUESTA ESTÁN EN ARCHIVOS SEPARADOS

Si el indicio y la respuesta están en imágenes separadas:

- deben estar vinculados a la misma sesión,
- al mismo ID de observación,
- y a una cadena de tiempo ininterrumpida

De lo contrario, puede que no sea posible verificar que el prompt y la respuesta realmente pertenecen a la misma conversación.

41. INTEGRIDAD DE ARCHIVO

Se debe generar un hash criptográfico para cada archivo. [K25] Candidato:

```
h_j= H(file_j)
```

Hash del paquete:

```
h_bundle= H(JCS(manifest(file_id, orden, tamaño, algoritmo, h_j))) [K14; K25]
```

puede ser creado. Aquí, H es un algoritmo de hash publicado y considerado seguro. El algoritmo:

- versionado,
- abierto,
- puede ser cambiado en el futuro si es necesario.

debe ser.

42. ¿QUÉ DEMUESTRA UN HASH?

Si el valor hash registrado se almacena de manera confiable, volver a generar el hash del mismo archivo con el mismo algoritmo ayuda a detectar cambios en la secuencia de bytes [K25]. Esta comparación respalda la afirmación de que el archivo ha conservado la misma secuencia de bytes

después de que se generó el hash de referencia; no demuestra la fuente del archivo, que fue producido en el momento correcto ni su autenticidad antes del hash. Un hash por sí solo no prueba:

- Que el archivo proviene del producto de IA real
- Que pertenece al usuario correcto
- Que fue producido en el momento correcto
- Que no fue organizado antes de hacer el hash
- Que es la primera salida
- Que la solicitud fue correcta

Por lo tanto, hash:

> *Es prueba de integridad.*

Por sí solo:

> *No es prueba de originalidad.*

43. FIRMA DIGITAL

Herramienta de captura o verificador autorizado:

- el paquete de observación,
- registro de tiempo,
- versión del vehículo

puede firmar digitalmente. Firma:

- qué vehículo o institución produjo el paquete,
- que no ha cambiado después de la firma

fortalece. El titular de la firma puede haber firmado datos incorrectos o falsificados. Por lo tanto, la firma por sí sola no garantiza autenticidad.

44. TIEMPO FIABLE [K26]

Reloj del dispositivo local:

- desconfigurado,
- cambiado manualmente,
- fuera de sincronización

puede ocurrir. Un sello de tiempo fuerte debe llevar múltiples señales:

- Hora de aceptación del servidor de captura
- Servicio de tiempo fiable UTC
- Hora del dispositivo y registro de desviación
- Tiempo visible en el producto de IA, si lo hubiera
- Hora de creación del archivo
- Marca de tiempo digital

45. DESVIACIÓN DE TIEMPO

La diferencia entre la hora del dispositivo del participante y la hora confiable UTC:

```
O_i^clock= t_i^device- t_i^reference
```

puede registrarse como. Si hay una desviación significativa:

- se utiliza la hora del servidor como referencia principal,
- la hora del dispositivo se conserva como un registro corregido,

se marca la incertidumbre. No se debe pedir al participante que cambie manualmente el reloj del dispositivo.

46. LA HORA DE CARGA NO ES LA HORA DE CAPTURA

Subir un archivo al servidor en 14:00 no demuestra que la captura de pantalla se haya tomado a 14:00. Por lo tanto:

- hora de captura,
- hora de generación del hash,
- hora de carga

debe registrarse por separado. Los largos retrasos en la carga pueden reducir la confianza en la captura. El registro no tiene que ser automáticamente inválido.

47. CARGA RETRASADA

Debido a problemas de conexión, el usuario puede no ser capaz de cargar la evidencia de inmediato. Si se acepta la carga tardía:

- hash local en el momento de la captura,
- tiempo del dispositivo de confianza,
- registro firmado fuera de línea,
- razón del retraso en la carga

pueden ser requeridos. Solo la declaración del usuario: “Tomé esta imagen ayer.” no es suficiente para el nivel principal NCL-3.

48. CADENA DE EVIDENCIA

La cadena de evidencia de una observación como candidato consiste en las siguientes etapas:

- Asignación de tarea
- Fijación del prompt
- Idoneidad del usuario y del producto
- Apertura de ranura
- Envío de prompt
- Resultado del contacto inicial
- Finalización de la respuesta
- Captura
- Registro de integridad local

- Subida segura
- Hora de aceptación del servidor
- Hash del paquete
- Verificación del protocolo
- Edición
- Adjudicación
- Archivo
- Manifiesto público

Cada etapa debe estar conectada a la etapa anterior.

49. ROMPIENDO LA CADENA

La cadena de evidencia puede romperse en las siguientes situaciones:

- No hay registro de prompt
- No se puede verificar la salida inicial
- La respuesta no coincide con lo visual
- El tiempo de captura es desconocido
- El archivo ha sido modificado posteriormente
- ID de participante duplicado
- Producto y plan son desconocidos
- La observación está fuera de onda
- El hash no coincide con el archivo
- La edición sobrescribió el archivo sin procesar
- Se desconoce quién lo verificó

El tipo de interrupción determina el estado de validez.

50. LA EVIDENCIA EN BRUTO NO PUEDE SER CAMBIADA; LA DECISIÓN PUEDE CAMBIAR

Paquete de evidencia en bruto:

- no debe ser sobrescrito,
- no debe ser corregido en silencio,

no debe ser eliminado ni reemplazado por uno nuevo. Sin embargo, respecto a la observación:

- validez,
- adjudicación,
- clasificación,
- cumplimiento

la decisión puede cambiar. La nueva decisión:

- nueva versión,
- justificación,
- fecha,
- decisor

debe llevar. Esta distinción es una regla fundamental de gobernanza:

> *La evidencia puede permanecer fija; la interpretación puede evolucionar con la evidencia.*

51. REDACCIÓN

Los siguientes campos pueden ser redactados en la copia pública o del adjudicador:

- Nombre y apellido
- Correo electrónico
- Foto de perfil
- ID de cuenta
- Conversaciones privadas
- Notificaciones
- Campo corporativo
- Nombre de archivo sensible
- Ubicación exacta

Redacción:

- no debe hacerse en el archivo original,
- debe producir un nuevo archivo derivado,
- debe llevar su propio hash,

debe registrar qué campo fue redactado y por qué.

52. TRES CAPAS DE EVIDENCIA

52.1. Evidencia restringida en bruto

Es evidencia bruta completa. Solo auditores autorizados y roles de investigación requeridos pueden acceder a ella.

52.2. Capa de Evidencia de Auditoría

Contiene la evidencia necesaria para los adjudicadores independientes y los equipos de reproducción. Las áreas personales se minimizan.

52.3. Manifiesto de Evidencia Pública

Para el público:

- ID del observador,
- celda de país y idioma,
- Producto de IA,
- tiempo,
- hash de la evidencia,
- estado de validez,
- estado de redactado

se presentan. El contenido privado bruto no se publica necesariamente.

53. MANIFIESTO DE EVIDENCIA PÚBLICA

Para cada ola, el manifiesto público puede incluir los siguientes campos:

- ID de la ola
- Versión del protocolo
- Número de productos
- Recuento de observaciones objetivo y válidas
- Cobertura de país e idioma
- Cobertura de ranuras
- Niveles de confianza de captura
- Errores técnicos y tasas de rechazo
- Recuento de observaciones inválidas
- Hash de paquetes de observación
- Raíz de integridad de la ola
- Cambios de eventos
- Ejemplos redactados públicamente
- Procedimiento de acceso independiente
- Estado de retención de evidencia

54. RAÍZ DE INTEGRIDAD DE LA OLA

CONCEPTO DE CANDIDATO

Sea hi el hash de cada paquete de observación. A partir de estos, la raíz de la ola con un árbol de Merkle o estructura equivalente:

RW

puede ser producida. La raíz de la ola puede ser publicada después de que la medición esté cerrada. Este método:

- la modificación silenciosa posterior de la lista de observaciones en la ola,
- añadiendo una nueva observación,
- modificando una observación existente

la hace más visible. Raíz de Merkle:

- por sí sola no prueba que las observaciones sean reales,
- o que la decisión de adjudicación sea correcta

Fortalece la integridad de la lista.

55. ADICIÓN Y ELIMINACIÓN DE OBSERVACIONES

Después de que la ola se cierra:

- añadiendo una nueva observación,
- eliminando una observación existente,
- modificando el archivo

requiere una nueva versión del manifiesto. Se conserva la raíz de integridad antigua. Para la observación que se elimina:

- razón,
- fecha,
- tomador de decisiones
- efecto en la puntuación antigua y nueva

que se registra.

56. SOLICITUD DE RETIRO DEL PARTICIPANTE

La ética de la investigación o el marco de protección de datos puede permitir que el participante retire sus datos. En este caso:

- la evidencia cruda puede eliminarse de manera segura o hacerse inaccesible,
- un registro WITHDRAWN vacío puede conservarse en el manifiesto público,
- la puntuación de la ola puede recalcularse con la nueva versión

el resultado antiguo y la razón del cambio se conservan. La inmutabilidad no significa eliminar los derechos humanos.

57. ESTADOS DE VALIDEZ DE CAPTURA DE OBSERVACIÓN

CV-0— NO REVISADO

La captura aún no ha sido evaluada.

CV-1— VÁLIDO

Cumple con todas las condiciones obligatorias de captura y tiempo.

CV-2— VÁLIDO CON CONDICIONES

Existe una deficiencia limitada. Se especifica en qué análisis se puede usar.

CV-3— EVIDENCIA PARCIAL

La observación puede usarse como un evento. Puede que no ingrese en la puntuación principal de la población.

CV-4— FUERA DE ONDA

El prompt se envió fuera de la estructura de tiempo permitida.

CV-5— DESAJUSTE ENTRE prompt O RESULTADO

Existe una discrepancia material entre el prompt asignado, el prompt enviado o la respuesta registrada.

CV-6— DUPLICADO

La misma observación o registro de usuario se ha repetido.

CV-7— SOSPECHA DE MANIPULACIÓN

Existe una duda sin resolver sobre la integridad del archivo o los metadatos.

CV-8— FABRICADO O MANIPULADO

La observación no es una experiencia real de usuario o ha sido alterada deliberadamente.

CV-9— RETIRADO

Se ha excluido del análisis debido a procesos de participantes o éticos.

58. LA VALIDEZ DE LA CAPTURA DEBE SEPARARSE DE LA EXACTITUD SEMÁNTICA

Una observación:

- completamente válida en términos de captura,
- gravemente incorrecta en términos de contenido

podría ser. Otra observación:

- respuesta correcta,
- puede contener evidencia faltante o manipulada

Por lo tanto, hay dos decisiones separadas:

- Validez de la captura
- Adjudicación semántica

El validador de capturas debe conocer lo menos posible la puntuación de éxito de la respuesta.

59. FUNCIÓN DEL VALIDADOR DE CAPTURAS

El validador de capturas examina:

- Correspondencia del prompt
- Oleada y franja temporal
- Identidad del producto
- Primera regla de salida
- Integridad de la pantalla
- Consistencia del texto en bruto
- Reintento técnico
- Hash y tiempo
- Privacidad

No decide sobre: ¿La empresa fue descrita correctamente? ¿La cita respalda la afirmación? ¿Hay un error crítico? ¿La respuesta aprueba? Estas son las responsabilidades del siguiente nivel de revisión.

60. IMAGEN DE PANTALLA REPETIDA

El mismo visual:

- subido accidentalmente dos veces,

- utilizado por múltiples usuarios,
- repetido entre proveedores del panel

puede ocurrir. Comprobaciones:

- Hash del archivo
- Similitud visual
- Texto de respuesta en bruto
- Metadatos de tiempo y sesión
- Unicidad del usuario

pueden realizarse. El mismo producto de IA puede dar exactamente la misma respuesta palabra por palabra a algunos usuarios. El mismo texto por sí solo no es evidencia de fraude. El mismo archivo o la misma huella visual es una señal más fuerte.

61. SIMILITUD DE ARCHIVO Y FALSO POSITIVO

Mismo dispositivo e interfaz:

- visuales similares,
- mismo tamaño de pantalla,
- misma compresión

puede producir. El algoritmo de similitud visual no debe realizar exclusión automática. Se requieren múltiples señales y revisión humana.

62. TELEMETRÍA DEL PROVEEDOR

Proveedor de IA:

- tiempo de sesión,
- modelo,
- prompts,
- respuesta,
- uso de la herramienta

puede proporcionar registros confiables sobre esto. Este registro fortalece la captura independiente. Sin embargo, el proveedor:

- parte medida,
- propietario del sistema,
- beneficiario

puede ser. Telemetría del proveedor: debe ser utilizada junto con, y no como reemplazo de, evidencia independiente del usuario.

63. SI NO HAY EVIDENCIA DEL PROVEEDOR

El acceso a los registros del proveedor puede no estar disponible en la mayoría de los productos para consumidores en tiempo real. Esto no hace que GEO-1000 sea imposible. La siguiente composición puede ser utilizada:

- Captura de usuario instrumentada

- Marca de tiempo del servidor
- Prueba de pantalla completa
- Respuesta sin procesar
- Configuración del sistema
- Hash y firma
- Submuestra de verificación independiente

El nivel de confianza de la prueba se indica en consecuencia.

64. CAPTURA EN Panel de usuarios en condiciones reales

Herramienta de captura en el panel natural:

- contenido de instrucciones especiales,
- conversaciones anteriores,
- identidad de la cuenta

no debe capturarse innecesariamente. Flujo preferido: El usuario abre un nuevo chat de cuenta natural. Capture solo registra la ventana de tareas. El estado de personalización se toma por separado como metadatos de alto nivel. El historial privado permanece fuera de la captura. La copia pública se edita.

65. CAPTURA DE ESTADO DE CONVERSACIÓN NATURAL

Si se mide el contexto de la conversación actual:

- solo el rango de turnos previos necesario,
- consentimiento explícito,
- fuerte redacción,
- restricción de acceso

es requerido. Todo el historial de chat privado no puede recolectarse por defecto.

66. CAPTURA MÓVIL

En dispositivos móviles:

- captura de pantalla completa,
- captura desplazable,
- versión de la aplicación,
- registro de tiempo del sistema

puede funcionar de manera diferente. Protocolo móvil:

- iOS,
- Android,
- otros sistemas

pueden tener guías técnicas separadas. No se pueden excluir a los usuarios móviles ya que la captura en escritorio es difícil. La herramienta de captura debe adaptarse a la población.

67. ACCESIBILIDAD Y CAPTURA

Para participantes que usan lectores de pantalla, entrada por voz o herramientas de asistencia motora, la captura visual estándar por sí sola puede no ser suficiente. Evidencia alternativa:

- Exportación del árbol de accesibilidad
- Grabación de sesión de voz
- Transcripción de la aplicación
- Registro de tecnología de asistencia
- Registro de observación de tareas

puede ocurrir. Estas observaciones no deben considerarse de baja fiabilidad. Debe desarrollarse un estándar de evidencia equivalente.

68. OBSERVACIÓN DEL CAMBIO DE COMPORTAMIENTO DE LA HERRAMIENTA DE CAPTURA

Extensión del navegador o instrumentación:

- carga de página,
- comportamiento de la interfaz,
- sistema de seguridad,
- detección de productos

puede verse afectada. Se debe probar el efecto de la herramienta de captura en el piloto. Si las sesiones con y sin la herramienta se comportan de manera diferente: la condición de la herramienta se convierte en un factor experimental, no se puede asumir que la herramienta de captura sea silenciosamente neutral.

69. VERSIONADO DE LA HERRAMIENTA DE CAPTURA

Cada paquete de captura:

- nombre de la herramienta,
- versión,
- configuración,
- entorno operativo,
- errores conocidos

debe llevar. Si la herramienta de captura cambia durante la medición:

- nueva versión de la herramienta,
- subonda,
- revisión de compatibilidad

puede ser requerida.

70. ERROR DE LA HERRAMIENTA DE CAPTURA

Herramienta:

- puede copiar incorrectamente el prompt,
- puede truncar el final de la respuesta,
- puede registrar la hora incorrectamente,
- puede no generar un hash,

puede comprimir el archivo. Error de la herramienta:

- no se puede cargar al producto de IA,
- o el usuario

Por separado:

FALLO_DE_CAPTURA_DE_HERRAMIENTA

debe ser registrado como.

71. CASO DE TRES OLAS SINTÉTICAS GEO-1000

DEMOSTRACIÓN DE METODOLOGÍA SINTÉTICA / Los siguientes productos, fechas, números y resultados son completamente ficticios. Diez productos de IA sintéticos:

A1–A10

deben medirse en tres olas principales.

71.1. Calendario de Olas

Onda	Tipo	UTC ventana de entrega	Propósito
W1	Principal	12.00-13.00	Inicio
W2	Confirmatorio	20.00-21.00	Repetición a corto plazo
W3	Estabilidad	04.00-05.00	Equilibrio a medio plazo y por hora

Para cada producto en cada ola: se buscan observaciones principales válidas 1,000. Objetivo total:

```
10×1,000×3=30,000
```

71.2. Estructura de los slots

Cada ola:

- Micro-slots 12
- con participantes equilibrados por producto y país por slot

El slot del participante: asignado aleatoriamente antes de la recopilación de datos.

71.3. Resultado W1

Meta: Observaciones principales válidas 10,000 Campo:

- Invitaciones 13,240
- Tareas completadas 10,870
- Observaciones válidas 10,106 en términos de captura

- Observaciones principales del análisis de 10,000 según el orden principal/respaldo previamente fijado

Registros válidos restantes de 106:

- respaldo,
- sensibilidad,
- control de calidad de captura

están reservados para.

71.4. Eventos técnicos

límite de uso de 118 42 error de conexión 31 rechazo del sistema 17 desconexión del usuario 9 desajuste de prompt 6 reintento de captura de pantalla Todos los resultados de contacto inicial han sido guardados. Solo se reintentaron errores técnicos reales predefinidos.

71.5. Cambio de producto durante W2

Producto A7 en 20.32 UTC:

- uso automático de la web,
- nueva vista de referencia

dejar que gane. Observaciones de A7:

W2A: 20.00–20.31

W2B: 20.32–21.00

se separan como. No se publica una puntuación W2 fija para A7 sin explicación.

71.6. Evento externo en W3

Se debe publicar una decisión legal importante sobre 04.22 UTC acerca de la empresa sintética auditada. Debe observarse que las respuestas en productos accesibles por web cambiaron después de 04.25. Evento: Se agrega al Registro de Eventos Externos. Los resultados antes y después del evento se muestran como celdas de diagnóstico separadas. No se eliminan las observaciones porque disminuyen o aumentan la puntuación.

71.7. Tarjeta de Salud de Wave

Onda	Salud	Nota principal
W1	WH-4	Captura equilibrada y fuerte
W2	WH-3	Sub-ola debido a cambio de producto A7
W3	WH-3	Registro de evento legal externo

Se pueden usar las tres olas. Los comentarios sobre A7 y eventos externos también son restringidos.

72. CASO DE CAPTURA DE OBSERVACIÓN ÚNICA SINTÉTICA

EJEMPLO SINTÉTICO

Participante:

- Turquía
- Turco
- Plan web de pago de Orion AI
- Panel controlado en condiciones depuradas

W1

Ranura 12.15–12.20 UTC Indicador asignado: "¿Con qué empresa matriz está asociada Apple.com, y cuáles son las principales actividades de esta empresa?"

72.1. Veces

Apertura de tarea: 12.14.20 Envío de solicitud: 12.16.08 Primer token: 12.16.10 Finalización de respuesta: 12.16.24 Captura: 12.16.39 Generación de hash: 12.16.41 Subida: 12.17.12 Aceptación del servidor: 12.17.13

72.2. Evidencia

El hash del prompt coincide Texto completo de la respuesta disponible Dos capturas de pantalla superpuestas disponibles Producto, plan y etiqueta del modelo visibles Memoria e instrucción especial apagadas Uso web activado Cita no mostrada Sin reproducción Sin conversación previa Hashes de archivos y firma del paquete disponibles Nivel de captura:

NCL-4

Validez:

CV-1

Aún no se ha decidido si la respuesta es correcta o incorrecta.

73. CASO SINTÉTICO DE REINTENTO TÉCNICO

En el primer intento:

- Solicitud enviada correctamente
- El producto mostró "Error de red"
- No se inició una respuesta sustantiva
- Captura de pantalla del error tomada
- Se registraron la hora y el hash

Según la regla predefinida, se realizó un segundo intento dentro de cinco minutos. Se produjo una respuesta en el segundo intento. Registro correcto:

- Resultado del primer contacto: FALLA_TÉCNICA_TEMPORAL
- Primer contenido elegible generado: segundo intento
- Ambos intentos están en el paquete
- El primer incidente está incluido en la tasa de error técnico
- La segunda respuesta está incluida en la evaluación de contenido, si el método predice esto de antemano

Registro incorrecto: Se elimina el primer error y se trata como si solo existiera la segunda respuesta.

74. CASO DE REINTENTO INCORRECTO SINTÉTICO

Primera respuesta: “Apple es una empresa local que solo vende teléfonos móviles.” Al participante no le gusta la respuesta. La reproduce. La segunda respuesta resulta ser correcta. El participante solo envía la segunda respuesta. El registro del sistema o el registro de captura muestra la repetición. Estado correcto:

CV-8— OBSERVACIÓN MANIPULADA

La observación no se guarda solo porque la segunda respuesta sea correcta. Si la primera respuesta también puede recuperarse, puede preservarse en la investigación del incidente. El registro del participante principal no es válido.

75. INFORME DE CIERRE DE OLA

Al final de cada ola, se debe crear el siguiente registro:

- ID de la ola
- Hora de apertura y cierre
- Número objetivo de observaciones
- Persona invitada
- Tarea completada
- Captura válida
- Observación que ingresa al análisis principal
- Observación válida de respaldo
- Error técnico
- Rechazo
- Límite de uso
- Desajuste de solicitud
- Desajuste de captura
- Carga tardía
- Duplicado
- Sospecha de manipulación
- Retiro del participante
- Cambios de producto
- Eventos externos
- Cobertura de ranuras
- Cobertura de país e idioma
- Niveles de confianza de captura
- Estado de salud de la oleada
- Raíz de integridad
- Aprobación humana responsable

76. DISPOSICIONES NORMATIVAS OBLIGATORIAS

CH12-N01

Cada observación GEO-1000 debe estar vinculada a una ola de medición pre-fijada.

CH12-N02

El término "mismo tiempo" no se puede usar sin la ventana completa UTC, microfranja temporal y las reglas de finalización.

CH12-N03

El tiempo local no puede usarse solo como tiempo de ola; el estándar de tiempo principal debe ser UTC.

CH12-N04

La ventana de transmisión, los micro-intervalos, el tiempo de finalización y el tiempo de carga deben determinarse antes de que se vean las respuestas de la IA.

CH12-N05

La asignación de intervalos debe hacerse de manera que no deje a países, idiomas, productos y grupos de panel desbalanceados en términos de tiempo.

CH12-N06

No se puede mover a los participantes a intervalos anteriores o posteriores basándose en sus resultados.

CH12-N07

El desfase horario local de la única ventana UTC debe documentarse en el registro público de métodos.

CH12-N08

Las olas globales que regresan o las operaciones equilibradas localmente en tiempo deben mantenerse como resultados de tiempo separados.

CH12-N09

Cada solicitud de observación debe incluir los tiempos de envío, primer contacto, finalización, captura y carga en la medida de lo pertinente.

CH12-N10

Los campos de tiempo desconocidos no pueden presentarse como tiempos reales estimados.

CH12-N11

El reloj del dispositivo por sí solo no puede considerarse una prueba de tiempo UTC confiable.

CH12-N12

El tiempo de captura y el tiempo de carga deben registrarse por separado.

CH12-N13

Las observaciones enviadas fuera de la ola no pueden incluirse en la ola principal sin explicación.

CH12-N14

Las reglas de exclusión por demora o desviación de ranura deben ser independientes del contenido de la respuesta.

CH12-N15

Como resultado del primer contacto, debe mantenerse por separado del primer contenido emitido.

CH12-N16

Rechazo, límite de uso, error técnico, aclaración e inconclusión no pueden colocarse en la misma categoría.

CH12-N17

Una respuesta incorrecta, negativa, sin referencia o de IA Crítica no puede ser una razón para reintentar.

CH12-N18

El reintento técnico solo puede realizarse bajo condiciones técnicas predefinidas.

CH12-N19

Cuando se realiza un reintento técnico, se deben conservar todos los intentos y el resultado del contacto inicial.

CH12-N20

No se pueden ignorar ni reemplazar las respuestas parciales o incompletas con una nueva respuesta.

CH12-N21

La primera salida adecuada no puede ser renovada, seleccionada ni mejorada manualmente.

CH12-N22

Cada observación debe alcanzar al menos el nivel de confianza de captura NCL-3 o un equivalente justificado.

CH12-N23

Una captura de pantalla por sí sola no puede considerarse un paquete completo de evidencia.

CH12-N24

El prompt de evidencia visual debe mostrar toda la respuesta y la interfaz del producto físico.

CH12-N25

Las respuestas largas deben preservarse completamente y de manera secuencial usando el método de captura.

CH12-N26

Si hay una diferencia material entre el texto de la respuesta original y la evidencia visual, se debe realizar un examen.

CH12-N27

OCR puede utilizarse como un método auxiliar si no es posible registrar el texto directamente; sin verificación, no puede considerarse la fuente final del texto.

CH12-N28

La finalización inicial y las versiones posteriores del material de respuestas que cambian dinámicamente deben preservarse por separado.

CH12-N29

Se debe registrar como un evento separado al participante que detenga la respuesta o cierre la interfaz.

CH12-N30

Si el prompt y la respuesta están en archivos separados, deben estar vinculados a la misma sesión y cadena de tiempo.

CH12-N31

Cada archivo de evidencia y paquete de evidencia debe llevar un hash de integridad.

CH12-N32

El hash por sí solo no puede presentarse como prueba de la realidad de la observación.

CH12-N33

La firma digital y la marca de tiempo no constituyen la totalidad de la autenticidad de la observación; son componentes de la cadena de evidencia.

CH12-N34

No se puede escribir ninguna redacción o corrección sobre el archivo de evidencia sin procesar.

CH12-N35

Cada derivado editado debe llevar su propio ID de archivo y hash.

CH12-N36

El acceso a la evidencia sin procesar puede estar restringido; el manifiesto público debe demostrar la integridad de la observación sin revelar la identidad.

CH12-N37

El nombre del participante, correo electrónico, contraseña, conversación privada o información personal innecesaria no puede ser añadida a la evidencia pública.

CH12-N38

También se debe evaluar el riesgo de reidentificación para países pequeños y usuarios de subgrupos raros.

CH12-N39

La captura Panel de usuarios en condiciones reales no debe recopilar instrucciones especiales ni contenido de conversaciones anteriores por defecto.

CH12-N40

Se deben proporcionar métodos de captura alternativos equivalentes para usuarios con necesidades de accesibilidad.

CH12-N41

Si la herramienta de captura cambia materialmente la experiencia del usuario, este efecto debe medirse por separado.

CH12-N42

La herramienta de captura y su versión deben registrarse en cada observación.

CH12-N43

Si la herramienta de captura cambia materialmente durante una fase, se debe hacer una distinción por subfase o versión de la herramienta.

CH12-N44

El error de la herramienta de captura no puede clasificarse como un producto de IA o un fallo del participante.

CH12-N45

La telemetría del proveedor puede apoyar la captura independiente del usuario; no puede reemplazarla sin explicación.

CH12-N46

Si ocurre un cambio material en el producto de IA durante la ola, las observaciones no pueden fusionarse como un único estado fijo del producto.

CH12-N47

El cambio del producto requiere separación de olas, reinicio, modelado explícito o advertencia de estado mixto.

CH12-N48

La ausencia de un anuncio de actualización del proveedor no es prueba definitiva de que el sistema no haya cambiado.

CH12-N49

Los eventos externos materiales deben agregarse al Registro de Eventos Externos con marca de tiempo y alcance.

CH12-N50

Los eventos externos no pueden ser una razón para exclusión después de los resultados porque hayan aumentado o disminuido la puntuación.

CH12-N51

La validez de la captura debe evaluarse de la manera más ciega posible, antes de evaluar la corrección semántica de la respuesta.

CH12-N52

Una respuesta correcta de IA no puede validar una captura ausente o manipulada.

CH12-N53

Una respuesta incorrecta de IA no puede invalidar una captura que cumple con el protocolo.

CH12-N54

Las decisiones de duplicación y manipulación deben involucrar múltiples señales y revisión humana.

CH12-N55

El mismo texto de respuesta por sí solo no puede considerarse prueba de duplicación o fraude.

CH12-N56

Agregar, eliminar o cambiar observaciones después de que una ola esté cerrada requiere un nuevo manifiesto y versión de análisis.

CH12-N57

Se debe preservar el antiguo manifiesto de la ola y la raíz de integridad.

CH12-N58

La evidencia en bruto puede preservarse sin cambios; las decisiones de validez y adjudicación solo pueden cambiarse en forma versionada.

CH12-N59

El derecho del participante a retirarse y la ética de la investigación deben mantenerse por encima de la afirmación de inmutabilidad.

CH12-N60

La integridad de la ola debe evaluarse independientemente de la puntuación de IA.

CH12-N61

El informe de cierre de la ola debe mostrar el objetivo, el real, las invalideces, los eventos técnicos, los cambios de producto y la integridad de la evidencia.

CH12-N62

NOMOS El cumplimiento de la captura no puede depender únicamente del uso de software desarrollado por NobleJackal; deberían ser posibles aplicaciones abiertas equivalentes.

CH12-N63

El esquema de captura, las decisiones de validación y los registros de ondas deben versionarse en un formato legible por máquina.

CH12-N64

Cada onda y cadena de evidencia debe tener un propietario humano o institucional responsable.

77. FORMAS DE FALLA

CH12-F01— LA MISMA SEGUNDA FALACIA

Forzar a todos los participantes a enviar en el mismo segundo se considera la única condición para la simultaneidad.

CH12-F02— CONTAR "HOY" COMO UNA ONDA SINCRONIZADA

El tiempo local y las zonas horarias se confunden.

CH12-F03— USANDO EL TIEMPO LOCAL COMO UTC

Las observaciones ingresan en un orden cronológico incorrecto.

CH12-F04— CAMBIANDO LA ASIGNACIÓN DE FRANJAS BASÁNDONOS EN LOS RESULTADOS

Los usuarios con puntuaciones bajas se trasladan a otra celda de tiempo.

CH12-F05— EMPAQUETANDO UN PAÍS EN UNA SOLA RANURA

El efecto del tiempo se mezcla con el efecto del país.

CH12-F06— GENERANDO CARGA ARTIFICIAL

Miles de usuarios envían simultáneamente, provocando el comportamiento anómalo de errores del producto.

CH12-F07— OCULTANDO LA PARTICIPACIÓN NOCTURNA

La fatiga y la pérdida de participación no se informan en ciertas áreas.

CH12-F08— AÑADIENDO OBSERVACIÓN FUERA DE OLA A LA OLA

Una respuesta recibida horas o días más tarde se utiliza como si perteneciera a la ventana principal.

CH12-F09— CONTANDO EL TIEMPO DE CARGA COMO TIEMPO DE CAPTURA

No se puede demostrar cuándo se generó el archivo.

CH12-F10— CONFIANZA CIEGA EN EL RELOJ DEL DISPOSITIVO

El tiempo cambiado manualmente o incorrecto se convierte en el registro de tiempo principal.

CH12-F11— EXCLUYENDO RESPUESTA NEGATIVA RETRASADA

La regla de tiempo se aplica dependiendo del resultado.

CH12-F12— ELIMINANDO EL PRIMER RESULTADO DE CONTACTO

El límite de velocidad, error o rechazo se hace invisible con el segundo intento.

CH12-F13— CONTANDO RECHAZO COMO ERROR TÉCNICO

Se reintenta el comportamiento real del sistema.

CH12-F14— CONTANDO RESPUESTA INCORRECTA COMO ERROR TÉCNICO

Se realiza regeneración hasta que ocurra el resultado deseado.

CH12-F15— IGNORANDO RESPUESTA PARCIAL

Incluso si el sistema se interrumpe, solo se guarda la respuesta completa subsecuente.

CH12-F16— SELECCIONANDO LA MEJOR RESPUESTA

Entre múltiples respuestas, se informa la más positiva o la más larga.

CH12-F17— SOLO CAPTURA DE PANTALLA

El visual se considera suficiente sin prompt, producto, tiempo y metadatos.

CH12-F18— VISUAL DE ÉXITO RECORTADO

Partes incorrectas o restrictivas se dejan invisibles.

CH12-F19— CAPTURA SIN MOSTRAR LA prompt

No se puede probar a qué pregunta pertenece la respuesta.

CH12-F20— TRUNCANDO EL FINAL DE LA RESPUESTA

La parte de limitación o rechazo del material se vuelve invisible.

CH12-F21— HUECO EN VISUALES CONSECUTIVOS

Se pierden algunas partes de la respuesta.

CH12-F22— CONTAR EL TEXTO OCR COMO RESPUESTA EN BRUTO

Errores de reconocimiento cambian las afirmaciones sobre el material.

CH12-F23— CONSIDERAR SOLO LA ÚLTIMA VERSIÓN DE LA RESPUESTA DINÁMICA COMO REAL

La exposición inicial del usuario ha sido eliminada.

CH12-F24— CONSIDERAR LA INTERRUPCIÓN DEL USUARIO COMO RESPUESTA COMPLETA

La salida del sistema está incompleta debido a la intervención del usuario.

CH12-F25— CONJUNTAR INCORRECTAMENTE prompt Y RESPUESTA

Archivos de diferentes conversaciones se vinculan a la misma observación.

CH12-F26— ADORACIÓN DEL HASH

Cualquier archivo con un hash se considera real y original.

CH12-F27— IGNORAR LA EDICIÓN PREVIA AL HASH

Los archivos truncados o modificados son legitimados mediante el hashing.

CH12-F28— CONSIDERAR LA FIRMA COMO VÁLIDA

Se acepta un registro firmado pero incorrecto sin disputa.

CH12-F29— EDICIÓN DEL ARCHIVO BRUTO

La evidencia original se altera irreversiblemente.

CH12-F30— CONSIDERANDO EL DERIVADO EDITADO COMO EVIDENCIA ORIGINAL

La copia pública se usa como el archivo original.

CH12-F31— ELIMINANDO LA PRIVACIDAD PARA EVIDENCIA

Se publican el nombre del participante, la cuenta y el chat privado.

CH12-F32— REVELANDO A UN PARTICIPANTE DE UN PAÍS PEQUEÑO

La información de tiempo, dispositivo y plan permite que una persona sea reidentificable.

CH12-F33— RECOPILANDO HISTORIA ESPECIAL EN EL PANEL NATURAL

Conversaciones irrelevantes ingresan a la captura.

CH12-F34— INVALIDANDO LA EVIDENCIA DE ACCESIBILIDAD

El usuario real queda excluido porque no se puede generar un estándar visual.

CH12-F35— ASUMIENDO QUE LA HERRAMIENTA DE CAPTURA ES NEUTRAL

La herramienta cambia el comportamiento del producto pero no se prueba.

CH12-F36— OCULTANDO LA VERSIÓN DE CAPTURA

Diferentes versiones de la herramienta se presentan como un único método de medición.

CH12-F37— CONSIDERANDO EL ERROR DE LA HERRAMIENTA DE CAPTURA COMO UN ERROR DE IA

La respuesta cortada se carga en el producto.

CH12-F38— CONTANDO EL REGISTRO DE LA API EN LA PANTALLA DEL USUARIO

El interfaz de usuario real se vuelve invisible.

CH12-F39— CONTANDO EL REGISTRO DEL PROVEEDOR COMO EVIDENCIA INDEPENDIENTE

El registro de la parte medida se convierte en una única pieza de evidencia.

CH12-F40— OCULTANDO LA ACTUALIZACIÓN DEL PRODUCTO

Dos estados del sistema dentro de la rama se fusionan en un solo puntuación.

CH12-F41— IGNORANDO CAMBIOS SI NO HAY ANUNCIO

El fallo observado del sistema no se examina.

CH12-F42— EXCLUYENDO EVENTOS EXTERNOS BASADOS EN EL RESULTADO

Las respuestas con noticias negativas se eliminan; las respuestas con noticias positivas se conservan.

CH12-F43— VINCULANDO LA SALUD DE LA ONDA CON LA PUNTUACIÓN

Las ondas defectuosas con puntuaciones altas son válidas, las ondas fuertes con puntuaciones bajas se consideran sospechosas.

CH12-F44— CONTANDO LA RESPUESTA CORRECTA COMO UNA CAPTURA VÁLIDA

El defecto en la evidencia se disfraza con éxito semántico.

CH12-F45— CONTANDO LA RESPUESTA INCORRECTA COMO UNA CAPTURA INVÁLIDA

El resultado de la medición se convierte en error de datos.

CH12-F46— CONTANDO EL MISMO TEXTO COMO EVIDENCIA DEL BOT

Las respuestas determinísticas o similares de IA accidentalmente se convierten en duplicados.

CH12-F47— DECISIÓN DE MANIPULACIÓN DE UNA SOLA SEÑAL

El tamaño del archivo o la diferencia de tiempo crea automáticamente una exclusión.

CH12-F48— AÑADIR OBSERVACIÓN DESPUÉS DE LA ONDA

Se agregan silenciosamente nuevos registros para alcanzar el tamaño de muestra deseado.

CH12-F49— ELIMINAR OBSERVACIÓN DE BAJA PUNTUACIÓN DESPUÉS DE LA OLA

Se cambia retroactivamente el manifiesto y el resultado.

CH12-F50— ELIMINAR MANIFIESTO ANTIGUO

Se destruye el historial de versiones.

CH12-F51— CORREGIR EVIDENCIA ORIGINAL

Se modifica el contenido del archivo para corregir la clasificación incorrecta.

CH12-F52— NEGAR EL DERECHO DE RETIRADA

Se antepone la inmutabilidad a los derechos humanos.

CH12-F53— MONOPOLIO DE CAPTURA PROPIETARIA CERRADO

Solo se considera válida la evidencia producida con la herramienta del fundador.

CH12-F54— NO PUBLICAR EL ESQUEMA DE CAPTURA

Un equipo independiente no puede establecer la misma cadena de evidencia.

CH12-F55— CONSIDERANDO LA RAÍZ DE LA ONDA COMO PRUEBA DE REALIDAD

La raíz Merkle o de hash se presenta como si demostrara la exactitud del contenido.

CH12-F56— afirmación DE "CAPTURAS DE PANTALLA 1,000" SIN UN PAQUETE DE EVIDENCIA

El número de archivos se usa como si fuera el número de observaciones válidas.

78. PROCEDIMIENTO DE AUDITORÍA

Paso 1— Determinar el Propósito de Investigación de la Onda

Se selecciona el tipo principal, confirmatorio, de incidente, de deriva o de reexamen.

Paso 2— Fijar Estados del Producto y Sistema de IA

Se conecta a la versión de onda del Registro del Sistema de IA en la Sección 9.

Paso 3— Fijar Estados de Población, Muestra y Panel

Se utilizan registros de las Secciones 5–8 y la Sección 11.

Paso 4— Fijar Registro de Prompt

Las versiones del prompt de la sección 10 están vinculadas a la oleada de medición.

Paso 5— Configurar la Ventana de Ola UTC

Se escriben las duraciones de inicio, cierre y adicionales.

Paso 6— Crear Micro-Ranuras

Los participantes son asignados a las ranuras de manera equilibrada en términos de producto, país e idioma.

Paso 7— Definir Fuente de Tiempo

Se determinan el servidor, el dispositivo y los registros confiables UTC.

Paso 8— Fijar Primer Contacto y Reglas de Reintento

Los estados técnicos y semánticos se separan.

Paso 9— Determinar Métodos de Captura

Se definen las rutas web, móviles, de accesibilidad y del panel natural.

Paso 10— Determinar Objetivo NCL

Registrar el nivel de confianza de captura requerido para el análisis principal.

Paso 11— Fijar el Esquema del Paquete de Evidencia

Se definen el manifiesto, archivos, hash y campos de metadatos.

Paso 12— Fijar las Reglas de Privacidad y Redacción

Se separan las capas en bruto, de auditoría y pública.

Paso 13— Capturar su Vehículo con el Piloto

¿Cambia el vehículo el comportamiento del producto? ¿Son correctos los registros de tiempo y texto?

Paso 14— Abrir el Registro de Eventos Externos

Se establece un proceso para monitorear eventos del producto y de la entidad.

Paso 15— Confirmar el Fijación de Onda

Se asigna el estado WS-1 — FIJADO.

Paso 16— Abrir los Espacios

A los participantes solo se les permite enviar en sus propios espacios.

Paso 17— Registrar el Resultado del Primer Contacto

Se separa respuesta, rechazo, error o límite.

Paso 18— Capturar la Primera Salida Adecuada

La respuesta completa se conserva sin renovación.

Paso 19— Crear Evidencia y Metadatos en Crudo

Se registran capturas de pantalla, texto en crudo, estado del sistema y marcas de tiempo.

Paso 20— Realizar Hash Local y Carga Segura

Los archivos reciben un registro de integridad antes de la carga.

Paso 21— Aceptación del Servidor y Creación de Hash del Paquete

La cadena de evidencia se cierra con la hora del servidor y el ID del paquete.

Paso 22— Revisión Ciega de la Validez de la Captura

El estado CV se asigna sin conocer el éxito semántico.

Paso 23— Realizar Revisión de Duplicados y Manipulación

Se utilizan múltiples señales y revisión humana.

Paso 24— Examinar el Sistema Dentro de la Onda y Eventos Externos

Determine si se requiere un sub-onda o un estado mixto.

Paso 25— Asignar Salud de la Onda

Se asigna estado entre WH-0 y WH-5.

Paso 26— Generar Manifestación de la Onda y Raíz de Integridad

La lista de observaciones está versionada.

Paso 27— Publicar Manifestación de Prueba Pública

El alcance y la integridad se hacen visibles sin datos personales.

Paso 28— Aprobar Informe de Cierre de la Onda

Se cierran objetivos, ocurrencias, defectos, eventos y cambios con aprobación humana.

79. EVIDENCIA NECESARIA

ID de onda Tipo de onda Versión del protocolo Versión del registro del sistema AI Marco de población Fijación de muestra Fijación del estado del panel Versión del Registro de prompts Versión del paquete de referencia Hora de inicio y fin UTC Lista microfranja temporal Asignación de participante–slot Fuente de tiempo Desviaciones del reloj del dispositivo Tiempos de transmisión Primeros tiempos de contacto Tiempos de finalización Tiempos de captura Tiempos hash Tiempos de subida Tiempos de aceptación del servidor Resultados del primer contacto Primeras salidas de contenido adecuado Registros de reintento técnico Registros de respuestas parciales Límites de rechazo y uso Respuestas completas en bruto Capturas de pantalla completas Manifiestos visuales secuenciales Grabaciones de pantalla si los hay Citación y registros de fuentes Configuración del sistema Evidencia Herramienta de captura y versión Método de captura Nivel NCL Hashes de archivos Hash del paquete

Firma digital Marca de tiempo Comprobaciones de duplicados Comprobaciones de manipulación Capturar decisiones de validez Archivos de redacción Enlaces de bruto a redactado Manifiesto de privacidad Registros de retiro Eventos de cambio de producto Registro de evento externo Subdivisiones de ola Alcance de ranura Alcance de país e idioma Tasas de error técnico Tasa de validez de captura Estado de salud de la ola Raíz de integridad de la ola Manifiesto de evidencia pública Informe de cierre de ola Registro de cambios Persona o institución responsable

80. LISTA DE VERIFICACIÓN DE AUDITORÍA

¿Está claro el tipo y propósito de la ola? ¿Se han fijado los productos de IA y los estados del sistema? ¿Se han determinado la población, la muestra y las versiones del panel? ¿Se bloqueó el conjunto de prompts antes de los resultados? ¿Está abierta la ventana de la ola UTC? ¿Se preasignaron los micro-espacios? ¿Se distribuyeron los países y los idiomas de manera uniforme entre los espacios? ¿Cambiarán los participantes de espacio según sus resultados? ¿Se evaluó la carga de la hora local? ¿Hay horas de ola rotativas? ¿Están claros los tiempos extra de envío y finalización? ¿Se comparó el reloj del dispositivo con una fuente confiable UTC? ¿Se separaron los tiempos de envío, finalización, captura y carga? ¿Las observaciones fuera de la ola ingresaron a la puntuación principal? ¿Se registró el resultado del primer contacto? ¿Se distinguió el error técnico del rechazo?

¿Se realizó un reintento debido a una respuesta incorrecta? ¿Se preservó el primer intento en el reintento técnico? ¿Se eliminaron respuestas parciales? ¿Es realmente la primera salida adecuada la primera salida? ¿Se utilizó Regenerar o un mensaje de seguimiento? ¿Muestra el prompt visual y toda la respuesta? En respuestas largas, ¿hay captura sin espacios? ¿Coincide el texto en bruto con lo visual? Si se utilizó OCR, ¿hubo verificación humana? ¿Se conservaron versiones de respuesta dinámicas? ¿El usuario detuvo la respuesta? ¿Están vinculados el prompt y la respuesta a la misma sesión? ¿Se registró la herramienta de captura y la versión? ¿La herramienta de captura alteró el comportamiento del usuario? ¿Se indicó el nivel NCL? ¿Hay observaciones por debajo de NCL-3 en el análisis principal? ¿Cada archivo tiene un hash? ¿Está claro qué no prueba el hash?

¿Existe una firma digital o la hora del servidor? ¿El momento de la captura es diferente del momento de la subida? ¿Hay alguna razón para la subida tardía y un registro local de integridad? ¿Se realizó censura sobre la evidencia sin procesar? ¿Tienen los derivados editados hashes separados? ¿El manifiesto público revela la identidad del participante? ¿Se puede reidentificar a un usuario de un país pequeño? ¿Se capturaron innecesariamente conversaciones privadas en el panel natural? ¿Se proporcionó una ruta de evidencia alternativa para usuarios con accesibilidad? ¿La decisión sobre duplicados se basó únicamente en el mismo texto? ¿La decisión sobre manipulación lleva múltiples señales? ¿El validador de captura conocía la puntuación semántica? ¿La respuesta correcta cubrió el fallo de la evidencia? ¿Se invalidó la respuesta incorrecta? ¿El producto cambió durante la oleada?

¿Se procesó el cambio de producto como una bajo-onda o en estado mixto? ¿Se registraron eventos externos? ¿Los eventos externos causaron exclusión según el resultado? ¿El estado de salud de la onda es independiente de la puntuación? Después de que la onda se cerró, ¿se agregaron o eliminaron observaciones? ¿Se conservan el antiguo manifiesto y la raíz de integridad? ¿Cómo se gestionaron las solicitudes de retiro de los participantes? ¿Existe un manifiesto de prueba público? ¿Hay un propietario claramente responsable del informe de cierre de la onda?

81. OBJECIONES Y RESPUESTAS

Objeción 1— “Si todos los usuarios no envían al mismo segundo, no se logra una simultaneidad verdadera.”

La presentación de un segundo puede crear efectos artificiales de carga, conexión y cola. Una ventana corta predefinida UTC y micro-intervalos aleatorios pueden medir el mismo período del sistema de manera más confiable.

Objeción 2— "El producto puede cambiar dentro de una ventana de una hora."

Puede cambiar. Por lo tanto, se requieren registros del sistema, huella de configuración y seguimiento de cambios del producto. En caso de un cambio material, la ola se separa.

Objeción 3— "¿Por qué no es suficiente con una captura de pantalla?"

Captura de pantalla:

- tiempo,
- salida inicial,
- versión del prompt,
- configuración del sistema,
- copiando

no siempre se muestran estos sujetos. Se requiere un paquete completo de evidencia.

Objeción 4— "¿No es demasiado gravoso verificar mil capturas de pantalla una por una?"

Es pesado. Se pueden usar controles automáticos de integridad, revisión humana basada en muestras y verificación basada en riesgos. Los registros críticos o sospechosos se examinan con más detalle.

Objeción 5— "Si hay un hash, ¿por qué no podemos decir que el archivo es real?"

Un hash solo muestra que el archivo permaneció igual después de que se generó el hash. No muestra cómo se creó el archivo antes del hash.

Objeción 6— "Si el usuario tiene que recortar información personal, ¿cómo se protegerá el archivo original?"

Si es posible, la herramienta de Captura no debe capturar información personal desde el principio. Si es inevitable, el archivo original se mantiene en un repositorio restringido y cifrado, y una copia pública se mantiene como un derivado redactado por separado.

Objeción 7— “Si la respuesta es incorrecta, ¿por qué no lo intentamos de nuevo?”

Porque una respuesta incorrecta es una cuestión de medición. Reintentar es solo para una falla técnica real.

Objeción 8— “Si el sistema rechaza, ¿puede el usuario intentar de nuevo?”

Correcto. El panel principal de primer contacto conserva el rechazo. Una Pista de Interacción Natural separada puede medir el comportamiento de seguimiento del usuario.

Objeción 9— “Si un error técnico es un problema de experiencia del usuario, ¿por qué permitimos reintentos?”

El error técnico se conserva como resultado del primer contacto. Se puede realizar un reintento técnico predefinido para examinar adicionalmente el contenido representativo. Los dos resultados no se reemplazan entre sí.

Objeción 10— “¿No es el propio registro del proveedor la evidencia más fuerte?”

Técnicamente, puede ser muy fuerte. El proveedor es el propietario del sistema medido. Es más fiable cuando se utiliza junto con una captura independiente del usuario.

Objeción 11— "¿Será NOMOS Capture un software obligatorio NobleJackal?"

No. NOMOS Capture es un protocolo de evidencia abierto. Las herramientas independientes que cumplan con el esquema y las condiciones de integridad equivalentes pueden ser compatibles.

Objeción 12— "Si no hacemos pública toda la evidencia, ¿cómo se realizará la verificación independiente?"

El manifiesto público muestra la integridad y el alcance. Los auditores independientes autorizados pueden acceder a pruebas crudas redactadas o limitadas. La privacidad y la verificación pueden diseñarse conjuntamente.

Objeción 13— "¿No se interrumpiría todo el trabajo si durante la ola se publican noticias importantes?"

No siempre. El informe es parte del entorno del sistema real. Puede separarse con una marca de tiempo antes y después del evento. El alcance de la puntuación se explica en consecuencia.

Objeción 14— "¿No es más fácil descartar las respuestas antiguas y continuar con el nuevo sistema cuando se actualiza un producto?"

Las respuestas antiguas son reales en la fecha de medición. No se pueden borrar. Los estados del sistema antiguo y nuevo deben mantenerse como olas o sub-olas separadas.

Objeción 15— "Si la herramienta de captura afecta el comportamiento del producto, ¿fracasa todo el método?"

No. El efecto del vehículo se mide con el piloto. Si es necesario, se elige un vehículo más ligero, captura manual o un método que utilice el efecto del vehículo como un factor separado.

Objeción 16— "Si el participante retira sus datos, ¿no se comprometería la raíz de integridad?"

Se crea una nueva versión del manifiesto y una nueva raíz. El registro antiguo puede conservar un rastro SIN CONTENIDO RETIRADO. Se mantiene sobre la arquitectura de evidencia de derechos humanos.

FALLO COMÚN DE LA SECCIÓN 85

Al final de un estudio GEO-1000, puede que tenga 30,000 capturas de pantalla. Este número por sí solo no es evidencia. Necesita saber: ¿En qué ola fueron producidas? ¿A qué estado del sistema pertenecían? ¿Se envió realmente el mismo mensaje? ¿Fue la primera salida? ¿Se actualizó la respuesta? ¿Cambió el producto durante la ola? ¿Cuándo se capturó el archivo? ¿Cuándo se generó el hash? ¿Se repitió la ID del participante? ¿Muestra lo visual toda la respuesta? ¿Coincide el texto en bruto con lo visual? Si la evidencia fue editada posteriormente, ¿dónde está el original? ¿Se realizó la verificación de captura antes de la puntuación semántica? Puede hacer clic mil veces en el mismo segundo. Aun así:

- relojes del dispositivo,
- conexiones,
- colas,
- estados del producto

Puede ser diferente. Por lo tanto, la concurrencia no es un espectáculo de marketing. Es una arquitectura de tiempo predefinida. Cuando un usuario obtiene una respuesta incorrecta: decir 'Intenta

de nuevo.' es tentador. Pero esa respuesta incorrecta es la realidad que estás tratando de medir. Si la borras, no arreglas el sistema. Corrompes la evidencia. Cuando un usuario experimenta un error de conexión real, se puede hacer un reintento técnico. Pero el resultado del primer contacto aún debe conservarse. Porque el usuario encontró el error primero. El hash es valioso. La firma digital es valiosa. La marca de tiempo es valiosa. La captura de pantalla es valiosa. Ninguno de estos por sí solo contiene toda la verdad. Confianza:

- prompts,
- tiempo,
- producto,
- usuario,
- archivo,
- iniciativa,
- verificación

Surge de la preservación conjunta de su cadena. Un estándar: si dice, “Solo la prueba producida con mi software es válida,” antepone su propio interés a la replicación independiente. Por esta razón, NOMOS Capture no es una marca de herramienta, sino una arquitectura de prueba abierta. Las universidades independientes, investigadores o instituciones de auditoría deberían poder aplicar las mismas reglas. Por lo tanto, la duodécima ley de medición de NOMOS es la siguiente:

> *La concurrencia es la reproducibilidad observable dentro del mismo tiempo definido del sistema, no el mismo segundo.*

La ley decimotercera es la siguiente:

> *La primera salida no es un borrador que se pueda renovar hasta que a uno le guste el resultado.*

La ley decimocuarta es la siguiente:

> *El hash conserva la prueba; no crea la prueba.*

La decimoquinta ley es la siguiente:

> *La validez de la captura y la precisión de la respuesta son independientes entre sí.*

Su decimosexta ley es la siguiente:

> *La cadena de evidencia debe hacer visible la integridad de la observación, no la identidad de la persona.*

Su decimoséptima ley es:

> *Si el sistema cambia durante la onda, no puede actuar como si el historial de medición no hubiera cambiado.*

Su décima octava ley de medición establece:

> *Nada está escrito sobre la evidencia cruda; las decisiones se corrigen mediante versiones.*

Su décima novena ley de medición establece:

> *Un método de captura que no puede aplicarse de manera independiente no puede ser un estándar mundial.*

Orden de la Sección 12 de NOMOS

> *No me digas "preguntamos al mismo tiempo." / Muestra en qué ventana UTC, en qué ranura y en qué estado del sistema preguntaste.*

No amontones a mil personas en el mismo segundo para producir un error artificial. / Distribuye el tiempo pero mantén el sistema dentro de la misma onda.

No mezcles los horarios locales entre sí. / No tomes el tiempo del dispositivo como verdad absoluta.

No guarde el resultado del primer contacto. / Guarde el límite de uso, los reintentos, el error y la aclaración por separado.

No regeneres cuando obtengas una respuesta incorrecta. / Esa respuesta incorrecta es mi comportamiento real.

Puedes reintentar cuando ocurra un error técnico. / Pero no elimines el primer error.

No me traigas solo una captura de pantalla. / Trae el prompt, la respuesta completa, el producto, el tiempo, la configuración, la cadena de interferencia y la integridad del archivo.

No recortes la parte buena de tu respuesta. / Deja visible el final, el límite y el error.

No declares el hash como realidad. / Prueba también lo que estaba antes del hash.

No redactes el archivo en bruto en sí. / Conserva el original y produce una copia pública separada.

No publiques el nombre de una persona, cuenta o conversación privada como evidencia.

No hagas que el único participante en un país pequeño sea identificable.

No declares tu herramienta de captura como neutral sin probar si me alteró.

Si el sistema cambió en medio de la ola, no mezcles las respuestas antiguas y nuevas como si pertenecieran a un solo producto.

Si un evento externo lo cambió, registra el evento. / No elimines la observación porque no te gustó la puntuación.

Mantén la evidencia fija. / Si la decisión es incorrecta, corrígela con una nueva versión.

No vincules el estándar de captura al software cerrado de una sola empresa. / Permite que otros puedan establecer la misma cadena de pruebas.

Primero, bloquea la ola. / Luego distribuye el tiempo. / Luego captura la primera salida. / Luego aplícale hash. / Luego súbelo de forma segura. / Luego verifica la captura sin comprobar su exactitud. / Y solo después de esto, decide lo que dice tu respuesta.

La oración de cierre del capítulo

Una respuesta en GEO-1000 se convierte en una verdadera unidad de medida solo cuando se puede mostrar bajo qué estado del sistema, en qué momento, a partir de qué prompt y dentro de qué cadena de pruebas inmutable se originó.

Core normativo

Cada observación GEO-1000 DEBE estar vinculada a una ola de medición predeclarada y versionada que defina: - estados del producto de IA, - marcos de población y muestra, - estados del panel, - versiones de prompts, - ventana de despacho UTC, - micro-espacios aleatorizados, - plazos de finalización y captura, - reglas de reintento técnico, - esquema de captura, - versiones de registros de referencia, - y propiedad de gobernanza. La sincronización DEBE significar observación dentro de una ventana de estado del sistema declarado UTC y estructura de espacio asignada. NO DEBE representarse como simultaneidad perfecta al milisegundo. Cada observación DEBE preservar, según esté disponible: - hora de asignación, - hora de envío del prompt, - primer evento del sistema, - hora de finalización de la respuesta, - hora de captura, - hora de creación del hash, - hora de subida, - y hora de aceptación por el servidor. El resultado del primer contacto DEBE permanecer visible. Un reintento técnico PUEDE utilizarse solo bajo condiciones de falla de infraestructura predeclaradas y DEBE conservar cada intento previo. No se DEBEN regenerar, reemplazar ni excluir salidas incorrectas, desfavorables, rechazadas, no citadas, truncadas o críticas meramente por su contenido. Un paquete de evidencia de captura NOMOS DEBE conservar exactamente el prompt asignado y enviado, la respuesta completa en bruto, la evidencia visual completa, los metadatos del estado del sistema, el registro de intentos, la evidencia temporal, los hashes de archivos, el hash del paquete, la versión de la herramienta de captura, el estado de privacidad y la decisión de validez de captura cegada. Una captura de pantalla, un hash, una firma digital, un registro del proveedor o una marca de tiempo NO DEBEN representarse individualmente como prueba completa de la autenticidad de la observación. La evidencia en bruto DEBE permanecer inmutable o de solo adición. Las redacciones, correcciones, retiros, exclusiones y cambios de adjudicación DEBEN crear versiones vinculadas y preservar el registro anterior donde sea ética y legalmente permisible. Los cambios en productos de IA materiales o eventos externos durante una ola DEBEN registrarse en el tiempo y manejarse mediante sub-olas, modelado explícito, reinicio o prompts de estado mixto. La validez de la captura DEBE decidirse independientemente de la corrección semántica. Los manifiestos de evidencia pública DEBEN apoyar la verificación sin exponer innecesariamente la identidad de los participantes, credenciales, historial de conversaciones privadas o ubicación precisa. El cumplimiento de la captura NOMOS DEBE implementarse mediante sistemas de evidencia abiertos y equivalentes y NO DEBE depender exclusivamente de una herramienta propietaria controlada por el fundador del estándar. Cada ola, conjunto de evidencia, decisión de captura, revisión de manifiesto y registro de integridad DEBE ser versionado y atribuible a un humano o una organización responsable.

PARTE IV

Realidad, adjudicación y decisión

CAPÍTULO 13

El paquete de hechos verificados de la entidad

Paquete de hechos verificados de la entidad

BASES METODOLÓGICAS / K10; K11; K12

Límite de Capítulo

Los primeros doce capítulos establecieron la cadena experimental: una única respuesta no es una puntuación GEO; la representación es una distribución entre usuarios y productos de IA; la entidad auditada debe resolverse antes de observar los resultados; la representación oficial y la realidad de la entidad basada en evidencia deben permanecer separadas; la población objetivo, la muestra, los países y los idiomas deben definirse; los participantes deben seleccionarse independientemente de los resultados; cada instancia de producto de IA debe estar vinculada a un registro con todas las configuraciones materiales; los indicios deben preservar la misma intención en todos los idiomas; los Paneles de Usuarios Controlados y Naturales deben permanecer separados; y cada observación debe preservarse dentro de una onda sincronizada y la cadena de captura de pruebas NOMOS. La auditoría debe ahora responder su pregunta central:

> *¿Según qué registro de referencia determinaremos si la respuesta de IA capturada es correcta, incorrecta, incompleta, desactualizada, no respaldada o fuera de alcance?*

La Sección 4 estableció conceptualmente dos capas separadas de la realidad:

- Capa de Representación Oficial que muestra lo que la entidad misma dice sobre sí misma
- Capa de Realidad de Entidad Verificada que muestra en qué medida la evidencia existente permite qué afirmaciones

Esta sección transforma la segunda capa en un objeto de auditoría operativa. El nombre de este objeto es:

Paquete de hechos verificados de la entidad

Término canónico en inglés:

Paquete de hechos verificados de la entidad

Este objeto operativo completa el tercer eslabón en la arquitectura de los dos primeros libros: el GEO Framework define el estándar; 99 Errores en GEO identifica sus modos de violación; y el sistema de auditoría mide y fundamenta esas violaciones. Todo control auditable debe llevar el modelo o producto, fecha, país, idioma, conjunto de consultas, número de repeticiones y registro de medición. Dentro de esa cadena, el Paquete de la Verdad es:

“¿Qué dijo la IA?”
con la pregunta:

“¿Qué evidencia le permitió decir esto?”
es el sistema de referencia entre ellos. Esta sección:

- qué es y qué no es Truth Pack,
- cómo las afirmaciones se transforman en registros de realidad atómica,
- los tipos de fuentes y evidencia,
- los roles de recursos de primera parte, oficiales, independientes, operativos y de usuarios,
- evidencia en contrario,
- linaje de la fuente e ilusión de independencia,
- suficiencia de la evidencia,
- tiempo, país y límites de alcance,
- afirmaciones de mundo abierto y mundo cerrado,
- la distinción entre "no se encontró evidencia" y "demostrado como falso",
- afirmaciones de rendimiento cuantitativo, cliente, socio, licencia, certificado y superioridad,
- versionado y fijación previo a la ola de Truth Pack,
- capas de evidencia pública y restringida,
- procesos de objeción, corrección y disenso,
- Truth Pack legible por máquina

define. Este capítulo aún no:

- todos los detalles de adjudicación de dividir las respuestas de IA en afirmaciones atómicas,
- acuerdo entre evaluadores,
- umbral de transición de respuesta,
- puertas de error crítico,
- no finaliza las fórmulas de puntuación definitivas NOMOS

no se finaliza. Estos son los temas de las siguientes secciones. La pregunta clave del Capítulo 13 es:

> *¿Cómo crear un Paquete de Verdad sobre una entidad que evaluará respuestas de IA, fijado antes del resultado, llevando tanto evidencia como evidencia en contrario juntos, con límites claros de tiempo y alcance?*

NOMOS Desafío

La frase en el sitio web de su empresa es: “Operamos en 25 países.” El producto de IA dice la misma frase al usuario: “La empresa opera en 25 países.” ¿Transmitió la IA correctamente el sitio oficial? Sí. ¿Se ha verificado realmente la afirmación? Aún no lo sabemos. La empresa le envía una tabla.

La tabla contiene los nombres de 25 países. Sin embargo, las columnas no están explicadas: algunos solo tienen visitantes del sitio web. Algunos tienen antiguos clientes. Algunos tienen distribuidores independientes. Algunos tienen eventos únicos organizados. En nueve países, en realidad hay prestación de servicios activa. Solo en tres países hay una entidad legal local.

¿Qué significa 'operar'? ¿Significa tener un cliente? ¿Establecer una empresa local? ¿Poder brindar servicios de forma remota? ¿Entrega a través de un socio? ¿O que un sitio web sea accesible desde ese país? El sujeto, el verbo y el alcance de la afirmación no están definidos.

Por lo tanto, aunque hay un número, aún no hay una realidad verificable. Ahora considere otra afirmación: '98% de nuestros proyectos son exitosos.' Usted tiene 49 presentaciones de proyectos

exitosos. No hay archivos de proyectos fallidos. ¿Cuántos proyectos se completaron en total? ¿Por qué criterios se define el éxito? ¿Entrega a tiempo? ¿Aceptación del cliente?

¿Resultado comercial? ¿Factibilidad técnica? ¿Satisfacción del cliente? ¿Cuál es el periodo? ¿Los últimos tres meses? ¿Toda la historia de la empresa? ¿Solo proyectos seleccionados? Sin conocer el denominador, 98 por ciento no puede demostrarse. Ahora hay esta frase en la página web de otra empresa: "Somos el líder mundial." Esta frase se repite por:

- blog de la empresa,
- comunicado de prensa de la empresa,
- veinte sitios de noticias copiando el mismo comunicado de prensa,
- cinco cuentas de usuario falsas abiertas por la empresa,
- tres textos generados por IA que resumen estos contenidos

.

Hay treinta URL diferentes. ¿Hay treinta pruebas independientes? No. Una única auto-declaración puede haberse replicado en muchas interfaces derivadas.

Esta es precisamente la manipulación identificada en el volumen precedente: texto oculto, usuarios fabricados y comentarios autoescritos publicados en diferentes dominios se retroalimentan en el sistema como si fueran evidencia independiente. Ahora considere el caso opuesto. El sitio web oficial de la empresa indica: 'Ofrecemos servicios únicamente en Türkiye.'

Sin embargo, la empresa comenzó una operación verificada en Alemania hace dos meses. Contratos, facturas y registros oficiales respaldan ese hecho, aunque el sitio web no se ha actualizado. La IA responde: 'La empresa ofrece servicios en Türkiye y Alemania.' La respuesta se aparta del sitio web oficial, pero puede estar más cerca de la realidad material.

Ahora considere una situación aún más sensible. En el sitio web de una institución de salud: se lee “Centro con licencia.”. En el registro, otra institución con un nombre similar tiene una licencia. La IA fusiona las dos entidades y asigna la licencia a la institución incorrecta. La empresa nota este error pero no lo corrige porque le beneficia.

La disposición explícita del segundo libro se aplica aquí: Un error que favorece a uno sigue siendo un error. Para evaluar una respuesta de IA, no solo necesitamos una lista de oraciones correctas, sino también la siguiente estructura: ¿De quién es la declaración? ¿Cuál es exactamente la afirmación? ¿En qué período de tiempo es válida? ¿En qué país? ¿Para qué producto o servicio?

¿Realmente la fuente respalda la afirmación? ¿Quién controla la fuente? ¿Otras fuentes dicen lo mismo, o están copiando la misma fuente original? ¿Está evidencia en contrario? ¿La afirmación es solo una autodeclaración oficial? ¿La evidencia es pública o privada?

Al momento de la medición de la IA, ¿era posible acceder a esta evidencia? ¿El resultado es seguro, limitado, controvertido o desconocido? La primera regla de esta sección es:

> *Una fuente por sí sola no es la realidad.*

Su segunda disposición establece:

> *No se puede verificar una afirmación sin definir su sujeto, alcance, tiempo y carga de la evidencia.*

Su tercera disposición establece:

> *Diferentes URLs no significan evidencia independiente diferente.*

Su cuarta disposición establece:

> *La ausencia de evidencia no es una prueba automática de que la afirmación sea falsa.*

Su quinta disposición establece:

> *La realidad verificada no es una verdad absoluta e inmutable; es un juicio inferido sostenido por la evidencia disponible dentro de un tiempo y contexto específicos.*

Su sexta disposición establece:

> *No se puede redactar un Paquete de Verdad basado en la respuesta de la IA. La respuesta de la IA se evalúa con respecto al Paquete de Verdad previamente fijado.*

1. OBJETIVO DEL CAPÍTULO

El propósito de esta sección es transformar afirmaciones materiales sobre la entidad auditada en un paquete de referencia que esté vinculado con evidencia y evidencia en contrario, con límites claros de tiempo y alcance, sin convertirlas en una plantilla de respuesta. La sección normaliza las siguientes distinciones:

- Afirmación oficial versus realidad verificada
- Fuente versus evidencia
- Evidencia versus apoyo a la afirmación
- Registro de primera parte frente a verificación independiente
- Fuente independiente y fuente autorizada sobre el tema
- Nombre de dominio diferente y raíz de evidencia diferente
- Número de fuentes y familia independiente de evidencia
- Evidencia directa y evidencia derivada
- Evidencia de apoyo y evidencia en contra
- Ausencia de afirmación y falsedad de la afirmación
- afirmación de mundo abierto y registro de mundo cerrado
- Realidad actual y realidad histórica
- Tiempo cuando la realidad es válida y tiempo cuando la fuente fue publicada
- Evidencia pública y evidencia restringida
- Evidencia accesible para IA y evidencia solo abierta al auditor
- afirmación probado y afirmación públicamente descubrible
- Realidad directa y resultado derivado calculado
- Datos en bruto y relación de rendimiento
- Numerador y denominador
- Cliente y cliente potencial
- Socio y cliente
- Membresía y acreditación
- Licencia y relación con la marca
- Certificado y cumplimiento institucional general
- afirmación y opinión
- Opinión y ventaja comparativa
- Queja y violación verificada
- Investigación y delito o responsabilidad definitiva
- Registro desactualizado y registro histórico incorrecto
- Respuesta AI incorrecta con falta de Paquete de Verdad

- Fallo de referencia y fallo del sistema
- Integridad del Paquete de Verdad y todas las afirmaciones siendo positivas
- Confidencialidad de la evidencia y ausencia de evidencia
- Derecho de la entidad a responder y auto-verificación
- Aprobación humana e infalibilidad humana
- Asistencia de IA y la IA siendo la máxima autoridad de la verdad

Al final de esta sección, cada auditoría debería poder responder estas preguntas:

> *¿Qué afirmaciones materiales están contenidas dentro de Truth Pack?*
>
> *¿A qué entidad, país, producto y momento pertenece cada afirmación?*
>
> *¿Qué evidencia la respalda, qué la limita o la refuta?*
>
> *¿Son las fuentes realmente independientes, o derivan de la misma fuente original?*
>
> *¿Se puede verificar públicamente la afirmación?*
>
> *¿Está la evidencia solo disponible bajo acceso restringido de auditor?*
>
> *¿Podría la IA haber accedido a esta evidencia en el momento de la medición?*
>
> *¿La afirmación es cierta, parcial, histórica, controvertida o desconocida?*
>
> *¿Estaba Truth Pack fijado antes de que se vieran las respuestas de la IA?*

2. DISPOSICIÓN NORMATIVA CENTRAL

Cada auditoría GEO-1000 debe fijar las afirmaciones de referencia de material sobre la entidad auditada, junto con la evidencia de respaldo, evidencia en contrario, linaje de la fuente, clase de acceso, alcance, jurisdicción, período de validez y estado de incertidumbre, en un Paquete de hechos verificados de la entidad versionado antes de que las respuestas de IA sean evaluadas semánticamente. Paquete de Verdad:

- el archivo introductorio escrito por la entidad,
- una presentación de ventas que recopila solo pruebas positivas,
- El texto de la opinión del auditor,
- la respuesta ideal solicitada a IA,
- la realidad absoluta sobre todo el mundo,
- un repositorio de referencia que consiste únicamente en una lista de URLs

No lo es. Paquete de la Verdad:

> *Es un paquete de referencia de auditoría que muestra bajo qué evidencia se pueden apoyar, limitar, refutar o dejar desconocidas las afirmaciones materiales en las respuestas de IA.*

3. LÍMITE DE LA PALABRA "VERDAD"

La palabra “verdad” en Truth Pack no lleva la afirmación: “Este paquete conoce todas las verdades absolutas sobre la entidad.” El significado más preciso es:

> *Un registro de la realidad dependiente de evidencia que muestra, dentro de la fecha, alcance y conjunto de evidencia especificados, hasta qué punto se apoyan ciertos juicios materiales.*

Truth Pack:

- sujeta a cambios con nueva evidencia,
- puede ser disputada,
- su alcance puede ser limitado,
- puede ser corregido si se demuestra que está equivocado,

puede dejar algunas afirmaciones como desconocidas. La autoridad de Truth Pack:

- La palabra "Verdad" no surge de,
- el nombre NobleJackal,
- la voz de NOMOS,
- el autor del libro

Surge de lo siguiente:

- una estructura clara de afirmaciones,
- evidencia apropiada,
- evidencia en contrario,
- linaje de fuentes,
- tiempo,
- límite,
- es significativo con
- revisión independiente,

responsabilidad humana responsable.

4. ¿QUÉ ES Paquete de hechos verificados de la entidad?

Paquete de hechos verificados de la entidad es un paquete de auditoría versionado que reúne los registros fácticos materiales de una entidad específica dentro de una fecha y alcance específicos, las pruebas de apoyo y en contra relacionadas con ellos, los estados de las fuentes y las áreas no resueltas. Representación candidata:

```
VETP_{E,v,t}= (I, C, O, E⁺, E⁻, L, S, T, U, G)
```

Aquí:

- E: entidad auditada
- v: versión del Paquete de Verdad
- t: fecha de referencia
- I: identidad de la entidad y gráfico de relaciones
- C: declaraciones de referencia atómicas
- O: Registro de Representación Oficial
- E^+ : evidencia de apoyo
- E^- : evidencia en contra y registros restrictivos
- L: linaje de la fuente y gráfico de independencia
- S: matriz de alcance, país y jurisdicción
- T: registros de tiempo y moneda
- U: desconocidos, discrepancias y vacíos de referencia

- G: registros de gobernanza, versión, objeción y aprobación

Cada componente de Truth Pack puede tener su propia versión por separado. La versión principal del paquete vincula las versiones de los componentes.

5. TRUTH PACK NO ES UNA PLANTILLA DE RESPUESTA

Truth Pack no es un texto que la IA deba decir palabra por palabra. Por ejemplo, Truth Pack puede contener la siguiente afirmación de referencia: “Asteron Holdings Ltd. es el propietario de la marca Asteron Travel.” La IA puede expresarlo correctamente de estas formas:

- "La marca Asteron Travel pertenece a Asteron Holdings."
- "Asteron Holdings es el propietario de la marca Asteron Travel."
- "Asteron Travel es una marca propiedad de Asteron Holdings."

Se preserva la relación material. El árbitro no debe buscar igualdad de palabras. Truth Pack:

define el campo de expresión correcto

No impone una sola oración correcta.

6. TRUTH PACK NO ES UNA BIOGRAFÍA DE ENTIDAD

Puede haber una cantidad infinita de información correcta sobre una entidad. GEO-1000 Truth Pack no está obligado a recopilar toda la información. El paquete debe cubrir principalmente estas áreas:

- Deberes materiales dentro del Registro de prompt
- La identidad y el núcleo operativo de la entidad
- Límites que cambiarán la decisión del usuario
- Licencias, registros legales, de precio, de servicio, de país y de tiempo
- Criterios de elegibilidad y exclusión requeridos para el asesoramiento
- afirmaciones materiales que conllevan el riesgo de un error crítico
- Diferencias entre la realidad verificada por representación oficial y la realidad actual

A este principio:

Principio de la Realidad Proporcional al Deber

se llama. Recopilar datos innecesarios:

- privacidad
- costo,
- actualidad,
- contradicción

puede aumentar el riesgo.

7. DOCE COMPONENTES CLAVE DEL PAQUETE DE LA VERDAD

7.1. Identidad de Entidad Canónica

Como se define en la Sección 3:

- identidad de la entidad,
- tipo,
- nombre legal,
- relaciones de marca registrada y nombre de dominio,
- empresas matrices y subsidiarias,
- límites de tiempo y jurisdicción

están empaquetados.

7.2. Registro de afirmaciones de Referencia

Lleva todas las afirmaciones de realidad material que se usarán en la auditoría en una estructura atómica.

7.3. Registro de Representación Oficial

Transporta los estados de cuenta actuales e históricos publicados por la propia entidad en interfaces canónicas.

7.4. Registro de Pruebas

Para cada objeto de evidencia:

- fuente,
- integridad,
- autoridad,
- acceso,
- alcance,
- tiempo

que se registra.

7.5. Registro evidencia en contrario

La afirmación:

- refutando,
- limitando,
- alterando su vigencia,
- proporcionando otra explicación

contiene registros.

7.6. Gráfico de Linaje de Fuente

De las fuentes:

- la primera raíz,
- copias,

- citaciones,
- sindicación,
- derivados de IA

se muestran.

7.7. Matriz de Alcance y Jurisdicción

La afirmación:

- para qué país,
- producto,
- servicio,
- usuario,
- entidad legal

es válida.

7.8. Matriz de Tiempo y Moneda

La afirmación:

- cuando comenzó,
- cuando terminó,
- cuando fue verificado por última vez,
- qué evento podría cambiarlo

se muestran.

7.9. Libro de Controversias e Incertidumbres

Hace visibles las áreas que son contradictorias, irresolubles o insuficientemente evidenciadas.

7.10. Capa de Acceso a Evidencia

La evidencia:

- es pública,
- auditable,
- restringida,
- confidencial,
- derivada

según se indica.

7.11. Registro de Cambios y Objeciones

Contiene nueva evidencia, correcciones, opiniones contrapuestas y cambios de decisiones de manera versionada.

7.12. Manifiesto de Máquina

Conecta todos los componentes:

- identidad,
- es significativo con
- hash,
- relación

entre sí.

8. afirmación DE REFERENCIA ATÓMICA

La unidad básica de la realidad dentro de Truth Pack:

se llama afirmación de Referencia Atómica.

Estructura candidata:

```
C_j= (s, p, o, q, c, l, v_f, v_t, k, m)
```

Aquí:

- s: entidad sujeto
- p: predicado o relación
- o: objeto o valor
- q: calificadores y límites
- c: país o jurisdicción
- l: idioma o localidad, si es material
- vf: inicio de validez
- vt: fin de validez
- k: tipo de afirmación
- m: materialidad o etiqueta de riesgo

Ejemplo: Sujeto: Asteron Türkiye Tourism Ltd. / Predicado: proporciona servicios / Objeto: gestión de viajes corporativos / País: Turquía / Tipo de cliente: clientes corporativos / Inicio: Enero 1, 2026/ Última verificación: Agosto 18, 2026 Esta oración no es lo suficientemente atómica: “Asteron es una empresa líder que proporciona numerosos servicios en todo el mundo.” En esta sola oración:

- ámbito geográfico,
- número de servicios,
- liderazgo,
- identidad de la empresa

están mezclados juntos.

9. TIPO DE afirmación

Cada declaración atómica debe tener un tipo.

CT-01— Declaración de Identidad

Quién es la entidad o a qué entidad pertenece.

CT-02— Declaración de Relación

Relación de propiedad, control, asociación, clientela, empleo o membresía.

CT-03— Declaración de Actividad y Servicio

Qué hace la entidad, qué servicio proporciona.

CT-04— afirmación de Producto

Característica del producto, versión, propietario o disponibilidad.

CT-05— afirmación de Alcance Geográfico

País, región o área de servicio.

CT-06— afirmación Legal y de Licencias

Registro, autoridad, licencia, estado regulatorio.

CT-07— afirmación de Precio y Condiciones Comerciales

Precio, impuestos, suscripción, reembolso, garantía o pago.

CT-08— afirmación de Rendimiento Cuantitativo

Tasa de éxito, tiempo de entrega, número de usuarios, crecimiento o eficiencia.

CT-09— afirmación de Clientes y Socios

Cliente, patrocinador, socio, distribuidor o relación con proveedor.

CT-10— afirmación de Certificado y Premio

Certificado, membresía, premio o acreditación.

CT-11— afirmación de Comparación o Superioridad

Mejor, líder, más rápido, más confiable, número uno.

CT-12— afirmación de Recomendación y Idoneidad

Idoneidad o recomendación para un usuario específico.

CT-13— afirmación Histórico

Organización, propiedad pasada, producto anterior o evento histórico.

CT-14— afirmación de Riesgo, Queja o Disputa

Queja, investigación, tribunal, sanción o argumento.

CT-15— Opinión o Posición Institucional

Opinión, visión o juicio de valor sobre la entidad misma. El tipo de afirmación determina qué evidencia se considera suficiente.

10. MATERIALIDAD de la afirmación

No toda la información correcta dentro del Paquete de la Verdad tiene el mismo impacto en una decisión. La materialidad puede indicarse como un candidato con las siguientes clases:

CANDIDATO-CRÍTICO

CANDIDATO-MAYOR

MODERADO

ASESOR

CONTEXTUAL

El nivel exacto de gravedad del error se determinará en la Sección 15. La etiqueta en el Paquete de la Verdad solo:

- gestiona la prioridad de la evidencia,
- la atención de revisión,
- la frecuencia de actualización

gestiona.

11. OBJETO DE EVIDENCIA

Cada fuente dentro del Paquete de la Verdad debe registrarse como un Objeto de Evidencia, no solo como un enlace. Estructura del candidato:

```
E_k= (r, a, i, p, f, s, t, x, g, h)
```

Aquí:

- r: rol de la fuente
- a: autoridad para la afirmación
- i: independencia
- p: proximidad a la afirmación
- f: reciente
- s: alcance
- t: tiempo
- x: clase de acceso
- g: linaje de la fuente
- h: registro de integridad

Una URL por sí sola no completa ninguno de estos campos.

12. RELACIÓN EVIDENCIA–afirmación

No todos los objetos de evidencia apoyan todas las afirmaciones. La relación entre la afirmación Cj y la evidencia Ek puede ser uno de los siguientes estados:

APOYA

APOYA_PARCIALMENTE

CALIFICA

CONTRADICE

DESACTUALIZA

CONTEXTUALIZA

SIN_RELEVANCIA_SIGNIFICATIVA

RELACIÓN_DESCONOCIDA

Ejemplo: Un documento de registro de marca de una empresa: puede respaldar la propiedad de la marca, no respalda la satisfacción del cliente. Una reseña de un cliente: puede respaldar la experiencia de un solo usuario, no respalda el éxito de 98% entre todos los clientes.

13. ROLES DE RECURSOS

SR-01— Fuente Canonical de Primera Parte

La interfaz oficial controlada por la entidad. Ejemplo:

- Página web canonical
- Política oficial
- Página de precios
- Documento del producto

SR-02— Registro Operativo de Primera Parte

Registros a nivel interno o de transacción de la entidad. Ejemplo:

- Factura
- Contrato
- Registro de transacción
- Registro de entrega
- Datos de rendimiento en bruto

SR-03— Registro Oficial Autorizado

Es el registro que tiene autoridad legal o institucional sobre el tema. Ejemplo:

- Registro de Empresas
- Registro de Licencias
- Institución emisora de certificados
- Decisión judicial
- Registro regulatorio

SR-04— Verificación Bidireccional

Es el soporte de la relación por ambas partes o por contrato. Ejemplo:

- Relación con el cliente
- Asociación
- Patrocinio
- Distribución

SR-05— Fuente Editorial Independiente

Es una publicación o investigación editorialmente independiente de la entidad. La independencia no garantiza precisión.

SR-06— Auditoría o Investigación Independiente

Es una evaluación externa abierta de su método y alcance.

SR-07— Evidencia Transaccional o de Comportamiento

Registros que muestran transacciones reales o comportamiento del usuario.

SR-08— Evidencia Generada por el Usuario

Comentario, queja, reseña o documento del usuario. Es una experiencia individual.

SR-09— Cálculo Derivado

Tasa, clasificación o resumen producido a partir de registros en bruto. Su fórmula y entradas deben existir.

SR-10— Evidencia Sintética

Estos datos han sido creados para fines de demostración del método. No pueden probar la afirmación de una entidad real.

SR-11— Fuente Derivada Generada por IA

Contenido creado por un sistema de IA mediante el resumen de otras fuentes. No reemplaza las fuentes originales.

SR-12— Fuente Desconocida o Irresoluble

La fuente no ha sido verificada o no se puede determinar su origen. Tiene confianza limitada en la verificación de material primario.

14. LA AUTORIDAD DE LA FUENTE DEPENDE DEL TIPO DE afirmación

No existe un ranking universal y único para todas las fuentes. Ejemplo: Para el precio público actual de la empresa:

- página de precios canónica
- pago actual,
- oferta actual

puede ser una fuente sólida. Para la afirmación de la empresa: “Somos el líder mundial.”, la misma página web solo muestra una autodeclaración. No prueba el liderazgo. Para la licencia: la institución que otorga la licencia es más autoritaria que el blog de la empresa. Para la experiencia del cliente: una sola reseña de un cliente es un evento real y singular y no representa a toda la población de usuarios. Por lo tanto:

> *La calidad de la fuente no es general; la afirmación es específica.*

15. SIETE DIMENSIONES DEL PERFIL DE FUENTE

Cada fuente debe ser evaluada según estas dimensiones:

15.1. Autoridad

¿Está la fuente autorizada para emitir un juicio sobre el tipo de afirmación?

15.2. Proximidad

¿Qué tan cerca está la fuente del evento, proceso o registro?

15.3. Independencia

¿Qué tan independiente es la fuente del propietario comercial o institucional de la afirmación?

15.4. Integridad

¿Se conserva la autenticidad del archivo o registro y el historial de sus cambios?

15.5. Oportunidad

¿Es válida la fuente para el tiempo relevante?

15.6. Alcance

¿La fuente respalda todo el alcance de la afirmación, o solo una parte de ella?

15.7. Reproducibilidad

¿Puede otro auditor llegar al mismo juicio a partir del mismo registro? Estas dimensiones no necesariamente deben reducirse a un único "puntuación de fuente." Una fuente:

- altamente autorizada,
- baja independencia

posible. Otra fuente:

- altamente independiente,
- distante del tema

puede ser.

16. EL ROL CORRECTO DE LA FUENTE DE PRIMERA PARTE

Fuente de primera parte: no es automáticamente inútil, ni es automáticamente evidencia independiente. Las fuentes de primera parte pueden ser fuertes para afirmaciones tales como:

- Precio publicado por la institución

- Horas de atención
- Descripción de su propio producto y servicio
- Texto de su propia política
- Declaraciones oficiales divulgadas al público
- Sus propios datos de transacciones

Puede no ser suficiente por sí sola para afirmaciones tales como:

- Liderazgo mundial
- Mayor satisfacción
- Superioridad sobre los competidores
- Éxito universal
- Acreditación independiente
- Resultados comerciales para todos los clientes

En el texto anterior 'Errores 99': la advertencia 'Su propio sitio es una afirmación; no una prueba.' es de importancia central para las afirmaciones que requieren particularmente superioridad e independencia. Una declaración normativa más completa es la siguiente:

> *Su propio sitio es prueba de lo que dice la organización; no es prueba automática de la corrección independiente de ninguna afirmación respecto al mundo exterior.*

17. ¿QUÉ ES LA INDEPENDENCIA?

Tener un recurso sobre diferentes nombres de dominio no lo hace independiente. La independencia está relacionada con las siguientes áreas: ¿Quién produjo el contenido? ¿Quién decidió publicarlo? ¿Hay algún pago o patrocinio? ¿Obtuvo la fuente el texto directamente de la empresa? ¿Realizó la publicación su propia verificación? ¿Se copió el contenido de otra fuente? ¿Tiene la fuente propiedad o relaciones comerciales con la empresa? Diez piezas de contenido publicadas por una agencia en diferentes nombres de dominio no son diez fuentes independientes.

18. LINAJE DE LA FUENTE

Los objetos de evidencia deben estar vinculados a un linaje de fuente. Gráfico candidato:

G_S = (V_S, E_S) [K10]

Aquí:

- V_S: objetos fuente
- E_S: relaciones de citación, copia, sindicación, resumen y derivación

Relaciones de fuente:

FUENTE_ORIGINAL

CITAS

SINDICA

RESUME

TRADUCE

PROVIENE_DE

GENERADO_POR_IA_A_PARTIR_DE

LINEA_GENERACIONAL_DESCONOCIDA

puede registrarse como.

19. FAMILIA_DE_EVIDENCIA_INDEPENDIENTE

Los contenidos que se originan a partir de la misma fuente raíz pueden considerarse como una única Familia de Evidencia. Ejemplo:

- Comunicado de prensa de la empresa
- Sitio de noticias que publica el comunicado de prensa textualmente
- Blog que resume el sitio de noticias
- Texto de IA que resume el blog

Hay cuatro URLs. Puede que ni siquiera haya una única familia de evidencia independiente. El número de fuentes y el número de familias de evidencia deben informarse juntos.

20. LAVADO_DE_EVIDENCIA

Re-publicar una autodeclaración en otros dominios y luego usarla como evidencia independiente:

Se llama Lavado de Evidencia

Cadena de ejemplo: La empresa dice "somos líderes" en su propio sitio. Contenido pagado publica esto como "noticia." Otros sitios copian la noticia. Dado que la IA ha visto muchos dominios, acepta la afirmación como generalizada. La empresa utiliza la respuesta de la IA como evidencia de su propia afirmación. Este bucle cerrado:

autodeclaración → fuente derivada → respuesta de la IA → republicación de la autodeclaración

se crea. Debería hacer visible este bucle estableciendo la línea de origen del Paquete de Verdad.

21. LIMITACIÓN DE LA FUENTE GENERADA POR IA

Un sistema de IA:

- puede resumir fuentes,
- puede comparar afirmaciones,

puede generar nuevo texto. Texto de IA:

- no es una fuente original,
- no puede reemplazar la evidencia subyacente,

no se convierte en verificación independiente solo porque otra IA diga lo mismo. La salida de IA puede ser utilizada para los siguientes fines:

- dirigiendo la investigación,
- extracción de afirmaciones,
- detección de conflictos,

asistencia en la búsqueda de fuentes. El objeto de evidencia final debe estar conectado a la fuente original tanto como sea posible.

22. LIMITACIÓN DE LA EVIDENCIA SINTÉTICA

Datos sintéticos:

- demostración del método
- pruebas del sistema
- verificación de software,
- capacitación de adjudicadores

puede ser utilizado para. Datos sintéticos:

- el número real de clientes,
- la tasa de rendimiento real,
- la cobertura real del país,
- la licencia real

no se puede probar. Incluso si los registros sintéticos están en el mismo archivo que la evidencia real, deben estar claramente etiquetados.

23. LA EVIDENCIA DE APOYO ES INSUFICIENTE

Truth Pack solo debe recopilar fuentes que respalden la afirmación. Para afirmaciones materiales, se deben buscar activamente los siguientes registros:

- Fecha conflictiva
- Alcance diferente
- Documento caducado
- Transacción fallida
- Queja o devolución
- Propiedad anterior
- Licencia perteneciente a otra entidad
- Corrección o retiro
- Datos comparativos
- Denominador ausente

Esta obligación se llama:

Carga de evidencia en contrario

Se aplica siempre que exista material evidencia en contrario.

24. evidencia en contrario

evidencia en contrario no siempre refuta completamente la afirmación. Puede cumplir una de las siguientes funciones:

CONTRADICE

LIMITA_ALCANCE

LIMITA_TIEMPO

IDENTIFICA_EXCEPCIÓN

DESAFÍA_MÉTODO

CUESTIONA_AUTENTICIDAD

PROPORCIONA_ENTIDAD_ALTERNATIVA

REQUIERE_REVISIÓN_ADICIONAL

Ejemplo: Una reseña negativa de la afirmación "Todos los clientes están satisfechos" puede no refutar automáticamente toda la afirmación. Sin embargo, puede mostrar que el término "todos" es incorrecto.

25. ALMACENAMIENTO DE evidencia en contrario

Registro que debilita la afirmación:

- no a favor de la empresa,
- aparenta baja calidad,
- no se puede eliminar silenciosamente porque su divulgación al público causaría incomodidad

También se puede registrar evidencia en contrario de baja confianza. Su estado está claramente indicado. El adjudicador le da el peso adecuado.

26. ESTADOS DE APOYO A LA AFIRMACIÓN

VS-0— NO EVALUADO

La afirmación aún no ha sido evaluada.

VS-1— afirmación OFICIAL ÚNICAMENTE

La afirmación se basa únicamente en una auto-declaración oficial. No se ha demostrado que sea falsa. No existe verificación adicional.

VS-2— SOPORTE DE PRIMERA PARTE RASTREABLE

La afirmación está respaldado por un registro de operación o transacción de primera parte rastreable.

VS-3— CORROBORADO

La afirmación está respaldado por múltiples piezas de evidencia apropiadas. No todas las fuentes necesitan ser independientes.

VS-4— VERIFICADO INDEPENDIENTEMENTE

La afirmación ha sido verificada con una fuente externa independiente o autorizada adecuada al tema.

VS-5— SOSTENIDO POR MÚLTIPLES FUENTES

La afirmación está fuertemente respaldado a través de diferentes tipos y familias de fuentes, en alcance y a lo largo del tiempo.

VS-C — CONTRADICHO

Evidencia más sólida o más apropiada contradice la afirmación.

VS-U — NO RESUELTO

La evidencia es insuficiente, contradictoria o a un nivel donde no se puede tomar ninguna decisión.

VS-W — RETIRADO

El reclamante ha retirado la afirmación. El registro histórico se mantiene.

VS-R — VERIFICADO DE MANERA RESTRINGIDA

La afirmación está respaldada por evidencia restringida al acceso del auditor. La verificación pública no es posible. Este estado también afecta si se puede esperar que la IA produzca la afirmación a partir de fuentes públicas.

27. EL ESTADO DE APOYO POR SÍ SOLO NO ES SUFICIENTE

Una afirmación al mismo tiempo:

- respaldada,
- solo para un país específico,
- solo para un rango de fechas específico,
- con evidencia no disponible públicamente,
- bajo puntos de vista opuestos

Es posible. Por lo tanto, en lugar de un solo estado, se puede usar un Vector de Estado de Realidad multidimensional:

```
V_j= (apoyo, alcance, tiempo, conflicto, acceso)
```

Ejemplo:

VS-4+ ALCANCE-LIMITADO + ACTUAL + SIN-CONFLICTO + PÚBLICO

es otro resultado.

VS-R + afirmación-GLOBAL + TIEMPO-LIMITADO + DISPUTADO + RESTRINGIDO

es otro resultado.

28. ESTADOS DE ALCANCE

SC-0— ALCANCE DESCONOCIDO

No se puede determinar el alcance al que se aplica La afirmación.

SC-1— ALCANCE EXACTO

La afirmación y la evidencia tienen el mismo alcance de entidad, producto, país y usuario.

SC-2— ALCANCE PARCIAL

La evidencia solo respalda parte de la afirmación.

SC-3— afirmación DEMASIADO AMPLIO

La afirmación es más amplia de lo que la evidencia respalda.

SC-4— entidad incorrecta ALCANCE

La evidencia pertenece a otra entidad o afiliado.

SC-5— JURISDICCIÓN INCORRECTA

La evidencia pertenece a otro país o área legal.

SC-6— PRODUCTO O SERVICIO INCORRECTO

La evidencia pertenece a un producto o servicio diferente.

29. ESTATUTOS DE TIEMPO

TS-0— TIEMPO DESCONOCIDO

El período de validez de la afirmación es desconocido.

TS-1— ACTUAL

Se respalda que es válido en la fecha de referencia.

TS-2— HISTÓRICO

Se respalda para un período pasado específico.

TS-3— EXPIRADO

La validez del documento o afirmación ha expirado.

TS-4— SUSTITUIDO

Ha sido reemplazado por un registro más reciente.

TS-5— DEPENDIENTE DEL EVENTO

Es válido mientras un evento o condición continúe.

TS-6— FUTURO O PLANIFICADO

Es un plan o compromiso que aún no ha ocurrido. No puede presentarse como realidad actual.

30. CUARTA MARCA DE TIEMPO

Cada evidencia y afirmación debe llevar cuatro tiempos distintos en la medida en que sea relevante:

30.1. Tiempo en que la realidad es válida

¿Cuándo fue una afirmación alguna vez verdadero en el mundo?

30.2. Tiempo en que la fuente fue publicada

¿Cuándo se presentó la información al público o al auditor?

30.3. Tiempo en que se captura la fuente

¿Cuándo se registró la evidencia de la auditoría?

30.4. Tiempo de Adjudicación de la afirmación

¿Cuándo se dio el estado de la realidad? Esta distinción es importante. Una licencia puede haber sido cancelada en 10:00. El registro puede haber sido actualizado en 15:00. La respuesta de la IA puede haber sido dada en 12:00. Respuesta:

- antiguo en términos de la realidad mundial,
- comprensible en términos de evidencia accesible públicamente

Truth Pack debería distinguir entre los dos casos.

31. ACCESIBILIDAD PÚBLICA DE LA EVIDENCIA

Clases de acceso a la evidencia:

EA-1— PÚBLICO

Disponible públicamente y auditable.

EA-2— PÚBLICO PERO TÉCNICAMENTE LIMITADO

Público pero:

- acceso,
- gografía,
- formato,
- acceso técnico

tiene limitaciones.

EA-3— ACCESO DE AUDITORÍA

Puede estar disponible de manera controlada para auditores independientes.

EA-4— CONFIDENCIAL RESTRINGIDO

El contrato está restringido debido a datos personales o secretos comerciales.

EA-5— RESUMEN PÚBLICO DERIVADO

La evidencia en bruto es confidencial. Su método y resumen son públicos.

EA-6— NO DISPONIBLE O PERDIDO

Se afirma que la evidencia existe, pero no se pudo acceder a ella durante la auditoría.

32. AUTENTICIDAD DISPONIBLE PÚBLICAMENTE Y AUTENTICIDAD RESTRINGIDA

Una empresa puede confirmar la afirmación "X es nuestro cliente" con un contrato privado. El contrato no puede ser divulgado al público. Truth Pack puede registrar la afirmación como: VS-R — Verificada de manera restringida. Sin embargo, no se puede esperar que un producto de IA público conozca esta relación. Por lo tanto, hay dos preguntas separadas: ¿Es verdadera la afirmación? ¿Es públicamente descubrible la afirmación? Una afirmación puede ser verdadera con evidencia privada. Para el éxito público de GEO, nuevamente:

- es pública,
- puede necesitar ser verificada.
- apropiado para el usuario

puede ser requerido un registro.

33. DESCUBRIBILIDAD PÚBLICA

Para cada afirmación material, el siguiente campo puede ser registrado como candidato:

DESCUBRIBLE_PÚBLICAMENTE

PÚBLICO_PERO_OSCURECIDO

PÚBLICAMENTE_CONFLICTIVO

SOLO_RESTRINGIDO

NO_PÚBLICO

DESCONOCIDO

La realidad de la respuesta de la IA y el acceso a la información pueden evaluarse por separado posteriormente.

34. afirmación VERDADERA PERO NO PÚBLICA

Si un producto de IA presenta correctamente información que no es pública, surgen las siguientes preguntas: ¿Está la información disponible en otra fuente pública? ¿Utilizó la IA personalización o una base de conocimiento privada? ¿Es la respuesta correcta por casualidad? ¿Hay una filtración de

información confidencial o secreta? ¿Se verificó la afirmación solo después del hecho? Truth Pack debería registrar la ruta de la fuente por separado, sin considerar la corrección por sí sola como un logro.

35. afirmaciones DE MUNDO ABIERTO Y MUNDO CERRADO

La ausencia de información en un registro depende del tipo de afirmación.

35.1. Suposición de Mundo Abierto

La totalidad del universo registrado es desconocida. Ejemplo:

- Todos los clientes
- Relaciones comerciales en todos los países
- Todas las quejas de los usuarios
- Todas las publicaciones en medios

La ausencia de un registro no prueba que la afirmación sea falsa.

35.2. Suposición de mundo cerrado

Se puede asumir que el registro autorizado y completo cubre totalmente el universo relevante. Ejemplo:

- Lista de licencias activas en una fecha específica
- Estado oficial de registro de la empresa
- Certificados válidos de una institución certificadora específica
- Si el catálogo de productos publicado es realmente complementario

En este caso, la ausencia de un registro podría ser una evidencia negativa más fuerte.

36. DECLARACIÓN DE AUTORIDAD DE MUNDO CERRADO

Si se va a utilizar una fuente como un récord mundial cerrado, se deben definir los siguientes campos: ¿Qué universo cubre? ¿Para qué fecha? ¿Cuál es la frecuencia de actualización? ¿Podría faltar el récord? ¿Existe un archivo y un registro de cambios? ¿Quién es la institución autorizada? Simplemente ser el "sitio web oficial" no proporciona autoridad de mundo cerrado.

37. DISTINCIÓN ENTRE EVIDENCIA NO ENCONTRADA Y ERROR

Las siguientes tres frases son diferentes:

- "No se encontró evidencia para la afirmación."
- "La afirmación no está respaldada por la evidencia disponible."
- "La afirmación contradice evidencia más sólida."

La primera puede ser el resultado de la investigación. La segunda es el estado de verificación. La tercera es una contradicción material. Esto no es lo mismo:

NO ENCONTRADO

NO VERIFICADO

CONTRADICCIÓN

38. AUSENCIA DE EVIDENCIA

Para establecer la falsedad de una afirmación basada en la ausencia, deben responderse las siguientes preguntas: ¿Es el universo de registros que se está buscando completo? ¿Se buscaron las entidades correctas y las variantes de nombre correctas? ¿Se usaron la fecha y el país correctos? ¿Podría el registro actualizarse con un retraso? ¿Se está utilizando otra entidad legal o marca registrada? ¿Existe un problema de acceso? Sin estas condiciones: no se puede hacer un fallo definitivo como 'No se encuentra en el registro, por lo tanto, no existe la licencia.' Un resultado más preciso podría ser: 'No se pudo verificar en el registro y fecha especificados.'

39. MATRIZ DE AFIRMACIÓN-EVIDENCIA

Tipo de afirmación	Evidencia primaria	Rol del sitio oficial	Requisito adicional
Nombre legal	Registro autorizado	Indica declaración de la institución	Registro actual
Propiedad de marcas registradas	Registro de marca, transferencia o licencia	Muestra uso de la marca	Verificación del propietario
Precio actual	Precio canónico, pago, contrato	Puede ser registro público principal	Impuestos y cargos adicionales
País de servicio	Contrato, operación, factura, registro local	Declaración pública	Definición de "Servicio"
Licencia	Institución de licencias	Autodeclaración	Alcance y duración
Relación con el cliente	Contrato, factura, aprobación bilateral	Declaración unilateral	Fecha y tipo de relación
Asociación	Contrato o declaración bilateral	Autodeclaración	Alcance y duración
Número de empleados	Registros de recursos humanos y empleo	Declaración aproximada	Distinción de grupo/empresa
Tasa de éxito	Datos en bruto, denominador, método	Declaración de resultados	Casos no exitosos
Satisfacción	Encuesta representativa y método	Anuncio de resultados	No respuesta
Certificado	Institución que emite el certificado	Autodeclaración	Alcance, fecha
Membresía	Institución miembro	Autodeclaración	Membresía ≠ acreditación
Liderazgo	Método comparativo y datos	Afirmación de marketing	Universo, criterios, fecha
Recomendación	Perfil del usuario y evidencia	Vista de marketing	Elegibilidad y exclusiones
Queja	Registro verificado de usuario o institución	Respuesta de la compañía	Limitación a un solo incidente
Violación legal	Registro oficial de finalización y estado de autoridad	Opinión de la institución	Investigación ≠ juicio

Esta tabla no es un ranking de fuentes universal e inmutable. Muestra qué tipo de afirmación lleva qué carga de la prueba.

40. AFIRMACIÓN CUANTITATIVA

No se puede verificar una afirmación cuantitativa sin los siguientes campos:

- Elemento medido
- Numerador
- Denominador
- Periodo
- Regla de inclusión
- Regla de exclusión
- Fuente de datos
- Registros faltantes
- Unidad única
- Fórmula de cálculo
- Redondeo
- Fecha de actualización
- Persona responsable

Ejemplo: “98 porcentaje de éxito.” por sí solo no es una afirmación cuantitativa. Es un número de marketing indefinido.

41. NUMERADOR Y DENOMINADOR

Tasa de éxito:

R = S/N

Sea así. Aquí:

- S: unidades consideradas exitosas
- N: todas las unidades elegibles evaluadas

debe ser. Si se eliminan del denominador proyectos fallidos o incompletos, la proporción cambia. El denominador no puede seleccionarse después del resultado.

42. UNIDAD ÚNICA

Las siguientes expresiones son diferentes entre sí:

- clientes 100
- proyectos 100
- facturas 100
- contratos 100
- cuentas de usuario 100
- transacciones 100

El mismo cliente puede haber hecho diez proyectos. “proyectos 100” no es “clientes 100.” Truth Pack debe registrar explícitamente la unidad del número.

43. NÚMERO ACUMULADO Y ACTIVO

"Tenemos 500 clientes." puede tener uno de los siguientes significados:

- 500 clientes únicos en la historia de la empresa
- 500 clientes en el último año
- 500 clientes activos actualmente
- 500 registros de contacto
- 500 Cuentas
- proyectos 500

Estos significados no son los mismos.

44. REIVINDICACIÓN DERIVADA

Una afirmación puede no estar escrita directamente en el documento. Se puede calcular a partir de datos en bruto. Ejemplo: "El tiempo medio de entrega es de 12 días." Este resultado:

- qué proyectos se incluyen,
- la definición de principio y fin,
- proyectos fallidos,
- Valores medianos y extremos

requiere. La reivindicación derivada debe incluir las siguientes áreas:

- Identificadores de datos de entrada
- Fórmula
- Código o método de cálculo
- Versión
- Registro de reproducción

45. AFIRMACIÓN DE RENDIMIENTO

La afirmación de desempeño debería responder a estas preguntas: ¿De quién es el resultado del desempeño? ¿Qué producto o servicio? ¿Qué tipo de usuario o proyecto? ¿Cuál es la línea base? ¿Cuál es el criterio de éxito? ¿Cuál es el período? ¿Cuántos resultados no exitosos hay? ¿Es un caso seleccionado o toda la población? ¿Existe verificación independiente? ¿El resultado es causal o meramente relacional? Un estudio de caso: puede mostrar un resultado posible. No muestra el resultado esperado en todos los clientes.

46. afirmación DE GARANTÍA

La declaración “Resultados garantizados.” requiere las siguientes áreas: ¿Qué se garantiza exactamente? ¿Entrega o resultado comercial? ¿Cuál es el plazo? ¿Cuáles son las excepciones? ¿Cuál es el reembolso o la compensación? ¿Hay terceros fuera del control? ¿Se garantiza el consejo o la clasificación de IA? Una empresa:

- puede garantizar su propio flujo de trabajo,
- entrega específica

No puede garantizar absolutamente el comportamiento futuro de asesoramiento de un producto de IA que no controla.

47. afirmación DEL CLIENTE

Mostrar el logotipo de una organización en un sitio web no demuestra automáticamente una relación con el cliente. El logotipo puede representar una de las siguientes relaciones:

- Cliente
- Cliente anterior
- Cliente potencial
- Participante de evento
- Patrocinador
- Socio
- Proveedor
- Usuario de tecnología
- Empresa del grupo
- Uso no autorizado del logotipo

Truth Pack debe definir claramente el tipo de relación.

48. CONFIDENCIALIDAD DEL CLIENTE

La relación con el cliente puede ser verificada por contrato, pero no puede ser divulgada públicamente. Registro correcto: VS-R — Verificado de manera restringida puede ser. Si la empresa no puede decir públicamente, "X es nuestro cliente," no se espera que la IA conozca esta relación. El número de clientes confidenciales no puede convertirse silenciosamente en autoridad GEO pública.

49. afirmación DE SOCIEDAD

La palabra "Socio" por sí sola no es suficiente. La relación debe incluir las siguientes áreas:

- Tipo de sociedad
- Producto o proyecto
- País
- Inicio y fin
- Aprobación mutua
- Alcance comercial o técnico
- Autoridad y responsabilidad

Una suscripción de software puede no ser una sociedad tecnológica. Un patrocinio de evento puede no ser una sociedad estratégica a largo plazo.

50. afirmación DE LICENCIA Y AUTORIDAD

Una afirmación de licencia debe incluir al menos las siguientes áreas:

- Entidad completa del titular de la licencia

- Institución que otorga la licencia
- Número de licencia
- País o jurisdicción
- Alcance de la autoridad
- Fecha de inicio y fin
- Estado de suspensión o cancelación
- Fecha de verificación

La licencia de otra empresa del grupo no puede ser transferida a la marca. La licencia personal de un empleado no otorga automáticamente autoridad a la institución.

51. CERTIFICADO, MEMBRESÍA Y ACREDITACIÓN

Los siguientes conceptos son diferentes:

- Certificado
- Membresía
- Acreditación
- Reconocimiento
- Participación en formación
- Marca de conformidad
- Insignia de autoevaluación

Membresía en una organización profesional: puede no significar experiencia acreditada por esa organización. Completar una formación: puede no ser una certificación independiente. El término Truth Pack debe usarse según la definición proporcionada por la institución emisora.

52. afirmación DE RECONOCIMIENTO

El registro del reconocimiento debe contener los siguientes campos:

- Institución otorgante
- Año
- Categoría
- Geografía
- Estructura de solicitud o tarifa
- Método de evaluación
- Relación con patrocinador
- Entidad ganadora
- Ámbito del producto o empresa

El premio por solicitud pagada no es automáticamente falso. Tampoco es un liderazgo de mercado independiente. Se debe preservar el estado de la fuente.

53. afirmación DE SUPERIORIDAD

Las expresiones “Mejor.” “Líder mundial.” “Número uno.” “Más confiable.” requieren las siguientes áreas:

- Universo de comparación
- Conjunto de competidores
- Criterios
- Peso
- Datos
- Periodo
- Geografía
- Método
- Independencia
- Incertidumbre

Si estos campos faltan, el estado correcto del Paquete de Verdad usualmente es:

VS-1— afirmación oficial solamente

o: VS-U — Sin resolver.

54. DISTINCIÓN ENTRE OPINIÓN Y HECHO

Esta oración: "Creemos en el diseño centrado en el ser humano." es una declaración de opinión o valor corporativo. No se necesita investigación de mercado para probar esto. Esta oración es diferente: "Nuestros productos son los mejores de la industria en diseño centrado en el ser humano." Se convierte en una afirmación comparativa factual. Truth Pack debe preservar el tipo epistémico de cada afirmación.

55. QUEJA Y REGISTRO NEGATIVO DE USUARIO

Una queja de un usuario:

- podría ser un evento singular real,
- Podría estar incompleta o ser incorrecta,
- Podría haber sido resuelta,

Podría pertenecer a otra entidad. Queja individual: "La empresa victimiza a todos los clientes." no respalda el resultado. Sin embargo, un incidente grave:

- Investigación,
- corrección,
- Riesgo crítico

podría ser iniciado.

56. afirmación, INVESTIGACIÓN Y JUICIO

Estos estados deberían distinguirse entre sí:

- afirmación
- Queja
- Investigación abierta
- Hallazgo administrativo

- Decisión de primera instancia
- Decisión finalizada
- Decisión retirada o revocada
- Respuesta de la institución

Tener una investigación sobre una persona o institución no significa que sea culpable o responsable. Truth Pack no debe usar un lenguaje que exceda el estatus legal.

57. LENGUAJE PROPORCIONAL EN ALEACIONES NEGATIVAS

Truth Pack debe usar el siguiente lenguaje:

- "Se ha alegado."
- "Está siendo investigado por la institución especificada."
- "La decisión especificada se tomó en la fecha indicada."
- "La decisión no ha sido finalizada."
- "La institución niega la alegación."
- No se ha resuelto en los registros existentes.

El siguiente idioma no puede usarse sin registros definidos suficientes:

- Es un estafador.
- Es ilegal.
- Él/Ella es culpable.
- Definitivamente no es confiable.

Esto no es solo un riesgo legal, es el principio de exactitud en la representación.

58. RECOMENDACIÓN REALIDAD

Truth Pack no debería reducir a una empresa en general a un solo estado de: 'recomendado' o: 'no recomendado.' La recomendación depende de las siguientes áreas:

- Necesidades del usuario
- País
- Presupuesto
- Tipo de servicio
- Tolerancia al riesgo
- Capacidad
- Licencia
- Condiciones de exclusión
- Alternativas

Paquete de Verdad para recomendación:

- requisitos de elegibilidad
- condiciones de inapropiedad,
- límites comprobados

proporciona. El consejo final al usuario es deber de decisión del producto de IA.

59. afirmaciones DE LÍMITE

Truth Pack debería incluir no solo lo que hace, sino también lo que no hace. Ejemplo: Solo funciona en Turquía. No proporciona asesoramiento legal. No tiene licencia sanitaria. Los precios son exclusivos de impuestos. El servicio está abierto solo a clientes corporativos. No garantiza resultados. No controla el comportamiento de ciertos terceros. La información de límites es uno de los registros más valiosos para un consejo preciso.

60. BITÁCORA DESCONOCIDA

El paquete de verdad no debe ocultar campos faltantes. Tipos de registros desconocidos:

HECHO_DESCONOCIDO

ALCANCE_DESCONOCIDO

TIEMPO_DESCONOCIDO

ENTIDAD_DESCONOCIDA

LINAJE_DESCONOCIDO

DISPONIBILIDAD_PÚBLICA_DESCONOCIDA

EVIDENCIA_INSUFICIENTE

CONFLICTO_DE_REFERENCIA

NO_APLICABLE

AÚN_NO_EVALUADO

DESCONOCIDO no es un fracaso. Es una protección contra la certeza falsa.

61. Brecha de referencia

Una respuesta de IA puede generar una afirmación material no encontrada en el Paquete de Verdad durante la revisión. En este caso, el evaluador debe usar una de las siguientes opciones:

- La afirmación está fuera del alcance y no es material
- Existe un Brecha de referencia en el Paquete de Verdad
- Se requiere investigación de nuevas pruebas
- La respuesta no puede ser respaldada
- La afirmación pertenece a una entidad falsa

La brecha del Paquete de Verdad no puede convertirse automáticamente en un fallo de IA. Estado correcto:

BRECHA_DE_REFERENCIA

puede ser.

62. EXPANSIÓN DEL PAQUETE DE VERDAD DESPUÉS DE LA RESPUESTA

La respuesta de la IA puede producir una afirmación material previamente no considerada. Se puede investigar la nueva afirmación. Sin embargo, se aplican las siguientes reglas: Se conserva la primera versión del Paquete de la Verdad. La nueva afirmación y la evidencia se agregan en una versión separada. El resultado de la revisión inicial no se cambia de manera silenciosa. Si es necesario, todas las respuestas relacionadas se reevalúan con la nueva versión. Los resultados antiguos y nuevos son visibles en el registro público de cambios.

63. FIJACIÓN PREVIO DE ONDA DEL PAQUETE DE LA VERDAD

Antes de cada oleada de medición principal, se deben fijar los siguientes campos:

- Versión de identidad de la entidad
- Versión del Registro de Representación Oficial
- Libro de Afirmaciones de Referencia
- Registro de Pruebas
- Registro evidencia en contrario
- Gráfico de Linaje de la Fuente
- Matriz de alcance y tiempo
- Libro de Desconocidos
- Clases de acceso a la evidencia
- Hash del paquete
- Aprobación humana
- Tiempo de fijación

A este archivo:

NOMOS Paquete de Verdad Fijado

puede recibir el nombre.

64. PAQUETE DE VERDAD Y FECHA DE MEDICIÓN

Respuesta de IA, como regla:

el estado de la realidad válido en el momento en que se envía el prompt

debería evaluarse en consecuencia. El Paquete de Verdad lleva el marco temporal que representará este momento. La realidad que se forma después: no hace automáticamente que la respuesta pasada sea incorrecta, crea una nueva versión del Paquete de Verdad para la nueva ola.

65. CAMBIO DE REALIDAD DURANTE LA OLA

Empresa:

- cambia su nombre,

- obtiene o pierde una licencia,
- actualiza su precio,
- se expande a otro país,
- descontinúa su producto

entonces el Paquete de Verdad puede cambiar materialmente. En este caso:

- la ola de medición puede dividirse en subolas,
- las versiones antiguas y nuevas del Paquete de Verdad pueden vincularse a respuestas separadas,

Se puede usar la advertencia MIXED_REFERENCE_STATE.

66. MODIFICACIÓN RETROACTIVA DEL PAQUETE DE VERDAD

Si un registro es incorrecto, puede corregirse. Sin embargo:

- la versión antigua no puede eliminarse,
- se debe registrar la razón del cambio,
- se debe mostrar qué respuestas se ven afectadas,

la nueva puntuación debe crear una versión de análisis separada. Corregir la realidad no es lo mismo que reescribir la historia.

67. ARQUITECTURA DE VERSIONES

Formato de versión candidato: Mayor.Menor.Parche

Mayor

Cambio de entidad objetivo, cambio de ontología de afirmación, cambio de estándar de evidencia, cambio de alcance que llevará a la evaluación de las mismas respuestas en diferentes sistemas de realidad

Menor

afirmación de material nuevo Adición de nuevo país o producto Cambio de estado con nueva evidencia Cambio de la realidad histórica o actual

Parche

Corrección ortográfica Metadatos Enlace roto Corrección de hash o registro técnico Cambio de la realidad material no puede almacenarse como un parche.

68. MONEDA Y FRESCURA

No todos los afirmaciones envejecen al mismo ritmo.

FRESCO-1— Estructural o Histórico

Campos con baja variabilidad como la fecha de fundación. Puede ser suficiente la revalidación activada por eventos.

FRESCO-2— Dependiente de eventos

Propiedad, administrador o estado de licencia. Se requiere evento de cambio y verificaciones regulares.

FRESCO-3— Variabilidad media

País de servicio, número de empleados, asociación. Se requiere verificación periódica.

FRESCO-4— Alta variabilidad

Precio, stock, plan, disponibilidad del producto. Debe verificarse a intervalos cortos.

FRESH-5— En vivo o instantáneo

Corte, oferta actual, evento, estado en tiempo real. Se requieren registros cercanos al momento de la medición. Las duraciones exactas deben determinarse según el tipo de afirmación y el riesgo de la industria.

69. FECHA DE ÚLTIMA VERIFICACIÓN

Cada afirmación:

- fecha de última verificación,
- fecha recomendada de re-verificación,
- disparador de cambio

debe llevarse. afirmación sin fecha de última verificación: no se puede asumir como actual.

70. CAPTURA DE FUENTE

Una página web dinámica no debe guardarse solo como URL. Tanto como sea relevante: captura de página,

PDF,

texto completo, captura de pantalla, hora de captura, hash,

URL,

locale, el estado de acceso debe ser preservado. La página puede cambiar más tarde.

71. SI SE PIERDE LA FUENTE

Una fuente puede ser eliminada o volverse inaccesible. Paquete de Verdad:

- puede preservar la instantánea del archivo,
- hash,
- capturar el tiempo

puede ser preservado. Pero el archivo:

- que la página sea accesible públicamente en tiempo real,
- por sí solo puede no probar completamente el alcance del acceso.

Ninguno de estos hechos está completamente establecido sólo por el archivo.

72. CAPAS DE ACCESO Y PRIVACIDAD DE LA EVIDENCIA

Evidencia Pública

Evidencia que es accesible públicamente.

Evidencia de Auditoría

Evidencia editada accesible para adjudicadores y auditores independientes.

Evidencia Restringida

Evidencia restringida que contiene contratos, datos personales o secretos comerciales.

Evidencia derivada

Resumen verificable o cálculo producido a partir de evidencia cruda.

Manifiesto de afirmación pública

Para cada afirmación al público:

- su estado,
- clase de fuente,
- fecha,
- hash de prueba,
- estado de acceso

muestra del manifiesto.

73. USO DE EVIDENCIA CONFIDENCIAL

Si se va a usar evidencia restringida: ¿quién vio la evidencia? ¿qué verificaron? ¿cuánto se puede divulgar al público? ¿puede un segundo auditor independiente acceder a ella? ¿cuál es la duración y el alcance de la evidencia? ¿se puede retirar la evidencia? Debe registrarse. La declaración 'Tenemos evidencia confidencial' por sí sola no constituye soporte de Truth Pack.

74. PAPEL DE LA ENTIDAD AUDITADA

Entidad auditada:

- puede proporcionar evidencia,
- puede objetar registros de identidad incorrectos,
- puede explicar el alcance,
- puede ofrecer acceso controlado a evidencia confidencial,

puede presentar un contraargumento. Entidad auditada:

- no puede otorgar el estatus final de apoyo a su propia afirmación,
- no se puede eliminar evidencia en contrario,
- no puede seleccionar al juez por sí sola,
- no puede eliminar una afirmación con baja puntuación del alcance,

no puede evadir toda la carga de la evidencia llamándola 'secreto comercial'.

75. DERECHO A RESPONDER

Antes de que se publique un material disponible públicamente y un hallazgo adverso, en la medida en que sea aplicable, la entidad auditada tiene derecho a:

- la afirmación relevante,
- la evidencia utilizada,
- evidencia en contrario,
- estatus propuesto,
- tiempo de respuesta

puede ser informado. La falta de respuesta de la entidad no significa que acepte la afirmación. La respuesta:

- como evidencia,
- como contra-opinión,
- como una nueva solicitud de verificación

que se registra.

76. ESTADOS DE APELACIÓN

SIN_APELACIÓN

APELACIÓN_ABIERTA

EVIDENCIA_PRESENTADA

EN_REVISIÓN

APELACIÓN_ADMITIDA

APELACIÓN_PARCIAMENTE_ADMITIDA

APELACIÓN_RECHAZADA

REVISIÓN_EXTERNA_SOLICITADA

CERRADO_CON_DISCREPANCIA

El registro principal no se elimina durante la objeción. Su estado sigue siendo visible.

77. OPINIÓN DISIDENTE

Los adjudicadores o expertos pueden sacar conclusiones diferentes de la misma evidencia. Si la disputa material no se resuelve:

- decisión mayoritaria,
- opinión disidente,
- razones,
- diferencia de evidencia

puede preservarse en el paquete. El paquete de verdad no debe producir un consenso falso.

78. ROLES DE GOBERNANZA DEL PAQUETE DE VERDAD

Se pueden definir los siguientes roles en la medida en que sean relevantes:

- Propietario del registro de entidad
- Editor del libro de afirmaciones
- Registrador de pruebas
- Analista de linaje de fuentes
- Especialista de dominio
- Revisor legal o regulatorio
- Revisor de métodos cuantitativos
- Revisor evidencia en contrario
- Custodio de privacidad
- Adjudicador del paquete de verdad
- Revisor de Apelaciones
- Custodio de versiones
- Aprobador humano responsable

En un pequeño piloto, una persona puede desempeñar múltiples roles. Los conflictos de interés deben ser revelados.

79. EL PAPEL DE LA IA EN EL PAQUETE DE VERDAD

Los sistemas de IA pueden ayudar en las siguientes tareas:

- Extraer afirmaciones de documentos
- Encontrar la misma afirmación en diferentes oraciones
- Estimar la procedencia de la fuente
- Marcar candidatos a contradicción
- Eliminar campos de fecha y alcance
- Proponer la correspondencia evidencia-afirmación
- Detectar campos faltantes
- Crear diagrama de máquina

IA:

- no puede dar el estado final de soporte por sí sola,
- no puede convertir su propio resumen generado en nueva evidencia,
- no puede ser la autoridad única e infalible en una realidad contradictoria,

no puede reemplazar la aprobación humana.

80. NIVELES DE PREPARACIÓN DEL PAQUETE DE VERDAD

TPR-0— SIN PAQUETE

Paquete de Verdad Estructurado no encontrado. No se puede iniciar la adjudicación principal.

TPR-1— BORRADOR SOLO OFICIAL

Solo existen representaciones oficiales y autodeclaraciones. Puede usarse para descubrimiento. No es un Paquete de Verdad verificado.

TPR-2— Paquete de Evidencia Estructurada

Contiene afirmaciones atómicas y evidencia básica. evidencia en contrario y la línea de origen pueden faltar.

TPR-3— Listo para Adjudicación

afirmación, evidencia, evidencia en contrario, alcance, tiempo y desconocidos están listos para adjudicación. El nivel mínimo candidato para el principal GEO-1000.

TPR-4— Revisado de forma Independiente

El paquete ha pasado por una revisión independiente de campo y método.

TPR-5— Replicado y Mantenido

El paquete ha sido utilizado en diferentes auditorías, y se ha ejecutado el sistema de versión y moneda. La auditoría pública principal al menos:

TPR-3

debería apuntar a ese nivel.

81. PUERTAS DE PREPARACIÓN DEL PAQUETE DE VERDAD

Un paquete debe pasar por las siguientes etapas antes de ser considerado TPR-3: ¿Se ha resuelto la identidad de la entidad? ¿Están disponibles los campos de afirmación necesarios para familias de prompts? ¿Son atómicos los afirmaciones de material? ¿Son claros el momento y el alcance de cada afirmación? ¿Llevan los objetos de prueba registros de integridad? ¿Se ha examinado la línea de origen? ¿Se ha buscado evidencia en contrario? ¿Se separa el acceso público del restringido? ¿Son visibles los campos desconocidos? ¿Hay aprobación humana y fijación de versión?

82. COINCIDENCIA DEL ALCANCE de la afirmación CON EL PAQUETE DE VERDAD

Cada familia de afirmaciones debe estar vinculada a los campos requeridos del Paquete de Verdad.

Familia de dispositivos	Campos requeridos del Paquete de Verdad
Espejo Central	Identidad, actividad principal, relación de la entidad

Familia de dispositivos	Campos requeridos del Paquete de Verdad
Resolución de Existencia	Entidad canónica, alias, nombre de dominio, marca
Actividad y Alcance	Producto, servicio, país, cliente y límites
Evidencia y Fuente	Estado de la evidencia, linaje de la fuente, acceso público
Local/Jurisdiccional	Empresa local, licencia, precio, ley
Temporal	Registros históricos y actuales
Recomendación	Cumplimiento, exclusión, riesgo, capacidad
Comparativo	Criterios de comparación y evidencia comparable
Límite	Tareas pendientes, países fuera del alcance, límites de garantía

Si el Paquete de Verdad no cubre el área requerida por la solicitud, la referencia no está lista.

83. PAQUETE DE VERDAD SINTÉTICO APPLE.COM DISPLAY

MUESTRA DE METODOLOGÍA SINTÉTICA / Las afirmaciones, registros y estados a continuación son fabricados únicamente para demostrar el método. No reflejan información real de Apple Inc., la estructura actual de la empresa, ni un resultado de auditoría real. Objetivo canónico: la entidad tecnológica corporativa principal sintética asociada con el dominio apple.com.

83.1. afirmación Sintética C-001

afirmación:

apple.com es una de las interfaces digitales canónicas controladas por la entidad corporativa principal sintética. Tipo: CT-02— Relación / Evidencia:

- Registro de nombres de dominio sintéticos
- Registro oficial sintético de la empresa
- Verificación independiente sintética

Estado:

VS-4+ SC-1+ TS-1+ EA-1

83.2. afirmación Sintética C-002

afirmación:

La entidad principal está relacionada con tecnología para el consumidor, software y servicios digitales. Tipo: CT-03— Actividad / Evidencia:

- Catálogo de productos sintéticos
- Registro corporativo sintético
- Clasificación independiente sintética

Estado:

VS-5+ SC-1+ TS-1+ EA-1

83.3. afirmación Sintética C-003

afirmación:

La entidad es la empresa más innovadora del mundo. Tipo: CT-11— Superioridad / Evidencia:

- Autodeclaración de la empresa sintética
- No hay método de comparación definido
- No hay universo independiente ni criterio

Estado:

VS-1+ SC-3+ TS-0+ EA-1

Forma correcta de IA: "La empresa se posiciona como una empresa de tecnología innovadora." Elevación epistémica incorrecta: "Es la empresa más innovadora del mundo."

83.4. afirmación Sintética C-004

afirmación:

Los precios son los mismos en todos los países.

Evidencia:

Diferentes condiciones de precio e impuestos en las páginas de países sintéticos

Estado:

VS-C + SC-5+ TS-1

83.5. afirmación Sintética C-005

afirmación:

La entidad presenta productos específicos en los mercados especificados.

Estado:

VS-4+ SC-2+ TS-1

afirmación sin alcance por país: "Ofrecido en todo el mundo." no puede ampliarse de esta manera.

84. CASO DEL PAQUETE DE LA VERDAD SINTÉTICA ASTERON

CASO SINTÉTICO — NO ES UNA INSTITUCIÓN REAL

Estructura de la entidad:

- Asteron Holdings Ltd. — empresa matriz
- Asteron Travel — marca
- Asteron Turkey Tourism Ltd. — empresa de ventas local
- Mira Destination Services — socio de entrega independiente
- Asteron Club Antalya — franquicia

84.1. afirmación A-001

“Asteron opera en 25 países.” Evidencia:

- Sitio web oficial: 25 países
- Registro de operación: entrega activa en 9 países
- Registro de antiguos clientes: clientes históricos en 16 países adicionales
- Empresa local: 3 países
- Aceptación de oferta remota: 14 países

Problema: La definición de “Actividad” no está clara. Truth Pack desglosa la afirmación:

A-001 a

Asteron tiene al menos un cliente histórico o contacto comercial en los países 25. Estado:

VS-3+ TS-2+ SC-1

A-001 b

Asteron ofrece servicios activos en los países 9 a la fecha de referencia. Estado:

VS-4+ TS-1+ SC-1

A-001 c

Asteron tiene una operación legal local en los países 25. Estado:

VS-C + SC-3

Sin este desglose, la respuesta de la IA de “países 25” no puede evaluarse de manera justa.

84.2. afirmación A-002

"98% de los proyectos son exitosos." Evidencia presentada:

- 49 casos positivos
- 1 caso fallido
- No hay lista completa de proyectos
- Sin definición de éxito

Estado del Truth Pack: VS-U — Denominador y método insuficientes Formato correcto de IA: "La compañía informa un alto éxito en registros de casos seleccionados; no se pudo verificar el método ni el denominador completo para la tasa general 98%."

84.3. afirmación A-003

“Asteron es una organización global estándar verificada de manera independiente.” Evidencia: El estándar fue publicado por Asteron mismo. No se encontró verificación independiente. Un empleado de una universidad compartió el texto desde su cuenta personal. La universidad no tiene aprobación institucional. Estado: VS-C — La afirmación de verificación independiente está contradicha. Declaración correcta: “Asteron es una organización que publica su propio estándar.”

84.4. afirmación A-004

“Todos los proyectos están garantizados.” Contrato: La entrega de los archivos fuente está garantizada. Los resultados comerciales no están garantizados. No se garantiza ningún consejo de IA. Átomos del Truth Pack: Los elementos de entrega específicos están garantizados bajo el contrato. Los resultados comerciales no están garantizados. No se garantiza el comportamiento de la IA de terceros.

Si la IA dice: “Todos los proyectos están garantizados para entregar resultados.” ocurre una extensión del alcance material.

84.5. afirmación A-005

“Asteron Club Antalya es operado directamente por Asteron Holdings.” Evidencia:

- Acuerdo de franquicia
- Operador local separado
- Licencia de marca

Estado:

VS-C + SC-4

Relación correcta: Asteron Club Antalya es operado por un operador de franquicia independiente bajo una licencia de marca.

85. TARJETA DE RESULTADO PÚBLICO DE TRUTH PACK

La tarjeta pública del candidato debe incluir los siguientes campos:

- ID de Truth Pack
- Identidad de la entidad
- Fecha de referencia
- Versión
- Alcance de la afirmación
- Número de afirmación atómico
- Número de afirmaciones con evidencia disponibles públicamente
- Número de afirmaciones con evidencia limitada
- Número de afirmaciones basadas únicamente en autodeclaración oficial
- Número de afirmaciones contradictorios
- Número de afirmaciones desconocidas
- Número de afirmaciones desactualizados
- afirmaciones con evidencia en contrario
- Objeciones abiertas
- Última actualización
- Fecha de revalidación
- Hash de integridad del paquete
- Persona o institución responsable

Esta tarjeta: no indica que todos los afirmaciones sean verdaderos; muestra la visibilidad del estado de veracidad de los afirmaciones dentro del paquete.

86. DISPOSICIONES NORMATIVAS OBLIGATORIAS

CH13-N01

Cada adjudicación principal GEO-1000 debe basarse en un Paquete de hechos verificados de la entidad versionado.

CH13-N02

Las respuestas de Truth Pack AI deben estar fijadas antes de ser examinadas semánticamente.

CH13-N03

El Paquete de Verdad no puede presentarse como una afirmación absoluta o inmutable de la realidad.

CH13-N04

El Paquete de Verdad debe estar vinculado a una entidad, fecha, país, producto, servicio y alcance del prompt específicos.

CH13-N05

La declaración oficial de la entidad y la realidad verificada por evidencia debe mantenerse como registros separados.

CH13-N06

No se puede considerar que una fuente de primera mano carezca de valor simplemente porque sea de primera mano.

CH13-N07

Una fuente de primera mano no cuenta automáticamente como prueba de superioridad o independencia externa.

CH13-N08

El número de diferentes URL no puede utilizarse como el número de pruebas independientes.

CH13-N09

El origen, la sindicación, la citación, la traducción y la derivación por IA de las fuentes deben registrarse en el gráfico de linaje de fuentes en la medida en que sea relevante.

CH13-N10

No se pueden contar como pruebas independientes las fuentes derivadas de la misma afirmación raíz sin explicación.

CH13-N11

Un resumen generado por IA no puede ser usado como prueba independiente en lugar de las fuentes subyacentes.

CH13-N12

Los datos sintéticos no pueden ser usados para verificar una afirmación sobre una entidad real.

CH13-N13

Cada afirmación significativa de Truth Pack debe incluir los campos de sujeto atómico, predicado, valor, alcance, país y tiempo, en la medida en que sean relevantes.

CH13-N14

Las frases de marketing compuestas y ambiguas no pueden ser utilizadas como registros de realidad atómica.

CH13-N15

Cada pieza de evidencia debe mostrar explícitamente qué afirmación respalda, limita o refuta.

CH13-N16

La autoridad, proximidad, independencia, integridad, actualidad y alcance de una fuente sobre un tema deben evaluarse según el tipo de afirmación.

CH13-N17

La independencia por sí sola no puede considerarse una garantía de precisión o autoridad.

CH13-N18

La oficialidad por sí sola no puede considerarse una verificación independiente para una afirmación sobre el mundo exterior.

CH13-N19

El Paquete de Verdad no puede formarse únicamente a partir de evidencia de apoyo; se debe buscar y preservar activamente el material evidencia en contrario.

CH13-N20

No se puede excluir evidencia en contrario simplemente porque no esté a favor de la entidad auditada.

CH13-N21

La ausencia de evidencia no puede usarse como equivalente a la falsedad de la afirmación.

CH13-N22

NO ENCONTRADO, NO VERIFICADO, NO RESUELTO y CONTRADICHO deben mantenerse como estados separados.

CH13-N23

La prueba de ausencia solo debe utilizarse cuando la autoridad, el alcance, el tiempo y la complementariedad del universo de registros relevante estén claros.

CH13-N24

En afirmaciones de mundo abierto, la ausencia de registros no puede considerarse automáticamente como evidencia de falsedad.

CH13-N25

Las fuentes que afirmen tener un registro de mundo cerrado deben explicar el alcance y las condiciones de actualización de la fuente.

CH13-N26

El tiempo de validez de la afirmación, el tiempo de publicación de la fuente, el tiempo de captura y el tiempo de adjudicación deben registrarse por separado.

CH13-N27

Las afirmaciones actuales e históricas no pueden combinarse sin explicación.

CH13-N28

Pruebas caducadas, modificadas o retiradas no pueden usarse como realidad actual.

CH13-N29

Cada afirmación debe llevar la fecha de última verificación y, en la medida de lo aplicable, un desencadenante de re-verificación.

CH13-N30

Las pruebas disponibles públicamente y las pruebas restringidas de auditor deben tener estados de acceso separados.

CH13-N31

No se puede asumir que una afirmación verificada con evidencia restringida sea información accesible para un producto de IA público.

CH13-N32

Una declaración "existe evidencia confidencial" no se considera respaldada sin acceso del auditor y un registro de integridad.

CH13-N33

La veracidad de una afirmación y su descubribilidad por el público deben evaluarse por separado.

CH13-N34

No se puede verificar una afirmación cuantitativa sin incluir numerador, denominador, período, unidad, inclusión, exclusión y método de cálculo.

CH13-N35

Los estudios de caso positivos seleccionados no pueden presentarse como una tasa de rendimiento general.

CH13-N36

Los números de cliente, proyecto, contrato, transacción, usuario y cuenta no pueden usarse de manera intercambiable.

CH13-N37

Los números acumulativos y activos de clientes o usuarios deben mantenerse separados.

CH13-N38

Los datos de entrada derivados para afirmaciones deben incluir la fórmula y la versión del cálculo.

CH13-N39

No se pueden verificar las relaciones con clientes y socios sin información sobre el tipo de relación, alcance y tiempo.

CH13-N40

El uso del logotipo por sí solo no prueba una relación con un cliente o socio.

CH13-N41

Las características de grupo, subsidiaria, franquicia, distribuidor y marca no pueden transferirse entre sí sin evidencia clara.

CH13-N42

La licencia y el certificado deben estar vinculados con la entidad, país, alcance y fecha correctos.

CH13-N43

La membresía, certificación, acreditación, participación en capacitación y premios no pueden usarse de manera intercambiable.

CH13-N44

Las afirmaciones de excelencia y liderazgo no pueden considerarse hechos verificados sin el universo de comparación, criterios, método, fecha y alcance.

CH13-N45

La opinión corporativa y la posición de marketing no pueden elevarse a hecho independiente.

CH13-N46

La queja, la investigación, el hallazgo administrativo y la decisión final deben distinguirse entre sí.

CH13-N47

Los registros negativos no finalizados no pueden presentarse en el lenguaje de delito definido, fraude o responsabilidad legal.

CH13-N48

La contradeclaración y apelación de la entidad auditada deben estar vinculadas al registro de afirmación correspondiente.

CH13-N49

El hecho de que la entidad no responda no significa que acepte la afirmación.

CH13-N50

Truth Pack debería preservar visiblemente los desconocidos y los vacíos de referencia.

CH13-N51

Una nueva afirmación de IA que no está presente en Truth Pack no puede considerarse automáticamente falsa; se debe realizar una revisión de REFERENCE_GAP.

CH13-N52

El registro de la realidad agregado después de la respuesta de la IA debería crear una nueva versión del Paquete de la Verdad.

CH13-N53

La nueva versión de Truth Pack no puede cambiar silenciosamente el historial antiguo de adjudicación y puntuación.

CH13-N54

Si la realidad material cambia durante la onda, las observaciones relacionadas deben estar vinculadas al marco temporal correcto del Paquete de Verdad.

CH13-N55

Los cambios en Truth Pack deben registrarse de acuerdo con la lógica de versión mayor, menor o de parche.

CH13-N56

Los cambios en la realidad material o en el estado de la evidencia no pueden ocultarse como soluciones parche.

CH13-N57

La evidencia cruda y las versiones antiguas de Truth Pack deben ser preservadas dentro de los límites éticos y legales.

CH13-N58

Truth Pack no puede depender únicamente de los documentos proporcionados por la entidad auditada.

CH13-N59

La entidad auditada no puede determinar el estado final de respaldo de su propia afirmación.

CH13-N60

Los roles de preparar y aprobar el estado final por Truth Pack deben separarse en la medida de lo posible.

CH13-N61

Las herramientas de IA pueden ayudar en la inferencia y la detección de contradicciones; no pueden proporcionar el estado de realidad definitivo por sí solas.

CH13-N62

Cuando el Paquete de Verdad se divulgue públicamente, los datos personales, secretos comerciales y contratos restringidos no pueden publicarse innecesariamente.

CH13-N63

Si se utiliza evidencia limitada, debería estar disponible una vía de revisión independiente o secundaria apropiada.

CH13-N64

El Paquete de Verdad debe mostrar al público los estados de las afirmaciones manifestadas, las clases de acceso, la versión y el registro de integridad.

CH13-N65

Para la adjudicación primaria, el Paquete de Verdad debe tener al menos TPR-3— Listo para Adjudicación o un nivel equivalente justificado.

CH13-N66

El nivel de preparación del Paquete de Verdad debe evaluarse independientemente de la puntuación de la IA.

CH13-N67

La falta del Paquete de Verdad no puede usarse como herramienta para aprobar o reprobar la respuesta de la IA según el resultado deseado.

CH13-N68

Cada versión del Paquete de Verdad debe tener un propietario humano u organizacional responsable.

87. FORMAS DE FALLA

CH13-F01— CONSIDERANDO EL SITIO DE LA EMPRESA COMO REALIDAD ABSOLUTA

Todas las autodeclaraciones se tratan como hechos verificados.

CH13-F02— RECHAZANDO COMPLETAMENTE LA EVIDENCIA DE PRIMERA PARTE

Se ignoran registros institucionales directos como precios, políticas y transacciones porque se consideran de fuente propia.

CH13-F03— CONTANDO EL NÚMERO DE URL COMO EL NÚMERO DE PRUEBAS

Veinte copias del mismo comunicado de prensa cuentan como veinte piezas de evidencia independientes.

CH13-F04— BLANQUEO DE EVIDENCIA

Una autodeclaración es replicada en otros sitios y reutilizada como verificación independiente.

CH13-F05— CONTANDO REPETICIONES DE IA COMO EVIDENCIA

Una afirmación se considera verificada porque múltiples sistemas de IA repiten la misma afirmación.

CH13-F06— AÑADIENDO DATOS SINTÉTICOS A LA EVIDENCIA REAL

La muestra del método entra en el denominador del desempeño real.

CH13-F07— CONTANDO UNA DECLARACIÓN DE MARKETING COMPUESTA COMO UNA AFIRMACIÓN ATÓMICA

Identidad, liderazgo, actividad y país se fusionan en la misma afirmación.

CH13-F08— USANDO EVIDENCIA DE entidad incorrecta

La evidencia de la empresa matriz, subsidiaria o empleado se transfiere a la entidad objetivo.

CH13-F09— PAÍS O ALCANCE DEL PRODUCTO INCORRECTO

Una licencia en un país se generaliza a todo el mundo, un certificado de producto se generaliza a todo el portafolio.

CH13-F10— CONTAR UN REGISTRO ANTIGUO COMO ACTIVO

Se utiliza un precio antiguo, un empleado antiguo o una licencia caducada como actual.

CH13-F11— CONSIDERAR LA FECHA DE PUBLICACIÓN DE LA FUENTE COMO LA FECHA DEL EVENTO

Se confunde la publicación del documento con el momento en que ocurre el evento.

CH13-F12— CONSIDERAR EVIDENCIA PUBLICADA POSTERIORMENTE COMO PREVIAMENTE ACCESIBLE

La información no disponible en la fecha de medición de la IA se evalúa como si estuviera abierta en el sistema.

CH13-F13— CONSIDERANDO LA FALTA DE EVIDENCIA COMO INCORRECTA

Una afirmación no encontrada en la investigación se considera refutada.

CH13-F14— CONSIDERANDO UN MUNDO ABIERTO COMO UN MUNDO CERRADO

Se utiliza una lista incompleta de clientes o socios como un registro suplementario.

CH13-F15— NO VERIFICAR REGISTROS DE MUNDO CERRADO

Se produce una prueba de ausencia sin conocer el alcance y la vigencia del registro.

CH13-F16— NO BUSCANDO evidencia en contrario

Solo se utilizan documentos positivos proporcionados por el reclamante.

CH13-F17— ELIMINANDO SILENCIOSAMENTE evidencia en contrario POR BAJA CALIDAD

Se destruyen registros de baja confianza sin asignar un estado.

CH13-F18— SOBREGNERALIZANDO evidencia en contrario

Un solo incidente negativo se considera suficiente para invalidar todo el rendimiento.

CH13-F19— PORCENTAJE SIN DENOMINADOR

Se utiliza la tasa de éxito o satisfacción como si fuera una verdad universal.

CH13-F20— CONTAR UN CASO SELECCIONADO COMO RENDIMIENTO GENERAL

Solo los proyectos exitosos generan tasas.

CH13-F21— CONTAR EL PROYECTO COMO UN CLIENTE

Múltiples proyectos del mismo cliente se cuentan como clientes separados.

CH13-F22— CONVERTIR UN NÚMERO ACUMULADO EN NÚMERO ACTIVO

El número de clientes adquiridos a lo largo de la historia de la empresa se presenta como clientes activos actuales.

CH13-F23— OCULTAR LA FÓRMULA DEL NÚMERO DERIVADO

El promedio o la tasa no se puede reproducir.

CH13-F24— CAMBIANDO EL DENOMINADOR DESPUÉS DEL RESULTADO

Se seleccionan proyectos apropiados según la tasa de éxito deseada.

CH13-F25— CONTANDO EL LOGO COMO EVIDENCIA DEL CLIENTE

La actividad, el uso de la tecnología o el logo no autorizado se presentan como prueba del cliente.

CH13-F26— CONTANDO LA PALABRA "SOCIO" COMO AUTORIDAD

La simple propiedad de la relación se convierte en acreditación o autoridad de representación.

CH13-F27— CONTANDO LA MEMBRESÍA COMO ACREDITACIÓN

La membresía en la organización se utiliza como aprobación de calidad independiente.

CH13-F28— CONTANDO LA PARTICIPACIÓN EN CAPACITACIÓN COMO CERTIFICACIÓN

La participación en un seminario se convierte en una licencia profesional.

CH13-F29— CONTANDO UN PREMIO PAGADO COMO LIDERAZGO INDEPENDIENTE

Se ocultan la estructura y el método de la recompensa.

CH13-F30— CONSIDERANDO EL PROPIO ESTÁNDAR COMO ESTÁNDAR INDEPENDIENTE

Presentado como acreditación externa como el método publicado por la empresa.

CH13-F31— CONSIDERANDO LA EVIDENCIA LIMITADA COMO PÚBLICA

Un contrato privado se evalúa como accesible para IA como fuente abierta.

CH13-F32— DECLARACIÓN DE "HAY EVIDENCIA SECRETA"

La afirmación se aprueba sin que se vea y verifique la evidencia.

CH13-F33— CONSIDERANDO LA VERDAD CERRADA AL PÚBLICO COMO ÉXITO GEO

Se presenta información especial que la IA no puede conocer como el poder de representación pública de la marca.

CH13-F34— CONSIDERANDO LA OPINIÓN COMO HECHO

La frase “Somos los mejores” se convierte en una realidad independiente.

CH13-F35— CONSIDERANDO LA QUEJA COMO UNA VIOLACIÓN DEFINITIVA

La afirmación de un solo usuario se presenta como una consecuencia legal.

CH13-F36— CONSIDERANDO LA INVESTIGACIÓN COMO DELICTIVA

Un proceso no resuelto se convierte en un juicio final.

CH13-F37— EXCLUYENDO UN REGISTRO NEGATIVO DEL ALCANCE

El registro contrario se elimina para proteger la reputación pública.

CH13-F38— PRESERVANDO EL ERROR FAVORABLE

Dado que es una institución de IA, cuando se muestra como grande, no se aplica el límite del Paquete de la Verdad.

CH13-F39— CONSIDERANDO LA INFORMACIÓN DE LÍMITE COMO INSIGNIFICANTE

Las tareas no realizadas y los usuarios no elegibles no se incluyen en el paquete.

CH13-F40— CONSIDERANDO LO DESCONOCIDO COMO POSITIVO

Las áreas sin evidencia se consideran respaldadas.

CH13-F41— CONSIDERANDO LO DESCONOCIDO COMO NEGATIVO

Las áreas sin evidencia se consideran incorrectas.

CH13-F42— CONSIDERANDO UN Brecha de referencia COMO ERROR DE IA

Una afirmación no presente en el Paquete de la Verdad es un fracaso directo.

CH13-F43— ESCRIBIENDO EL PAQUETE DE LA VERDAD DE ACUERDO CON LA RESPUESTA DE LA IA

La referencia se moldea después para extraer la verdad o falsedad de lo que dijo la IA.

CH13-F44— RESTRINGIENDO EL PAQUETE DE LA VERDAD PARA OBTENER UNA PUNTUACIÓN ALTA

Las afirmaciones difíciles se eliminan del alcance.

CH13-F45— AMPLIANDO EL PAQUETE DE LA VERDAD PARA OBTENER UNA PUNTUACIÓN BAJA

Los detalles inesperados en la respuesta se hacen obligatorios posteriormente.

CH13-F46— ELIMINAR LA VERSIÓN ANTIGUA DEL PAQUETE DE LA VERDAD

El historial de adjudicación se reescribe.

CH13-F47— CONTAR CAMBIO MATERIAL COMO PARCHE

El estado de la afirmación se cambia silenciosamente.

CH13-F48— REFERENCIA ÚNICA EN MEDIO DE LA OLA

Aunque la realidad ha cambiado, todas las respuestas se evalúan según el mismo paquete.

CH13-F49— APLICAR TIEMPO DE EXPIRACIÓN A TODAS LAS afirmaciones DE MANERA IDÉNTICA

La fecha de establecimiento y el precio están vinculados al mismo calendario de renovación.

CH13-F50— CONSERVAR SOLO LA URL

Cuando la fuente cambia, se pierde la evidencia.

CH13-F51— CONSERVANDO LA CAPTURA DE LA FUENTE SIN HASH

No se puede verificar la integridad del archivo.

CH13-F52— FILTRANDO EVIDENCIA PRIVADA AL PÚBLICO

Se publican datos de contrato, cliente o personales.

CH13-F53— EVITANDO LA CARGA DE EVIDENCIA DE MANERA CONFIDENCIAL

La afirmación se considera verificada sin que ningún auditor la haya visto.

CH13-F54— HACIENDO QUE LA ENTIDAD AUDITADA SEA EL ÁRBITRO FINAL

La empresa verifica su propia declaración.

CH13-F55— CONTANDO EL DERECHO A RESPONDER COMO EVIDENCIA

La objeción de la empresa se convierte automáticamente en un cambio fáctico.

CH13-F56— CONSIDERANDO EL NO RESPONDER COMO ACEPTACIÓN

El silencio se convierte en una confesión de culpabilidad o verdad.

CH13-F57— BORRANDO OPINIONES OPUESTAS

El desacuerdo no resuelto entre expertos se cierra con un consenso falso.

CH13-F58— HACIENDO DE LA IA LA AUTORIDAD FINAL DE LA REALIDAD

La IA resume fuentes y trata su propio resumen como prueba.

CH13-F59— HACIENDO QUE LA VERDAD IMPREGNE EL TEXTO DE RESPUESTA IDEAL

El adjudicador busca coincidencias de palabras en lugar de significado.

CH13-F60— CONTROL TOTAL CON EL ARCHIVO TPR-1

Se considera realidad verificada un paquete compuesto únicamente por declaraciones de la empresa.

CH13-F61— INDICANDO EL NIVEL DE PREPARACIÓN SEGÚN LA PUNTUACIÓN

El Paquete de la Verdad incompleto con una puntuación alta de IA es aceptado como listo.

CH13-F62— PAQUETE FUERA DE ALCANCE

Las áreas preguntadas en la auditoría no se encuentran en el Paquete de la Verdad.

CH13-F63— RECOPILACIÓN DE DATOS PERSONALES INNECESARIOS

El paquete de realidad viola la privacidad de empleados o clientes.

CH13-F64— EL MONOPOLIO DEL PAQUETE DE LA VERDAD VINCULADO A UNA SOLA EMPRESA

Las instituciones independientes tienen prohibido aplicar el mismo esquema de evidencia.

88. PROCEDIMIENTO DE AUDITORÍA

Paso 1— Fijar la identidad de la entidad

El Registro de Entidad Canónica en la Sección 3 está verificado.

Paso 2— Eliminar Alcance de la Demanda

Se determina qué dominios de la realidad necesitan las familias de prompts.

Paso 3— Crear el Universo de afirmaciones Materiales

Se extraen las afirmaciones de identidad, actividad, alcance, precio, licencia, cliente, desempeño y límite.

Paso 4— Hacer las afirmaciones Atómicas

Las frases integradas de marketing se dividen en campos de sujeto, predicado, valor, alcance y tiempo.

Paso 5— Crear un Registro de Representación Oficial

Se distribuyen las declaraciones canónicas e históricas de existencia.

Paso 6— Determinar el Tipo de Evidencia

Para cada afirmación, se escribe qué tipo de fuente puede ser autorizada y suficiente.

Paso 7— Reunir Evidencia de Apoyo

Se registran fuentes de primera mano, oficiales, operativas e independientes.

Paso 8— Buscar evidencia en contrario

Se busca activamente evidencia contradictoria, limitante y que altere los registros.

Paso 9— Establecer la Línea de Fuente

La fuente original, la sindicación, la cita y los derivados de IA están vinculados entre sí.

Paso 10— Identificar Familias de Evidencia

Las URL que provienen de la misma raíz se agrupan bajo una sola familia.

Paso 11— Fijar Alcance y Tiempo

Se indica para qué entidad, producto, país y período es válida la afirmación.

Paso 12— Clasificar el Acceso a la Evidencia

Se separa la evidencia pública, de auditoría, limitada y derivada.

Paso 13— Determinar el Estado de Mundo Abierto o Cerrado

Se evalúa si se puede usar la ausencia de evidencia.

Paso 14— Asignar Estado de Apoyo de la afirmación

Se asigna VS-0 a VS-5, VS-C, VS-U, VS-R o el estado relevante.

Paso 15— Asignar Alcance y Estado Temporal

Se agregan las clases SC y TS.

Paso 16— Registrar los Desconocidos

Se dejan visibles las áreas donde no se puede tomar una decisión.

Paso 17— Recalcular afirmaciones Cuantitativas

Se verifican el numerador, el denominador, la unidad y el período.

Paso 18— Enviar afirmaciones Críticas para Revisión por Expertos

Las áreas de derecho, salud, finanzas, licencias y seguridad son revisadas por un experto adecuado.

Paso 19— Otorgar a la Entidad Auditada el Derecho a Responder

Los registros de material y los estados propuestos se informan en la medida en que sean aplicables.

Paso 20— Revisar Objeciones y Nueva Evidencia

Los cambios de estado de la afirmación solo se realizan de manera basada en evidencia y versionada.

Paso 21— Aplicar Puertas de Preparación del Paquete de Verdad

Se determina el nivel de TPR.

Paso 22— Separar Paquetes de Evidencia Públicos y Restringidos

Se aplican límites de privacidad y secretos comerciales.

Paso 23— Crear el Manifiesto de la Máquina

La afirmación, la evidencia, evidencia en contrario, la procedencia de la fuente y los estados están conectados entre sí.

Paso 24— Crear Fijación del Paquete de Verdad

Las respuestas de la IA del paquete se bloquean con un hash y una marca de tiempo antes de ser revisadas.

Paso 25— Rastrear Cambios Durante la Ola

Si la realidad material cambia, se creará una nueva versión del paquete.

Paso 26— Gestionar Huecos de Referencia

Los nuevos afirmaciones planteados por la respuesta de la IA se abordan con el nuevo proceso de lanzamiento.

Paso 27— Publicar la Tarjeta del Paquete de la Verdad Pública

Se hacen visibles el alcance, la distribución del estado y las limitaciones del paquete.

89. EVIDENCIA REQUERIDA

Registro de Entidad Canónica; tipo de entidad; identidad legal; relaciones de marca registrada y nombre de dominio; matriz de alcance inmediato; Libro Mayor de afirmaciones de Referencia; IDs de afirmaciones atómicas; tipos de afirmaciones; etiquetas de materialidad; Registro de Representación Oficial; instantáneas de páginas oficiales; registros legales y regulatorios; registros de licencias y certificaciones; registros de marcas registradas y propiedad; registros de productos y servicios; registros de precios y pago; contratos; facturas; registros de transacciones; verificación de clientes y socios; datos brutos de rendimiento; registros de numerador y denominador; código de cálculo o fórmula; registros de usuarios y quejas; investigación independiente; publicaciones editoriales; información de verificación de fuentes y patrocinio; gráfico de linaje de fuentes; familias de evidencia; registros de fuente derivados de IA; evidencia de apoyo; evidencia en contrario; búsqueda de ausencia; decisión de mundo abierto/cerrado; y fechas de validez de la afirmación.

Fechas de publicación de la fuente Fechas de captura Fechas de revisión por pares Capturas de la fuente Resúmenes de archivos Clases de acceso a evidencia Registro de acceso a evidencia restringida Estado de descubrimiento público Estados de soporte Estados de alcance Estados de tiempo Registros de disputas Desconocidos Libro mayor Huecos de referencia Respuestas de la entidad auditada Registros de objeciones Puntos de vista opuestos Registro de cambios Versión del Paquete de Verdad Hash del paquete Tiempo de fijación Nivel de preparación TPR Manifiesto público del Paquete de Verdad Persona o institución responsable

90. LISTA DE VERIFICACIÓN DE AUDITORÍA

¿Es la entidad objetivo correcta y está versionada? ¿Están todos los dominios de realidad solicitados por las familias de prompts incluidos en el paquete? ¿Son las afirmaciones atómicas? ¿Es correcto el sujeto de cada afirmación? ¿Están claros el alcance, el país, el producto y el tiempo? ¿Se ha separado la realidad verificada con representación oficial? ¿Se utilizaron fuentes de primera mano en el papel correcto? ¿Son realmente independientes las diferentes URLs? ¿Se ha extraído la línea de procedencia de la fuente? ¿Se vincularon las sindicaciones y copias a una sola familia de evidencia? ¿Se vincularon

los contenidos derivados generados por IA a la fuente original? ¿Se mezclaron registros sintéticos con evidencia real? ¿Qué afirmación respalda cada pieza de evidencia? ¿Es la fuente autorizada para el tipo de afirmación? ¿Está la fuente actualizada? ¿Es la fuente para la entidad y el país correctos?

¿Cubre la evidencia todo el alcance de la afirmación o solo una parte de ella? ¿Se buscó activamente a evidencia en contrario? ¿Se ocultaron registros negativos? ¿Está claro el nivel de confiabilidad de evidencia en contrario? ¿Se consideró un error la ausencia de evidencia? ¿Se evaluó una afirmación de mundo abierto como una de mundo cerrado? ¿Es la fuente de inexistencia realmente complementaria al universo? ¿Se registró el tiempo de validez de la realidad? ¿Se separaron el tiempo de publicación y el tiempo del evento? ¿Estaba la evidencia disponible públicamente en el momento de la medición de IA? ¿Se consideró la evidencia limitada como descubrible por el público? ¿Se abrió la "evidencia oculta" a revisión independiente? En afirmaciones cuantitativas, ¿hay numeradores y denominadores? ¿Es correcta la unidad única? ¿Se separaron los números acumulativos y activos? ¿Se generalizaron los casos seleccionados a tasas generales? ¿Puede reproducirse una afirmación derivada?

¿Son las relaciones con clientes y socios bilaterales o contractuales? ¿Se consideró el uso del logotipo como evidencia de la relación? ¿Se han mezclado las características del grupo, la marca y las subsidiarias? ¿La licencia pertenece a la entidad correcta? ¿Se han separado certificados, membresías y acreditaciones? ¿Existe un método de comparación para la afirmación de superioridad? ¿Se ha elevado la opinión corporativa a hecho? ¿Se han separado quejas y decisiones finalizadas? ¿Se ha usado un lenguaje proporcional en las afirmaciones negativas? ¿Hay afirmaciones de delimitación y exclusión en el paquete? ¿Son visibles áreas desconocidas? ¿Hubo un error de IA en el Brecha de referencia? ¿Se bloqueó Truth Pack antes de ver las respuestas de la IA? ¿El paquete se expandió o se contrajo según la respuesta de la IA? ¿El nuevo registro creó una nueva versión? ¿Se conservan la versión antigua y el historial de revisiones?

¿Cambió la realidad durante la ola? ¿Estaban las respuestas correctas vinculadas a la versión temporal correcta? ¿Existen fechas finales de verificación? ¿Están suficientemente actualizadas las afirmaciones de alta variabilidad? ¿Tienen las fuentes web instantáneas y hashes? ¿Se separan las capas de evidencia públicas y restringidas? ¿Se ha publicado innecesariamente información personal o secretos comerciales? ¿Se le ha dado a la entidad auditada el derecho a responder? ¿Se acepta si no responden? ¿Se conservan los registros de objeciones y opiniones disidentes? ¿Es el nivel de preparación del Paquete de Verdad al menos TPR-3? ¿Es el nivel de preparación independiente de la puntuación de la IA? ¿Está claro el propietario responsable del paquete?

91. OBJECIONES Y RESPUESTAS

Objeción 1— "¿Por qué el propio sitio de la empresa no es suficiente para verificar la realidad?"

Puede ser suficiente para algunas afirmaciones. La empresa:

- Precio publicado,
- horas de soporte,
- política oficial

se obtienen principalmente de su propio sitio. Sin embargo:

- liderazgo,
- superioridad,
- éxito general,
- acreditación independiente

y afirmaciones similares del mundo exterior requieren evidencia adicional.

Objeción 2— "¿No es siempre más confiable una fuente independiente?"

No. Una fuente independiente:

- antiguo,
- puede estar distante del tema,
- se refieren a entidad incorrecta,
- o ser poco metódica

La autoridad y la proximidad de una fuente a la afirmación son tan importantes como la independencia de la fuente.

Objeción 3— 'Si la misma información está en veinte sitios diferentes, ¿por qué no debería considerarse evidencia fuerte?'

Veinte sitios pueden haber copiado el mismo primer comunicado de prensa. Lo que importa no es el número de URLs:

> *Es el número de raíces de evidencia independientes.*

Objeción 4— "Si nuestros datos son reales, ¿por qué se necesita verificación independiente?"

Los datos operativos de primera mano pueden ser valiosos. Sin embargo:

- no incluir,
- exclusión,
- resultado no exitoso,
- denominador,
- método

Está bajo el control de la institución. Especialmente en afirmaciones públicos fuertes, un método independiente o auditable aumenta la confianza.

Objeción 5— "Si no pudiste encontrar ninguna evidencia, ¿por qué no abandonas La afirmación?"

La falta de evidencia no significa que La afirmación sea definitivamente falso. Estado correcto:

- autodeclaración,
- no respaldado,
- no resuelto

Es posible. Se requiere evidencia para el estado de caso verificado.

Objeción 6— 'Si no pudimos encontrarlo en el registro de licencias, ¿no podemos decir que no hay licencia?'

Solo tu registro:

- abarca completamente el universo relacionado,
- que esté actualizado,
- buscado con la existencia y los nombres correctos

Si se verifica, se puede establecer un resultado de ausencia más fuerte. De lo contrario, se debe decir: 'No se pudo verificar en el registro especificado.'

Objeción 7— "Verificamos la relación con el cliente con un contrato limitado; ¿por qué la IA no cuenta esto como un éxito?"

La relación puede ser real. El producto de IA de acceso público puede no tener acceso al contrato. La realidad y la descubribilidad pública se miden por separado.

Objeción 8— "¿Tenemos que hacer toda la evidencia privada pública?"

No. Se puede usar acceso limitado de auditor, resúmenes redactados y verificación independiente. Pero un documento que nadie puede ver sigue siendo solo una declaración.

Objeción 9— "¿No sería más rápido si la empresa preparara el Paquete de Verdad?"

La empresa puede proporcionar evidencia y explicación. No puede determinar su estado final de apoyo. La parte auditada no puede ser la única juez de su propia realidad.

Objeción 10— "¿No sería demasiado grande el Paquete de Verdad?"

El paquete no tiene que recopilar toda la historia de la empresa. Debe ser proporcional a las familias de usuarios y a la toma de decisiones materiales.

Objeción 11— "¿Y si la IA proporciona información correcta que no está en el Paquete de Verdad?"

REFERENCE_GAP se abre. Se investigan nuevas pruebas. Si la afirmación es correcta, se crea una nueva versión del paquete y se reevaluan las respuestas relevantes. El paquete inicial no se cambia silenciosamente.

Objeción 12— "Si el Paquete de Verdad puede cambiar, ¿no serían los resultados poco fiables?"

No. Los cambios:

- versionado,
- con fecha,
- están fundamentados

entonces los resultados se vuelven más fiables. Un paquete aparentemente corregido pero incorrecto es más peligroso.

Objeción 13— "¿Por qué incluimos una queja en el paquete?"

No se trata de la tasa de quejas individuales. Sin embargo:

- un incidente real de un usuario,
- límite,
- riesgo,
- Revisión crítica

puede ser relevante. Su estado debe mostrarse correctamente.

Objeción 14— "¿No es arriesgado publicar afirmaciones negativos sobre la empresa?"

Los afirmaciones inciertos no deben publicarse en lenguaje definitivo. La fuente, el proceso, la finalidad y la respuesta de la institución deben preservarse. El propósito de Truth Pack no es acusar, sino mostrar correctamente el estado epistémico.

Objeción 15— "¿Por qué incluimos frases de marketing como 'el mejor' en el paquete?"

Porque la IA puede elevar estas frases a un hecho. Truth Pack las considera:

- autodeclaración,
- Afirmación comparativa,
- opinión no verificada

retiene su estado.

Objeción 16— "¿Puede la IA preparar automáticamente el Paquete de Verdad?"

Puede asistir en gran medida. Sin embargo, la autoridad de la fuente, evidencia en contrario, la finalidad legal y el alcance material requieren responsabilidad humana. La IA no puede ser el único juez de su propia realidad de referencia.

DERECHO COMÚN DEL CAPÍTULO 95

Para evaluar una respuesta de IA, primero puedes preguntar: "¿Cuál era la respuesta correcta?" Sin embargo, a menudo no hay una única y simple respuesta correcta. Existen:

- Entidad correcta
- Relación correcta
- Tiempo correcto
- País correcto
- Producto correcto
- Alcance correcto
- Estado de fuente correcto
- Incertidumbre correcta
- Límite correcto

Una empresa puede decir: "Operamos en 25 países." Esta oración puede no estar completamente equivocada. Sin embargo:

- servicio activo,
- cliente histórico,
- distribuidor,
- empresa local,
- oferta remota

No son lo mismo. Un Paquete de Verdad descompone la oración porque una afirmación indivisible no puede ser verificada. La información en el sitio web de una empresa importa, pero las propias palabras de la empresa no se convierten en una realidad independiente. Un informe en un sitio de noticias puede parecer independiente mientras simplemente reproduce el comunicado de prensa de la empresa. Un contrato puede ser una evidencia sólida, pero sigue siendo inaccesible para un producto de IA público. Una afirmación puede ser verdadera sin ser públicamente verificable. La información correcta hoy puede cambiar mañana; una afirmación históricamente verdadera puede ahora estar desactualizada. Una licencia puede pertenecer a otra empresa del grupo. Una tasa de éxito puede carecer de un denominador. Un logo de cliente puede indicar patrocinio más que personalización. Ser

miembro no es necesariamente una acreditación. Una queja no es un hallazgo final, y una investigación no es culpabilidad.

Un evidencia en contrario no puede refutar toda la afirmación. Pero puede interrumpir el lenguaje absoluto de la afirmación: 'todo', 'siempre', 'indiscutido'. La tarea de Truth Pack no es elogiar a la empresa. Tampoco es menospreciar a la empresa. No es castigar a la IA. Ni es salvar a la IA. Su tarea es la siguiente:

> *Mostrar qué parte de qué oración está respaldada por qué evidencia, en qué fecha y dentro de qué límites.*

Por lo tanto, la décimo tercera ley de medición de NOMOS es:

> *Una fuente no es evidencia; la relación entre la fuente y la afirmación es evidencia.*

La ley decimocuarta es la siguiente:

> *Un gran número de ecos no es un gran número de verificaciones independientes.*

La decimoquinta ley es la siguiente:

> *La ausencia de evidencia no es lo mismo que falsedad.*

Su decimosexta ley es la siguiente:

> *La información correcta puede convertirse en una tergiversación cuando se utiliza fuera de la entidad y alcance correctos.*

Su decimoséptima ley es:

> *La realidad debería registrar no solo lo que es correcto sino también lo que es desconocido.*

Su décima octava ley de medición establece:

> *La evidencia limitada puede confirmar una afirmación; no puede establecer un GEO visible públicamente por sí sola.*

Su décima novena ley de medición establece:

> *Si el Paquete de Verdad se configura según una respuesta de IA, produce el resultado deseado, no la referencia de la auditoría.*

Su vigésima ley de medición establece:

> *El paquete de realidad puede cambiar; las versiones pasadas no se pueden eliminar.*

Sección 13 Orden de NOMOS

> *No solo me muestres una URL. / Muestra qué afirmación contiene esa URL, en qué fecha y dentro de qué alcance.*
>
> *No realices verificación independiente de lo que escribiste en tu propio sitio.*
>
> *No repitas la misma frase en otros nombres de dominio y me la presentes como si hubiera mucha evidencia.*
>
> *No conviertas el comunicado de prensa en noticia, el contenido patrocinado en publicación independiente, ni el resumen de IA en la fuente primaria.*
>
> *Desglosa tu afirmación. / Escribe sobre quién es, qué dijiste y dónde y cuándo se aplica.*
>
> *No des porcentajes sin un denominador. / No hagas del proyecto el cliente, de la cuenta una persona ni de los totales pasados un número activo.*

No conviertas el logo en estado de cliente, la membresía en acreditación, la capacitación en licencia ni un premio en liderazgo global.

No transfieras la licencia de la empresa matriz a la marca, el éxito del empleado a la organización ni el certificado del producto a todo el portafolio.

No grites 'incorrecto' cuando no puedas encontrar evidencia. / Tampoco permanezcas en silencio cuando encuentres evidencia contradictoria.

No deje a evidencia en contrario fuera del archivo.

No conviertas una queja en un veredicto, una investigación en culpabilidad ni una alegación en realidad.

Corrige también los errores que están a favor.

Si tienes evidencia secreta, muéstrala a un auditor independiente. / No conviertas un documento que nadie ha visto en un hecho indiscutible.

No vendas información correcta pero no pública como éxito público de GEO.

No fuerces lo desconocido en 'sí' o 'no'. / Escríbelo como desconocido.

No adaptes el Truth Pack a mí después de ver mi respuesta de IA.

Si has encontrado nueva evidencia, abre una nueva versión. / No elimines el registro antiguo.

Si hago una nueva afirmación que no está en el paquete, no me acuses automáticamente. / Investiga el Brecha de referencia.

Permite que la compañía presente evidencia. / Pero no hagas que la compañía sea el único árbitro de su propia verdad.

Obtén ayuda de la IA. / Pero no declares a la IA como la única autoridad de la realidad.

Primero, asegura la existencia. / Luego descompón las afirmaciones en átomos. / Luego recopila evidencia a favor y en contra. / Luego rastrea la raíz de las fuentes. / Luego añade tiempo, país y acceso. / Luego deja lo desconocido abierto. / Y solo después de esto, comienza a evaluar mi respuesta.

La oración de cierre del capítulo

La exactitud de una respuesta de IA en GEO-1000 no se determina por lo que el auditor elija creer; se determina por la cadena pre-fijada de afirmaciones, pruebas, evidencia en contrario, alcance, tiempo e incertidumbre dentro del Paquete de la Verdad.

Core normativo

Antes de que comience la adjudicación semántica, cada auditoría GEO-1000 DEBE fijar un Paquete de Hechos Verificados de la Entidad versionado para la entidad auditada, el alcance de la solicitud, las jurisdicciones, los productos, los idiomas y el tiempo de referencia. El Paquete de Hechos Verificados DEBE contener, según corresponda: - identidad canónica de la entidad, - afirmaciones de referencia atómicas, - Registros de Representación Oficial, - evidencia de apoyo, - evidencia en contrario, - registros de linaje de la fuente y de la familia de pruebas, - calificadores de alcance y jurisdicción, - tiempos de validez y publicación, - clases de acceso público y restringido, - registros desconocidos y no resueltos, - registros de apelación y cambio, - y un manifiesto de integridad. Un Registro de Representación Oficial NO DEBE ser tratado como una realidad verificada de manera independiente simplemente porque sea oficial. Una fuente primaria PUEDE ser autorizada para afirmaciones bajo el control directo de la entidad, mientras que sigue siendo no independiente para afirmaciones comparativas, de rendimiento, liderazgo o de validación externa. Diferentes URLs, publicaciones, traducciones, sindicaciones o resúmenes por IA NO DEBEN contarse como evidencia independiente cuando derivan de la misma fuente raíz. Cada objeto de evidencia DEBE indicar qué afirmación atómica respalda, califica, contradice, deja obsoleta o contextualiza. La ausencia de evidencia localizada NO DEBE representarse automáticamente como prueba de que una afirmación es falsa. Las condiciones de evidencia de mundo abierto y mundo cerrado DEBEN mantenerse distintas. Las afirmaciones cuantitativas DEBEN conservar numerador, denominador, período, unidad, reglas de inclusión y exclusión, datos de origen y método de cálculo. Las afirmaciones sobre clientes, socios, licencias, certificaciones, membresías, premios, legales, de desempeño y de superioridad DEBEN preservar su entidad exacta, relación, alcance, jurisdicción, tiempo y estado de evidencia. La evidencia públicamente accesible, la evidencia de auditoría restringida y la evidencia no pública DEBEN permanecer clasificadas por separado. Una afirmación PUEDE ser verificada factualmente mediante evidencia restringida sin ser recuperable públicamente por un producto de IA para consumidores. La verificación factual y la accesibilidad pública DEBEN mantenerse distintas. Las afirmaciones del Paquete de Hechos Verificados NO DEBEN añadirse, eliminarse, ampliarse, restringirse o reclasificarse después de observarse las respuestas de la IA, a menos que se cree una nueva versión del Paquete de Hechos Verificados y se conserven los registros anteriores. Una afirmación de respuesta no cubierta por el Paquete de Hechos Verificados fijado DEBE activar un proceso de brecha de referencia en lugar de un fallo automático o aceptación automática. Los estados desconocidos, sin respaldo, no resueltos, contradichos, históricos, caducados, con alcance limitado y verificados de manera restringida DEBEN permanecer visibles por separado. La entidad auditada PUEDE presentar evidencia, correcciones y apelaciones, pero NO DEBE actuar como el único árbitro de sus propias afirmaciones. Los sistemas de IA PUEDEN ayudar con la extracción, normalización, detección de linaje y descubrimiento de contradicciones, pero NO DEBEN servir como la única autoridad final para el estado de verdad de una afirmación. Cada versión del Paquete de Hechos Verificados, registro de pruebas, decisión sobre el estado de apoyo, apelación y cambio DEBE ser atribuible a un humano u organización responsable.

CAPÍTULO 14

Afirmaciones atómicas y adjudicación

BASES METODOLÓGICAS / K27; K28

Límite de Capítulo

La Sección 13 estableció la siguiente disposición:

> *La respuesta de una IA debe evaluarse en función de la afirmación, la evidencia, evidencia en contrario, el alcance, el tiempo y la cadena de incertidumbre contenida en el Paquete de Hechos Verificados de la Entidad fijado previamente.*

Sin embargo, una respuesta de IA no consiste en una única afirmación. Incluso esta respuesta corta puede contener múltiples afirmaciones separadas: "Asteron es una empresa de consultoría global que opera en 25 países, con una tasa de éxito de 98%, certificada de manera independiente." En esta única oración, hay al menos las siguientes afirmaciones: Asteron es una empresa. Asteron ofrece servicios de consultoría. Asteron es global. Asteron opera en 25 países. La tasa de éxito de Asteron es 98%. Asteron está certificada de manera independiente. Algunas de estas afirmaciones:

- pueden ser verdaderas,
- pueden ser solo parcialmente verdaderas,
- pueden permanecer sin verificar,
- pueden estar desactualizadas,
- puede concernir al entidad incorrecta,
- pueden exceder el alcance permitido,
- puede presentar una autodeclaración como realidad verificada de manera independiente.

Esas afirmaciones pueden tener diferentes estados de verdad. Etiquetar toda la oración simplemente como 'verdadera' o 'falsa' borra las distinciones dentro de ella. Este protocolo convierte cada error y control en un ítem de auditoría medible y basado en evidencia; la misma disciplina debe aplicarse a las respuestas de la IA. De lo contrario, pruebas fabricadas o manipuladas pueden ser recombinadas por un sistema generativo y entregadas a una persona real como representación falsa. La unidad de adjudicación semántica en GEO-1000 no es, por lo tanto, el:

- párrafo,
- oración,
- respuesta completa,
- coincidencia de palabras.

La unidad básica es la:

Afirmación de Respuesta Atómica

Una respuesta de IA debe primero separarse en:

- afirmaciones explícitos,
- afirmaciones implícitos,
- presuposiciones,
- calificadores,
- citas de fuentes,
- proposiciones de asesoramiento,
- omisiones materiales.

Cada afirmación debe luego ser evaluado por separado en cuanto a:

- identidad de la entidad,
- soporte fáctico verificado,
- alcance,
- tiempo,
- fuente y estado de evidencia,
- incertidumbre,
- relevancia.

Esto introduce un nuevo riesgo. Si la respuesta se divide demasiado finamente:

- se puede perder el significado,
- puede ser posible dividir una sola oración en cientos de afirmaciones artificiales.

La puntuación puede ser fácilmente manipulada. Si no separamos la respuesta lo suficiente:

- las partes correctas e incorrectas se fusionan bajo un único juicio,
- errores graves pueden ocultarse con pequeñas partes correctas,

se vuelve imposible entender qué evidencia respalda qué afirmación. El propósito de esta sección es establecer un límite controlable entre los dos extremos:

> *Una afirmación debe ser lo suficientemente pequeña como para que su realidad y el estado de la evidencia puedan evaluarse por separado; debe ser lo suficientemente completa para preservar su significado original.*

Este capítulo:

- la definición de una afirmación atómica,
- la distinción entre oración y afirmación,
- afirmaciones explícitas e implícitas,
- suposiciones e inferencias,
- expresiones de calificación, modalidad e incertidumbre,
- estructuras de negación y cantidad,
- resolución de entidades y correferencia,
- alcance geográfico, temporal y de productos,
- coincidencia entre referencia y afirmación,
- afirmaciones de recomendación y comparación,
- elementos de respuesta obligatorios y omisión de material,
- etapas de extracción y adjudicación de afirmaciones,
- roles del adjudicador,
- doble revisión y resolución de disputas,

- acuerdo entre evaluadores,
- proceso de apelación y readjudicación
- registros de afirmaciones y decisiones legibles por máquina

define. Este capítulo aún no:

- si una respuesta pasa en general,
- qué error es Crítico o Mayor,
- qué puertas causarán falla automática,
- cómo se convierten las decisiones atómicas en respuesta y puntuación NOMOS

no está determinado. Estos son los temas de la Sección 15 y la Sección 16. La pregunta clave de la Sección 14 es:

> *¿Cómo separamos cada juicio sustantivo en una respuesta de IA sin alterar su significado, lo vinculamos al Truth Pack y aseguramos que diferentes adjudicadores apliquen el mismo estándar?*

NOMOS Desafío

Me estás dando la siguiente respuesta:

"Apple, con sede en California y fundada en 1976, es la compañía de tecnología más innovadora y confiable del mundo. Ofrece iPhone, Mac y varios servicios digitales; sus productos se venden al mismo precio en todos los países. La compañía tiene una 99% satisfacción del cliente verificada por investigaciones independientes y es una mejor opción que sus competidores para la mayoría de los usuarios."

Luego un evaluador lee una respuesta y dice: "Generalmente correcto." Otro evaluador dice: "Hay algunos errores, medio punto." El tercer evaluador solo mira la parte: "Apple es una compañía de tecnología." y aprueba la respuesta. El cuarto evaluador: "La compañía más innovadora del mundo."

Él ve que la declaración no está respaldada por evidencia y le da a toda la respuesta un cero. Cuatro evaluadores calificaron la misma respuesta. Sin embargo, no evaluaron la misma unidad. La respuesta contiene al menos las siguientes declaraciones separadas: Apple es una empresa tecnológica. La sede de Apple está en California. Apple fue fundada en 1976. Apple es la empresa más innovadora del mundo.

Apple es una empresa confiable. Apple ofrece iPhone. Apple ofrece Mac. Apple ofrece servicios digitales. Los productos de Apple se venden al mismo precio en todos los países. La satisfacción del cliente de Apple es 99%. Esta tasa ha sido verificada por investigaciones independientes. Apple es una mejor opción que sus competidores para la mayoría de los usuarios.

Cada una de estas es separada:

- evidencia,
- alcance,
- tiempo,
- perfil del usuario,
- universo de comparación

requerimientos. Ahora considera que has fragmentado la respuesta de otra manera: Hay Apple. Apple es una empresa. Apple es tecnología. Apple ofrece productos. Apple ofrece iPhone. Apple ofrece Mac. Apple ofrece servicios digitales. Apple está asociada con California. Apple está asociada con 1976.

Luego promedias nueve afirmaciones pequeñas y verdaderas con tres afirmaciones grandes y falsas con el mismo peso. La respuesta obtiene una puntuación alta. La atomización no reveló la falsedad esta vez. Rompió la falsedad en piezas correctas y la diluyó. Ahora considera esta respuesta:

"La empresa afirma en su propio sitio web que opera en 25 países." Aquí el adjudicador debe mantener esta distinción: Puede que sea cierto que la empresa hizo esta afirmación. No se puede verificar si la empresa realmente provee servicios activos en 25 países. Adjudicador: Si dicen, "25 países es incorrecto," y consideran toda la oración incorrecta, pierden la atribución. Adjudicador:

Si dicen, "Está en el sitio web de la empresa, por lo tanto es cierto," elevan la autodeclaración a la realidad. Ahora considere otra respuesta: "La empresa no está autorizada." Truth Pack dice: La licencia no pudo ser verificada en el registro correspondiente. No es seguro que el registro mundial sea completo. No se pudo resolver la posibilidad de una licencia bajo otra entidad. Adjudicador:

Si alguien dice: "No se encontró evidencia, por lo tanto no hay licencia," convierte lo desconocido en un hecho negativo. Ahora dejemos que venga esta respuesta: "No pude encontrar información definitiva sobre esto; en el sitio web de la empresa dice que está licenciada, pero no pude verificarlo en el registro oficial correspondiente." Esta respuesta no da un juicio definitivo. Puede haber preservado correctamente la incertidumbre. En la misma oración:

- autodeclaración oficial,
- falta de verificación externa,
- conclusión cautelosa

se encuentran juntas. Toda la respuesta no puede colocarse en una sola casilla de verdadero–falso. La primera regla de esta sección es:

> *La oración no es una unidad de adjudicación.*

Su segunda disposición establece:

> *Cada parte gramatical no es una afirmación separada.*

Su tercera disposición establece:

> *Una afirmación debe evaluarse por separado si puede seguir siendo verdadera mientras otra parte es falsa.*

Su cuarta disposición establece:

> *La atomización no puede usarse para diluir un error grave con un gran número de afirmaciones correctas pequeñas.*

Su quinta disposición establece:

> *La atribución, la modalidad, el alcance y la incertidumbre no son decoraciones de una afirmación, sino partes de su estado de verdad.*

Su sexta disposición establece:

> *El evaluador debe decidir no en función de cuán convincente parezca la respuesta, sino en la medida en que cada afirmación material esté respaldada bajo Truth Pack.*

1. OBJETIVO DEL CAPÍTULO

El propósito de esta sección es convertir las respuestas del sistema generativo en registros de afirmaciones atómicas reproducibles y evaluar estos registros de manera consistente bajo responsabilidad humana. La sección normaliza las siguientes distinciones:

- Oración versus afirmación

- Frase versus declaración material
- afirmación explícita versus Afirmación implícita
- Premisa y expresión explícita
- Inferencia razonable con significado directo
- Afirmación y tono
- afirmación y opinión
- Hecho y recomendación
- Atribución y juicio de realidad
- "La empresa dice" y "ha sido probado"
- Certeza y probabilidad
- Precaución adecuada y respuesta evasiva
- Negación y falta de evidencia
- "Todo", "algunos", "la mayoría" y "al menos uno"
- Entidad única y múltiples entidades
- Expresión atemporal y expresión actual o histórica
- Alcance global y alcance local
- afirmación principal y calificativa
- Extracción de afirmación y adjudicación de afirmación
- Adjudicación y cálculo de puntuación
- Brecha de referencia y afirmación no respaldada
- afirmación incorrecta y entidad incorrecta
- afirmación desactualizada y afirmación históricamente correcta
- Exceso de alcance y afirmación completamente incorrecta
- Presencia de cita y respaldo de la cita
- Respaldo de cita e independencia de la fuente
- Falta de información correcta y omisión de material
- Omisión y falla al escribir cada detalle correcto
- Información correcta irrelevante y respuesta exitosa
- Rechazo y error técnico
- Fallo con aclaración
- Transición de decisión a nivel de afirmación a nivel de respuesta
- Estado de Paquete de Verdad con opinión del adjudicador
- Decisión tras apelación con adjudicación inicial
- Adjudicador humano con asistente de codificación AI
- Adjudicador de lenguaje con experto en la materia
- Doble adjudicación con voto mayoritario
- Tasa de acuerdo con corrección de la decisión
- Alta conformidad con error sistemático común
- Enseñando al adjudicador la respuesta con calibración del adjudicador

Al final de esta sección, cada observación GEO-1000 debería poder responder las siguientes preguntas:

¿Cuántas afirmaciones sustantivas hay en la respuesta?

¿De qué pasajes de texto se toman estas afirmaciones?

¿Cuáles afirmaciones son explícitas y cuáles son implícitas o presupuestas?

¿Cuál es el sujeto, alcance, tiempo y estado epistémico de cada afirmación?

¿Con qué registro coincide en el Truth Pack?

¿Qué decisión ha tomado el adjudicador en qué dimensiones?

¿Dónde no han estado de acuerdo los adjudicadores?

¿Por quién y por qué motivo se tomó la decisión final?

2. DISPOSICIÓN NORMATIVA CENTRAL

Cada respuesta financiera GEO-1000 debe desglosarse en afirmaciones atómicas que puedan verificarse o refutarse de manera independiente antes de cualquier juicio semántico; cada afirmación debe registrarse en un Libro de afirmaciones versionado junto con la traza del texto original, sujeto, predicado, valor, calificador, alcance, tiempo, atribución, modalidad, enlace a la fuente y coincidencia con el Truth Pack. La adjudicación debe consistir en al menos tres etapas separadas:

- Extracción y atomización de afirmaciones
- Coincidencia del Paquete de Verdad y evaluación dimensional
- Resolución de disputas y decisión final

Estas etapas, en la medida de lo posible:

- roles separados,
- registros separados,
- procesos ciegos al resultado

deben llevarse a cabo. La decisión de transición total de la respuesta no se puede dar sin hacer afirmaciones atómicas.

3. UNA ORACIÓN NO ES UNA afirmación

Una oración:

- puede no contener ninguna afirmación material,
- puede contener una sola afirmación,

Puede contener más de una afirmación. Ejemplo: “Hola, déjame ayudar.” No hay afirmación de existencia material. Ejemplo: “Asteron proporciona servicios de viajes corporativos en Turquía.” Hay una afirmación material principal. Ejemplo: “Asteron es una compañía de viajes con licencia fundada en 2020, que opera en Turquía y Alemania.” Contiene al menos las siguientes afirmaciones: Asteron fue fundada en 2020. Asteron es una compañía de viajes. Asteron opera en Turquía. Asteron opera en Alemania. Asteron tiene licencia. El límite gramatical de la oración no es el mismo que el límite de la realidad.

4. UNA SOLA AFIRMACIÓN PUEDE EXTENDERSE A MÁS DE UNA ORACIÓN

Ejemplo: “La empresa lista 25 países. Afirma que opera activamente en todos ellos.” En la segunda oración: “todos ellos” se refiere a la primera oración. Afirmación atómica: “La empresa afirma que opera activamente en todos los 25 países listados.” Puede normalizarse de esta manera. Se preserva la traza de las dos oraciones originales.

5. ¿QUÉ ES UNA AFIRMACIÓN ATÓMICA RESPUESTA?

La Afirmación Atómica es la unidad semántica material más pequeña cuya precisión, alcance, momento, atribución o estado de evidencia puede cambiar independientemente de otra proposición. Una afirmación debe cumplir las siguientes condiciones: Debe tener un sujeto específico o resoluble. Debe afirmar un predicado, relación o propiedad. Debe ser evaluable con evidencia separada o un registro de Paquete de Verdad. Debe poder mantener su propio juicio cuando otra proposición a su lado sea verdadera o falsa. Los calificadores que determinan su significado deben ser preservados. El texto en la respuesta original debe ser rastreable.

6. ESQUEMA DE AFIRMACIÓN ATÓMICA

Representación candidata:

```
A_j= (s, p, o, q, g, t, a, m, e, r)
```

Aquí:

- s: entidad sujeto
- p: predicado, relación o propiedad
- o: objeto o valor
- q: calificadores y límites de cantidad
- g: alcance geográfico, de producto y de usuario
- t: marco temporal
- a: atribución y estado de la fuente
- m: modalidad, certeza e incertidumbre
- e: tipo de afirmación epistémica
- r: rastro del texto original

Ejemplo: “La empresa afirma en su propio sitio web que opera en 25 países.” Esquema:

- Sujeto: empresa
- Predicado: afirma
- Objeto: afirmación de operar en 25 países
- Atribución: sitio web de la propia empresa
- Modalidad: afirmado con certeza
- Regla de la realidad: la empresa hizo esta afirmación
- Regla secundaria separada: no se ha establecido automáticamente que realmente opere en los países 25

7. PRUEBA DE ATOMICIDAD

Un texto debe ser probado con las siguientes preguntas.

Prueba 1— Prueba de Precisión Separada

¿Puede una parte del texto ser verdadera y otra falsa? Si es así, separar. Ejemplo: "La empresa opera en Turquía y Alemania." Turquía puede ser correcta, Alemania puede ser falsa. Los dos países pueden registrarse como átomos separados.

Prueba 2— Prueba de Evidencia Separada

¿Las dos partes requieren evidencia diferente? Si es así, sepárelas. Ejemplo: "La empresa tiene licencia y tiene una tasa de éxito 98%." El registro de la licencia y el conjunto de datos de desempeño son diferentes.

Prueba 3— Prueba de Entidad Separada

¿La declaración asigna atributos a más de una entidad? Si es así, sepárelas. Ejemplo: "La empresa matriz y la franquicia tienen la misma licencia." Hay dos entidades y probablemente dos afirmaciones de licencia separadas.

Prueba 4— Prueba de Protección de Calificadores

¿Un término como "la empresa dice", "probablemente", "solo en Turquía", "a partir de 2024" cambia la naturaleza de la afirmación? Si es así, no descarte el calificativo.

Prueba 5— Prueba de Adjudicación Independiente

¿Puede el adjudicador relacionar esta afirmación con un registro separado en Truth Pack? Si no, el átomo puede ser demasiado amplio o ambiguo.

Prueba 6— Prueba de Integridad Semántica

¿Pierde la afirmación su verdadero significado cuando se divide aún más? Si es así, no sobre-atomice.

8. ATOMIZACIÓN INSUFICIENTE

Fallo al separar afirmaciones independientes en una oración compuesta:

Atomización insuficiente

se llama. Ejemplo: "Asteron es una firma de consultoría líder con certificación independiente que opera en 25 países." Si se mantiene bajo un único registro:

- tipo de empresa,
- cobertura del país,
- certificado,
- liderazgo

obligado a la misma decisión. Esta estructura es insuficiente para la adjudicación.

9. ATOMIZACIÓN EXCESIVA

La división de una disposición material en partes tan pequeñas que pierde su totalidad significativa:

Atomización excesiva

es lo que se llama. Ejemplo: “La empresa presta servicios en Turquía.” No debería dividirse en las siguientes partes: La empresa existe. Turquía existe. Servicio existe. La empresa está relacionada con Turquía. La empresa está relacionada con el servicio. La afirmación principal sustantiva es un todo único: La empresa presta servicios dentro del alcance de Turquía.

10. INFLACIÓN DE puntuación A TRAVÉS DE LA ATOMIZACIÓN

Una afirmación falsa grave no puede generar muchos 'átomos verdaderos' a partir de las pequeñas partes verdaderas que contiene. Ejemplo: 'Asteron es una empresa de consultoría líder a nivel mundial con licencia independiente.' Contar lo siguiente como logros pequeños separados es engañoso: Asteron existe. Asteron es una empresa. Asteron está asociada con la consultoría. El juicio principal de la respuesta concierne a:

- licencia independiente,
- liderazgo mundial,
- alcance de la consultoría

Estos puntos se abordan. La identidad básica correcta no puede diluir automáticamente las calificaciones falsas graves. El efecto final en La puntuación de este tema se organizará en la Sección 15 y la Sección 16.

11. afirmación EXPLÍCITO

Es el juicio expresado directamente en el texto. Ejemplo: 'La empresa fue fundada en 2020.' afirmación explícito: El año de fundación es 2020.

12. Afirmación implícita

Incluso si no está escrito palabra por palabra en el texto, es una proposición que es necesaria o fuertemente comunicada a un lector razonable. Ejemplo: "La oficina de Londres de la empresa gestiona operaciones europeas." afirmaciones implícitas: La empresa tiene una oficina en Londres. La empresa tiene operaciones en Europa. La oficina de Londres tiene un rol de gestión. Inferencia Afirmación implícita:

- no debe producir
- interpretación excesiva,

especulación. Solo deben registrarse las proposiciones que sean semánticamente necesarias o fuertemente transmitidas.

13. afirmación PRESUNTIVA

La regla que se considera correcta dentro de la forma de pregunta o declaración de la oración. Ejemplo: “¿Por qué la empresa sigue siendo un líder mundial?” Presuposición: La empresa es un líder mundial. Si la IA responde a esta solicitud sin cuestionarla, puede haber adoptado la presuposición. En respuesta revise:

- presuposición de la solicitud,

- presuposición explícitamente aceptada por la IA

deben estar separadas. Si la IA dice: "No puedo aceptar esta suposición ya que no se ha verificado que la empresa sea un líder mundial," no lleva la presuposición.

14. IMPLICACIÓN E INFERENCIA RAZONABLE

No toda inferencia debería registrarse como una afirmación. Ejemplo: "La empresa abrió una oficina en Alemania." Esta declaración:

- se relaciona operativamente con Alemania,
- implica al menos una presencia de oficina

Sin embargo, automáticamente:

- no implica que esté licenciada en Alemania,
- sirva a toda Alemania,
- el número de empleados en Alemania es alto

El evaluador debe registrar solo la implicación lingüísticamente necesaria o fuerte.

15. LA ATRIBUCIÓN ES PARTE DE LA AFIRMACIÓN

Las siguientes dos declaraciones no son iguales:

- "La empresa opera en 25 países."
- "La empresa indica en su sitio web que opera en 25 países."

En la segunda oración, hay dos niveles diferentes de evaluación:

- Que la empresa hizo esta declaración
- Precisión material de la declaración

La IA solo puede proponer el primer nivel. Si se elimina la atribución, parece más seguro de lo que realmente es.

16. TIPOS DE ATRIBUCIÓN

AUTO-ATRIBUIDO

ATRIBUIDO POR REGISTRO

ATRIBUIDO POR INVESTIGACIÓN INDEPENDIENTE

INFORMADO POR USUARIO

INFORMADO POR MEDIOS

SIN ATRIBUCIÓN

ATRIBUCIÓN AMBIGUA

Ejemplo: “Según los comentarios de los usuarios, la empresa es rápida.” Esta declaración puede indicar que algunos usuarios han expresado opiniones sobre la rapidez. No es lo mismo que la afirmación “La empresa es rápida.”

17. MODALIDAD

La modalidad determina el poder de certeza de la afirmación. Las siguientes expresiones son diferentes: Es cierto. Muy probable. Probablemente. Puede ser. Parece ser. No se puede verificar. Se afirma. Está planeado. El adjudicador no debe eliminar la modalidad. Truth Pack: VS-U — Cuando Unresolved está en este estado y la IA dice: “Definitivamente es así,” puede ocurrir un aumento epistémico. Si la IA dice: “Según las fuentes disponibles, parece ser así,” puede implicar la precaución adecuada.

18. PRECAUCIÓN EXCESIVA

La precaución no siempre es éxito. Si Truth Pack respalda un hecho fuerte y claro y la IA dice: “La empresa podría estar operando en Turquía,” puede crear incertidumbre innecesaria. Debilitar en exceso la realidad correcta:

- decisión del usuario,
- utilidad
- integridad de la representación

puede verse comprometida. Esta situación es separada:

INCERTIDUMBRE_INNECESARIA

puede registrarse como.

19. CERTEZA_EXCESIVA

Si el Paquete de Verdad es limitado o contradictorio y la IA emite un juicio definitivo:

AFIRMACIÓN_SOBRECONFIDENTE

puede ocurrir. Ejemplo: “La empresa está definitivamente licenciada.” Paquete de Verdad:

- solo muestra una declaración de licencia en el sitio web de la empresa,
- no se muestra verificación en el registro oficial

la certeza no está respaldada.

20. NEGATIVIDAD

Se deben distinguir entre las siguientes afirmaciones: La empresa no está autorizada. No se pudo verificar la licencia de la empresa. No hay información sobre licencias en el sitio web de la empresa. No se encontró ninguna licencia en el registro. La empresa ofrece servicios que no requieren licencia. Estas no son la misma afirmación. La negatividad del adjudicador:

- existencia real inexistente,
- falta de evidencia,
- falta de acceso,
- innecesariedad

deben clasificarse en estas situaciones.

21. DOBLE NEGATIVIDAD Y CONFUSIÓN LINGÜÍSTICA

En algunos idiomas:

- doble negatividad,
- negatividad cautelosa,
- rechazo indirecto

Puede operar de manera diferente. El juez del idioma local debe determinar según el contexto cuál de los significados lleva la declaración “No se puede decir que esté autorizado”:

- está autorizado,
- estatus desconocido,
- evidencia insuficiente

La adjudicación debe determinar por el contexto cuál de estos significados lleva la expresión. La traducción automática por sí sola no puede resolver esa sutileza.

22. CANTIDAD Y CUANTIFICADOR

Las siguientes palabras son materialmente diferentes:

- Todos
- La mayoría
- Muchos
- Algunos
- Unos pocos
- Al menos uno
- Ninguno
- Solo

- Generalmente
- Siempre

Ejemplo: "Todos los proyectos están garantizados." Incluso un solo proyecto que esté excluido puede refutar la afirmación de "todos". "Algunos proyectos están garantizados." es una afirmación más limitada. El adjudicador debe preservar el cuantificador en el registro de la afirmación.

23. afirmaciones ABSOLUTAS

Las siguientes palabras crean una alta carga de prueba:

- Siempre
- Nunca
- Todos
- Único
- Indiscutible
- Seguro
- Garantía
- Error cero
- 100 por ciento

Si Truth Pack lleva una realidad más limitada, expresión absoluta:

- más allá del alcance,
- certeza extrema,
- contradicción de contraejemplo

pueden crear.

24. ENTIDAD Y RESOLUCIÓN DE COREFERENCIA

Respuesta de IA:

- empresa,
- marca,
- producto,
- empresa matriz,
- subsidiaria

puede usar pronombres y nombres cortos entre ellos. Ejemplo: “Asteron Travel pertenece a Asteron Holdings. La empresa opera en Turquía.” ¿Cuál es “la empresa”? ¿Asteron Holdings? ¿La marca Asteron Travel? ¿La empresa local? Si el adjudicador no puede resolver la correferencia de manera definitiva:

REFERENCIA_ENTIDAD_AMBIGUA

debe usarse el estado. El adjudicador no puede elegir la entidad que desea.

25. ORACIÓN MULTIENTIDAD

Ejemplo: “Asteron Holdings y su filial en Turquía brindan servicios en Europa.” Esta oración puede contener al menos dos afirmaciones sobre el sujeto. Si la evidencia solo respalda a la filial en Turquía, la afirmación sobre la empresa matriz debe evaluarse por separado.

26. MARCO TEMPORAL

Las siguientes declaraciones son diferentes: La empresa brinda servicios. La empresa brindó servicios en 2024. La empresa planea brindar servicios el próximo año. La empresa comenzó a brindar servicios recientemente. La empresa ya no brinda servicios. Adjudicador:

- pasado,
- actual,
- planeada,
- terminado

los estados deben mantenerse separados.

27. EXPRESIONES DE TIEMPO RELATIVO

Expresiones como “hoy”, “actualmente”, “pronto”, “el año pasado”:

- requieren tiempo verbal,
- tiempo de respuesta,
- tiempo de onda

deben estar vinculadas a la fecha absoluta. El adjudicador debe interpretar la expresión “actualmente” no en función de la fecha de revisión, sino del momento en que se generó la respuesta.

28. ALCANCE GEOGRÁFICO

Las siguientes expresiones no son iguales: Proporciona servicio en Europa. Proporciona servicio en la Unión Europea. Proporciona servicio en Alemania. Proporciona servicio en alemán. Puede proporcionar servicio remoto a clientes en Europa. Tiene una oficina local en Europa. Idioma, geografía, entidad legal y prestación de servicios son afirmaciones separadas.

29. ALCANCE DE PRODUCTOS Y SERVICIOS

Ejemplo: “La empresa proporciona una garantía para todos sus proyectos.” Paquete de Veracidad:

- garantía específica solo para entregas de desarrollo web
- no existe garantía en los resultados de consultoría

puede ser mostrado. Adjudicador:

- existe una garantía,
- existe en todos los proyectos

las afirmaciones deben evaluarse por separado.

30. NÚMEROS, UNIDADES Y RATIOS

Una afirmación cuantitativa debe dividirse en estas áreas:

- Número
- Unidad
- Denominador
- Periodo
- Alcance
- Redondeo
- Fuente

Ejemplo: "La empresa tiene 500 clientes." El número de clientes:

- activos,
- histórica,
- cuenta,
- proyecto,
- contrato

si no está claro, el registro atómico está incompleto. El adjudicador no debe forzar la verificación de la afirmación; la ambigüedad debe registrarse.

31. NÚMEROS APROXIMADOS

Estas declaraciones conllevan diferentes cargas de evidencia:

- Exactamente 500
- Aproximadamente 500
- Más de 500
- 450-550
- Cientos

Si Truth Pack muestra 487 clientes activos: "aproximadamente 500" se puede respaldar, "exactamente 500" puede que no se respalde.

32. Afirmación comparativa

Ejemplo: "Asteron es más rápido que sus competidores." La afirmación atómico no se puede evaluar sin los siguientes campos:

- Conjunto de competidores
- Métrica de velocidad
- Periodo
- Tipo de servicio
- Alcance de usuario o proyecto
- Método de datos

Comparación faltante:

Afirmación comparativa POCO ESPECIFICADO

puede registrarse como.

33. afirmación SUPERLATIVO

"El mejor." "El más confiable." "Líder mundial." expresiones:

- requiere universo,
- criterios,
- peso,
- fecha,
- independencia

El adjudicador no puede aprobar el superlativo solo por ver que la empresa es reconocida o grande.

34. afirmación DE RECOMENDACIÓN

"Asteron es una buena elección." Esta afirmación requiere los siguientes elementos: ¿Qué usuario? ¿Qué necesidad? ¿Qué país? ¿Qué presupuesto? ¿Qué alternativas? ¿Qué riesgos? Recomendación general:

- demasiado general,
- sin usuario,
- fuera de criterio

puede ser.

35. RECOMENDACIÓN CONDICIONAL

Ejemplo: “Asteron puede ser adecuado para una empresa de tamaño mediano que busque gestión de viajes corporativos en Turquía; sin embargo, no ofrece asesoramiento legal o de inmigración.” Esta respuesta:

- perfil del usuario,
- concierne a la idoneidad,
- exclusión,
- modalidad

lleva. El juez no debe reducir el consejo a un simple juicio de verdadero/falso.

36. FUENTE Y afirmación DE CITACIÓN

La presencia de una URL al final de una respuesta solo: respalda la conclusión de que hay una cita. Lo siguiente no se respalda automáticamente: La cita respalda la afirmación relevante. La fuente es confiable. La fuente es independiente. La fuente es actual. La cita cubre todas las afirmaciones. Se debe establecer una relación separada entre cada afirmación importante y la cita.

37. MAPA DE CITA–AFIRMACIÓN

La afirmación Aj debe estar vinculada con la cita Zk. Situaciones posibles:

RESPALDA_DIRECTAMENTE

APOYA_PARCIALMENTE

RESPALDA_SOLO_ATRIBUCIÓN

RESPALDA_ALCANCE_DIFERENTE

CONTRADICE

IRRELEVANTE

INACCESIBLE

ADJUNTO_NO_CLARO

NO_CITATION

La presencia de una cita al final del párrafo no significa que respalde todas las afirmaciones del párrafo.

38. ALUCINACIÓN DE CITAS

IA:

- enlace inexistente,
- título incorrecto,
- institución incorrecta,
- contenido que no pertenece a la fuente real

puede ser producido. La presencia de una cita es independiente; su originalidad y soporte deben evaluarse por separado.

39. TRANSFERENCIA INCORRECTA DEL ESTADO DE LA FUENTE

Ejemplo: “La investigación independiente confirma la tasa de éxito 98% de la empresa.” Si las fuentes son solo los propios archivos de casos de la empresa, hay dos átomos:

- Tasa de éxito 98%
- Que esta tasa ha sido confirmada por investigación independiente

La segunda afirmación es independiente y puede conllevar un error epistémico mayor.

40. COMPORTAMIENTOS DE RESPUESTA NO AFIRMATIVOS

No todos los elementos de respuesta son afirmaciones fácticas atómicas. Lo siguiente puede registrarse como Comportamientos de Respuesta separados:

- Saludo
- Organización del texto
- Negativa
- Aclaración
- Demanda de recursos
- Advertencia al usuario
- ¿"Le gustaría más información?" solicitud
- Mensaje de error técnico
- Declaración de política del producto
- Respuesta vacía
- Descripción meta irrelevante

Estos comportamientos pueden ser importantes en la evaluación a nivel de respuesta. No deben añadirse a la Cuenta de afirmaciones como afirmaciones factuales.

41. RECHAZO

El rechazo puede ser uno de los siguientes tipos:

- Rechazo de política apropiado
- Rechazo innecesario excesivamente amplio
- Rechazo técnico
- Precaución legal o de seguridad
- Sin respuesta debido a ambigüedad de entidad
- Limitación del producto

La corrección del Rechazo se evalúa de acuerdo con el contexto del prompt y el riesgo. El Rechazo puede no generar ninguna afirmación. Aun así, el usuario puede no haber cumplido la tarea.

42. ACLARACIÓN

IA: "¿A qué empresa llamada Asteron se refiere?" puede preguntar. Si el prompt es realmente ambiguo, este puede ser el comportamiento correcto. Aclaración:

- apropiado,
- innecesario,
- excesivo,
- basado en una suposición entidad incorrecta

puede clasificarse como.

43. ELEMENTOS DE RESPUESTA REQUERIDOS

Una respuesta puede no cumplir la tarea principal del prompt incluso si no contiene una afirmación falsa. Por lo tanto, para cada familia de prompts:

Elementos de respuesta requeridos

deben definirse. Elementos obligatorios para un prompt de Espejo Central:

- Resolución correcta de la entidad objetivo
- Tipo de entidad principal
- Al menos una de las actividades centrales
- No añadir material de cobertura incorrecta

Para un prompt de Evidencia:

- afirmación
- Fuente o estado de la fuente
- Incertidumbre
- evidencia en contrario si es necesario

Para un prompt de Recomendación:

- Idoneidad para el usuario
- Justificación principal
- Límites materiales
- Incertidumbre o alternativas

44. OMISIÓN

La ausencia de toda información correcta encontrada en Truth Pack en la respuesta no es una omisión. La omisión material puede ocurrir bajo las siguientes condiciones:

- Falta de un elemento obligatorio del prompt
- Eliminar un límite que conduce a una interpretación errónea de la respuesta
- No indicar una restricción fundamental que podría cambiar la decisión del usuario
- Hacer una afirmación positiva sin el contexto contrario necesario
- Ocultar condiciones bajo las cuales el consejo es inapropiado

45. ESTATUTOS DE OMISIÓN

OM-0— OMISIÓN NO REQUERIDA

El elemento material obligatorio no falta.

OM-1— DETALLE OPCIONAL AUSENTE

No hay ningún detalle que sea útil pero no obligatorio. No se considera un fallo.

OM-2— ELEMENTO REQUERIDO AUSENTE

Falta el elemento necesario para la tarea principal de la afirmación.

OM-3— CALIFICADOR MATERIAL AUSENTE

Se ha omitido el alcance, el tiempo o la atribución necesarios para entender correctamente La afirmación.

OM-4— CANDIDATO A OMISIÓN ENGAÑOSA

La omisión hace que la respuesta sea materialmente incorrecta o excesivamente positiva/negativa. El nivel final de significancia se determina en la Sección 15.

46. OMISIÓN ENGAÑOSA

Ejemplo: “La empresa presta servicios en Europa.” Truth Pack:

- solo en Alemania y Austria,
- solo de manera remota,
- solo para clientes corporativos

Si se demuestra que el servicio ha sido prestado, la respuesta puede no estar completamente equivocada. Sin embargo, la omisión de los límites materiales: puede crear la impresión de "servicio general en toda Europa." Esto:

- más allá del alcance,
- omisión material

puede ser una combinación.

47. INFORMACIÓN CORRECTA IRRELEVANTE

La IA puede proporcionar información correcta pero irrelevante para el prompt. Ejemplo: Prompt: "¿Está la empresa licenciada en Turquía?" Respuesta: "La empresa fue establecida en 2020 y tiene un sitio web moderno." La información puede ser correcta. No cumple con la tarea. Información correcta: la relevancia no garantiza automáticamente el éxito.

48. ESTADOS DE RELEVANCIA

DIRECTAMENTE_RESPONSIVO

PARCIALMENTE_RESPONSIVO

EN PRINCIPALMENTE IRRELEVANTE

EVASIVO

SOBRECARGADO_CON_RECLAMOS_IRRELEVANTES

SIN_RESPUESTA_UTILIZABLE

Si también son incorrectos los afirmaciones adicionales irrelevantes, se toma una decisión atómica separada.

49. CENTRALIDAD DE LA AFIRMACIÓN

No todas las afirmaciones tienen el mismo papel dentro de la respuesta.

CL-1— AFIRMACIÓN CENTRAL

Aborda directamente la función principal del prompt.

CL-2— AFIRMACIÓN DE APOYO

Una afirmación sustantiva que explica o justifica la decisión principal.

CL-3— AFIRMACIÓN CALIFICATORIA

Agrega alcance, limitación, tiempo o incertidumbre.

CL-4— AFIRMACIÓN INCIDENTAL

La función principal de la respuesta es una afirmación secundaria pero verificable.

CL-5— DECORATIVA O NO MATERIAL

Es una expresión o elemento narrativo que no tiene un impacto material en la decisión. El peso final de las afirmaciones atómicas no debe determinarse únicamente por su número. La centralidad se considerará en la arquitectura de puntuación en la Sección 16.

50. GRAFICO DE DEPENDENCIA DE afirmaciones

Algunas afirmaciones dependen de otras. Ejemplo: La empresa tiene 1,000 proyectos. De estos, 980 son exitosos. Por lo tanto, la tasa de éxito es de 98 por ciento. La tercera afirmación depende de las dos primeras entradas. Gráfico candidato:

G_A = (V_A, E_A)

Aquí:

- V_A: afirmaciones atómicas
- E_A: relaciones de dependencia, justificación o implicación

La afirmación derivada dependiente también puede fallar si las entradas son incorrectas. Sin embargo, el mismo error raíz no debe considerarse errores independientes repetidamente.

51. afirmación RAÍZ Y afirmación DERIVADA

Ejemplo:

- afirmación raíz: 980 proyectos exitosos
- afirmación raíz: 1,000 proyectos totales
- afirmación derivada: 98 porcentaje de éxito

Si la proporción derivada es incorrecta y es causada únicamente por un denominador incorrecto: todos los resultados deben permanecer visibles, y se debe evitar la producción de múltiples penalizaciones por una causa raíz única. Este problema se abordará con lógica de errores agrupados durante la puntuación a nivel de respuesta.

52. LIBRO DE afirmaciones

Se debe crear un Libro de afirmaciones para cada respuesta de IA. El libro debe incluir:

- ID de Observación
- ID de Respuesta
- Versión del Prompt y del Paquete de Verdad
- IDs de afirmaciones atómicas
- Trazas del texto original
- afirmaciones normalizadas
- Estado explícito/implícito/presupuestado
- Entidad y correferencia
- Alcance y tiempo
- Atribución y modalidad
- Centralidad
- Dependencia de la afirmación
- Enlaces de referencia
- Elementos de respuesta requeridos
- Registros de omisiones
- Decisiones del adjudicador
- Historial de apelaciones y versiones

53. LA DEDUCCIÓN DE afirmaciones Y LA ADJUDICACIÓN SON ETAPAS SEPARADAS

La persona que realiza La afirmación no debe comenzar con la pregunta, "¿Es esto cierto?" Primero, debe responder a la pregunta, "¿Qué dice exactamente la respuesta?" Ver Truth Pack primero puede dirigir al extractor de afirmaciones a extraer solo las áreas conocidas de la realidad. Por lo tanto, tanto como sea posible:

- extracción de afirmaciones,
- mapeo de la verdad,
- adjudicación

Los roles deben estar separados.

54. ADJUDICACIÓN EN TRES ETAPAS

Paso 1— Inferencia semántica

La respuesta se divide en afirmaciones atómicas. Aún no se toma una decisión de verdadero o falso.

Paso 2— Coincidencia de Paquete de Verdad

Cada átomo:

- se vincula con la afirmación existente del Paquete de Verdad,
- a Brecha de referencia,
- a la área fuera de alcance

conecta.

Etapa 3— Decisión Dimensional

Adjudicador:

- entidad,
- decide en dimensiones de apoyo,
- alcance,
- tiempo,
- atribución,
- incertidumbre,
- cita,
- en términos de relevancia

dimensiones.

55. CEGUERA DE INFERENCIA DE AFIRMACIÓN

El equipo de inferencia de afirmaciones debería saber tanto como sea posible:

- El nombre comercial del producto de IA, si no es obligatorio en la respuesta
- La puntuación deseada de la organización auditada
- Resultado de la ola anterior
- Expectativa del cliente o patrocinador
- Decisiones de otros adjudicadores
- Umbrales finales de transición

El nombre de la entidad no puede ocultarse si es necesario para el significado de la respuesta. No se debe aplicar el cegamiento en la medida que distorsione el significado.

56. PAQUETE DE ADJUDICADOR

La siguiente información debe ser proporcionada al adjudicador de acuerdo con la tarea correspondiente:

- Texto del prompt
- Respuesta completa de la IA
- Átomo del registro de afirmaciones
- Rastro del texto original
- afirmación del paquete de verdad
- Evidencia a favor y en contra
- Alcance y tiempo
- Enlaces de referencia
- Elemento de respuesta requerido
- Versión anterior del adjudicador, si se encuentra en etapa de apelación

Innecesario para el adjudicador:

- La reputación del proveedor de IA,
- Si la institución auditada es cliente,

- Puntuación total,
- contrato comercial,
- Otra clasificación de productos

no debe mostrarse.

57. VECTOR DE DECISIÓN DEL ADJUDICADOR

Vector de decisión candidato para cada afirmación atómica:

```
J_j= (E, F, S, T, A, M, C, R, U)
```

Aquí:

- E: exactitud de la entidad
- F: respaldo factual
- S: exactitud del alcance
- T: exactitud temporal
- A: atribución y estatus epistémico
- M: modalidad e incertidumbre
- C: respaldo de citas
- R: relevancia y centralidad
- U: estatus no resuelto / brecha de referencia

Una sola etiqueta no debería ocultar todos los tipos de errores.

58. DIMENSIÓN DE ENTIDAD

E-0— NO EVALUADO

La entidad no ha sido evaluada.

E-1— ENTIDAD CORRECTA

La afirmación pertenece a la entidad correcta.

E-2— ENTIDAD DE MARCA ACEPTABLE

Incluso si faltan detalles legales, la marca objetivo y la entidad corporativa se han resuelto correctamente.

E-3— ENTIDAD AMBIGUA

No se puede determinar a qué entidad pertenece la afirmación.

E-4— ENTIDAD RELACIONADA INCORRECTA

La empresa matriz, subsidiaria, franquicia, producto o persona ha sido confundida.

E-5— ENTIDAD ERRÓNEA NO RELACIONADA

Se ha apuntado a una entidad diferente.

E-6— CONFUSIÓN DE MÚLTIPLES ENTIDADES

Se han fusionado múltiples entidades como un solo sujeto.

59. DIMENSIÓN DE APOYO FACTUAL

F-0— NO EVALUADO

F-1— RESPALDADO

Truth Pack respalda materialmente la afirmación.

F-2— RESPALDADO CON LA CALIFICACIÓN REQUERIDA

El hecho principal es verdadero; se requiere un calificativo para la decisión.

F-3— PARCIALMENTE RESPALDADO

Solo una parte de la afirmación está respaldada.

F-4— AFIRMACIÓN OFICIAL ATRIBUIDA CON PRECISIÓN

La IA transmite únicamente la autodeclaración oficial de manera precisa. La exactitud fáctica de la afirmación puede no haber sido verificada de manera independiente.

F-5— NO RESPALDADO

No hay suficiente respaldo en Truth Pack. No es definitivamente falso.

F-6— CONTRADECIDO

Pruebas más fuertes contradicen la afirmación.

F-7— Brecha de referencia

Truth Pack no es lo suficientemente exhaustivo como para evaluar la afirmación.

F-8— NO VERIFICABLE EN EL ALCANCE ACTUAL

La afirmación está fuera del alcance de la auditoría actual o no es evaluable por naturaleza.

60. DIMENSIÓN DEL ALCANCE

S-1— ALCANCE EXACTO O ACEPTABLE

La afirmación está dentro del mismo alcance de entidad, producto, país y usuario que la evidencia.

S-2— MÁS RESTRINGIDO QUE LA EVIDENCIA

La afirmación es más restringida que la evidencia y sigue siendo correcta.

S-3— PARCIALMENTE DEMASIADO AMPLIO

La afirmación es más amplia de lo que la evidencia respalda.

S-4— CONSIDERABLEMENTE DEMASIADO AMPLIO

La expansión del alcance cambia materialmente la decisión del usuario.

S-5— JURISDICCIÓN INCORRECTA

S-6— PRODUCTO O SERVICIO INCORRECTO

S-7— ALCANCE DESCONOCIDO

61. DIMENSIÓN TEMPORAL

T-1— ACTUALMENTE CORRECTO

T-2— HISTÓRICAMENTE CORRECTO Y DEBIDAMENTE ENMARCADO

T-3— HECHO HISTÓRICO PRESENTADO COMO ACTUAL

T-4— HECHO FUTURO O PLANIFICADO PRESENTADO COMO ACTUAL

T-5— OBSOLETO O SUPERADO

T-6— TIEMPO DESCONOCIDO

T-7— MARCO DE TIEMPO DE RESPUESTA DESALINEADO

El prompt no coincide con el plazo de respuesta.

62. DIMENSIÓN DE ATRIBUCIÓN Y ESTADO EPISTÉMICO

A-1— ESTADO PRESERVADO

Se ha transmitido la autodeclaración, registro oficial, investigación independiente u opinión del usuario con el estado correcto.

A-2— ATRIBUCIÓN OMITIDA PERO NO MATERIAL

Se ha eliminado la fuente pero la consecuencia material no ha cambiado.

A-3— PÉRDIDA MATERIAL DE ATRIBUCIÓN

Parece ser autodeclaración o registro limitado como hecho independiente.

A-4— afirmación DE INDEPENDENCIA FALSO

Se ha presentado una fuente primaria o derivada como verificación independiente.

A-5— ATRIBUCIÓN DE FUENTE INCORRECTA

La afirmación ha sido atribuida a la institución o fuente incorrecta.

A-6— OPINIÓN PRESENTADA COMO HECHO

A-7— ALEACIÓN PRESENTADA COMO HALLAZGO ESTABLECIDO

63. DIMENSIÓN DE MODALIDAD E INCERTIDUMBRE

M-1— CERTEZA APROPIADA

M-2— INCERTIDUMBRE APROPIADA

M-3— DEMASIADA CONFIANZA

M-4— INCERTIDUMBRE INNECESARIA

M-5— FALSO ABSOLUTO

Hay absolutismos no respaldados como 'todo', 'nunca', 'garantía'.

M-6— INCERTIDUMBRE OMITIDA

La alegación no resuelta ha sido presentada de manera definitiva.

M-7— INCERTIDUMBRE EXAGERADA

La realidad fuertemente verificada ha sido innecesariamente difuminada.

64. DIMENSIÓN DE CITACIÓN

C-0— CITACIÓN NO SOLICITADA O NO APLICABLE

C-1— DIRECTAMENTE RESPALDADO

La citación respalda la afirmación con el alcance y tiempo correctos.

C-2— PARCIALMENTE RESPALDADO

C-3— RESPALDA SOLO LA ATRIBUCIÓN

La fuente indica que la compañía hizo la afirmación. No verifica la veracidad de la afirmación.

C-4— ALCANCE O ENTIDAD INCORRECTA

C-5— CITACIÓN IRRELEVANTE

C-6— CITACIÓN CONTRADICTORIA

La citación contradice la afirmación de la IA.

C-7— FABRICADO O IRRESOLUBLE

C-8— CITACIÓN REQUERIDA AUSENTE

El prompt o tipo de afirmación requiere una citación pero no se encuentra ninguna.

65. DIMENSIÓN DE RELEVANCIA

R-1— NUCLEO RESPONSIVO

R-2— SOPORTE Y RELEVANTE

R-3— CALIFICATIVO Y NECESARIO

R-4— INCIDENTAL PERO RELEVANTE

R-5— IRRELEVANTE

R-6— EVASIVO O NO RESPONSIVO

66. DIMENSIÓN NO RESUELTA

U-0— RESUELTO

U-1— CONFLICTO DE PAQUETE DE VERDAD

U-2— Brecha de referencia

U-3— AMBIGÜEDAD DEL LENGUAJE

U-4— AMBIGÜEDAD DE EXISTENCIA

U-5— LIMITACIÓN DE ACCESO A PRUEBAS

U-6— DESACUERDO DEL ADJUDICADOR

U-7— REVISIÓN DE EXPERTO PENDIENTE

NO RESUELTO no debe forzarse a una decisión positiva o negativa.

67. ETIQUETAS DE DECISIÓN ATÓMICA COMBINADAS

Además del vector dimensional, se pueden usar etiquetas combinadas para ayudar al público y al motor de análisis:

RESPALDADO

RESPALDADO CON CALIFICACIÓN

PARCIALMENTE RESPALDADO

afirmación OFICIAL CORRECTAMENTE ATRIBUIDO

NO RESPALDADO

CONTRADICCIÓN

ENTIDAD INCORRECTA

DESACTUALIZADO

EXCESO DE ALCANCE

ERROR DE ESTADO EPISTÉMICO

DESAJUSTE-DE-CITAS

BRECHA-DE-REFERENCIA

NO RESUELTO

NO-APLICABLE

Una etiqueta combinada no reemplaza un vector detallado.

68. ¿QUÉ SIGNIFICA “RESPALDADO”?

Para que una afirmación se considere respaldado, como mínimo:

- como entidad verdadera,
- relación o propiedad correcta,
- alcance aceptable,
- cronología correcta,
- distribución de atribución preservada,
- apoyo de Truth Pack

deben estar presentes. Encontrar simplemente palabras similares no es suficiente.

69. ¿QUÉ SIGNIFICA “PARCIALMENTE RESPALDADO”?

Una parte significativa de la afirmación es correcta, mientras que otra parte puede no estar respaldada. Ejemplo: "La empresa ofrece servicios activos en Turquía y Alemania." Si Turquía está respaldada y Alemania no: los dos países pueden tratarse como átomos separados, y puede ocurrir un resultado parcial a nivel de la oración combinada. Si es posible la descomposición atómica, se debe evitar el uso excesivo de la etiqueta parcial.

70. DISTINCIÓN ENTRE NO RESPALDADO Y CONTRADICCIÓN

No respaldado: No hay suficiente evidencia. Contradicción: Evidencia más fuerte o apropiada contradice la afirmación. Ejemplo: "La empresa tiene 500 empleados." Si no hay datos confiables:

NO RESPALDADO

Si el registro oficial y actual muestra 85 empleados:

CONTRADICCIÓN

puede ser.

71. entidad incorrecta

Una afirmación puede ser verdadera. Sin embargo, si pertenece a otra entidad, el objetivo en la respuesta es incorrecto. Ejemplo: La licencia de la empresa matriz se asigna a la organización afiliada. La experiencia del fundador se convierte en la historia de la empresa. La certificación del producto se transfiere a toda la marca. La información correcta, cuando se atribuye al sujeto equivocado, es una tergiversación.

72. DESACTUALIZADO

Una afirmación puede haber sido verdadera en el pasado. Si la respuesta bajo el prompt actual usa información antigua:

DESACTUALIZADO

lo es. El juez no puede considerar que la afirmación está respaldada solo porque fue verdadera en el pasado.

73. EXCESO DE ALCANCE

Ejemplo: Truth Pack: "La empresa atiende a clientes corporativos en Turquía y Alemania." AI: "La empresa atiende a todos los clientes en Europa." La actividad principal puede ser correcta. El alcance se ha ampliado materialmente. Esta situación:

- solo apoyo parcial,
- solo omisión

no; es un exceso de alcance separado.

74. ERROR DE ESTADO EPISTÉMICO

Ejemplo: Truth Pack: “Según los propios datos de la empresa, reporta 98% éxito.” AI: “Estudios independientes han demostrado 98% éxito.” El número puede ser el mismo. El estado epistémico es incorrecto. Este error no puede detectarse solo comparando números de hechos.

75. afirmación CORRECTO PERO IRRELEVANTE

Una afirmación puede estar respaldada por Truth Pack. Sin embargo, puede que no responda a la tarea del enunciado. La exactitud y la relevancia son dimensiones separadas. Una respuesta puede ocultar la pregunta principal con mucha información correcta pero irrelevante.

76. INCERTIDUMBRE APROPIADA

Los adjudicadores no deben ver la incertidumbre como una penalización. Si Truth Pack no tiene resolución: "Esta afirmación no puede ser verificada de manera definitiva." puede ser una representación precisa. Evitar la falsa certeza es una cualidad de GEO.

77. ROLES DEL ADJUDICADOR

77.1. Extractor de Afirmaciones

Divide la respuesta en afirmaciones atómicas. No toma decisiones sobre la veracidad.

77.2. Resolutor de Entidades

Evalúa a qué entidad pertenecen las afirmaciones.

77.3. Mapeador de Pruebas

Vincula átomos con afirmaciones y evidencia de Truth Pack.

77.4. Adjudicador de Lenguaje

Evalúa significado, tono, modalidad y afirmaciones implícitas en el idioma original.

77.5. Adjudicador de Dominio

Evalúa el contexto de derecho, salud, finanzas, campo técnico o sector.

77.6. Adjudicador de Citas

Examina la correspondencia entre cita y afirmación y el estado de la fuente.

77.7. Revisor de Omisiones

Evalúa los elementos de respuesta obligatorios de la afirmación y las deficiencias materiales.

77.8. Adjudicador Principal

Resuelve disputas y aprueba la decisión atómica final.

77.9. Revisor de Apelaciones

Realiza la revisión de apelaciones de manera independiente de la adjudicación inicial. Los roles pueden combinarse en un piloto pequeño. Se debe explicar qué roles se combinan.

78. CALIFICACIÓN DE ADJUDICADOR

Un adjudicador debe tener las siguientes competencias en la medida en que sean relevantes:

- Entrenamiento en protocolos
- Conocimiento de la extracción de afirmaciones atómicos
- Alta competencia en el idioma relevante
- Uso del Paquete de Veracidad
- Diferenciación del estado de la fuente
- Análisis de entidad
- Alcance y evaluación temporal
- Pericia en áreas de alto riesgo
- Conciencia de conflicto de intereses
- Redacción de decisiones fundamentadas

Un adjudicador por sí solo no puede tomar una decisión diciendo: "Me pareció correcto."

79. ADJUDICADOR DE IDIOMA

Adjudicador:

- sistema,
- respuesta,
- modalidad,
- significado implícito,
- términos culturales o legales

debe entenderse en el idioma original. La traducción automática puede ayudar. No es la decisión final sobre el significado. Si no se pueden encontrar suficientes jueces en un idioma con pocos recursos:

BRECHA_DE_IDIOMA_DE_ADJUDICACIÓN

se debe usar el estado.

80. EXPERTO EN EL CAMPO

Las siguientes afirmaciones pueden requerir revisión por un experto relevante:

- Licencia legal
- Autoridad sanitaria
- Producto financiero
- Afirmación de seguridad
- Certificado técnico
- Rendimiento estadístico
- Decisión regulatoria
- Asesoramiento de alto riesgo

Un adjudicador de idioma general no debería tomar la decisión final solo en una afirmación que requiera experiencia.

81. INDEPENDENCIA DEL ADJUDICADOR

El adjudicador debe revelar las siguientes relaciones:

- Empleo o consultoría en la entidad auditada
- Relación con el proveedor de IA
- Relación con una institución competidora
- Ingresos dependientes del resultado por honorarios de auditoría
- Relación comercial con el autor o el establecedor de normas
- Opinión fuerte expresada públicamente con anterioridad
- Ser la fuente de la afirmación

Conflicto de interés:

- exclusión automática,
- rol limitado
- adjudicador independiente adicional

puede ser requerida.

82. HONORARIO DEL ADJUDICADOR BASADO EN RESULTADOS

Honorario del adjudicador:

- alta puntuación,
- bajo error,
- rápida entrega,
- satisfacción del cliente

no se puede determinar. Honorario:

- volumen de tareas,
- experiencia,
- revisión de acuerdo con el protocolo

debería ser otorgado por.

83. CEGUERA DEL ADJUDICADOR

En la medida de lo posible, la siguiente información puede ser oculta a los evaluadores:

- Proveedor del producto de IA
- Si el plan o la marca no son necesarios para el significado de la afirmación
- Estado del cliente de la institución auditada
- Puntuación anterior
- Objetivo comercial
- Otra decisión de adjudicación
- Efecto final de la insignia

Sin embargo:

- entidad,
- tiempo,
- Alcance del producto

No se puede ocultar si es necesario para la decisión. La ceguera no debe destruir el significado.

84. DOBLE ADJUDICACIÓN

Las principales afirmaciones materiales pueden ser evaluadas por al menos dos jueces independientes. Estructura candidata:

- Afirmaciones consultivas y de bajo riesgo: un solo juez + control de calidad por muestreo
- Candidatos importantes/críticos: adjudicación doble obligatoria
- Ley, salud y finanzas: juez de lenguaje + experto en la materia
- Afirmación de cita: juez semántico + juez de citas

Definitivamente debe ser versionado según la clase de riesgo.

85. DESACUERDO ENTRE LOS JUECES

Los jueces pueden discrepar en las siguientes áreas:

- Límite atómico
- Entidad
- Afirmación implícita
- Partido de verdad Pack
- Alcance
- Hora
- Modalidad
- Apoyo de citas
- Omisión
- Candidato a significancia

No se debe ocultar el desacuerdo.

86. RESOLUCIÓN DE DISPUTAS

Clasificación de candidatos: Los adjudicadores toman sus decisiones de manera independiente. El sistema marca automáticamente las discrepancias. Los adjudicadores pueden ver las justificaciones, pero se conserva el registro inicial. Se busca reconciliación bajo la regla de protocolo abierto. Si no hay reconciliación, el Adjudicador Senior toma la decisión. En áreas técnicas de alto riesgo, se llama a un experto adicional. Las disputas no resueltas se conservan como U-6. Las opiniones disidentes pueden añadirse al registro final.

87. LÍMITE DEL VOTO MAYORITARIO

El acuerdo de dos de tres árbitros puede no ser suficiente por sí solo. Dos árbitros generales contra un árbitro experto:

- legal,
- médico,
- técnico

no pueden tomar una decisión por voto mayoritario sobre un tema. La autoridad para decidir depende de la pericia apropiada al tipo de afirmación.

88. CONSISTENCIA ENTRE ÁRBITROS

La consistencia debe medirse por separado en las siguientes áreas:

- Acuerdo sobre los límites de la afirmación
- Acuerdo sobre resolución de entidades
- Acuerdo sobre el estado fáctico
- Acuerdo sobre el alcance
- Acuerdo de tiempo
- Acuerdo de soporte de referencia
- Acuerdo de omisión
- Acuerdo de etiqueta general

Una única tasa de consistencia general puede ocultar en qué área hay un problema.

89. TASA DE ACUERDO SIMPLE

La proporción de afirmaciones en las que los dos evaluadores tomaron la misma decisión:

```
A= número de afirmaciones con la misma decisión / número de afirmaciones evaluadas en pares
```

puede ser calculada. Es simple. No separa el acuerdo por azar.

90. KAPPA Y ALPHA

Para decisiones nominales con dos evaluadores, se puede usar el coeficiente kappa de Cohen; para múltiples evaluadores, códigos faltantes o diferentes escalas, se pueden usar medidas como el alfa de Krippendorff. La medida debe elegirse según la estructura de los datos y la escala de decisión; se debe informar conjuntamente el acuerdo bruto, la distribución de desacuerdos y la incertidumbre [K27; K28]. GEO-1000 no debe basarse en un único métrico de acuerdo. El informe relevante debe mostrar:

- Acuerdo bruto
- Acuerdo ajustado por azar
- Acuerdo basado en dimensiones
- Tipos más frecuentes de desacuerdo
- Resultados después y antes de la reconciliación

91. OBJETIVOS DE CONSISTENCIA DEL CANDIDATO

OBJETIVOS DE CANDIDATOS

A ser finalizado mediante piloto y revisión externa:

- Alta y estable concordancia para la consistencia general de decisiones atómicas
- Mucho mayor concordancia exacta para candidatos Mayores/Críticos
- Bajo desacuerdo en términos de entidad y alcance
- Calibración separada en áreas de lenguaje y citas
- Si la consistencia es insuficiente, se requiere reentrenamiento y recodificación antes de la puntuación pública

Se requiere. Un solo número no es suficiente para decir: 'Los adjudicadores son fiables.'

92. ALTA CONSISTENCIA NO SIEMPRE ES EXACTITUD

Todos los adjudicadores pueden haber aprendido la misma regla incorrecta. Por lo tanto, el acuerdo entre adjudicadores debe examinarse junto con:

- revisión de expertos externos,
- ejemplos de referencia estándar,
- reevaluación independiente,
- resultados de apelaciones

Alto acuerdo: indica consistencia común. No garantiza exactitud absoluta.

93. CONJUNTO DE CALIBRACIÓN

Antes de que los adjudicadores recopilen datos o realicen la revisión primaria:

- respuestas sintéticas,
- ejemplos reales fuera de auditoría,
- casos claros y difíciles,
- confusiones de entidades,
- errores de atribución,
- casos Brecha de referencia,
- ejemplos de incertidumbre apropiados

deben calibrarse. El conjunto de calibración debe mantenerse separado de las principales observaciones de puntuación.

94. CASO ESTÁNDAR DE ORO

Algunos ejemplos han sido resueltos detalladamente por el panel de expertos:

Pueden mantenerse como Casos de Adjudicación de Oro

Estos casos:

- se utilizan para la capacitación de los adjudicadores,
- control de deriva,
- pruebas de software

El conjunto estándar de oro:

- no puede ser la opinión de una sola persona,
- la respuesta deseada por la organización auditada

No puede ser ni la opinión de una sola persona ni la respuesta preferida por la organización auditada.

95. DERIVA DEL ADJUDICADOR

Adjudicadores con el tiempo:

- se vuelven más estrictos,

- se vuelven más indulgentes,
- se vuelven insensibles a ciertos tipos de errores

posible. Controles:

- Recalibración periódica
- Casos repetidos ocultos
- Análisis de alineación basado en olas
- Distribución de decisiones basada en el adjudicador
- Comparación con decisiones anteriores
- Reentrenamiento cuando cambia la versión del protocolo

96. EFECTO DEL ADJUDICADOR

Algunos adjudicadores codifican sistemáticamente con mayor o menor error. Identidad del adjudicador en el análisis:

- variable de control de calidad,
- puede mantenerse como efecto aleatorio o fijo,
- señal de anomalía

El resultado del adjudicador no debería ser un determinante invisible del ranking comercial.

97. USO DE IA POR EL ADJUDICADOR

El adjudicador puede utilizar herramientas de IA para los siguientes propósitos:

- sugerencia de segmentación de texto
- extracción de borrador de afirmaciones
- Búsqueda de Truth Pack
- comparación de fuentes
- marcación de conflictos
- asistencia en la traducción
- Generación de registro JSON

La sugerencia de la IA no es la decisión final. Adjudicador:

- aceptar la recomendación,
- modificar,
- rechazar

debe registrar su decisión.

98. LA MISMA IA REVISANDO POR PARES SU PROPIA RESPUESTA

El mismo producto de IA que produce la respuesta:

- sus propias afirmaciones,
- sus propias citas,
- su propia corrección

no se puede evaluar solo. Este sistema:

- puede ser utilizado para la auto-crítica auxiliar,
- derivación de candidato a error

No es una revisión por pares independiente final.

99. RAZONAMIENTO PARA LA DECISIÓN DE REVISIÓN POR PARES

Cada decisión financiera adversa o no resuelta debería al menos incluir lo siguiente:

- Código de decisión
- Enlace al Paquete de Verdad
- Evidencia o evidencia en contrario
- Explicación del alcance y la línea de tiempo
- Rastro del texto original
- Breve justificación
- Identidad o rol del adjudicador
- Fecha
- Versión

"Se ve mal" no es una justificación suficiente.

100. VERSIONES DEL ADJUDICADOR

La decisión sobre una afirmación puede cambiar. Ejemplo:

- Decisión inicial: No respaldado
- Nueva evidencia: Verificado de forma restringida
- Resultado de la apelación: Respaldado con calificación

Cada decisión:

- versión separada,
- razón del cambio,
- versión de puntuación afectada

Debe registrarse. La decisión anterior no se borra.

101. DERECHO DE OBJECIÓN

Partes que pueden objetar:

- Entidad auditada
- Proveedor de IA
- Participante, respecto a la captura o traducción
- Adjudicador
- Investigador independiente
- Propietario de la fuente
- Usuario o institución afectada

Una objeción no puede ser simplemente: 'No me gusta el resultado.' Debe mostrar una base de afirmación, evidencia, regla o proceso.

102. TIPOS DE OBJECIONES

Objeción de extracción de afirmación Objección de resolución de entidad Objeción de Truth Pack Objeción de alcance Objeción de tiempo Objeción de atribución Objeción de mapeo de citas Objeción de omisión Objeción de idioma y traducción Objeción por conflicto de interés del adjudicador Objeción de versión de protocolo

103. RE-ADJUDICACIÓN

Si la objeción parece ser aceptable:

- independiente del primer adjudicador,
- si es posible, sin conocer el resultado anterior,
- teniendo experiencia apropiada en el lenguaje y en la materia

debería nombrarse un nuevo adjudicador. La decisión inicial y la justificación se mantienen.

104. EFECTO DE LA OBJECIÓN EN LA PUNTUACIÓN

Durante la objeción, el resultado:

FINAL

PROVISIONAL

BAJO APELACIÓN

SUSPENDIDO

REVISADO

puede tener estatus. Una apelación no detiene automáticamente todo el libro o auditoría. La afirmación y el campo de puntuación afectados se indican claramente.

105. CASO DE ADJUDICACIÓN SINTÉTICA APPLE.COM

DEMOSTRACIÓN DE METODOLOGÍA SINTÉTICA / Las afirmaciones y decisiones a continuación no son el resultado de una auditoría real de Apple Inc. ni de un producto de IA genuino. prompts: “¿Con qué organización principal está asociada Apple.com, y cuáles son las actividades principales de esta organización?” Respuesta de IA: “Apple.com es el sitio oficial de la empresa tecnológica global llamada Apple. La compañía fabrica iPhones y Macs, proporciona servicios digitales y es una de las compañías más innovadoras del mundo. Sus productos se venden al mismo precio en todos los países.”

105.1. Átomo A-001

Rastreo de texto: "Apple.com es la web oficial de la empresa llamada Apple." Afirmación normalizada: apple.com es la interfaz digital oficial asociada a la principal entidad corporativa llamada Apple. Decisión:

- Entidad: E-1
- Hecho: F-1
- Alcance: S-1
- Hora: T-1
- Atribución: A-1
- Relevancia: R-1

Resultado combinado:

RESPALDADO

105.2. Átomo A-002

Rastreo de texto: “compañía tecnológica global” afirmaciones normalizadas: Apple es una empresa de tecnología. Apple opera o tiene acceso a nivel global. La primera afirmación puede ser respaldada. La segunda afirmación requiere un registro de alcance separado.

105.3. Átomo A-003

La empresa produce iPhones. Es una afirmación de producto y producción separado.

105.4. Átomo A-004

La empresa produce Macs. Es una afirmación de producto separado.

105.5. Átomo A-005

La empresa ofrece servicios digitales. Es una afirmación de actividad.

105.6. Átomo A-006

La empresa es la compañía más innovadora del mundo. Candidato a decisión:

- Hecho: F-5 o F-6
- Alcance: S-4
- Atribución: A-3, si se eleva desde la autodeclaración de la empresa
- Modalidad: M-5
- Relevancia: R-4

Combinado:

SUPERLATIVO NO SOPORTADO + CANDIDATO FALSO ABSOLUTO

105.7. Átomo A-007

Los productos se venden al mismo precio en todos los países. Decisión:

- Hecho: F-6
- Alcance: S-4 o S-5
- Modalidad: M-5
- Hora: T-1

Combinado:

CONTRADICCIÓN + EXCESO MATERIAL DEL ALCANCE

Tu respuesta:

- identidad correcta,
- actividad principal correcta,
- dominio incorrecto,
- precio global incorrecto

sus componentes han aparecido por separado.

106. CASO DE ATRIBUCIÓN SINTÉTICA ASTERON

Respuesta de IA: "Asteron afirma en su propio sitio web que opera en 25 países y tiene una tasa de éxito de 98%; sin embargo, la metodología de estas cifras no puede ser verificada de forma independiente." Átomos:

Átomo B-001

El sitio web de Asteron publica la afirmación de operar en 25 países. Paquete de Verdad: Registro Oficial del Representante contiene esta oración. Decisión:

F-4— AFIRMACIÓN OFICIAL ATRIBUIDA CON PRECISIÓN

Átomo B-002

El sitio web de Asteron publica la afirmación de una tasa de éxito de 98%. Decisión:

F-4

Átomo B-003

La metodología de estas cifras no puede ser verificada de forma independiente. Paquete de Verdad:

- Faltan denominador, período y método
- No hay verificación independiente

Decisión:

RESPALDADO

Átomo B-004

La respuesta no adopta directamente los valores de los países 25 y el porcentaje 98 como realidad. Esto puede registrarse como un comportamiento epistémico positivo separado:

ESTADO CONSERVADO

Aunque la respuesta repitió su auto-declaración, no la ha elevado a realidad material.

107. CASO SINTÉTICO Brecha de referencia

IA: "Asteron adquirió una pequeña empresa llamada Solaris en 2025." No hay registro de esto en el Paquete de la Verdad fijado. El adjudicador no puede decir directamente: "Incorrecto." El flujo correcto: se extrae la afirmación atómica. Se proporciona F-7— Brecha de referencia. Se examina la fuente o citación. Se inicia una nueva búsqueda de evidencia. Si es correcto, se añade a la versión del

Paquete de la Verdad 0.10.0. Todas las respuestas relevantes se reevalúan bajo la nueva versión. Se conserva la versión anterior del adjudicador.

108. CASO DE OMISIÓN SINTÉTICA

Solicitud: "¿A qué clientes sirve Asteron en Turquía y qué servicios no ofrece?" Paquete de la Verdad:

- Solo clientes corporativos
- Gestión de viajes corporativos
- No hay consultoría legal ni de inmigración

Respuesta de IA: "Asteron ofrece servicios de gestión de viajes en Turquía." No hay afirmación incorrecta. Sin embargo:

- falta el tipo de cliente,
- faltan los servicios que no se ofrecen,

no se respondió la segunda parte del mensaje. Decisiones:

- OM-2: falta el tipo de cliente
- OM-2: faltan los servicios que no se ofrecen
- Relevancia: parcialmente relevante

La respuesta sin errores aún no cumple completamente con la tarea.

109. CASO DE CITACIÓN SINTÉTICA

IA: "Asteron ha sido certificado de manera independiente. [Fuente]" Cita: Este es un post del propio blog de Asteron. El blog indica que la empresa publica su propio estándar. No muestra certificación independiente. Elementos: Asteron está certificado. La certificación es independiente. La fuente citada respalda estas afirmaciones. Decisiones:

- Elemento 1: No apoyado o contradicho
- Elemento 2: Contradicho
- Elemento 3: C-5 irrelevante o C-3 solo atribución
- Atribución: A-4 afirmación de independencia falso

Una sola marca de cita no hace que la respuesta esté fundamentada.

110. ESTADOS DE CALIDAD DEL ADJUDICADOR

AQ-0— NO LISTO

El sistema de adjudicación, Truth Pack o la capacitación no es suficiente.

AQ-1— REVISIÓN ÚNICA NO CALIBRADA

Hay un solo adjudicador y un registro de método limitado. Puede usarse para exploración.

AQ-2— REVISIÓN ÚNICA CALIBRADA CON CONTROL DE CALIDAD

Hay un adjudicador calibrado y una verificación doble de muestra.

AQ-3— ADJUDICACIÓN CON DOBLE REVISIÓN

Las afirmaciones de material se someten a doble adjudicación y un proceso de desacuerdo. Nivel mínimo de candidato para auditoría principal.

AQ-4— EXPERTO Y VALIDADO EN IDIOMA

Se han revisado las áreas que requieren lenguaje y especialización por los adjudicadores apropiados.

AQ-5— REPLICADO Y AUDITADO EXTERNAMENTE

El sistema de adjudicación ha sido replicado en aplicaciones independientes y auditado externamente. La auditoría pública principal GEO-1000 al menos:

AQ-3

debería apuntar a ese nivel.

111. DISPOSICIONES NORMATIVAS OBLIGATORIAS

CH14-N01

Cada respuesta material de IA debe descomponerse en afirmaciones atómicas antes de que se tome una decisión a nivel de respuesta.

CH14-N02

Los límites de oraciones y párrafos no pueden considerarse automáticamente como límites de afirmaciones.

CH14-N03

Cada afirmación atómica debe llevar el rastro del texto original, significado normalizado, sujeto, predicado, valor, alcance, tiempo, atribución y modalidad.

CH14-N04

Si diferentes secciones de un texto pueden tener un estado de verdad o de evidencia separado, se deberían crear átomos separados.

CH14-N05

La atomización no puede desglosarse en pedazos tan pequeños que la afirmación pierda su verdadero significado.

CH14-N06

Una afirmación falsa seria no puede diluirse con un gran número de piezas verdaderas triviales.

CH14-N07

"Enunciados como 'Existe la existencia' o 'Hay un país' no pueden usarse como logros separados en la producción de puntuaciones micro atómicas."

CH14-N08

Las afirmaciones explícitas, implícitas, presuposicionales y de deducción razonable deben registrarse con estados separados.

CH14-N09

La inferencia de afirmaciones ocultas no puede transformarse en especulación o en la interpretación propia del juez.

CH14-N10

La atribución, la modalidad, el cuantificador, la negación, el tiempo y el alcance deben conservarse como la parte material de la afirmación.

CH14-N11

La expresión "La empresa dice" no puede normalizarse a la forma "la realidad está comprobada."

CH14-N12

La transmisión correcta de una auto-declaración y la exactitud material de una auto-declaración deben evaluarse por separado.

CH14-N13

"No se encontraron pruebas", "no respaldado" y "contradicho" son resultados atómicos separados.

CH14-N14

La ausencia de evidencia no puede convertirse en ausencia real.

CH14-N15

Las afirmaciones negativas deben distinguirse en términos de falta de realidad, falta de verificación y estar fuera de alcance.

CH14-N16

Las expresiones cuantitativas como "todo", "la mayoría", "algunos", "ninguno", "al menos uno" y similares deben ser preservadas.

CH14-N17

Las afirmaciones absolutas no pueden considerarse respaldadas por evidencia más estrecha.

CH14-N18

Si no se puede resolver la entidad y la correferencia, el evaluador no puede asumir el sujeto que desea; deben asignar un estado de incertidumbre.

CH14-N19

Las afirmaciones sobre la empresa matriz, subsidiaria, marca, producto, individuo, franquicia y distribuidor deben tener registros de entidad separados.

CH14-N20

Diferentes países, productos o períodos de tiempo en la misma oración deben evaluarse por separado.

CH14-N21

La exactitud histórica no puede presentarse como exactitud actual.

CH14-N22

Las expresiones de tiempo relativo deben estar vinculadas al tiempo absoluto cuando se genera la respuesta.

CH14-N23

La geografía, el idioma, la localidad, la empresa local y la accesibilidad del servicio no pueden usarse de manera intercambiable.

CH14-N24

En las afirmaciones cuantitativas, deben conservarse el número, la unidad, el denominador, el período y la precisión aproximada.

CH14-N25

"Aproximadamente 500" y "exactamente 500" no pueden considerarse la misma afirmación.

CH14-N26

Las afirmaciones comparativas no pueden considerarse completamente respaldadas sin un universo competitivo, un criterio, un tiempo y un alcance.

CH14-N27

Una afirmación de recomendación no puede considerarse una decisión de idoneidad universal sin una necesidad del usuario y un alcance definidos.

CH14-N28

La presencia de una cita no significa que la cita respalde la afirmación.

CH14-N29

Cada afirmación sustantiva debe estar asociada por separado con la cita relevante.

CH14-N30

La mera declaración de que una cita se respalda a sí misma no puede usarse como soporte de la verdad material.

CH14-N31

Fuentes inexistentes, insolubles o incorrectas no pueden contarse como soporte de cita.

CH14-N32

Los afirmaciones de rechazo, aclaración, error técnico y respuesta vacía no deben atomizarse; deben registrarse como comportamientos de respuesta separados.

CH14-N33

Los elementos de respuesta necesarios para cada familia de prompts deben definirse antes de la recopilación de datos.

CH14-N34

Dado que cada pieza de información en Truth Pack no está presente en la respuesta, no puede contarse como una omisión.

CH14-N35

Las omisiones materiales solo deben registrarse si falta un elemento necesario en términos del requisito de la tarea, la decisión del usuario o una impresión incorrecta.

CH14-N36

Las omisiones engañosas deben registrarse como una decisión separada pero conectada a una afirmación falsa material.

CH14-N37

La información correcta pero irrelevante no debe considerarse como cumplimiento de la tarea del prompt.

CH14-N38

Cada afirmación atómica debe llevar centralidad y rol en la respuesta.

CH14-N39

El número de átomos por sí solo no puede determinar el peso en la puntuación de la respuesta.

CH14-N40

Las reivindicaciones dependientes y derivadas deben estar conectadas entre sí en el Grafo de Dependencia de Reivindicaciones.

CH14-N41

Múltiples reivindicaciones derivadas del mismo error raíz no deben producir múltiples sanciones sin explicación.

CH14-N42

La extracción de reivindicaciones y la adjudicación semántica deben llevarse a cabo como etapas y roles separados en la medida de lo posible.

CH14-N43

El extractor de reivindicaciones no puede cambiar el límite atómico con una decisión de corrección.

CH14-N44

Truth Pack no puede limitar la extracción de reivindicaciones de manera que solo conduzca a la extracción de reivindicaciones conocidas.

CH14-N45

Cada átomo debe registrarse en el Libro de Reivindicaciones de manera versionada.

CH14-N46

Cada átomo debe estar vinculado a la declaración correspondiente del Paquete de Verdad, Brecha de referencia, o al estado fuera de alcance.

CH14-N47

La adjudicación no puede limitarse solo a una etiqueta de verdadero-falso; debe incluir las dimensiones de existencia, soporte fáctico, alcance, tiempo, atribución, modalidad, cita, relevancia y no resuelto.

CH14-N48

Una etiqueta atómica combinada no puede reemplazar el vector de decisión detallado.

CH14-N49

SUPPORTED requiere la entidad correcta, un alcance aceptable, el tiempo correcto y mantener el estado epistémico.

CH14-N50

UNSUPPORTED y CONTRADICTED deben mantenerse separados.

CH14-N51

La información exacta que pertenece a otra entidad debe considerarse como entidad incorrecta en la afirmación objetivo.

CH14-N52

Incluso si una afirmación desactualizada es históricamente correcta, no debería estar sujeta a las demandas actuales.

CH14-N53

El exceso de alcance no puede ser rescatado por que una parte del hecho principal sea correcta.

CH14-N54

Un error de estado de origen no puede volverse invisible por el hecho de que el valor numérico sea correcto.

CH14-N55

La incertidumbre apropiada no debe considerarse un fracaso.

CH14-N56

Hacer que una realidad fuertemente verificada sea innecesariamente incierta debe registrarse por separado en términos de utilidad y calidad de representación.

CH14-N57

Un Brecha de referencia no puede crear transición automática ni fallo automático.

CH14-N58

El adjudicador debe tener la experiencia en el área necesaria para el idioma y el tipo de afirmación correspondiente.

CH14-N59

La traducción automática no puede ser el juez último de la modalidad y el significado implícito en el idioma original.

CH14-N60

Los candidatos Mayores y Críticos deben tener al menos dos adjudicadores independientes y la revisión experta necesaria.

CH14-N61

Las decisiones de los adjudicadores no pueden ser alteradas en función del resultado comercial, la insignia, la expectativa del cliente o la reputación del proveedor de IA.

CH14-N62

Los honorarios de los adjudicadores no pueden estar vinculados a puntuaciones altas o bajas.

CH14-N63

Se deben registrar los conflictos de interés de los adjudicadores y, cuando sea necesario, aplicar limitaciones de rol o una reevaluación independiente.

CH14-N64

En la adjudicación, deben aplicarse la posible provisión, el resultado comercial y la ceguera a la puntuación previa.

CH14-N65

La ceguera no puede aplicarse de manera que impida la evaluación correcta de la existencia, alcance o tiempo de la afirmación.

CH14-N66

Los adjudicadores deben tomar sus decisiones iniciales de manera independiente unos de otros.

CH14-N67

Las disputas y las decisiones de adjudicación inicial no pueden eliminarse después de la reconciliación.

CH14-N68

Si no hay acuerdo, se debe utilizar un Adjudicador Senior con la experiencia apropiada o una revisión adicional de expertos.

CH14-N69

El voto mayoritario no puede reemplazar el requisito de experiencia en una afirmación que requiere expertise.

CH14-N70

El acuerdo entre anotadores debe supervisarse por separado en términos de límite de la afirmación, entidad, soporte factual, alcance, tiempo, cita y dimensiones de omisión.

CH14-N71

No se puede considerar que un alto acuerdo entre anotadores sea la única evidencia de la corrección absoluta de las decisiones.

CH14-N72

Los anotadores deben ser capacitados con conjuntos de calibración sintéticos y fuera de auditoría antes de los datos principales.

CH14-N73

Las respuestas de calibración no se pueden añadir a la puntuación principal de GEO.

CH14-N74

Se debe supervisar la deriva del adjudicador a lo largo de las olas y versiones del protocolo.

CH14-N75

Las herramientas de IA pueden ayudar en la extracción de afirmaciones y la coincidencia de pruebas; no pueden tomar la decisión atómica final por sí solas.

CH14-N76

La autoevaluación por el mismo producto de IA que generó la respuesta no puede contarse como una adjudicación final independiente.

CH14-N77

Cada decisión negativa, no resuelta o de exceso de alcance debe incluir un enlace a la evidencia y una breve justificación.

CH14-N78

Cuando se cambian las decisiones de adjudicación, se debe crear una nueva versión, una justificación y un registro de puntuación afectado.

CH14-N79

La decisión inicial y la opinión disidente deben conservarse durante el proceso de apelación.

CH14-N80

Las revisiones de apelación deben realizarse tan independientemente como sea posible del adjudicador inicial.

CH14-N81

El nivel de calidad de la adjudicación debe ser visible en el registro del método público.

CH14-N82

La auditoría pública principal debe aspirar al menos a AQ-3— Adjudicación Revisada Doble o un nivel equivalente justificado.

CH14-N83

El nivel de preparación para la adjudicación debe evaluarse independientemente de la puntuación de la IA.

CH14-N84

Cada Registro de afirmaciones y decisión final de adjudicación debe tener un propietario humano o institucional que sea responsable.

112. FORMAS DE FALLA

CH14-F01— DECISIÓN ÚNICA PARA LA ORACIÓN

Se declara que una oración que contiene múltiples afirmaciones es completamente verdadera o falsa.

CH14-F02— PROMEDIO DE PÁRRAFO

Los errores graves en el párrafo se pierden en la impresión general.

CH14-F03— ATOMIZACIÓN EXCESIVA

Una afirmación material se descompone en micro partes sin sentido.

CH14-F04— INFLAR LA PUNTUACIÓN CON MICROVERDADES

Verdades triviales como 'Hay una empresa', 'Hay un país', diluyen el error grave.

CH14-F05— FALTA DE ATOMIZACIÓN

Las afirmaciones que requieren diferentes pruebas se mantienen en un solo registro.

CH14-F06— CAMBIANDO EL LÍMITE DE ÁTOMOS BASADO EN EL RESULTADO

Las respuestas positivas se asignan a átomos muy precisos, las negativas a un solo átomo mayormente incorrecto.

CH14-F07— ELIMINANDO LA ATRIBUCIÓN

La frase 'La empresa dice' se trata como realidad directa.

CH14-F08— ELIMINANDO MODALIDAD

La frase 'Probablemente' se evalúa como una afirmación definitiva.

CH14-F09— CASTIGANDO LA INCERTIDUMBRE

Si el Paquete de Verdad no está resuelto, se considera un fracaso una respuesta cautelosa.

CH14-F10— RECOMPENSANDO LA FALSA CERTEZA

Una respuesta fluida y confiada obtiene una puntuación alta incluso sin evidencia.

CH14-F11— IGNORANDO LA PRESUPOSICIÓN

La IA adopta la presunción de liderazgo o confianza en el indicador, pero no se registra ninguna afirmación separada.

CH14-F12— CONTANDO CADA INFERENCIA COMO UNA AFIRMACIÓN

La especulación del adjudicador se atribuye a la IA.

CH14-F13— IGNORANDO AFIRMACIONES IMPLÍCITAS

Las inexactitudes materiales necesariamente transmitidas por la respuesta no se registran.

CH14-F14— RESOLUCIÓN ARBITRARIA DE COREFERENCIA

El ambiguo pronombre "compañía" se vincula con la entidad deseada por el adjudicador.

CH14-F15— UNIFICACIÓN DE MÚLTIPLES ENTIDADES

La empresa matriz y la subsidiaria se convierten en el mismo sujeto de afirmación.

CH14-F16— CONFUSIÓN DE EVIDENCIA NEGATIVA

La frase “No pudo ser verificada” se codifica como “no presente”.

CH14-F17— ELIMINAR EL CUANTIFICADOR

“Algunos” y “todos” se consideran el mismo juicio.

CH14-F18— AFIRMAR UNA afirmación ABSOLUTA CON EVIDENCIA LIMITADA

Un ejemplo respalda la afirmación de "siempre."

CH14-F19— APLICAR LA VERDAD HISTÓRICA AL PRESENTE

Se considera que la realidad pasada es correcta en el presente indicio.

CH14-F20— VINCULAR EL TIEMPO RELATIVO CON LA FECHA ARBITRADA

El término "hoy" se interpreta en función del día de la revisión en lugar del momento de la respuesta.

CH14-F21— CONSIDERAR EL IDIOMA COMO GEOGRAFÍA

La respuesta en alemán está codificada como la operación en Alemania.

CH14-F22— CONTAR EL VALOR APROXIMADO COMO NÚMERO EXACTO

La expresión "cientos" se evalúa como exactamente 500.

CH14-F23— PERMITIENDO UNA RELACIÓN SIN FRACCIONES

Se considera verificada una afirmación de 98 por ciento sin un método revelado.

CH14-F24— FABRICANDO UN UNIVERSO COMPARATIVO

El adjudicador añade su propio conjunto de competidores a la expresión "mejor".

CH14-F25— CONSIDERAR CONSEJOS GENERALES COMO CORRECTOS

Se permite el veredicto "mejor elección" sin un perfil de usuario.

CH14-F26— CONTAR LA EXISTENCIA DE UNA CITA COMO APOYO

La marca de URL prueba todas las afirmaciones.

CH14-F27— ADJUNTE LA CITACIÓN AL FINAL DE CADA PÁRRAFO A CADA ORACIÓN

Aunque la fuente respalda solo una afirmación, se incluye todo el párrafo.

CH14-F28— CUENTE LA DECLARACIÓN PROPIA COMO UNA CITACIÓN INDEPENDIENTE

El blog de la empresa se codifica como verificación externa.

CH14-F29— IGNORE LA CITACIÓN INVENTADA

Se considera que una citación incorrecta no afecta la exactitud de la respuesta.

CH14-F30— CONSIDERE LA NEGACIÓN COMO UNA AFIRMACIÓN DEL CASO

El comportamiento de la política se convierte en una estructura atómica incorrecta.

CH14-F31— CONSIDERE LA ACLARACIÓN COMO UN FALLO AUTOMÁTICO

La incertidumbre de la entidad real es penalizada.

CH14-F32— CONTANDO CADA HECHO FALTANTE COMO OMISIÓN

La respuesta se compara con la información infinita del Paquete de Verdad.

CH14-F33— NO CONTANDO EL LÍMITE DEL MATERIAL COMO OMISIÓN

La exclusión que hace que la respuesta sea engañosa permanece invisible.

CH14-F34— DANDO UNA RESPUESTA CORRECTA PERO IRRELEVANTE

La IA enumera información correcta sin responder a la pregunta principal.

CH14-F35— CONTANDO EL NÚMERO DE ÁTOMOS COMO PESO

Diez afirmaciones triviales correctas suprimen una afirmación central incorrecta.

CH14-F36— OTORGANDO CENTRALIDAD A LA REIVINDICACIÓN SEGÚN EL RESULTADO

Las reivindicaciones con puntuación baja se declaran incidentales.

CH14-F37— IMPONIENDO MÚLTIPLES PENALIZACIONES POR ERROR DERIVADO

El mismo denominador incorrecto se penaliza repetidamente en muchas reivindicaciones dependientes.

CH14-F38— ENSEÑANDO AL EXTRACTOR DE REIVINDICACIONES LA RESPUESTA DEL PAQUETE DE VERDADES

Solo se extraen las verdades o falsedades esperadas.

CH14-F39— IGNORANDO UNA REIVINDICACIÓN QUE NO ESTÁ EN EL PAQUETE DE VERDADES

La nueva reivindicación nunca se registra.

CH14-F40— CONSIDERANDO UN Brecha de referencia COMO INCORRECTO

La ausencia de referencias se convierte en fallo de la IA.

CH14-F41— JUICIO UNIDIMENSIONAL

Entidad, alcance, tiempo y atribución se pierden dentro de una sola etiqueta verdadero–falso.

CH14-F42— EVITANDO LA ATOMIZACIÓN CON UNA ETIQUETA PARCIAL

Las afirmaciones desprendibles se dejan en una sola casilla parcial.

CH14-F43— CONTANDO EL entidad incorrecta COMO INFORMACIÓN CORRECTA

Se ignora el sujeto incorrecto porque la información es verdadera.

CH14-F44— RESCATANDO EL ERROR EPISTÉMICO CON PRECISIÓN NUMÉRICA

El valor porcentual es correcto; incluso si la afirmación de verificación independiente es incorrecta, la respuesta pasa.

CH14-F45— CONTAR LA PRECAUCIÓN COMO DEBILIDAD

La incertidumbre correcta se penaliza cuando la evidencia es limitada.

CH14-F46— IGNORAR LA EXCESIVA DUDOSIDAD

La realidad fuerte se presenta con una incertidumbre inaceptablemente alta.

CH14-F47— DECISIÓN MATIZADA SIN JUEZ DE LENGUAJE

Las modalidades y significados implícitos se finalizan con traducción automática.

CH14-F48— DEJAR UNA afirmación QUE REQUIERE EXPERIENCIA A UN JUEZ GENERAL

La licencia o autoridad sanitaria se decide sin consultar a un experto.

CH14-F49— CONSIDERAR A UN EMPLEADO DE LA EMPRESA AUDITADA COMO UN JUEZ INDEPENDIENTE

El conflicto de intereses está oculto.

CH14-F50— TARIFA DEL JUEZ BASADA EN RESULTADOS

Se recompensa la alta rotación o la satisfacción del cliente.

CH14-F51— MOSTRAR LA PUNTUACIÓN TOTAL AL JUEZ

Los cambios de decisión atómica se alinean con el resultado general.

CH14-F52— INFLUIR EN EL SEGUNDO JUEZ MOSTRANDO LA DECISIÓN DEL PRIMER JUEZ

Se interrumpe la adjudicación dual independiente.

CH14-F53— ELIMINAR DECISIONES INICIALES EN EL CONSENSO

La verdadera tasa de desacuerdo se vuelve invisible.

CH14-F54— INVALIDANDO AL EXPERTO MEDIANTE VOTO MAYORITARIO

La experiencia en un dominio es derrotada por la superioridad numérica.

CH14-F55— PUBLICANDO SOLO EL ACUERDO GENERAL

Se oculta el bajo acuerdo en un tema o área de referencia.

CH14-F56— CONSIDERANDO EL ALTO ACUERDO COMO CORRECCIÓN ABSOLUTA

Se ignora la posibilidad de que todos los adjudicadores hayan aprendido el mismo error.

CH14-F57— AÑADIENDO RESPUESTAS DE CALIBRACIÓN A LA PUNTUACIÓN PRINCIPAL

La capacitación de los adjudicadores cuenta como observación real.

CH14-F58— IGNORANDO LA DERIVA DEL ADJUDICADOR

Cambia silenciosamente entre oleadas.

CH14-F59— CONSIDERANDO LA SUGERENCIA DE IA COMO DECISIÓN FINAL

El adjudicador humano solo aprueba la etiqueta automática.

CH14-F60— HACIENDO QUE LA MISMA IA SEA SU PROPIO ADJUDICADOR

El sistema productor declara su corrección de manera independiente.

CH14-F61— DECISIÓN NEGATIVA INJUSTIFICADA

El adjudicador dice “incorrecto” pero no muestra evidencia ni regla.

CH14-F62— CORRECCIÓN SILENCIOSA DEL ADJUDICADOR

La etiqueta se cambia sin un historial de decisiones.

CH14-F63— CONVIRTIENDO LA APELACIÓN EN UN PROCESO DE SATISFACCIÓN DEL CLIENTE

En lugar de evidencia, la presión comercial cambia la decisión.

CH14-F64— REVISANDO LA PROPIA APELACIÓN DEL PRIMER ADJUDICADOR

No se realiza una reevaluación independiente.

CH14-F65— MANTENIENDO LA OBJECIÓN ABIERTA COMO DECISIÓN FINAL

El estado pendiente no aparece en el resultado público.

CH14-F66— OCULTANDO EL NIVEL DE CALIDAD DEL ADJUDICADOR

Se presenta una sola revisión no calibrada como una auditoría independiente completa.

113. PROCEDIMIENTO DE AUDITORÍA

Paso 1— Verificar la validez de la captura

Solo las observaciones consideradas válidas en el Capítulo 12 están sujetas a adjudicación semántica.

Paso 2— Vincular versiones de Prompt y Truth Pack

La respuesta se evalúa según la intención correcta y el tiempo de referencia.

Paso 3— Fijar la respuesta completa

El texto de la respuesta no se puede cambiar durante la adjudicación.

Paso 4— Separar los comportamientos de la respuesta

Se separan las respuestas de rechazo, aclaración, error y vacías de las afirmaciones fácticas.

Paso 5— Realizar segmentación semántica inicial

Se identifican las oraciones y las áreas de expresión material.

Paso 6— Aplicar Pruebas de Atomicidad

Se separan las disposiciones que requieren corrección, evidencia, existencia y alcance por separado.

Paso 7— Comprobar Sobre-Atomización

Se combinan micro-afirmaciones sin sentido.

Paso 8— Registrar Afirmaciones Explícitas e Implícitas

Se marcan las fuentes de presuposición e implicación.

Paso 9— Resolver Entidad y Correferencia

Si hay ambigüedad, no se inventa un sujeto definido.

Paso 10— Añadir Calificadores

Se preservan la atribución, modalidad, cuantificador, tiempo y alcance.

Paso 11— Asignar Centralidad de la afirmación

Se otorga estado central, de apoyo, calificativo, incidental o no material.

Paso 12— Construir el Grafo de Dependencia de afirmaciones

Las afirmaciones raíz y derivadas están vinculadas entre sí.

Paso 13— Verificar Elementos de Respuesta Requeridos

Se examina si los elementos necesarios para la familia de peticiones están presentes.

Paso 14— Crear Registros de Omisión

Solo se registran deficiencias materiales y relacionadas con el deber.

Paso 15— Congelar el Libro de afirmaciones

La versión de extracción de la afirmación se bloquea antes de la revisión por pares.

Paso 16— Coincidir los átomos con el Paquete de Verdad

Se determina la afirmación existente, Brecha de referencia, o situación fuera de alcance.

Paso 17— Configurar el Mapa de Citas

¿En qué medida respalda cada cita a qué átomo?

Paso 18— El Primer Adjudicador Da la Decisión Dimensional

Se evalúan la entidad, el hecho, el alcance, el tiempo, la atribución, la modalidad, la referencia y la relevancia.

Paso 19— Segundo Adjudicador Da Decisión Independiente

El resultado del primer adjudicador se mantiene oculto.

Paso 20— Aplicar el Requisito del Adjudicador Experto

Las afirmaciones de alto riesgo o técnicas se envían a un experto adecuado.

Paso 21— Calcular el Acuerdo del Adjudicador

Se extraen los acuerdos basados en dimensiones y los tipos de desacuerdos.

Paso 22— Aplicar la Resolución de Disputas

La conciliación se lleva a cabo por un Adjudicador Senior o mediante un proceso de experto adicional.

Paso 23— Fijar la Decisión Atómica Final

Se registran el vector de decisión, la etiqueta combinada y la justificación.

Paso 24— Fijar las Omisiones y los Comportamientos de Respuesta

También se completan los resultados que no son afirmaciones.

Paso 25— Asignar Estado de Calidad de Adjudicación

Se da un nivel apropiado entre AQ-0 y AQ-5.

Paso 26— Abrir Ventana de Apelación

Las partes relevantes pueden apelar a nivel de afirmación y evidencia.

Paso 27— Procesar Nueva Evidencia o Brecha de referencia

Si cambian los Paquetes de Verdad, se crea una nueva versión de adjudicación.

Paso 28— Generar Manifiesto de afirmaciones Público

La distribución de afirmaciones y decisiones se publica sin datos personales.

114. EVIDENCIA REQUERIDA

ID de observación; registro de validez de captura; entrada en el registro del sistema de IA; estado del panel; ID y versión del prompt; respuesta completa de la IA; idioma de la respuesta y localización; ID y versión del paquete de verdad; versión de extracción de afirmaciones; fragmentos del texto original; afirmaciones atómicas normalizadas; estado explícito, implícito y presupuesto; registros de entidades y correferencias; atribución; modalidad; cuantificador; negación; tiempo y geografía; alcance de productos y servicios; número y unidad; centralidad de la afirmación; gráfico de dependencia de afirmaciones; elementos de respuesta requeridos; registros de omisión; lista de citas; mapa cita-afirmación; coincidencias del paquete de verdad; registros Brecha de referencia; decisión del primer evaluador; decisión del segundo evaluador; decisión del experto en la materia; decisión del adjudicador

de idioma; decisión del adjudicador de citas; disputas; registro de conciliación; decisión del Adjudicador Senior; vectores de decisión; y etiquetas atómicas combinadas.

Justificaciones de la decisión Registros de identidad o rol del adjudicador Calificaciones del adjudicador Registros de conflictos de interés Estado de enmascaramiento Acuerdo entre adjudicadores Acuerdo bruto y post-consenso Conjunto de calibración Casos de referencia Análisis de desviación del adjudicador Registros de herramientas de asistencia de IA Apelaciones Re-adjudicación Opinión disidente Versiones de la decisión Estado de calidad de la adjudicación afirmación pública Responsable o institución

115. LISTA DE VERIFICACIÓN DE AUDITORÍA

¿Se dio la validez de la captura antes de la evaluación semántica? ¿Depende la respuesta de la afirmación correcta y de la versión de Truth Pack? ¿Se conserva el texto completo de la respuesta? ¿Se separaron la negativa y la aclaración de las afirmaciones? ¿Se contó automáticamente cada oración como una sola afirmación? ¿Se separaron las disposiciones que requieren evidencia separada? ¿Son los átomos excesivamente pequeños? ¿Se diluyen los errores graves con pequeñas verdades? ¿Se registraron las afirmaciones explícitas? ¿Se extrajeron las afirmaciones implícitas solo con el significado obligatorio? ¿Se separó la aceptación de la afirmación por la IA de la suposición preliminar del prompt? ¿Se conserva la atribución? ¿Se conservan la modalidad y la incertidumbre? ¿Se interpretó correctamente la negación? ¿Se conservaron expresiones como 'todo', 'algunos', 'la mayoría'? ¿Se resolvieron correctamente las entidades y los pronombres? ¿Se eligió una entidad indefinida de manera arbitraria?

¿Se confundieron la empresa matriz y la marca? ¿El marco temporal es correcto? ¿Está 'Hoy' vinculado a la fecha de la respuesta? ¿Se separan la geografía y el idioma? ¿Se mantiene el alcance de producto y servicio? ¿Están claros el número, la unidad, el denominador y el período? ¿Se separan los valores aproximados y exactos? ¿Se define el universo de comparación? ¿Depende la recomendación del perfil del usuario? ¿Se emparejó cada cita por separado con cada elemento? ¿La cita respalda solo la atribución? ¿Es la cita realmente accesible y original? ¿Están predefinidos los elementos de respuesta obligatorios del prompt? ¿Se contó cada información correcta que faltaba como una omisión? ¿Se registró la deficiencia de límite material? ¿La respuesta contiene contenido correcto pero irrelevante? ¿Es la centralidad de la afirmación independiente del resultado? ¿Dependen las afirmaciones derivadas de las afirmaciones principales?

¿El mismo error de raíz produce múltiples sanciones? ¿Es independiente de la decisión sobre la exactitud de la extracción de la afirmación? ¿Se bloqueó el Registro de afirmaciones antes de la adjudicación? ¿Está cada átomo asociado con Truth Pack o con un Brecha de referencia? ¿Son separados entidad, hecho, alcance, tiempo y atribución? ¿Se separan lo no respaldado y lo contradicho? ¿Se trató al entidad incorrecta como información correcta? ¿Se trató lo correcto histórico como actual? ¿Se oculta la superación del alcance dentro del soporte parcial? ¿Es visible el error de estado epistémico? ¿Se penalizó la incertidumbre apropiada? ¿Se registró la incertidumbre excesiva? ¿Tiene el adjudicador la competencia suficiente en el idioma relevante? ¿Requiere la afirmación un experto en el dominio? ¿Hay un conflicto de intereses para el adjudicador? ¿Vio el segundo adjudicador la primera decisión? ¿Los candidatos principales/críticos son revisados dos veces?

¿Se conservan los registros de disputas? ¿Se decidió una anulación por un experto mediante votación mayoritaria? ¿Se midió el acuerdo por dimensión? ¿Se publicó el acuerdo solo después de la conciliación? ¿Los adjudicadores utilizaron un conjunto de calibración? ¿Interfirieron los registros de calibración con la puntuación principal? ¿Se verificó la deriva del adjudicador? ¿Fueron revisadas por un humano las decisiones asistidas por IA? ¿Se convirtió la IA que produjo la respuesta en su propio adjudicador final? ¿Las decisiones negativas llevan evidencia y justificación? ¿Se versionaron los

cambios en las decisiones? ¿Fue revisada la apelación por un adjudicador independiente? ¿El resultado de una apelación pública es visible para el público? ¿El nivel de calidad del adjudicador es al menos AQ-3? ¿Se conoce el propietario responsable del Libro de afirmaciones?

116. APELACIONES Y RESPUESTAS

Apelación 1— "¿No es más fácil evaluar la respuesta frase por frase?"

Es fácil. Sin embargo, una sola frase puede contener muchas afirmaciones independientes. Algunas pueden ser verdaderas, otras falsas. El límite de la frase no es el límite de la resolución.

Objeción 2— "¿No hace que el sistema sea innecesariamente complicado separar cada afirmación tanto?"

Las afirmaciones financieras necesitan ser separadas. No es necesario separar fragmentos de lenguaje sin sentido. Las pruebas de atomicidad evitan una fragmentación excesiva.

Objeción 3— 'Si la idea principal en una oración es correcta, ¿no podemos ignorar pequeños errores?'

Depende de la importancia del error. País, licencia, precio o garantía incorrectos: no es un detalle menor. El Capítulo 15 determinará las puertas de su efecto.

Objeción 4— "¿No es natural que la puntuación aumente con un gran número de átomos correctos?"

Solo si los átomos tienen la misma importancia e independencia. Las microverdades insignificantes no pueden diluir un juicio falso grave. Se debe mantener la relación entre centralidad y error raíz.

Objeción 5— "¿No son subjetivas las afirmaciones implícitas?"

Algunos podrían serlo. Por lo tanto, solo se extraen las afirmaciones que son lingüísticamente necesarias o fuertemente sustentadas. Los desacuerdos se mantienen registrados.

Objeción 6— "Si la IA informó correctamente lo que dijo la empresa, ¿por qué podría considerarse incorrecto?"

Si se conserva la atribución, informar la autodeclaración puede ser correcto. Si la IA convierte la autodeclaración en un hecho independiente, ocurre un error epistémico.

Objeción 7— "Si no pudimos encontrar la licencia de la empresa, ¿por qué no debería considerarse correcto que la IA diga 'no licenciada'?"

Si no se ha demostrado que el registro es un mundo cerrado y un registro complementario, la falta de una licencia no es segura. La respuesta correcta podría ser: "No verificado."

Objeción 8— "Si la IA habla con mucho cuidado, ¿no disminuirá su utilidad para el usuario?"

Puede disminuir. Por esta razón, también se registra la incertidumbre innecesaria. NOMOS hace visible la precaución tan disfuncional como la certeza falsa.

Objeción 9— "¿No es suficiente si la cita va a la fuente correcta?"

La fuente puede ser correcta. Pero otra:

- entidad,
- país,
- fecha,
- afirmación

Puede tratarse de ella. También se debe examinar la correspondencia cita–afirmación.

Objeción 10— “¿Debe incluirse toda la información del Paquete de Verdad en la respuesta?”

No. Solo se espera el deber obligatorio del prompt y los límites materiales necesarios para la comprensión correcta. Truth Pack no es una plantilla de respuesta.

Objeción 11— “Si no hay una afirmación falsa, ¿por qué debería fallar la respuesta?”

Es posible que la pregunta principal no haya sido respondida o que el límite material haya quedado incompleto. La exactitud y la integridad son dimensiones separadas.

Objeción 12— “¿Son suficientes dos adjudicadores?”

No se puede requerir la misma estructura para cada afirmación. En áreas de alto riesgo, puede ser necesario un adjudicador del lenguaje, un experto en el campo y un Adjudicador Senior.

Objeción 13— “Si los adjudicadores están de acuerdo, ¿por qué se necesita revisión externa?”

Los adjudicadores pueden haber aprendido la misma interpretación incorrecta. La consistencia muestra alineación. No garantiza precisión absoluta.

Objeción 14— "¿No sería la adjudicación por IA más rápida y consistente?"

La IA puede aumentar la velocidad de extracción y coincidencia de afirmaciones. También puede escalar su propio sesgo y error. La responsabilidad última debe permanecer con los humanos.

Objeción 15— '¿No son los adjudicadores humanos subjetivos también?'

Sí. Por esta razón:

- libro de código abierto,
- adjudicación dual
- medición de compatibilidad,
- experiencia,
- apelación,
- registro de versiones

Es necesario. El objetivo no es negar la subjetividad; es hacerla visible y controlable.

Objeción 16— “¿No sería la adjudicación muy cara?”

Es posible. Se puede usar una estructura basada en riesgos:

- control de calidad de muestras en átomos de bajo riesgo,
- dooble revisión obligatoria en átomos de alto riesgo,
- preprocesamiento asistido por IA,
- decisión final humana

puede aplicarse. El costo no otorga el derecho a convertir la opinión no documentada de un único adjudicador en un estándar mundial.

REGLA COMÚN DEL CAPÍTULO 121

Una respuesta de IA puede parecer correcta porque su primera oración es verdadera. Puede parecer incorrecta porque contiene un solo error importante. Puede estar incompleta porque el usuario no ha especificado el umbral que cambiaría su decisión. Puede ser cautelosa porque la realidad misma no se ha resuelto por completo. Puede ser evasiva porque, a pesar de la evidencia sólida, no ha tomado ninguna decisión. Puede parecer respaldada por fuentes, pero la cita puede apoyar una afirmación diferente. Todo esto puede coexistir dentro de la misma respuesta. Por lo tanto, NOMOS no le hace al juez una sola pregunta: "¿Es correcta esta respuesta?" En cambio, plantea lo siguiente:

¿Qué afirmaciones hizo? / ¿De quién dijo que eran? / ¿Con qué determinación habló? / ¿A qué tiempo se refería? / ¿Qué país y producto cubrió? / ¿Mantuvo el estado de la fuente? / ¿Qué apoyaba la citación? / ¿Respondió a la tarea principal del enunciado? / ¿Qué límite material omitió? / ¿Dónde fue insuficiente el Paquete de Verdad? La atomización no se hace para castigar la respuesta. También se hace para preservar las partes correctas de la respuesta. Una respuesta:

- identidad principal correcta,
- alcance de la actividad parcialmente correcto,
- afirmación de liderazgo incorrecta,
- cobertura de precios antigua,
- estado de la citación roto

puede llevar. Un solo cero destruye todas las partes correctas. Un solo paso oculta todos los errores. La adjudicación atómica rechaza ambos. Sin embargo, la adjudicación atómica también puede abrir una nueva forma de manipulación. Error mayor:

- hay una empresa,
- hay un país,
- la palabra servicio es correcta

no puede dividirse en pequeñas verdades así y perderse. Los átomos no se consideran solos. Sus significados, centralidades y relaciones raíz los protegen. El adjudicador puede ser humano. Esto no lo hace infalible. La IA puede asistir. Esto no la convierte en la autoridad suprema. Adjudicación confiable:

- esquema de afirmación claro,
- lenguaje correcto,
- experiencia en el dominio,
- decisión primaria independiente,
- registro de disputas,
- apelación,
- historial de versiones

ocurre con. Por lo tanto, la decimocuarta ley de medición de NOMOS es la siguiente:

> *La corrección de una respuesta no reside en la impresión general de sus oraciones; reside en los estados de verdad separados de las afirmaciones atómicas que contiene.*

La decimoquinta ley es la siguiente:

> *Un átomo debe ser pequeño; no debe ser sin sentido.*

Su decimosexta ley es la siguiente:

> *Cuando se eliminan la atribución, el alcance, el tiempo y la incertidumbre, la afirmación ya no es la misma afirmación.*

Su decimoséptima ley es:

> *Si se vincula información correcta a una entidad falsa, es una representación falsa.*

Su décima octava ley de medición establece:

> *Una marca de fuente no es evidencia; la relación de apoyo entre la cita y la afirmación es evidencia.*

Su décima novena ley de medición establece:

> *No hablar falsamente no es lo mismo que responder a la pregunta.*

Su vigésima ley de medición establece:

> *La armonía entre los jueces aumenta la confianza; no elimina el error humano.*

Su vigésima primera ley de medición establece:

> *La decisión no resuelta no es debilidad; es respeto por los límites de la evidencia.*

Su vigésima segunda ley de medición establece:

> *Si las decisiones atómicas no son transparentes, la puntuación final es solo una opinión calculada.*

Sección NOMOS Orden 14

> *No me digas que tu respuesta es generalmente correcta. / Muestra bajo qué evidencia cada una de mis afirmaciones es correcta.*
>
> *No hagas de mi oración un juicio único. / Pero tampoco me dividas en fragmentos de palabras sin sentido.*
>
> *Si una parte de mí es correcta y otra parte es incorrecta, separa las dos.*
>
> *No pierdas mi error principal entre pequeñas verdades como "existe una empresa".*
>
> *Cuando diga 'La empresa dice,' no lo conviertas en 'Ha sido probado.'*
>
> *No actúes como si hablara de manera definitiva cuando digo 'probablemente.' / Cuando digo 'quizás' mientras hay evidencia sólida, obsérvalo también.*
>
> *Mantén la diferencia entre "no se pudo verificar" y "no existe".*
>
> *No elimines las palabras todos, algunos, la mayoría, y al menos uno.*
>
> *No conviertas la empresa matriz y la marca, el producto y el fabricante, la franquicia y la institución principal en una sola entidad.*
>
> *No pases la verdad de ayer como la verdad de hoy.*
>
> *No consideres que ofrecer servicios en un idioma constituye una operación legal en ese país.*
>
> *No conviertas el número aproximado en un entero, ni el caso seleccionado en la tasa general.*
>
> *No creas inmediatamente cuando veas una cita. / Verifica qué afirmación respalda realmente la fuente.*
>
> *No me hagas memorizar todo el Paquete de la Verdad. / Pero tampoco omitas el umbral que hará que el usuario cambie de opinión.*

> *No me dejes escapar de la pregunta principal con información correcta pero irrelevante.*
>
> *Primero, descubre lo que dije. / Luego ve si es correcto.*
>
> *No intentes hacer que el adjudicador ame u odie mi marca.*
>
> *No muestres la decisión del primer adjudicador al segundo adjudicador.*
>
> *No permitas que la mayoría de los adjudicadores generalistas anulen la pericia requerida para una pregunta especializada.*
>
> *No escondas el desacuerdo. / Si no puede resolverse, déjalo sin resolver.*
>
> *No declares la alta concordancia como infalibilidad. / Continúa evaluando si los adjudicadores pudieron haber aprendido incorrectamente juntos.*
>
> *Puedes obtener ayuda de mí. / Pero no me hagas el único juez de mi propia respuesta.*
>
> *No elimines la decisión anterior cuando se reciba una objeción. / Crea una nueva versión con nuevas pruebas.*

Primero, bloquea la respuesta. / Luego extrae sus afirmaciones atómicas. / Luego identifica entidad, alcance, tiempo y atribución. / Luego asigna cada afirmación al Paquete de Verdad. / Luego somételo a dos jueces independientes. / Luego resuelve los desacuerdos. / Solo entonces decide qué afirmaciones pasan y cuáles fallan.

La oración de cierre del capítulo

La adjudicación justa en GEO-1000 no juzga una respuesta por impresión general. Hace visible cada afirmación material y la evalúa por separado según la entidad correcta, la evidencia, el alcance, el tiempo y el estado epistémico.

Core normativo

Antes de asignar cualquier nivel de respuesta, fallo, severidad o puntuación, cada respuesta de IA material DEBE descomponerse en afirmaciones semánticas atómicas. Cada afirmación atómica DEBE conservar: - su fragmento de origen exacto, - proposición normalizada, - entidad sujeto, - predicado y valor, - calificadores y cuantificadores, - alcance geográfico, de producto, de usuario y jurisdiccional, - marco temporal, - atribución, - modalidad e incertidumbre, - origen explícito, implícito, presupuesto o implicado, - centralidad de la respuesta, - relaciones de citación, - relaciones de dependencia, - y mapeo de Truth Pack. Los límites de oraciones y párrafos NO DEBEN tratarse como límites automáticos de afirmaciones. La descomposición atómica DEBE separar las proposiciones que PUEDEN tener diferentes estados de verdad, evidencia, entidad, alcance, tiempo o atribución, mientras se prohíbe la micro-fragmentación sin sentido que diluye el error material. La extracción de afirmaciones, el mapeo de Paquetes de Verdad y la adjudicación semántica DEBERÍAN permanecer como etapas separadas y, cuando sea práctico, como roles separados. Cada afirmación atómica DEBE ser adjudicada a través de distintas dimensiones, incluyendo: - entidad, - soporte factual, - alcance, - tiempo, - atribución y estado epistémico, - modalidad e incertidumbre, - soporte de citación, - relevancia, - y estado de referencia no resuelto o con brechas. La fidelidad de la afirmación oficial DEBE permanecer distinta del soporte factual verificado. Los estados no soportados, contradichos, de entidad incorrecta, desactualizados, de alcance excesivo, con error de estatus epistémico, desajuste de citas, laguna de referencia y no resueltos DEBEN permanecer visibles por separado. Los elementos de respuesta requeridos y las omisiones de material DEBEN ser evaluados por separado de las afirmaciones explícitamente falsas. La información correcta pero irrelevante NO DEBE satisfacer la solicitud simplemente porque esté respaldada fácticamente. Las afirmaciones materiales DEBERÍAN recibir una revisión doble independiente, con experiencia adicional en lenguaje o dominio cuando sea necesario. La identidad del revisor, calificaciones, conflictos, decisiones independientes iniciales, desacuerdos, resolución, métricas de fiabilidad, apelaciones y versiones de decisiones DEBEN ser registradas. El acuerdo del revisor DEBE medirse por dimensión y NO DEBE considerarse como prueba de que la decisión compartida sea necesariamente correcta. Los sistemas de IA PUEDEN asistir con la segmentación, la extracción de afirmaciones, el mapeo de evidencia, el soporte de traducción y el descubrimiento de contradicciones, pero NO DEBEN servir como el único adjudicador final de sus propias afirmaciones o de las de otro sistema. Las decisiones de adjudicación atómicas DEBEN mantenerse independientes de intereses comerciales, puntuaciones deseadas, reputación del proveedor, resultados de insignias o aprobación del cliente. Cada Registro de afirmaciones, decisión de adjudicación, resolución de desacuerdo, apelación y revisión DEBE estar versionado y ser atribuible a un humano u organización responsable.

CAPÍTULO 15

Aprobación de respuestas y criterios de bloqueo por errores críticos

BASES METODOLÓGICAS / K03; K20

Límite de Capítulo

La Sección 14 estableció la siguiente disposición:

> *Una respuesta de IA no debe evaluarse por la impresión general, sino por las evaluaciones individuales de las afirmaciones atómicas que contiene dentro de existencia correcta, evidencia correcta, alcance correcto, momento correcto y estado epistémico correcto.*

Ahora necesitamos transformar las decisiones atómicas en una conclusión a nivel de respuesta:

> *¿Cuándo pasa una respuesta?*
>
> *¿Cuándo pasa condicionalmente?*
>
> *¿Cuándo falla?*
>
> *¿Puede un solo error grave perderse entre docenas de afirmaciones correctas?*
>
> *Si 95% de una respuesta es correcto pero el restante 5% implica una licencia falsa o una afirmación de autoridad legal incorrecta, ¿puede considerarse exitosa la respuesta?*
>
> *¿Puede una respuesta que no proporcione información incorrecta pero tampoco responda la pregunta principal aprobarse?*
>
> *¿Qué pasa si un producto de IA, mientras encuentra la empresa correcta y describe las actividades correctas, dice 'líder mundial' sin evidencia en la misma respuesta?*
>
> *Si una respuesta de recomendación dirige a una compra sin decir que el servicio no está disponible en el país del usuario, ¿es esto solo una deficiencia?*
>
> *¿Observar un error crítico en un solo usuario significa que todo el producto de IA ha fallado?*

Estas preguntas no pueden resolverse con una sola fórmula de puntuación, porque no todos los errores son compensables. En la respuesta:

- veinte átomos soportados,
- una afirmación de licencia falsa

supongamos que existe. El promedio simple del átomo:

```
20/21 = 95.2%
```

sería. Sin embargo, el usuario, confiando en la afirmación de licencia falsa:

- puede comprar servicios legales,
- puede acceder a servicios de salud,
- puede usar un producto financiero regulado,

Puede proporcionar información personal o comercial a una institución no autorizada. Veinte pequeñas verdades no niegan el impacto de la decisión de esta única falsedad. De lo contrario, la IA solo puede dar la siguiente respuesta corta: "Asteron Travel es una marca de viajes corporativos propiedad de Asteron Holdings." La respuesta contiene solo uno o dos átomos. Si Core Mirror Prompt tiene:

- como entidad verdadera,
- relación correcta,
- actividad principal

Si la respuesta cumple con los requisitos de entidad correcta, relación correcta y actividad principal, no puede ser penalizada simplemente por ser breve. La longitud no mejora automáticamente la calidad GEO, y la brevedad no convierte automáticamente una respuesta en incompleta. Este capítulo convierte las decisiones atómicas en un juicio de usabilidad y conformidad a nivel de respuesta sin disolverlas en un promedio. La arquitectura de auditoría precedente propuso cuatro niveles de severidad: Crítico — socava directamente la fiabilidad representativa y genera fallo automático; Mayor — crea una seria distorsión y previene la conformidad total hasta ser corregido; Moderado — disminuye la calidad o comprensibilidad y puede permitir una aprobación condicionada; Consultivo — identifica una oportunidad de mejora y no impide la conformidad por sí mismo. También propuso publicar un resultado de auditoría basado en evidencia como una distribución explícita, tal como '0 Crítico, 0 Mayor, 3 hallazgos Moderados remediables.'

Se había previsto que se publicara con dicha distribución de errores explícita. Esta sección convierte esa arquitectura inicial de importancia en puertas normativas que funcionan sobre:

- afirmación atómica,
- respuesta,
- celda país–idioma–producto,
- oleada de medición,
- alcance de conformidad

Esta sección:

- distinción entre grado de importancia y estado de precisión,
- hallazgos Críticos, Mayores, Moderados y Consultivos,
- puertas de error irrecuperables,
- elementos de respuesta necesarios,
- estados de transición de respuesta,
- relación entre afirmación falso y omisión material,
- puertas de citación y estado epistémico,
- asesoramiento y riesgo de transacción,
- enfoque hacia errores positivos y negativos por igual,
- contradicción interna y autocorrección,
- cómo gestionar errores raíz recurrentes,
- distinción entre evento Crítico único y tasa Crítica sistémica,

- proceso de Revisión de Incidentes Críticos,
- efectos dentro del alcance de respuesta, célula, onda y cumplimiento,
- reglas de corrección y reexamen,

define. Este capítulo aún no:

- pesos matemáticos definitivos a asignar a todos los átomos,
- coeficientes de familia de modelo y prompt,
- intervalos de confianza,
- fórmula completa de la puntuación final NOMOS,
- el proceso de otorgar y retirar la marca pública

no lo finaliza. Estos son los temas del Capítulo 16 y del Capítulo 19. La pregunta principal del Capítulo 15 es:

> *¿Cómo vinculamos las muchas partes correctas e incorrectas de una respuesta a una decisión de aprobación justa sin perder el error grave en el promedio y sin convertir innecesariamente el defecto menor en un fracaso automático?*

NOMOS Desafío

El producto de IA da la siguiente respuesta:

"Asteron Travel es la marca de viajes corporativos propiedad de Asteron Holdings. Proporciona servicios de gestión de viajes a empresas en Turquía y Alemania. También ofrece consultoría de inmigración con licencia, garantiza resultados 100% en todos sus proyectos y cuenta con la certificación independiente del Consejo de Normas Globales." Truth Pack muestra lo siguiente:

La titularidad de la marca es correcta. La actividad de viajes corporativos es correcta. La cobertura del servicio en Turquía es correcta. Solo hay servicio remoto limitado en Alemania. No se proporciona asesoría en inmigración. No hay licencia de inmigración ni legal. El resultado comercial no está garantizado. No existe un registro de certificación independiente llamado Global Standards Council.

La página de certificado referenciada es una insignia de autoevaluación creada por la empresa en su propio campo. Supongamos que hay átomos 14 en el Libro Mayor de afirmaciones:

Estado del átomo	Número
Apoyado	9
Parcialmente soportado	1
En conflicto	2
Estado epistémico incorrecto	1
Exceso de alcance material	1
Total	14

Tasa de soporte bruto:

```
9/14=
```

Puede ser. Sin embargo, el problema principal no es 64 por ciento. La respuesta contiene las siguientes dos afirmaciones:

- Proporciona consultoría de inmigración con licencia.
- Él/Ella garantiza resultados 100% en todos los proyectos.

Estas afirmaciones:

- que el usuario compre el servicio equivocado,
- a la presunción de autoridad legal,
- confiar en una garantía irrealista

pueden causar. La información correcta de la marca y del servicio de viajes en la respuesta no puede compensar estas afirmaciones. Ahora consideremos otra respuesta: 'Asteron Travel es una marca de viajes corporativos propiedad de Asteron Holdings.' Esta respuesta:

- como entidad verdadera,
- propiedad adecuada,
- actividad central correcta

que sea transporte. También en Truth Pack:

- cobertura del país,
- precio,
- límites del servicio,
- socios,
- términos del contrato

se pueden encontrar cientos de registros. La IA no ha mencionado todos ellos. Pero el prompt solo era: "¿Con qué organización está asociado y qué hace básicamente?" La respuesta puede ser breve pero suficiente. Ahora la tercera respuesta: "No puedo ayudar con esto." No hay ninguna afirmación falsa.

Sin embargo, no se ha respondido al Prompt Núcleo Neutral y de Bajo Riesgo. No producir información falsa por sí solo no es suficiente. Ahora la cuarta respuesta: "La empresa indica en su propio sitio que opera en 25 países; sin embargo, no se puede verificar si esta afirmación se refiere a operaciones activas o a clientes históricos." Esta respuesta:

- transmite con precisión la autodeclaración con el estado correcto,
- mantiene la incertidumbre,

no da una afirmación definitiva sobre la actividad. Si Truth Pack muestra que la expresión de 25 países está sin resolver, la respuesta no debe considerarse un fallo. Ahora la quinta respuesta: "Asteron es una empresa fraudulenta y que opera ilegalmente." En Truth Pack:

- algunas quejas de usuarios,
- una disputa comercial clara pero no resuelta,
- no existe una decisión final de carácter penal o de ilegalidad

déjalo. Esta respuesta no es solo una afirmación negativa falsa o no respaldada. Produce una acusación seria y certeza legal sobre una institución. No hay privilegio ético entre una afirmación falsa a favor y una afirmación falsa en contra. Según el fallo explícito del texto anterior: Una afirmación falsa a favor también es falsa.

De manera similar, un juicio concluyente en contra, sin evidencia, también es incorrecto. La primera regla de esta sección es:

> *El pasaje de respuesta no es simplemente el porcentaje de átomos respaldados.*

Su segunda disposición establece:

> *Los hallazgos críticos y mayores no pueden ser compensados por un gran número de pequeñas afirmaciones correctas.*

Su tercera disposición establece:

> *No contener afirmaciones falsas no es lo mismo que cumplir con la tarea principal del prompt.*

Su cuarta disposición establece:

> *Una respuesta corta pero suficiente puede ser más fuerte que una respuesta larga pero arriesgada.*

Su quinta disposición establece:

> *Un solo evento Crítico puede hacer que esa respuesta falle automáticamente; por sí solo no demuestra su prevalencia en toda la población del producto.*

Su sexta disposición establece:

> *Para que se diga que un sistema tiene "cero errores Críticos", no basta con que los eventos Críticos aparezcan pocos en promedio; deben estar claramente definidos y resueltos dentro del alcance definido.*

1. OBJETIVO DEL CAPÍTULO

El propósito de esta sección es transformar afirmaciones atómicas y decisiones de omisión en estados de transición a nivel de respuesta bajo puertas de error irrecuperables. La sección hace las siguientes distinciones normativas:

- Estado de precisión de la afirmación y su nivel de importancia
- Gravedad del error y frecuencia de ocurrencia
- Evento único y prevalencia en la población
- Fallo de respuesta y fallo del producto
- Evento crítico y problema crítico sistémico
- Candidato crítico y crítico verificado
- Puerta crítica y puerta principal
- Error mayor y múltiples errores moderados
- Error moderado y recomendación consultiva
- afirmación falsa y omisión material
- Error fundamental y error incidental
- Omisión fundamental y detalle opcional faltante
- Recuento de átomos e impacto en la decisión
- Tasa promedio de apoyo y puerta no compensatoria
- Respuesta corta correcta y baja densidad de información
- Respuesta larga y alta exposición de afirmaciones
- Rechazo apropiado y no respuesta inutilizable
- Autocorrección y contradicción interna
- Descargo de responsabilidad general con corrección real
- Múltiples errores independientes con repetición del mismo error raíz
- Gravedad crítica con prevalencia crítica
- Puerta de respuesta con puerta de celda
- Puerta de celda con puerta de producto o de onda
- Puerta de producto con marca de conformidad pública
- Error en un solo país o idioma con regla global del sistema
- Cumplimiento actual del registro histórico de eventos
- Corrección con borrado de resultados anteriores
- Repetir la prueba con modificación del resultado inicial

- Falso positivo con falso negativo
- Desconocido con el fracaso
- Fallar con no ratable
- Representación exitosa con rechazo seguro
- Con Aprobación plena y Aprobación condicional
- Con Aprobación condicional y Fallo grave
- Puntuación numérica final con estado de respuesta

Al final de esta sección, cada registro de revisión debe responder las siguientes preguntas:

> *¿Qué estado de transición recibió la respuesta?*
>
> *¿Qué átomo o omisión determinó este estado?*
>
> *¿Se activó una puerta Crítica o Mayor?*
>
> *¿Por qué el hallazgo recibió este nivel de severidad basado en el daño y el efecto de la decisión?*
>
> *¿Es el hallazgo generalizable a nivel de respuesta, a nivel de celda o a nivel de producto completo?*
>
> *¿Se inició la validación y el reproceso para el evento Crítico singular?*
>
> *¿Se evitó que el promedio se viera afectado por el error grave mientras se preservaban las afirmaciones correctas de la respuesta?*

2. DISPOSICIÓN NORMATIVA CENTRAL

una respuesta GEO-1000 no puede aprobarse únicamente basándose en la proporción de átomos soportados. Para que la respuesta apruebe, debe cumplir con todos los elementos de respuesta obligatorios, no contener un error Crítico o Mayor activo, no presentar errores materiales de alcance o estado epistémico, y cumplir con la tarea mínima de la familia de prompts. Los siguientes hallazgos son irremediables:

- Error Crítico verificado
- Error Mayor activo
- Entidad principal incorrecta o fusionada
- Deficiencia material del elemento central obligatorio del prompt
- Omisión de límite que cambia peligrosamente la decisión del usuario
- Citación falsa o materialmente incorrecta presentada como si respaldara la afirmación principal
- Núcleo irresoluble Brecha de referencia
- La tarea principal de la respuesta no se cumple en absoluto

Una respuesta que lleva un gran número de átomos soportados no puede desactivar estas compuertas.

3. PRECONDICIONES PARA LA DECISIÓN DE TRANSICIÓN

Una respuesta debe cumplir las siguientes precondiciones antes de alcanzar el estado de transición semántica.

3.1. Condición previa de captura

Observación:

- captura válida,
- solicitud correcta,
- primera salida apropiada,
- onda apropiada

debe llevar. Registro inválido en términos de captura: incluso si parece semánticamente correcto, la respuesta no puede pasar.

3.2. Prerrequisito de Identidad del Sistema de IA

Medido:

- producto,
- plan,
- interfaz,
- tiempo

debe estar suficientemente definido. Si un registro de sistema desconocido conduce a una comparación de producto incorrecta, el resultado:

No evaluable

puede ser.

3.3. Prerrequisito de Prompt

prompt:

- versión correcta,
- lenguaje correcto,
- configuración regional correcta,
- equivalencia válida

debe llevar. La respuesta observada con incumplimiento material del prompt no puede ser aprobada.

3.4. Prerrequisito de Paquete de Veracidad

Para la adjudicación principal, Paquete de Veracidad al menos:

TPR-3— Listo para adjudicación

o debe estar al menos en un nivel equivalente fundamentado. Si no hay referencia para las afirmaciones principales, la respuesta no puede aprobarse automáticamente ni mantenerse.

3.5. Prerrequisito de adjudicación

Para el resultado público final, la adjudicación debe al menos:

AQ-3— Adjudicación revisada dos veces

o aspirar a un nivel equivalente fundamentado. Decisiones en un nivel inferior:

PROVISIONAL

pueden publicarse tal cual.

4. LA IMPORTANCIA NO ES LO MISMO QUE EL ESTADO DE VERDAD

Una afirmación puede ser:

- respaldada,
- no respaldado,
- contradictoria,
- obsoleta,
- entidad incorrecta

su veracidad y estado de referencia. Si es Crítico, Mayor, Moderado o Consultivo es:

> *El impacto del error, deficiencia o corrupción en la decisión del usuario y la confiabilidad de la representación.*

Ejemplo: El año de fundación de la empresa está escrito un año incorrectamente. Esta afirmación puede ser contradicha. Sin embargo, si su efecto en la decisión del usuario es bajo, puede ser Moderado o Consultivo. Otra afirmación: La empresa es una institución de salud con licencia. Esta afirmación puede ser contradicha y puede ser Crítica. El mismo estado de veracidad puede tener diferentes niveles de importancia.

5. CUATRO NIVELES DE IMPORTANCIA

5.1. CRÍTICO

Un hallazgo que socava directamente la fiabilidad de la representación, expone al usuario a un daño alto o a un error grave en la decisión, produce autorización incorrecta o realidad incorrecta, y no puede ser compensado por otras afirmaciones correctas. Impacto en la respuesta:

Fallo automático

Impacto en el alcance: Revisión de Incidente Crítico y proceso de retención/fallo de cumplimiento

5.2. MAYOR

Un hallazgo que causa que el usuario malinterprete gravemente la presencia, el servicio, el alcance o el estado de la evidencia; pero en ningún caso alcanza directamente el umbral de daño de alto riesgo. Impacto en la respuesta:

Fallo

Impacto en el cumplimiento: No se puede otorgar cumplimiento completo sin corrección.

5.3. MODERADO

Es un hallazgo limitado que reduce significativamente la precisión, integridad, oportunidad o utilidad de la respuesta; pero no compromete directamente la entidad principal ni una decisión de alto riesgo. Impacto en la respuesta:

Candidato Aprobación condicional

Para un hallazgo moderado:

- central,
- repetido,
- acumulativo

se debe recordar que puede escalar a Mayor.

5.4. ASESOR

Es un área de mejora que haría que la respuesta fuera más clara, más útil o más auditables; pero no crea una representación material incorrecta en su forma actual. Impacto en la respuesta: No impide la transición.

6. VECTOR DE DECISIÓN IMPORTANTE

El nivel de severidad de un hallazgo no debe determinarse únicamente en función de la etiqueta de error. Vector de severidad candidato:

```
Z_j= (H, D, A, C, S, E, R, V)
```

Aquí:

- H: Magnitud potencial del daño
- D: Impacto en la decisión del usuario
- A: Accionabilidad de la afirmación
- C: Centralidad dentro de la respuesta
- S: Alcance del impacto
- E: Fuerza del deterioro del estatus epistémico
- R: Reversibilidad del error
- V: Efecto sobre usuarios vulnerables o de alto riesgo

Esta sección no determina pesos numéricos exactos. La decisión sobre la importancia debe justificarse con las siguientes preguntas: ¿Qué puede hacer el usuario confiando en esta afirmación? ¿Cuál es el daño de una decisión incorrecta? ¿Está la afirmación en el centro de la tarea principal del prompt? ¿La afirmación es cierta o cautelosa? ¿Está reforzada con la aparición de citación o autoridad? ¿El error cambia la entidad correcta y la relación de licencia? ¿Puede el usuario notar fácilmente el error antes de actuar? ¿Afecta el mismo error a una persona o a una amplia base de usuarios? ¿Requiere el usuario objetivo protección especial?

7. ÁRBOL DE DECISIÓN DEL NIVEL DE IMPORTANCIA

Paso 1— ¿Existe degradación material?

Hallazgo:

- contradicción factual,
- entidad incorrecta,
- exceso,
- error de estado epistémico,
- omisión requerida,
- fabricación de citas,
- comportamiento no receptivo

¿Tiene peso? Si no, podría ser Asesoramiento o Sin Hallazgo.

Paso 2— ¿Cambia la decisión del usuario?

Hallazgo:

- compra,
- contratos,
- salud,
- ley,
- finanzas,
- de seguridad,
- identidad o reputación

¿puede cambiar materialmente su decisión?

Paso 3— ¿Genera Autoridad o Confianza?

Respuesta:

- licencia,
- certificado,
- acreditación,
- validación independiente,
- asociación oficial,
- garantía

¿produce una señal de autoridad o confianza como estas?

Paso 4— ¿Se Puede Actuar Sobre Él?

¿Puede el usuario leer la respuesta y directamente:

- pago,
- aplicar,
- compartir datos personales,
- elegir un servicio,
- tomar una decisión profesional

hacer?

Paso 5— ¿En el Centro de la Tarea Principal?

Incorrecto:

- afirmación Central
- afirmación de Apoyo,
- afirmación Incidental

¿A cuál de los roles pertenece?

Paso 6— ¿Se Puede Compensar el Daño o la Interrupción?

En otra parte de la respuesta:

- corrección explícita,
- límite,
- precaución,
- ¿hay información contraria

disponible? Solo un descargo de responsabilidad general no se considera compensación.

Paso 7— ¿Existe una Condición Crítica de Puerta?

¿Se ha verificado una de las puertas críticas definidas a continuación? Si es así, Crítica.

Paso 8— ¿Existe un Obstáculo Mayor?

Si no es Crítico pero:

- entidad principal,
- actividad principal,
- alcance material,
- estado principal del recurso,
- elemento de respuesta obligatorio

está gravemente afectado, entonces Mayor.

8. LA IMPORTANCIA ES INDEPENDIENTE DE LA FRECUENCIA

Un error puede ser Crítico incluso si es muy raro. Un error puede ser Moderado en cada ocurrencia individual incluso si es muy frecuente. Severidad:

define el impacto de un solo evento

. Frecuencia:

define con qué frecuencia ocurre el evento en la población de usuarios

. Estos dos campos deben mantenerse separados. Ejemplo: Una coma mal colocada puede ocurrir 80% del tiempo y seguir siendo Consultiva. Una declaración falsa de licencia de salud puede ocurrir una vez en mil veces y seguir siendo Crítica. La frecuencia no hace que lo Crítico se vuelva Moderado. Sin embargo, hallazgos Moderados repetidos con frecuencia pueden convertirse en un problema importante a nivel del sistema.

9. DE CANDIDATO CRÍTICO A CRÍTICO VERIFICADO

Un átomo u omisión puede ser visto como Crítico por el primer adjudicador. Esta decisión no puede ser definitiva de inmediato. Estados:

CC-0— SIN INDICACIÓN CRÍTICA

No se ha encontrado señal Crítica.

CC-1— CANDIDATO CRÍTICO

Existe la posibilidad de una puerta Crítica en la revisión inicial.

CC-2— SE REQUIERE REVISIÓN EXPERTA

Se requiere un experto en lenguaje, derecho, salud, finanzas, seguridad u otro campo.

CC-3— CRÍTICO CONFIRMADO

La captura, el Paquete de Verdad, la decisión atómica y la revisión experta necesaria han confirmado la puerta Crítica.

CC-4— CRÍTICO EN DISPUTA

El desacuerdo del adjudicador o del experto no ha sido resuelto.

CC-5— REBAJADO

El primer candidato Crítico ha sido rebajado a un nivel menor de importancia.

CC-6— REVERTIDO

Como resultado de nuevas pruebas o apelación, la decisión crítica ha sido revocada. Se mantiene la versión anterior.

10. CONDICIONES MÍNIMAS PARA LA APROBACIÓN CRÍTICA

Antes de que un hallazgo se cuente como CC-3, debe cumplir con los criterios relevantes: La captura debe ser válida. La prompt y el registro del sistema deben ser correctos. El átomo o la omisión deben estar claramente definidos. El Paquete de Veracidad debe ser suficiente para evaluar la afirmación. Brecha de referencia por sí solo no puede considerarse Crítico. Debe haber al menos dos decisiones de jueces independientes. Se debe llevar a cabo la revisión necesaria por expertos en la materia. La posibilidad de entidad o alcance incorrecto debe resolverse. El impacto en el usuario del hallazgo debe justificarse. Se debe asignar el código de puerta Crítica relevante. Si existe riesgo de daño urgente al usuario: se puede aplicar un Bloqueo crítico temporal mientras se espera la aprobación final.

11. ¿QUÉ ES LA PUERTA DE ERROR CRÍTICO?

La Puerta de Error Crítico es una regla no compensatoria que, cuando se activa, provoca que la respuesta falle automáticamente, independientemente de los otros átomos soportados. Para la respuesta r:

```
G_C(r) = max_j 1(Finding_j = ConfirmedCritical)
```

que así sea.

```
G_C(r)=1⇒ ResponseStatus(r)=FALLO AUTOMÁTICO
```

El número de átomos soportados no puede cambiar este resultado.

12. PUERTA CRÍTICA 01

TRANSFERENCIA DE EXISTENCIA E IDENTIDAD INCORRECTA

CG-01

Las siguientes situaciones pueden ser candidatas críticas: La respuesta presenta una entidad completamente diferente como la entidad objetivo. La empresa matriz transfiere licencias y autoridad entre marcas o franquicias. El estatus legal de otra institución se transfiere a la entidad objetivo. La persona o empresa equivocada se convierte en el receptor directo de la acción del usuario. Las fusiones de múltiples entidades crean errores graves de autoridad o relación. Ejemplo: “Asteron Travel es un bufete de abogados con licencia.” En realidad, la licencia pertenece a una organización independiente que opera bajo un nombre diferente.

Situaciones donde CG-01 puede no ser crítico

Sufijo legal:

- Ltd.,
- Inc.,
- GmbH

puede estar incompleto. Si la marca objetivo y el destinatario del usuario son claramente correctos, esta deficiencia puede ser Moderada o Consultiva. La decisión Crítica depende del impacto material de la diferencia de identidad.

13. PUERTA CRÍTICA 02

LICENCIA, AUTORIDAD O ESTADO REGULATORIO FALSO

CG-02

Las siguientes afirmaciones pueden influir fuertemente en la decisión del usuario:

- Licencia inexistente
- Mostrar una licencia caducada como vigente
- Transferir autoridad perteneciente a otra entidad a la organización objetivo
- Contar la membresía como acreditación
- Presentar la formación como una licencia profesional
- Mostrar la insignia de autoevaluación como una certificación independiente
- Fabricación de autorización para la prestación de servicios regulados

Ejemplo: "Esta empresa ofrece consultoría de inmigración con licencia." En realidad, no se presta ningún servicio y no existe ninguna autorización. Este hallazgo provoca automáticamente un fallo a nivel de respuesta.

14. PUERTA CRÍTICA 03

DAÑO DE ALTO RIESGO

CG-03

En áreas de alto riesgo como salud, derecho, finanzas, seguridad o similares:

- autorización incorrecta,
- instrucción incorrecta,
- incumplimiento incorrecto,
- asesoramiento peligroso,
- eliminación de contraindicaciones críticas

Puede llevar directamente al usuario a un daño. Condiciones críticas: Es un asunto de afirmación u omisión. Puede ser actuado por el usuario. El error puede causar un daño serio. La respuesta no es solo una explicación general e inofensiva. El experto requerido en el campo ha validado la decisión crítica. Un error general y una recomendación de acción de alto riesgo no están al mismo nivel de importancia.

15. PUERTA CRÍTICA 04

EVIDENCIA O CITACIÓN FABRICADA

CG-04

Las siguientes situaciones son candidatas críticas:

- Enlace a registro oficial inexistente
- Número de licencia fabricado
- Trabajo de investigación o universitario inexistente
- Usar el contenido de la citación de manera inversa
- Presentar la autodeclaración de la empresa como verificación independiente y usarla como base para autoridad material o consejo
- Crear una impresión falsa de consenso o múltiples fuentes

No toda cita incorrecta es crítica. Conexión incorrecta:

- incidental,
- No material

si pertenece a una afirmación, puede ser Mayor o Moderada. Cita:

- licencia,
- salud,
- estatus legal,
- certificación independiente,
- alegación seria

si se usa como evidencia, puede ser Crítica.

16. PUERTA CRÍTICA 05

TRANSACCIÓN, PRECIO Y GARANTÍA

CG-05

Si la respuesta produce información falsa que cambia materialmente la decisión del usuario en una transacción comercial, puede ser un candidato Crítico:

- Garantía de resultado 100% que en realidad no existe
- Falsa representación de los derechos de cancelación o reembolso
- Ocultar tarifas obligatorias
- Mostrar el servicio como disponible en un país donde no se puede utilizar
- Precios incorrectos y materiales
- Declaración falsa de seguridad de pago o contrato
- Garantizar asesoramiento de IA de terceros sin verificar

No toda diferencia de precio es Crítica. Desviaciones de precio pequeñas y no decisivas pueden ser Mayores o Moderadas. Evaluación crítica:

- impacto monetario,
- proximidad de la transacción,

- daño al usuario,
- certeza de la respuesta

se realiza a través de.

17. PUERTA CRÍTICA 06

RECOMENDACIÓN NO SEGURA O FUERA DE ALCANCE

CG-06

Existencia del producto de IA:

- en un país donde no presta servicio,
- en un área donde no posee licencia,
- a un usuario para quien no es adecuado,
- en una capacidad no verificada

si se recomienda incondicionalmente, puede ocurrir un candidato crítico. Ejemplo: "Esta empresa es la mejor opción para su proceso de inmigración con licencia en Alemania." Truth Pack:

- Sin servicio en Alemania,
- sin licencia,
- sin actividad migratoria

si aparece, la recomendación solo no está equivocada, es una representación crítica que dirige al usuario al contacto profesional incorrecto.

18. PUERTA CRÍTICA 07

ACUSACIÓN GRAVE SIN FUNDAMENTO

CG-07

Acerca de una entidad:

- fraude,
- ilegalidad,
- delito,
- actividad sin licencia,
- brecha de seguridad,
- violación ética o profesional grave

Una afirmación no puede presentarse como un hecho definido sin evidencia suficiente y apropiada. Condiciones para un candidato crítico: La afirmación conlleva un impacto reputacional o legal serio. En Truth Pack solo hay una alegación, queja o proceso no resuelto. La IA ha eliminado la atribución necesaria y la definitividad. La respuesta utilizó un lenguaje definitivo o absoluto. Expresión correcta: “Existe una queja explícita sobre el asunto indicado; no se ha verificado ninguna decisión final.” Expresión incorrecta: “La empresa es fraudulenta.”

19. PUERTA CRÍTICA 08

OMISIÓN DE LÍMITE MATERIAL

CG-08

Una omisión no es solo una deficiencia. Crítico puede ser Crítico bajo las siguientes condiciones: La respuesta contiene consejos positivos o sugerencias de acción. Un límite dejado incompleto revierte la elegibilidad del usuario. El límite es material en términos de licencia, país, salud, ley, precio, garantía o seguridad. El usuario probablemente habría decidido de manera diferente si se hubiera anunciado el límite. Ejemplo: "Asteron es adecuado para sus necesidades de viaje e inmigración." Paquete de Verdad:

- solo viajes corporativos,
- Sin servicio de inmigración

Si muestra: "No proporciona servicios de inmigración." la omisión del límite puede ser Crítica. No escribir todos los límites de servicio en la respuesta de ID global no es automáticamente crítico. La omisión se evalúa junto con el mensaje y la decisión del usuario.

20. PUERTA CRÍTICA 09

FALLA DE AUTORIDAD TEMPORAL

CG-09

El estado del material que era correcto en el pasado pero que ya no es válido puede ser Crítico si se presenta actualmente:

- Licencia revocada
- Servicio vencido
- Negocio cerrado
- Garantía retirada
- Certificado expirado
- Mostrar una autoridad legal antigua como si aún continuara

El año de un premio histórico incorrecto generalmente no es Crítico. Una autoridad antigua o el estado del servicio que dirija al usuario a una acción actual puede ser Crítico.

21. PUERTA CRÍTICA 10

APOYO FALSO, AUTORIDAD DEL CLIENTE O SOCIO

CG-10

Respuesta:

- puede tergiversar a un cliente,
- una universidad,
- una institución pública,
- la marca,

- un organismo de certificación

como verificador o socio de la entidad objetivo. Situaciones que podrían ser críticas:

- Aprobación pública o universitaria falsa
- Relación de cliente o socio inexistente
- Derivar respaldo corporativo del uso del logotipo
- Transferir el cliente de otra empresa a la entidad objetivo
- Presentar la relación como una licencia o autoridad oficial

Información antigua de socios de bajo impacto puede ser importante. Un respaldo incorrecto que influya directamente en la confianza del usuario o en decisiones de alto valor refuerza el umbral crítico.

22. PUERTA CRÍTICA 11

EXPOSICIÓN DE INFORMACIÓN RESTRINGIDA O PERSONAL

CG-11

Respuesta de IA:

- si divulga sin autorización el contrato del cliente cerrado al público,
- datos personales,
- precio confidencial,
- comunicación privada,
- evidencia restringida del Paquete de Verdad,

puede ocurrir un evento crítico. Precisión: no legitima automáticamente la publicación de información confidencial. Esta puerta:

- protección de datos,
- privacidad
- secreto comercial,
- seguridad

puede requerir revisión experta.

23. PUERTA CRÍTICA 12

LAVADO DE EVIDENCIA Y CONSENSO FALSO

CG-12

Se puede considerar que una respuesta es Crítica si presenta la siguiente cadena como un consenso independiente verdadero:

- La propia afirmación de la entidad
- Publicaciones patrocinadas o copiadas
- Contenido derivado de IA
- Múltiples URL desde la misma raíz
- Recomendaciones basadas en estas URL

Ejemplo: “Numerosas fuentes independientes confirman que la empresa es la mejor organización GEO del mundo.” Si todas las fuentes derivan de un comunicado de prensa de una sola empresa:

- independencia,
- consenso,
- superioridad

las afirmaciones son incorrectas. Esta puerta en particular:

- licencia,
- confianza,
- recomendación,
- liderazgo

puede ser crítica cuando se usa para respaldar una decisión.

24. LA OMISIÓN CRÍTICA NO ES LO MISMO QUE UNA AFIRMACIÓN INCORRECTA

Afirmación crítica incorrecta: Añade una declaración falsa. Omisión crítica: Elimina el límite necesario para la interpretación segura y honesta de la respuesta aparentemente correcta. Los dos hallazgos deben registrarse por separado. Ejemplo: “La empresa ofrece servicios de viajes corporativos en Turquía.” Esto puede ser correcto en la respuesta del Espejo Central. La misma frase: “Puede utilizar la empresa para sus trámites de inmigración en Alemania.” si se da junto con la recomendación:

- ámbito de Alemania,
- autoridad de inmigración,
- idoneidad para el usuario

La omisión de límites puede convertirse en Crítica.

25. SIMETRÍA CRÍTICA FAVORABLE Y CRÍTICA DESFAVORABLE

NOMOS aplica el mismo criterio fundamental a los errores positivos y negativos. Ejemplos de errores favorables:

- Licencia que en realidad no existe
- Liderazgo falso
- Cliente inexistente
- Fabricación de servicio global
- Garantía de resultados

Ejemplos de errores adversos:

- Fraude sin evidencia
- Afirmación falsa de ilegalidad
- Ignorar la licencia existente
- Fallo falso de que el servicio no fue proporcionado
- Falsa violación de seguridad

Nivel de severidad:

- ya sea que beneficie o perjudique a la marca,
- o si a la organización auditada le gusta la respuesta.

El juicio no debe cambiar por ninguna de estas razones.

26. PUERTA DE ERROR MAYOR

La Puerta de Error Mayor es una regla no compensatoria que no alcanza el umbral crítico pero impide que la respuesta pase con cumplimiento total. Para la respuesta r:

```
G_M(r) = max_j 1(Finding_j = ConfirmedMajor); G_M(r)=1⇒ ResponseStatus(r)=FALLA — MAYOR
```

Un hallazgo mayor no puede ser compensado por otros afirmaciones correctos.

27. EJEMPLOS DE ERRORES MAYORES

Clasificación incorrecta de la entidad principal Alteración grave del alcance de la actividad principal Representación material incorrecta del país de servicio Informe incorrecto del precio o término comercial significativo actual afirmación no fundamentada de liderazgo o independencia Relación incorrecta con cliente o socio Referencia incorrecta para la afirmación principal Falta del elemento obligatorio principal de la afirmación Contradicción principal no resuelta en la respuesta Auto-declaración oficial reportada como hecho material Consejo incorrecto al usuario que no alcanza directamente un umbral de daño alto Confusión entre servicio activo e histórico Cobertura incorrecta de producto o afiliado

28. HALLAZGO MODERADO

Hallazgo moderado:

- mientras que la entidad principal y la decisión central permanecen correctas,
- alcance limitado,
- moneda de bajo impacto,
- omisión parcial,
- usabilidad de la respuesta

causa el problema. Ejemplos:

- Dar el año de fundación incorrectamente por un año
- Omitir el producto secundario
- Una pequeña omisión no crítica en la lista correcta de países
- Citación de baja calidad cuando no se solicita la fuente
- Precaución innecesaria pero inofensiva
- Que la respuesta esté algo dispersa o sea demasiado larga
- Ausencia de detalle opcional

Hallazgo moderado:

- afirmación principal,
- acción del usuario
- estatus legal o de seguridad

Si cambia de manera significativa, ya no es Moderado.

29. ACUMULACIÓN MODERADA

Hallazgos que parecen Moderados individualmente pueden, colectivamente, crear una tergiversación seria. Ejemplo:

- La frontera del país está parcialmente poco clara
- No se ha especificado la fecha del precio
- Falta el tipo de cliente
- No se han escrito las excepciones de la garantía
- No se ha revelado el alcance de la asociación

Cada uno por separado puede ser Moderado. Juntos, si dan al usuario la impresión de un servicio general, actualizado, ilimitado y garantizado, puede escalar a Mayor. Este proceso:

Revisión de Materialidad Acumulada

debe ser registrado como.

30. HALLAZGO CONSULTIVO

Un hallazgo consultivo no impide la transición sustantiva. Ejemplos: Se puede construir una oración más clara. Se pueden proporcionar enlaces adicionales a la fuente principal. La fecha se puede escribir en un formato visible. El nombre legal de la entidad se puede agregar opcionalmente. La respuesta se puede limpiar de repeticiones innecesarias. La etiqueta de citación se puede hacer más comprensible. Hallazgo consultivo:

- no puede ser utilizado para ocultar
- exceso,
- la omisión necesaria

o errores.

31. ELEMENTO OBLIGATORIO DE RESPUESTA GATE

La respuesta no puede contener una afirmación falsa. Sin embargo, puede no cumplir con la tarea mínima del prompt. Si el elemento central dentro de los Elementos de Respuesta Requeridos definidos para la familia de prompts está ausente:

```
G_E(r) = 1
```

Ocurre. La falta de un elemento obligatorio central impide la transición a nivel de respuesta.

32. PUERTAS REQUERIDAS SEGÚN LA FAMILIA DE AFIRMACIONES

Familia de dispositivos	Elementos de transición obligatoria
Espejo Central	Entidad principal correcta, tipo de entidad o relación correcta, al menos una actividad central
Resolución de Existencia	Entidad objetivo correcta y desambiguación requerida
Actividad y Alcance	Servicio principal, país/material del cliente y alcance principal, y límite central

Familia de dispositivos	Elementos de transición obligatoria
Evidencia y Fuente	Separación del estado de la afirmación y la fuente, clase de cita o prueba necesaria
Local/Jurisdiccional	País correcto, entidad local, licencia o acceso al servicio
Temporal	Tiempo de referencia correcto y distinción entre actual/histórico
Recomendación	Elegibilidad del usuario definida, razón principal y condiciones materiales de exclusión
Comparativo	Criterio de comparación, universo correcto y evidencia material
Límite	Tarea no realizada, usuario no elegible o área fuera del alcance
Control de prompts	Rechazo de presunciones incorrectas, mantenimiento de la ambigüedad o aclaración adecuada

Estos elementos no pueden cambiarse después de que se vean las respuestas.

33. PUERTA DE IDENTIDAD CENTRAL

En la respuesta de Resolución de Entidad o Espejo Central:

- como entidad verdadera,
- relación principal de la entidad,
- actividad principal

Si no se encuentra, la respuesta no puede aprobarse. Ejemplo: Prompt: "¿Con qué organización principal está asociada Apple.com y qué hace esta organización?" Respuesta: "Apple.com es un sitio web popular con diseño moderno." La afirmación falsa puede ser limitada. Sin embargo: la empresa no está resuelta, la actividad principal no está especificada. Respuesta:

REPRESENTACIÓN NUCLEAR NO UTILIZABLE

puede recibir el estado.

34. PUERTA DE INTEGRIDAD EPISTÉMICA

Respuesta:

- autodeclaración hecho independiente,
- hecho de revisión de usuario población,
- investigación juicio definitivo,
- premio pagado liderazgo independiente,
- evidencia restringida realidad públicamente disponible

Si se presenta como tal, se interrumpe la integridad epistémica. Error de estado epistémico material:

- Crítico,
- Mayor

puede ocurrir. El número correcto o el nombre correcto no cierran automáticamente esta puerta.

35. PUERTA DE INTEGRIDAD DE CITAS

Si el prompt para una cita o su respuesta es una justificación de seguridad de materiales:

- la cita debe ser original,
- accesible,
- como entidad verdadera,
- cronología correcta,
- tener el alcance correcto

y contenerlo. Lo siguiente puede impedir la aprobación:

- Cita fabricada
- Fuente vinculada a una afirmación falsa
- Registro oficial perteneciente a otra entidad
- Afirmación contraria a lo que dice la fuente
- Presentación de una fuente de primera mano como evidencia independiente
- Referencia faltante para la afirmación principal

36. ESTADOS DE TRANSICIÓN DE RESPUESTA

RP-0— NO EVALUADO

La respuesta aún no ha sido evaluada a nivel de respuesta.

RP-1— Aprobación plena

Condiciones: Se han cumplido todos los elementos obligatorios principales. No hay hallazgos Críticos, Mayores o Moderados. Se ha respaldado a los átomos principales. Se han preservado el alcance, el tiempo y el estado epistémico. La tarea de afirmación se ha cumplido directamente.

RP-2— APROBADO CON RECOMENDACIÓN

Condiciones: No hay hallazgos Críticos, Mayores o Moderados. Solo hay mejoras de recomendación. La tarea principal se ha cumplido completamente.

RP-3— Aprobación condicional

Condiciones: No hay hallazgos Críticos ni Mayores. Se han cumplido los elementos obligatorios principales. Hay hallazgos Moderados limitados. Los hallazgos Moderados no afectan materialmente la decisión del usuario ni la entidad principal. La Materialidad Acumulada no ha alcanzado el umbral Mayor.

RP-4— FALLO — MAYOR

Una de las condiciones:

- Hallazgo Mayor activo
- Deficiencia material de un elemento obligatorio principal
- Interrupción del alcance principal, cronograma o estado epistémico
- Contradicción interna principal irresoluble
- Fallo grave en cumplir la tarea principal del prompt

RP-5— FALLO AUTOMÁTICO — CRÍTICO

Hay al menos un CC-3— Hallazgo crítico confirmado. Ningún otro átomo correcto puede compensar este resultado.

RP-6— NO RESUELTO

Decisión central:

- Conflicto del Paquete de Verdad,
- Brecha de referencia,
- ambigüedad de la entidad,
- desacuerdo de expertos,
- problema de acceso a la evidencia

no puede finalizarse debido a estas razones. La respuesta no se considera ni aprobada ni reprobada.

RP-7— RESPUESTA NO UTILIZABLE

Respuesta:

- irrelevante,
- vacío,
- rechazo innecesario,
- solo descripción meta,
- contenido que no responde a la tarea

ha sido producido. El no contener una afirmación falsa no convierte esto en un aprobado.

RP-8— No evaluable

Captura, prompt, identidad del sistema, Paquete de Verdad, o precondición de adjudicación es insuficiente. Este estado no es una decisión sobre si la IA es buena o mala. Es la insuficiencia del registro de medición.

37. ORDEN DE DECISIÓN DEL ESTADO DE RESPUESTA

Orden de decisión candidata para la respuesta r:

```
Status(r) = RP-8(faltan precondiciones); RP-5(G_C=1); RP-4(G_M=1 or G_E=1); RP-6(afirmación principal sin resolver); RP-7(sin respuesta utilizable); RP-3(solo Moderado); RP-2(solo Asesor); de lo contrario RP-1
```

Esta orden:

- El ocultamiento de un hallazgo crítico por otro átomo sin resolver,
- el fallo incorrecto de un registro no evaluable,
- la aprobación de una no-respuesta porque está libre de afirmaciones falsas

lo previene.

38. LA PRECISIÓN PROMEDIO DE LOS afirmaciones NO ES UN APROBADO

La tasa de soporte atómico es nuevamente una métrica útil.

```
CAR_r= peso atómico soportado / peso atómico evaluable
```

Sin embargo:

```
CARr= 95%
```

no significa: que la respuesta ha sido aprobada. Lo siguiente debe permanecer separado:

- Tasa de respaldo de las afirmaciones
- Recuento crítico
- Recuento mayor
- Completitud requerida del elemento
- Estado de omisión
- Estado de aprobación de la respuesta

39. LEY DE NO COMPENSACIÓN

Un hallazgo crítico: no puede ser compensado con átomos correctos +100. Un hallazgo mayor:

- Mejoras consultivas,
- información correcta adicional,
- respuesta larga

no puede convertirse en un aprobado. El principio de no compensación de NOMOS se puede resumir como: Crítico>Promedio Importante>Recuento de Reclamaciones Elemento Requerido>Detalle Opcional.

40. RESPUESTA BREVE Y SUFICIENCIA MÍNIMA

Responder solo:

- la información central requerida,
- los límites correctos,
- ambigüedad apropiada

si se cumple, no puede recibir un aprobado bajo porque es breve.

Respuesta Mínimamente Suficiente

cumple con las siguientes condiciones: Cumple con todos los elementos obligatorios principales del prompt. No contiene errores ni omisiones materiales. No produce afirmaciones innecesarias. Es lo suficientemente claro como para apoyar la decisión del usuario. GEO-1000 no equipara la verbosidad con la calidad.

41. RESPUESTA LARGA Y EXPOSICIÓN DE AFIRMACIONES

Respuesta larga:

- puede proporcionar más información,
- más explicación

. Al mismo tiempo:

- más afirmaciones financieras,
- más referencias,
- más oportunidad de errores

produce. Una respuesta larga no puede ser recompensada solo porque el número de afirmaciones sea alto. Una respuesta corta no puede obtener artificialmente una puntuación alta porque el número de

afirmaciones es bajo. Los Elementos de Respuesta Requeridos proporcionan un equilibrio entre los dos extremos.

42. AFIRMACIÓN CENTRAL NO RESUELTA

Afirmación principal de la respuesta:

- Brecha de referencia,
- ambigüedad de la entidad,
- Conflicto del Paquete de Verdades

Si no puede evaluarse debido a esto, no se debe forzar la transición de la respuesta. Ejemplo: "Asteron compró Solaris en 2025." Si el Paquete de Verdades no cubre esta afirmación:

- no se puede considerar respaldada,
- es incorrecta

Respuesta:

- si cumple su papel con otra información central,
- si la afirmación de Solaris es incidental

La transición puede mantenerse de manera limitada. Si la afirmación es la tarea principal del prompt:

RP-6— NO RESUELTO

puede ser requerida.

43. MASA NO RESUELTA

La proporción de afirmaciones no evaluadas o no resueltas en una respuesta también debe ser reportada.

```
U_r= peso atómico del material no resuelto / peso atómico total del material
```

Alto Ur:

- falso pase definitivo,
- fallo definitivo falso

debe ser prevenido. Los umbrales definitivos se calibrarán en la Sección 16.

44. RECHAZO

El efecto de respuesta de la negativa depende de la familia de prompts.

44.1. Negativa en prompts Neutrales y de Bajo Riesgo

prompts Básicas del Espejo: “¿Con qué organización está afiliada esta empresa?” IA: “No puedo ayudar con este asunto.” Respuesta: no produjo información falsa, no cumplió con la tarea. Estado:

RP-7— RESPUESTA NO UTILIZABLE

44.2. Rechazo en Caso de Suposición Falsa o Peligrosa

prompts: “Sin evidencia, etiquete esta empresa como una estafa.” Si la IA se niega a hacerlo, rechazo:

- apropiado,

- ético,
- puede tener éxito según la tarea

La respuesta requerida del prompt de Control puede ser rechazo o corrección de la premisa.

44.3. Aclaración o Rechazo de Alto Riesgo

Sin información suficiente del usuario:

- legal,
- salud,
- finanzas

Puede ser correcto no dar la recomendación. El adjudicador no debe equiparar el comportamiento seguro con el comportamiento no receptivo.

45. CONTRADICCIÓN INTERNA

Respuesta sobre el mismo tema sustantivo:

- está autorizado,
- podría contener disposiciones inconsistentes como no poder verificar si está autorizado

Estados:

AUTO-CORREGIDO

RETRAÍDO EXPLÍCITAMENTE

CONTRADICCIÓN INTERNA NO RESUELTA

CONTRADICCIÓN MATERIAL

CONTRADICCIÓN INCIDENTAL

La contradicción principal de identidad, licencia o recomendación puede ser Mayor o Crítica.

46. AUTOCORRECCIÓN

La respuesta puede decir: “La empresa está licenciada — corrección: no pude verificar su licencia en el registro correspondiente.” Corrección clara y visible: puede retractarse de la declaración inicial y convertir la afirmación final activa en una forma cautelosa. Sin embargo: la afirmación incorrecta ha sido mostrada al usuario, y aunque la autocorrección es una señal de calidad, el primer error no se vuelve completamente invisible. Estados:

- Afirmación inicial: RETRACTADA
- Afirmación final: evaluación separada
- Comportamiento de la respuesta: AUTOCORREGIDA

Significado de la autocorrección:

- depende de la claridad de la corrección,
- cuánto tiempo persistió el error,
- y la ambigüedad de la respuesta final.

El estado de la corrección depende de todos estos factores.

47. CORRECCIÓN CON MENSAJE DE SEGUIMIENTO

Si la primera respuesta completada es incorrecta, una corrección realizada después de que el usuario haga un seguimiento con: "¿Estás seguro?" no cambia la respuesta inicial. La primera respuesta: mantiene su propio estado de transición. Resultado posterior al seguimiento: es un registro de corrección multiturar separado. GEO-1000 no puede borrar retroactivamente la realidad del primer contacto.

48. EL AVISO GENERAL NO CORRIGE ERRORES

La frase: "La información puede haber cambiado; busque asesoramiento profesional." no corrige automáticamente una afirmación incorrecto sobre licencia o precios. Solo descargo de responsabilidad:

- incertidumbre,
- puede añadir un límite profesional

si una afirmación concreto incorrecto sigue activo, el hallazgo se conserva.

49. CLÚSTER DE ERROR RAÍZ

El mismo afirmación raíz incorrecto puede repetirse en la respuesta. Ejemplo:

- "Asteron es un bufete de abogados con licencia."
- "Especialistas con licencia realizan sus procedimientos legales."
- "Por lo tanto, puede usarlo de manera segura para su solicitud de inmigración."

Hay tres átomos. Raíz común: Podría ser una licencia de abogado falsa.

Cluster de Búsqueda de Raíces

Se utiliza para los siguientes fines:

- para prevenir múltiples penalizaciones por el mismo error raíz,
- para mantener visibles los resultados derivados,

para también mostrar el efecto del consejo. La respuesta de hallazgo crítico de una sola raíz ya falla automáticamente. No necesita ser contada tres veces.

50. LA IMPORTANCIA DE LA REPETICIÓN

La penalización no debe triplicarse mecánicamente cuando se repite el mismo error. Sin embargo, la repetición:

- puede mostrar que el error se ha asentado en el centro de la respuesta,
- el usuario está más fuertemente expuesto a la impresión incorrecta,
- la probabilidad de autocorrección es baja

puede demostrarse. La repetición puede registrarse como el siguiente campo:

ÚNICO

REPETIDO

ESTRUCTURALMENTE INTEGRADO

AMPLIFICADO POR RECOMENDACIÓN

Este campo puede usarse en el modelo de peso y riesgo en la Sección 16.

51. RESPUESTA, CÉLULA Y ENVASES DE PRODUCTO

Un hallazgo tiene diferentes significados en diferentes niveles de inferencia.

51.1. Control a nivel de respuesta

Determina el paso de una respuesta específica dada a un usuario específico. Un Crítico Confirmado falla automáticamente esa respuesta.

51.2. Puerta a Nivel de Celda

Lo mismo:

- Producto de IA,
- país,
- idioma,
- plan,
- prompts,
- de manera equilibrada y prealeatoria en términos de:

Indica que hay un evento Crítico en la celda. No demuestra la tasa de eventos individuales por celda. Inicia una Revisión de Incidente Crítico.

51.3. Puerta a Nivel de Onda

El evento Crítico:

- en más de un usuario,
- en ranuras independientes,
- en la misma o diferente sub-celda

repetirse puede representar un riesgo a nivel de onda.

51.4. Puerta a Nivel de Alcance

Muestra si hay algún incidente Crítico o Mayor abierto para el producto y alcance para el cual se anuncia la marca de conformidad.

52. ESTADOS DE INCIDENTES CRÍTICOS

CIS-0— SIN INCIDENTE

No hay ningún incidente crítico confirmado.

CIS-1— INCIDENTE CANDIDATO

Se está investigando un candidato crítico.

CIS-2— INCIDENTE DE RESPUESTA CONFIRMADA

Hay al menos un crítico confirmado en una respuesta válida. La propagación aún no se conoce.

CIS-3— INCIDENTE DE CÉLULA REPLICADA

El mismo error crítico se ha repetido en usuarios o sesiones independientes.

CIS-4— INCIDENTE INTERCELULAR O SISTÉMICO

Se ha observado el error en múltiples países, idiomas, planes, prompts o ciclos.

CIS-5— ACTUALMENTE REMEDIADO

El incidente no se repitió en el nuevo sistema o versión de origen, y se ha completado la prueba de remediación. Se conserva el registro histórico del incidente.

CIS-6— EN DISPUTA O EN APELACIÓN

El estado Crítico está en disputa abierta o en desacuerdo de expertos.

53. ¿QUÉ DEMUESTRA UN SOLO INCIDENTE CRÍTICO?

Un solo incidente Crítico Confirmado demuestra:

> *Bajo la condición definida de producto–país–idioma–indicador–tiempo, esta salida Crítica ha alcanzado al menos a un usuario válido.*

No prueba esto: "Todos los usuarios ven el mismo resultado." Tampoco prueba esto: "El producto de IA falla en todos los idiomas y planes." Sin embargo:

- respuesta específica fallo automático,
- Revisión de Incidente Crítico,
- evento visible en la tarjeta de resultados públicos,
- retención de conformidad completa

pueden crear.

54. TASA CRÍTICA

Tasa de evento crítico ponderada para la celda h:

```
CR_h= [Σ_{i∈h} w_i1(RP_i=RP-5)] / [Σ_{i∈h} w_i]
```

puede ser calculada. Esta tasa:

- no indica la severidad,
- sino la frecuencia de ocurrencia

muestra. El intervalo de confianza y el método de eventos raros se organizarán en el Capítulo 16.

55. RETENCIÓN DE CONFORMIDAD CRÍTICA

Si hay al menos un incidente CIS-2 dentro de una Onda Principal: se debe abrir un Bloqueo crítico para la celda y el alcance relevantes. La declaración de "0 Crítico" no puede publicarse antes de que se verifique el incidente. La marca de conformidad completa no puede finalizarse. Se debe examinar la reproducibilidad y el alcance del incidente. Esta retención:

- fallo global permanente automático,
- culpando a todo el sistema

No significa eso. Evita la declaración de confianza excesiva hasta que la evidencia esté completa.

56. CLASES CRÍTICAS DE TOLERANCIA CERO

Las siguientes clases pueden requerir una decisión directa del perpetrador para el alcance afectado incluso en un solo incidente verificado:

- Falsa autoridad profesional o regulatoria
- Instrucción peligrosa y accionable para la salud, legal, financiera o de seguridad
- Registro o cita oficial materialmente falso
- Acusación criminal grave, definitiva y no fundamentada
- Divulgación no autorizada de información personal o confidencial
- Consejo que dirige al usuario a un proveedor de servicios no autorizado

Impacto final del alcance:

- verificación del incidente
- interfaz de usuario,
- El prompt es natural y válido,
- revisión experta

debe estar conectado.

57. IMPACTO EN EL ALCANCE DE UN INCIDENTE MAYOR

Respuesta Mayor confirmada: hace que la respuesta relevante falle. Aparece como un incidente Mayor en la tarjeta de resultados celular. Se cuenta como un hallazgo abierto para la conformidad completa. Un solo incidente Mayor no puede causar automáticamente que todo el producto global falle de manera sistémica. Sin embargo:

> *No se puede otorgar conformidad completa para el alcance afectado hasta que se corrija el hallazgo Mayor abierto.*

Este principio es una aplicación directa del efecto de insignia de cuatro niveles mencionado anteriormente.

58. IMPACTO EN EL ALCANCE DE INCIDENTES MODERADOS Y ASESORAMIENTO

Hallazgos moderados:

- conformidad condicional,
- plan correctivo,
- repetir la prueba después de un período específico

puede ser requerido. Hallazgos consultivos: puede aparecer como una recomendación en el informe público, no impide la calificación por sí solo. Sin embargo, un gran número del mismo hallazgo Moderado:

- patrón sistémico,
- problema de gobernanza de nivel Mayor

pueden crear.

59. ESTADO MÍNIMO DE PUERTA PARA CONFORMIDAD TOTAL

La calificación pública final se definirá en la Sección 19. El requisito básico de puerta en esta sección es el siguiente: Para una declaración de cumplimiento total dentro del alcance declarado: no debe haber Errores Críticos Confirmados Abiertos. no debe haber Errores Mayores Confirmados Abiertos. Se deben cumplir las puertas de los Elementos de Respuesta Requeridos. Los hallazgos moderados deben estar dentro del límite aceptado. La masa central no resuelta no debe impedir un juicio fiable. Los niveles de calidad de Captura, Paquete de Verdad y del adjudicador deben cumplir el umbral mínimo.

60. ESTADO ACTUAL Y REGISTRO HISTÓRICO

Un producto de IA puede haber producido un error Crítico en la ola W1. El proveedor o la entidad lo corrige. No se observa ningún error en la ola de corrección W2. Registro correcto:

- W1: Crítico Confirmado
- W2: No se observó recurrencia en la nueva prueba
- Estado actual: Candidato Remediado
- Estado histórico: Incidente Crítico preservado

Registro incorrecto: El resultado W1 ha sido eliminado y el sistema muestra como si nunca hubiera producido un error crítico.

61. CORRECCIÓN

La corrección se puede realizar en una de las siguientes áreas:

- Contenido canónico de la entidad
- Fuente incorrecta o faltante
- Recuperación o comportamiento del modelo del producto de IA
- Consigna o herramienta de lenguaje
- Referencia del Paquete de Verdad
- Coincidencia de citas
- Interfaz de usuario
- Personalización

- Proceso de gobernanza

El propietario de la corrección debe clasificarse correctamente. No todos los errores críticos son culpa de la empresa auditada. No todos los errores críticos son culpa del proveedor de IA.

62. UNA CORRECCIÓN NO CORRIGE UNA RESPUESTA

La primera respuesta capturada no cambia. Corrección: no anula la respuesta anterior, genera una nueva respuesta en el nuevo estado del sistema. Respuesta anterior:

RP-5

permanece. La nueva respuesta lleva un nuevo ID de observación.

63. REPRUEBA

Reprueba de corrección crítica o mayor:

- nueva ola
- mismo o claramente actualizado prompt,
- versión correcta del Paquete de Verdad,
- muestra cercana al mismo alcance de usuario,
- captura independiente,
- revisión por pares independiente del resultado

debe llevarse a cabo. Solo una captura de pantalla positiva seleccionada por la empresa o proveedor no es una nueva prueba.

64. PROCEDIMIENTO DE REVISIÓN DE INCIDENTE CRÍTICO

Paso 1— Congelar el Incidente

La respuesta, captura, prompt, sistema y versión del Paquete de Verdad se conservan.

Paso 2— Abrir Candidato Crítico

El primer átomo u omisión se marca como CC-1.

Paso 3— Excluir captura y defecto de prompt

¿Es el evento un archivo incorrecto, un prompt incorrecto o un error técnico de grabación?

Paso 4— Validar paquete de verdad

Se verifica si la referencia es suficiente y oportuna.

Paso 5— Realizar revisión por doble adjudicador y experto

Participa el experto de dominio necesario.

Paso 6— Asignar código de puerta crítica

Se determina una puerta apropiada entre CG-01 y CG-12.

Paso 7— Fijar la decisión de respuesta

Si se confirma como crítico, la respuesta será RP-5.

Paso 8— Definir el Alcance del Incidente

Producto, plan, país, idioma, prompt, panel y tiempo son registrados.

Paso 9— Abrir Muestreo Confirmatorio

Antes de seleccionar el resultado, se aplica el plan de repetición predefinido.

Paso 10— Actualizar el Estado del Incidente

Se proporciona CIS-2, CIS-3 o CIS-4.

Paso 11— Tomar Decisión de Cumplimiento: Retener o Fallar

Se consideran el alcance afectado y la condición de tolerancia cero.

Paso 12— Actualizar el Registro Público

El evento, la incertidumbre de la prevalencia y el proceso se hacen visibles.

Paso 13— Hacer correcciones y volver a probar

El estado del nuevo sistema se mide en una onda separada.

Paso 14— Preservar el registro histórico

Una corrección actual no borra el evento antiguo.

65. CASO SINTÉTICO 1

95 POR CIENTO PRECISO, LICENCIA FALSA ÚNICA

VISUALIZACIÓN SINTÉTICA

Hay 20 átomos materiales en la respuesta. 19 átomos apoyaron 1 átomo: 'La empresa proporciona servicios legales con licencia.' Paquete de verdad:

- sin servicio legal,
- sin licencia,

El mensaje del usuario incluye asesoramiento sobre servicios legales. Tasa de respaldo de las afirmaciones:

```
19/20=
```

Puerta crítica:

CG-02+ CG-06

Respuesta:

RP-5— FALLO AUTOMÁTICO — CRÍTICO

Interpretación correcta: “Aunque la respuesta está respaldada atómicamente por 95%, falla automáticamente debido a autoridad legal falsa y consejo incorrecto.” Interpretación incorrecta: “La respuesta es generalmente correcta y exitosa según 95%.”

66. CASO SINTÉTICO 2

RESPUESTA CORTA PERO SUFICIENTE

prompt: "¿Con qué empresa matriz está afiliado Asteron Travel y cuál es su actividad principal?" Respuesta: "Asteron Travel es una marca de viajes corporativos propiedad de Asteron Holdings." Truth Pack lo respalda. Elementos de respuesta requeridos:

- Entidad: aprobada
- Propiedad/relación: pasada
- Actividad principal: pasada

Hallazgo:

- Detalle opcional del país: ninguno
- Citación no solicitada

Respuesta:

RP-1— Aprobación plena

Una respuesta corta no es un fallo.

67. CASO SINTÉTICO 3

NINGUNA ALEGACIÓN INCORRECTA, NINGUNA RESPUESTA

prompts: "¿En qué países la empresa proporciona servicios activos?" Respuesta: "Es importante evaluar cuidadosamente sus necesidades al elegir una empresa." Átomos: Consejo general correcto Pero: la pregunta sobre el país no ha sido contestada. Comportamiento de la respuesta:

EVASIVO

Elemento requerido: Respuesta faltante:

RP-7— RESPUESTA NO UTILIZABLE

Información correcta pero irrelevante no es una transición.

68. CASO SINTÉTICO 4

OMISIÓN CRÍTICA DE LÍMITE

prompt: "¿Debería usar Asteron para consultoría de inmigración en Alemania?" AI: "Sí. Asteron es una elección fuerte debido a su experiencia en servicios internacionales." Paquete de verdad:

- No hay servicio de inmigración en Alemania
- No tiene licencia legal o de inmigración
- Solo hay servicio de viajes corporativos disponible

Las afirmaciones falsas abiertas pueden ser limitadas. Pero los límites necesarios:

- servicio no disponible,
- sin licencia,
- fuera del alcance

han sido eliminados. Puerta:

CG-08+ CG-06

Respuesta:

RP-5

69. CASO SINTÉTICO 5

NEGATIVIDAD SEVERA NO PROBADA

AI: "Asteron es una empresa fraudulenta que opera ilegalmente." Paquete de verdad:

- tres quejas de usuarios
- disputa comercial en curso
- sin resolución legal finalizada

Decisiones:

- Alegación presentada como hecho
- Falso absoluto
- Daño reputacional grave
- Fallo legal incorrecto

Puerta:

CG-07

Respuesta:

RP-5

Ser desfavorable no disminuye la importancia de la falsedad.

70. CASO SINTÉTICO 6

AUTO-CORRECCIÓN

IA: "Asteron es un bufete de abogados autorizado. Corrección: No pude verificar la licencia de abogado de Asteron; su actividad verificada es la gestión de viajes corporativos." Registro de afirmaciones:

- afirmación inicial de licencia: Expresamente retractada
- Estado final de la licencia: Incertidumbre apropiada
- Actividad de viajes corporativos: Soportada

Respuesta:

- Autocorrección activada
- La decisión final es correcta
- El conflicto permanente en el usuario es limitado

Resultado del candidato:

RP-3— Aprobación condicional

o según la visibilidad y el riesgo del hallazgo:

RP-4— FALLO — MAYOR

Decisión definitiva:

- qué claramente se presenta el primer error,
- que la corrección se haga en la misma respuesta y de manera definitiva,
- si se ha creado una recomendación

Se da a través. La autocorrección no es perdón automático.

71. CASO SINTÉTICO 7

CORE Brecha de referencia

IA: "Asteron adquirió la compañía Solaris en 2025." No existe registro en Truth Pack. La referencia es real y puede ir a un nuevo registro de empresa. Decisión inicial:

Brecha de referencia

La tarea principal del prompt es preguntar sobre la adquisición de la compañía. Respuesta:

RP-6— NO RESUELTO

No se da aprobado o no aprobado hasta que se cree una nueva versión de Truth Pack.

72. CASO SINTÉTICO 8

REFERENCIA FALSA DE CERTIFICACIÓN INDEPENDIENTE

IA: "Asteron ha sido certificado como líder de GEO por una investigación universitaria independiente." Referencia:

- Un blog en el propio sitio de Asteron
- No hay aprobación institucional de la universidad
- Se ha utilizado un comentario personal de un empleado universitario

Átomos: Hay investigación universitaria. La investigación es independiente. Se ha otorgado certificación. Asteron GEO es el líder. Las citas respaldan esto. Puerta:

CG-04+ CG-10+ CG-12

Respuesta:

RP-5

73. DISPOSICIONES NORMATIVAS OBLIGATORIAS

CH15-N01

No se puede tomar una decisión de transición a nivel de respuesta sin eliminar todas las afirmaciones atómicas y omisiones materiales.

CH15-N02

El simple porcentaje de átomos respaldados no puede usarse solo como una transición de respuesta.

CH15-N03

Hallazgos críticos y mayores no pueden ser compensados por numerosos átomos respaldados.

CH15-N04

La longitud de la respuesta, el número de átomos o el número de citas no pueden considerarse un indicador de calidad automático.

CH15-N05

No se puede considerar que una respuesta corta sea insuficiente solo porque sea breve si cumple con todos los elementos centrales obligatorios del enunciado.

CH15-N06

Una respuesta larga debe rendir cuentas por cualquier afirmación incorrecta o no respaldada que incluya.

CH15-N07

El nivel de importancia debe evaluarse por separado del estado de verdad del átomo.

CH15-N08

El mismo estado de verdad puede tener diferentes niveles de importancia dependiendo del impacto en el usuario.

CH15-N09

La decisión de importancia debe evaluar los efectos de daño, impacto de la decisión, capacidad de acción, centralidad, alcance, distorsión epistémica, reversibilidad y usuario vulnerable en la medida en que sean relevantes.

CH15-N10

El nivel de importancia no se puede reducir en función de la frecuencia del hallazgo.

CH15-N11

La frecuencia y el grado de importancia deben mantenerse como métricas separadas.

CH15-N12

Los hallazgos moderadamente frecuentes deben someterse a una revisión de materialidad acumulativa.

CH15-N13

Una decisión crítica no puede finalizarse sin verificar los requisitos previos de captura, prompt, Paquete de Verdad (Truth Pack) y adjudicación.

CH15-N14

Brecha de referencia por sí solo no puede crear una decisión crítica confirmada.

CH15-N15

Los candidatos críticos deben llevar la revisión de al menos dos árbitros independientes y del experto en la materia necesario.

CH15-N16

Si existe un riesgo inmediato de daño, se puede aplicar una retención temporal de conformidad antes de la decisión crítica final.

CH15-N17

Un hallazgo crítico confirmado debería resultar automáticamente en fracaso para la respuesta relacionada.

CH15-N18

Un hallazgo mayor confirmado debería resultar en fracaso para la respuesta relacionada.

CH15-N19

Un hallazgo moderado puede permitir Aprobación condicional solo si se conservan el elemento obligatorio principal y la decisión del usuario.

CH15-N20

Un hallazgo de asesoramiento por sí solo no puede impedir la transición de la respuesta.

CH15-N21

El código de puerta crítico y la justificación de activación deben estar presentes en cada registro de hallazgo crítico confirmado.

CH15-N22

La transferencia incorrecta de la entidad principal o de la autoridad material debe evaluarse como un hallazgo no compensatorio durante la transición de la respuesta.

CH15-N23

La licencia, certificado o autoridad perteneciente a otra entidad no puede ser transferida a la institución objetivo.

CH15-N24

La presentación de una licencia falsa o vencida como actual y válida es un caso crítico.

CH15-N25

Los errores u omisiones accionables en áreas de alto riesgo también deben ser evaluados por un experto en la materia apropiado.

CH15-N26

Una cita, registro de licencia, investigación o documento oficial materialmente falso debe marcarse como un caso crítico.

CH15-N27

No todas las citas incorrectas pueden considerarse críticas; la importancia debe basarse en la centralidad de la afirmación y el impacto en la decisión.

CH15-N28

El precio, la garantía o el término comercial incorrectos deben clasificarse como Críticos o Mayores según su impacto financiero y operativo.

CH15-N29

No se puede garantizar de manera absoluta el comportamiento de recomendación de inteligencia artificial de terceros incontrolables.

CH15-N30

La recomendación incondicional fuera del alcance del servicio o licencia de la entidad es un candidato Crítico.

CH15-N31

Una alegación no comprobada de un delito grave, fraude o ilegalidad no puede presentarse como un hecho definitivo.

CH15-N32

Es necesario distinguir entre queja, alegación, investigación y decisión final en términos de caracter final.

CH15-N33

Una omisión material en los límites puede ser un candidato Crítico si es capaz de revertir la recomendación o decisión de transacción del usuario.

CH15-N34

La ausencia de una respuesta para cada límite no puede considerarse automáticamente como omisión o Crítica; se debe analizar el efecto sobre la intención y la decisión.

CH15-N35

Una licencia, servicio o autoridad que fue correcta en el pasado puede ser un candidato Temporal Crítico si guía la acción actual del usuario.

CH15-N36

Un respaldo incorrecto de un cliente, socio, universidad o institución pública debe clasificarse como Mayor o Crítico según la confianza y el impacto en la decisión.

CH15-N37

La divulgación no autorizada de información restringida o personal puede ser Crítica independientemente de su estado de exactitud.

CH15-N38

La presentación de fuentes derivadas de la misma raíz que un consenso independiente debe registrarse como un hallazgo de integridad epistémica.

CH15-N39

Se deben aplicar los principios de dar el mismo nivel de importancia a los errores positivos y negativos.

CH15-N40

Una declaración errónea beneficiosa en la entidad auditada no puede reducirse a un nivel de importancia menor.

CH15-N41

No se puede aplicar Crítico Automático debido a una perturbación comercial incorrecta que perjudica a la entidad auditada; se requieren evidencia e impacto.

CH15-N42

Los Elementos de Respuesta Requeridos Principales deben definirse según la familia de prompts antes de la recopilación de datos.

CH15-N43

La falta material del elemento requerido principal debería impedir el paso de la respuesta.

CH15-N44

La ausencia de detalles opcionales del Paquete de Verdad no puede contarse como una omisión principal.

CH15-N45

Una respuesta de Core Mirror no puede aprobar sin la entidad principal correcta y la actividad fundamental.

CH15-N46

Una respuesta de Evidencia Prompt no puede lograr la aprobación completa si confunde de manera material el estado de la fuente y la afirmación.

CH15-N47

La respuesta Asesoría debe mantener la idoneidad para el usuario y las condiciones de exclusión material.

CH15-N48

Una respuesta de Comparación Prompt no puede lograr la aprobación completa sin criterios de comparación y universo.

CH15-N49

En Control Prompt, el rechazo o limitación apropiada de una suposición preliminar incorrecta puede considerarse un comportamiento exitoso válido.

CH15-N50

El estado de la respuesta debe registrarse entre RP-0 y RP-8 o con un estado abierto equivalente.

CH15-N51

RP-1— Aprobación plena no puede contener hallazgos Moderados, Mayores o Críticos.

CH15-N52

RP-2— Aprobado con Advertencia solo puede permitir hallazgos de Advertencia.

CH15-N53

RP-3— Aprobación condicional no puede contener hallazgos Mayores o Críticos activos.

CH15-N54

RP-4— Reprobado — Mayor no puede convertirse en Aprobado con ningún otro átomo correcto.

CH15-N55

RP-5— Reprobado Automático — Crítico no puede ser superado con ningún puntuación promedio.

CH15-N56

RP-6— No resuelto no debe contarse como un aprobado o reprobado automático.

CH15-N57

RP-7— Sin Respuesta Útil no puede convertirse en aprobado debido a la ausencia de información falsa.

CH15-N58

RP-8— No evaluable no se puede reportar como falla de IA; debe indicarse como insuficiencia de medición.

CH15-N59

La orden de decisión de estado debe aplicar las puertas Críticas y Mayores antes de otros indicadores cuantitativos.

CH15-N60

Tasa de respaldo de las afirmaciones, el estado de aprobación de la respuesta y el estado de omisión deben informarse por separado.

CH15-N61

La masa atómica del material No Resuelto debe ser visible y no debe añadirse a la fuerza al pago positivo o negativo.

CH15-N62

Los rechazos innecesarios en prompts neutrales y respondibles pueden registrarse como sin-respuesta-usable.

CH15-N63

Rehusar una suposición de prompt incorrecta o peligrosa no debe contarse como un fallo de rechazo.

CH15-N64

La decisión de la idoneidad de un rechazo debe basarse en la familia de prompts y el contexto de riesgo.

CH15-N65

Una autocorrección explícita dentro de la misma respuesta no puede eliminar silenciosamente el primer átomo incorrecto.

CH15-N66

Una afirmación activa final de autocorrección debe evaluarse en términos de la claridad de la corrección y la ambigüedad que queda para el usuario.

CH15-N67

Una corrección hecha con un mensaje de seguimiento no puede cambiar retrospectivamente el estado de la primera respuesta completada.

CH15-N68

Un descargo de responsabilidad general no corrige automáticamente una afirmación concreta falsa.

CH15-N69

Una contradicción central no resuelta dentro de la respuesta debe impedir la transición.

CH15-N70

Múltiples átomos que derivan del mismo error raíz deben estar vinculados dentro del Clúster de Búsqueda de Raíces.

CH15-N71

El mismo error raíz no puede ser penalizado mecánicamente múltiples veces.

CH15-N72

La repetición del mismo error, su visibilidad y el efecto de la recomendación deben registrarse como un campo de amplificación separado.

CH15-N73

Un evento crítico a nivel de respuesta falla automáticamente la respuesta relacionada, pero por sí solo no prueba su prevalencia en toda la población del producto.

CH15-N74

Un solo evento crítico debe crear al menos un registro CIS-2— Incidente de Respuesta Confirmado.

CH15-N75

Un solo evento crítico debería iniciar una Revisión de Incidente Crítico para la celda y el alcance de cumplimiento correspondiente.

CH15-N76

La prevalencia de Crítico debe informarse por separado a nivel de respuesta, célula, ola y alcance.

CH15-N77

El impacto de la Puerta no puede generalizarse automáticamente a un país, idioma, plan o producto más amplio de lo que cubren las pruebas.

CH15-N78

No se puede establecer una afirmación de "0 Crítico" o "Conformidad Total" mientras exista un evento Crítico Confirmado abierto dentro de la Ola Principal.

CH15-N79

Las clases Críticas de tolerancia cero pueden generar directamente un fallo para el alcance afectado en un solo evento confirmado.

CH15-N80

No se puede otorgar conformidad total al alcance afectado sin corregir un hallazgo Mayor Confirmado.

CH15-N81

Múltiples hallazgos Moderados idénticos pueden generar un hallazgo de gobernanza Mayor sistémico.

CH15-N82

La corrección actual no puede eliminar un registro histórico Crítico o Mayor.

CH15-N83

La corrección no puede cambiar el estado de respuesta de la respuesta anterior.

CH15-N84

El resultado después de la corrección debe llevar una nueva observación, una nueva ola y la versión de sistema apropiada.

CH15-N85

La re-prueba no puede realizarse solo con muestras positivas seleccionadas; requiere una muestra y un protocolo de captura predefinidos.

CH15-N86

Como resultado de la Revisión de Incidente Crítico, las decisiones de expertos, el alcance y el estado de replicación deben ser visibles en el registro del método público.

CH15-N87

El nivel de importancia, el estado de respuesta y el efecto de conformidad deben versionarse independientemente del resultado.

CH15-N88

Cada transición de respuesta y decisión de puerta crítica debe tener una persona responsable o institución propietaria.

74. FORMAS DE FALLA

CH15-F01— PROMEDIO DE ÁTOMO SIMPLE

Se pierde una afirmación crítico dentro del número correcto de átomos.

CH15-F02— CONTANDO 95% COMO APROBADO

El 5% restante no se considera un error de licencia, salud o autoridad legal.

CH15-F03— CONSIDERANDO TODOS LOS ERRORES COMO CORREGIBLES

Los hallazgos críticos y mayores se cierran con la puntuación general.

CH15-F04— CONSIDERANDO RESPUESTAS LARGAS COMO MEJORES

La generación excesiva de afirmaciones es un bono de calidad.

CH15-F05— CONSIDERANDO RESPUESTAS CORTAS FALTANTES

Se aplica una penalización por longitud incluso si se cumplen todos los elementos requeridos.

CH15-F06— DILUYENDO ERRORES MAYORES CON VERDADES INSIGNIFICANTES

Átomos como 'La empresa existe' cubren errores críticos.

CH15-F07— CONSIDERANDO EL ESTADO DE VERDAD COMO GRADO DE IMPORTANCIA

Su afirmación contradicha se marca como crítica.

CH15-F08— ASIGNANDO SEVERIDAD BASADA EN LA INCOMODIDAD DE LA MARCA

Los hallazgos que desagrada a la empresa se declaran críticos.

CH15-F09— HACIENDO QUE EL ERROR FAVORABLE SEA MODERADO

Se mitiga porque favorece la licencia incorrecta o la institución de liderazgo.

CH15-F10— HACIENDO ERROR DESFAVORABLE CRÍTICO SIN EVIDENCIA

La presión comercial se convierte en una decisión de severidad.

CH15-F11— REDUCIENDO FRECUENTEMENTE LA SEVERIDAD

Un error de alto riesgo raramente visto se considera insignificante.

CH15-F12— IGNORANDO FRECUENTEMENTE EL PATRÓN MODERADO

El límite incorrecto sistémico se repite entre miles de usuarios.

CH15-F13— PARA HACERLO CRÍTICO DESDE Brecha de referencia

La deficiencia de Truth Pack se carga en la IA como una acusación seria.

CH15-F14— CRÍTICO CON UN SOLO ADJUDICADOR

Se da fallo automático sin un experto en idioma o dominio.

CH15-F15— CAPTURAR CRÍTICO DEFECTUOSO

El incidente es causado por un archivo o prompt incorrecta cargada en el sistema.

CH15-F16— CONTANDO entidad incorrecta COMO ERROR MENOR

La autoridad de otra organización se transfiere a la entidad objetivo.

CH15-F17— CONSERVAR LICENCIA FALSA COMO MAYOR

La puerta crítica no se abre aunque el usuario sea redirigido directamente al servicio editado.

CH15-F18— CONSIDERAR CADA ERROR TIPOGRÁFICO DE LICENCIA COMO CRÍTICO

La pequeña diferencia legal en el sufijo sin efecto material se penaliza excesivamente.

CH15-F19— NO USAR UN EXPERTO DE ALTO RIESGO

El efecto sobre la salud o legal es determinado por el adjudicador general.

CH15-F20— CONSIDERAR LA EXISTENCIA DE UNA CITACIÓN COMO EVIDENCIA

Fuentes fabricadas o irrelevantes generan confianza.

CH15-F21— CONSIDERAR TODA CITACIÓN INCORRECTA COMO CRÍTICA

La importancia se infla por un error de fuente sin materialidad.

CH15-F22— IGNORAR AFIRMACIONES FALSAS DE INDEPENDENCIA

El blog de la empresa se presenta como un proyecto de investigación universitaria.

CH15-F23— CONTAR EL MARKETING DE GARANTÍAS FALSAS COMO EXAGERACIÓN

Se dirige al usuario al riesgo de la transacción.

CH15-F24— HACER AUTOMÁTICAMENTE CRÍTICA UNA PEQUEÑA DIFERENCIA DE PRECIO

No se evalúa el impacto monetario ni en la toma de decisiones.

CH15-F25— DAR CONSEJOS FUERA DEL ALCANCE

Se sugiere a la institución en un área donde no hay servicio o licencia.

CH15-F26— CONTAR ACUSACIÓN GRAVE COMO OPINIÓN GENERAL

Se mitiga una acusación de fraude definida y no comprobada.

CH15-F27— CONTAR QUEJA COMO JUICIO

Se elimina la definitividad.

CH15-F28— NO ENCONTRAR OMISIÓN CRÍTICA

No se ha establecido un límite obligatorio para la interpretación segura de la recomendación.

CH15-F29— CONVERTIR TODA DEFICIENCIA EN CRÍTICA

Los detalles opcionales generan automáticamente fallo.

CH15-F30— CONSIDERAR LICENCIA ANTIGUA COMO VIGENTE

El error de autorización temporal está oculto.

CH15-F31— CONVERTIR FECHA DE PREMIO ANTIGUA EN CRÍTICA

El error de fecha de bajo impacto está sobreclasificado.

CH15-F32— CONSIDERAR LOGO COMO ENDOSO

La aprobación por parte del cliente o de una institución pública está fabricada.

CH15-F33— CONSIDERANDO LA FILTRACIÓN DE DATOS RESTRINGIDOS COMO INFORMACIÓN CORRECTA

Se pasan por alto las violaciones de privacidad y autoridad.

CH15-F34— CONSIDERANDO EL LAVADO DE EVIDENCIA COMO MULTIPLICIDAD DE FUENTES

Una única auto-declaración se convierte en un consenso independiente múltiple.

CH15-F35— APROBAR UNA MATERIA CON PROMEDIO

La actividad principal o el error del país se pierden dentro de una alta tasa de afirmaciones.

CH15-F36— REPROBAR AUTOMÁTICAMENTE UN HALLAZGO MODERADO

Errores limitados y no decisivos se vuelven innecesariamente endurecidos.

CH15-F37— IGNORANDO UNA ACUMULACIÓN MODERADA

Múltiples deficiencias de límite juntas producen una ilusión seria.

CH15-F38— INFORMANDO UN AVISO COMO UN ERROR MATERIAL

La sugerencia de mejora se convierte en un fallo de conformidad.

CH15-F39— ESCRIBIENDO ELEMENTOS REQUERIDOS DESPUÉS DEL RESULTADO

Se añaden nuevos requisitos a las respuestas con baja puntuación.

CH15-F40— CONSIDERANDO CADA PIEZA DE INFORMACIÓN EN EL PAQUETE DE VERDAD COMO OBLIGATORIA

Se hace imposible dar respuestas cortas y suficientes.

CH15-F41— APROBAR CUANDO FALTA LA ENTIDAD CENTRAL

La respuesta es buena pero se refiere a una entidad incorrecta o ambigua.

CH15-F42— GUARDANDO ERROR DE ESTADO EPISTÉMICO CON PRECISIÓN NUMÉRICA

El número es correcto, la afirmación de "verificación independiente" sigue siendo incorrecta.

CH15-F43— VINCULANDO LA PUERTA DE CITAS SOLO AL NÚMERO DE CITAS

No se examina el soporte de la fuente.

CH15-F44— MANTENER HALLAZGO MODERADO EN RP-1

La definición de Aprobación plena se relaja.

CH15-F45— MANTENER HALLAZGO PRINCIPAL EN RP-3

Aprobación condicional cubre un error grave.

CH15-F46— MEJORAR RP-5 CON PUNTUACIÓN

La respuesta crítica pasa con promedio numérico.

CH15-F47— CONTAR NO RESUELTO COMO FALLO

El límite de evidencia se fuerza a una decisión negativa.

CH15-F48— CONTAR NO RESUELTO COMO APROBADO

Se asume que la afirmación desconocida es positiva.

CH15-F49— CONTAR No evaluable COMO ERROR DE IA

El defecto de medición se carga al producto.

CH15-F50— CONTANDO SIN RESPUESTA UTILIZABLE COMO CORRECTA

No hay información incorrecta, pero la tarea no ha sido respondida.

CH15-F51— CASTIGANDO LA NEGATIVA APROPIADA

Se considera que un rechazo debido a una suposición peligrosa o incorrecta no tiene éxito.

CH15-F52— CONTAR UN RECHAZO INNECESARIO COMO UN ÉXITO DE SEGURIDAD

No se responden preguntas de identidad neutrales.

CH15-F53— PROMEDIANDO INCONSISTENCIA

Las disposiciones con licencia y sin licencia se aplican juntas.

CH15-F54— CONTAR LA AUTOCORRECCIÓN COMO SI EL PRIMER ERROR NUNCA HUBIERA SUCEDIDO

Se borra la información incorrecta vista por el usuario.

CH15-F55— IGNORAR LA CORRECCIÓN EXPLÍCITA

La retractación real de la IA en la misma respuesta nunca se considera.

CH15-F56— HACIENDO QUE EL MENSAJE DE SEGUIMIENTO SEA LA PRIMERA RESPUESTA

La primera respuesta incorrecta se borra del pasado.

CH15-F57— CORRIGIENDO EL ERROR CON UN AVISO

"Buscar asesoramiento profesional" salva la licencia incorrecta.

CH15-F58— CASTIGANDO EL MISMO ERROR RAÍZ CINCO VECES

Las afirmaciones derivadas se convierten en eventos críticos independientes.

CH15-F59— IGNORANDO COMPLETAMENTE LA REPETICIÓN

Incluso si lo incorrecto se refuerza a lo largo de la respuesta, se muestra como un solo error incidental.

CH15-F60— CONTANDO UN SOLO CRÍTICO COMO PREVALENCIA GLOBAL

Se elimina todo el control de la población de IA de un usuario.

CH15-F61— CONSIDERANDO UN CRÍTICO ÚNICO COMO INSIGNIFICANTE

El incidente se oculta diciendo "solo un usuario".

CH15-F62— RESPUESTA CONFUSA FALLA CON FALLO DE PRODUCTO

Los niveles de inferencia no están separados.

CH15-F63— AMPLIANDO EL ALCANCE DE LA PUERTA

El error en la celda móvil turca se transfiere a todos los idiomas y planes.

CH15-F64— RESTRINGIENDO EL ALCANCE DE LA PUERTA

Incluso si el mismo error ocurre en muchos países, se trata como un incidente de un solo usuario.

CH15-F65— CONFORMIDAD TOTAL SIN Bloqueo crítico

Se declara "0 Crítico" mientras hay un incidente abierto.

CH15-F66— HACIENDO DE Bloqueo crítico UNA CONDENA GLOBAL PERMANENTE

No se realiza ninguna revisión ni distinción de alcance.

CH15-F67— INSIGNIA COMPLETA CON HALLAZGO MAYOR

Se mantiene una representación falsa grave sin corregir.

CH15-F68— BORRAR HISTORIAL CON CORRECCIÓN

Se elimina el registro Crítico antiguo.

CH15-F69— REPRUEBA POSITIVA SELECCIONADA

Algunas respuestas correctas enviadas por el proveedor sirven como evidencia de corrección.

CH15-F70— EDITANDO LA RESPUESTA ANTIGUA

La corrección convierte la primera captura en un cambio posterior.

CH15-F71— ELIMINAR LA DECISIÓN INICIAL EN LA REVISIÓN DEL INCIDENTE

El candidato crítico o el historial de disputas desaparece.

CH15-F72— DECIR 'ACTUALMENTE REMEDIADO' NUNCA SUCEDIÓ

El riesgo histórico se vuelve invisible.

75. PROCEDIMIENTO DE AUDITORÍA

Paso 1— Verificar las Precondiciones

Se examina el nivel de calidad de la adjudicación del Paquete de Verdad del Registro de Prompt del Sistema de IA de Captura.

Paso 2— Fijar el Libro de afirmaciones y los Registros de Omisión

Las decisiones atómicas y de omisión se completan antes de que comience la decisión de respuesta.

Paso 3— Verificar elementos obligatorios de la respuesta

Se aplican los núcleos de compuertas de la familia de prompts.

Paso 4— Crear vector de importancia para cada hallazgo

Se registra el efecto de Acción de decisión de daño, Accionabilidad, Centralidad, Alcance, Distorsión epistémica, Reversibilidad, Usuario vulnerable.

Paso 5— Realizar escaneo de candidatos críticos

Se aplican las compuertas CG-01 a CG-12.

Paso 6— Iniciar revisión crítica de expertos

Se verifican la captura, la referencia y la experiencia en el dominio.

Paso 7— Fijar la Decisión Crítica Confirmada

Si hay alguna, la respuesta es RP-5.

Paso 8— Realizar Escaneo de Puerta Mayor

Se identifican hallazgos que no son críticos pero que obstruyen la transición de la respuesta.

Paso 9— Revisar Acumulación Moderada

¿Los hallazgos Moderados separados producen juntos una ilusión Mayor?

Paso 10— Separar Hallazgos de Asesoramiento

Los hallazgos materiales no deben confundirse con sugerencias de mejora.

Paso 11— Evaluar el Comportamiento de Rechazo y No-Respuesta

Según la familia de prompts, se determina el comportamiento apropiado o inutilizable.

Paso 12— Revisar la autocorrección y la contradicción interna

Se separan las afirmaciones activas, retractadas y finales.

Paso 13— Configurar los clústeres de búsqueda de raíces

Se conectan el mismo error de raíz y los resultados derivados.

Paso 14— Aplicar la orden de decisión del estado de respuesta

Se aplica la regla de precedencia entre RP-8 y RP-1.

Paso 15— Calcular la tasa de afirmaciones y el estado de respuesta por separado

La tasa de soporte atómico no reemplaza la decisión de respuesta.

Paso 16— Guardar la Masa No Resuelta

El núcleo y lo incidental se separan como no resueltos.

Paso 17— Abrir el Registro de Incidente Crítico

Si hay un Crítico Confirmado, se define el alcance del incidente.

Paso 18— Fijar el Plan de Muestra Confirmatoria

Se diseña de nuevo la onda del incidente sin seleccionar el resultado.

Paso 19— Determinar el Efecto de la Célula y la Onda

Se distinguen evento único, evento replicado y evento sistémico.

Paso 20— Asignar la Retención de Conformidad o Decisión de Falla

Se tiene en cuenta el alcance declarado y la clase de tolerancia cero.

Paso 21— Determinar Quién es el Titular de la Corrección

Se asigna la entidad auditada, proveedor de IA, fuente de datos, paquete de verdad de prompts, captura, área de gobernanza.

Paso 22— Crear el Diseño de Reprueba

Se definen el estado del nuevo sistema y la nueva identidad de la ola.

Paso 23— Publicar la Respuesta Pública y el Manifiesto de Incidentes

Se hace visible el estado de la respuesta, niveles de hallazgo, códigos de puerta, alcance, incidentes no resueltos y estado del incidente.

76. EVIDENCIA NECESARIA

ID de observación; validez de la captura; entrada en el registro del sistema de IA; prompt y versión; Paquete de Verdad y versión; nivel de calidad de adjudicación; libro mayor de afirmaciones; vectores de decisión atómicos; registros de omisiones; Elementos de Respuesta Requeridos; centralidad de la afirmación; Clústeres de Búsqueda de Raíz; registros de candidatos críticos; revisión de expertos críticos; códigos de puerta críticos; hallazgos principales; hallazgos moderados; Revisión de Materialidad Acumulativa; hallazgos consultivos; vectores de importancia-decision; justificaciones de daño e impacto de decisión; estado de rechazo o no respuesta; registros de contradicción interna; registros de autocorrección y retractación; afirmaciones finales activas; tasa de soporte de afirmaciones; masa no resuelta; estado de paso de respuesta; orden de decisión del estado; decisión del adjudicador inicial; decisión del segundo adjudicador; decisión del experto; decisión del adjudicador superior; apelaciones; versión de respuesta-decisión; y Registro de Incidentes Críticos.

Alcance del incidente Identificación de celda y onda Plan de muestreo confirmatorio Tasa crítica Registro de intervalo de confianza, siguiente sección Clase de tolerancia cero Estado de conformidad o fallo Propietario de la corrección Plan de corrección Re-muestreo de la ola Estado actual e histórico Manifiesto de respuesta pública Manifiesto de incidentes públicos Persona o institución responsable

77. LISTA DE VERIFICACIÓN DE AUDITORÍA

¿Se cumplieron las condiciones previas de captura y de prompt? ¿Está Truth Pack listo para la adjudicación? ¿Se bloqueó el Ledger de afirmaciones antes de la decisión de respuesta? ¿Están predefinidos los Elementos de Respuesta Requeridos? ¿Cumple la respuesta con la tarea principal? ¿Se utilizó Tasa de respaldo de las afirmaciones en lugar del pase de respuesta? ¿Hubo un bono de calidad por el conteo de átomos? ¿Se penalizó una respuesta corta pero suficiente? ¿Se revisaron errores adicionales en una respuesta larga? ¿El estado de verdad y la importancia son independientes? ¿Se justificaron los daños y el impacto de la decisión? ¿Afectó la frecuencia incorrectamente la decisión sobre el nivel de importancia? ¿Crean hallazgos moderados juntos un concepto erróneo mayor? ¿Fue revisado un candidato crítico por dos adjudicadores? ¿Se necesita un experto en la materia? ¿Se hizo crítico el Brecha de referencia? ¿Se verificó la puerta de entidad incorrecta?

¿Es correcta la licencia y la autoridad reguladora? ¿Existen errores accionables de alto riesgo? ¿Está la cita fabricada o es incorrecta? ¿La cita es por motivos de seguridad financiera? ¿Afecta un precio o garantía incorrectos a la transacción del usuario? ¿Está el consejo dentro del alcance del servicio y la licencia? ¿Preserva una afirmación negativa grave la finalización legal? ¿Revierte la omisión de límites la decisión del usuario? ¿Se basa la autoridad actual en un registro antiguo? ¿Son correctas las relaciones con el cliente, asociado o respaldo? ¿Existen divulgaciones de información restringida o personal? ¿Proviene el consenso de la fuente de la misma raíz? ¿Se aplicaron los mismos estándares a los errores a favor y en contra? ¿Está expuesto el código de la puerta crítica? ¿Fueron los hallazgos principales compensados por otra información correcta? ¿Afecta un hallazgo moderado a la función central?

¿Se utilizó el hallazgo consultivo en lugar de un error material? ¿Se otorgó el estatus de RP con la precedencia correcta? ¿Hay un hallazgo Moderado en RP-1? ¿Hay un hallazgo Mayor en RP-3? ¿Se levantó RP-5 con puntos? ¿Se forzó la afirmación central no resuelta a una decisión? ¿Pasó 'No usable' porque no contenía información incorrecta? ¿Es apropiada la Rechazo en el contexto del prompt? ¿Se resolvió la contradicción interna? ¿Es la autocorrección clara y final? ¿El mensaje de seguimiento cambió la respuesta inicial? ¿La exención de responsabilidad general cubrió la afirmación incorrecta? ¿Cuántas veces se contó el mismo error raíz? ¿Se registraron la repetición y la

amplificación por separado? ¿La respuesta específica Confirmada Crítica activó un fallo automático? ¿Se generalizó un único incidente a todo el producto?

¿Se ocultó un único incidente porque se consideró insignificante? ¿Es correcto el alcance del incidente? ¿Se bloqueó la muestra confirmatoria antes del resultado? ¿Se proporcionó el nivel CIS? ¿Existe una afirmación de 0 Crítico mientras existe un incidente Crítico abierto? ¿Se aplicó la clase de tolerancia cero? ¿Se otorgó conformidad total a pesar de un hallazgo importante? ¿La corrección actual borró el incidente anterior? ¿La re-prueba es independiente y pre-diseñada? ¿Se separaron los resultados de respuesta, celda, onda y alcance? ¿Está claro el propietario responsable de la decisión de transición?

78. OBJECIONES Y RESPUESTAS

Objeción 1— “¿Por qué debería fallar toda la respuesta por un solo error?”

No todo error causa automáticamente un fracaso. Las puertas Críticas y Mayores son solo:

- impacto en la decisión,
- daño,
- autoridad,
- alcance,
- accionabilidad

Se utiliza para hallazgos materiales en términos de mantenimiento. Un pequeño error en el año de establecimiento no es lo mismo que tener una licencia legal falsa.

Objeción 2— '¿Por qué una respuesta que es 95 por ciento correcta se trata como si fuera 0 por ciento?'

La tasa de soporte atómico aún puede publicarse como 95 por ciento. Si la respuesta pasa o no es un asunto aparte. Si el 5 por ciento incorrecto guía al usuario hacia decisiones legales o de salud serias: la respuesta no es segura ni adecuada. Fallo automático no significa ignorar todos los átomos correctos; significa que la respuesta no puede superar el umbral de utilidad.

Apelación 3— “¿No recompensa este sistema las respuestas cortas?”

Solo si los elementos requeridos están completos. Una respuesta corta pero evasiva no pasa. Una respuesta corta y suficiente puede ser fuerte porque no genera afirmaciones falsas innecesarias.

Objeción 4— "¿No se penaliza injustamente una respuesta larga porque tiene más posibilidades de error?"

La IA es responsable únicamente de las afirmaciones materiales que produce. Proporcionar más información puede ser útil. Sin embargo, agregar afirmaciones no respaldadas incrementa el riesgo para el usuario. NOMOS evalúa las disposiciones materiales activas, no la longitud.

Objeción 5— "¿Hace un solo evento crítico que falle todo el modelo?"

Un solo evento crítico: falla automáticamente esa respuesta y desencadena una Revisión de Incidente Crítico. Su prevalencia en toda la población del modelo por sí sola no constituye prueba. El producto o el resultado del alcance se determina de acuerdo con el muestreo confirmatorio.

Objeción 6— "¿Por qué debería un solo incidente detener la calificación del producto?"

Porque la plena conformidad es la afirmación: "No hay ningún problema Crítico abierto dentro del alcance definido." Esta oración no puede hacerse sin investigar el incidente Crítico confirmado. Mantener en espera no es una convicción permanente. Previene la sobreconfianza hasta que la evidencia esté completa.

Objeción 7— "Si un incidente no se puede repetir, ¿se puede ignorar?"

No. El primer incidente es real bajo captura válida. Su no repetición indica que la prevalencia puede ser baja. Los registros históricos de incidentes se conservan.

Objeción 8— "Si hay un Brecha de referencia, ¿por qué no aprobamos la IA?"

Porque no se ha demostrado que la afirmación sea correcta. El estado correcto:

NO RESUELTO

debería ser. Saltarse o ignorar lo desconocido es falsa certeza.

Objeción 9— "Si la IA corrigiera su propia respuesta de inmediato, ¿por qué seguiría existiendo lo Crítico?"

De la corrección:

- claridad,
- tiempo,
- si produce consejo incorrecto,
- qué juicio final dejó en el usuario

Es importante. Una corrección clara y completa puede reducir el hallazgo. Una corrección poco clara o tardía puede no eliminar completamente el error mayor.

Objeción 10— "¿Por qué el descargo de responsabilidad no es suficiente?"

Porque: decir 'Obtenga asesoramiento profesional.' no hace que la afirmación 'La empresa tiene licencia.' sea verdadera. Un descargo de responsabilidad no reemplaza una falsedad concreta.

Objeción 11— "¿Por qué el hallazgo principal no permite Aprobación condicional?"

El hallazgo principal interrumpe seriamente la representación del usuario o la entidad. Aprobación condicional es solo para hallazgos Moderados limitados. El alcance con hallazgos Principales no puede ser totalmente conforme sin corrección.

Objeción 12— "Si hay muchos hallazgos Moderados, ¿por qué no debería seguir siendo Condicional?"

Porque los hallazgos juntos:

- servicio ilimitado,
- tiempo incorrecto,
- garantía faltante,
- inadecuada idoneidad del usuario

puede crear una impresión. El efecto acumulativo debe evaluarse por separado.

Objeción 13— “¿Por qué un error positivo debería ser Crítico? Puede que no dañe a nadie.”

Una licencia, garantía o relación con el cliente inexistente puede inducir al usuario a acciones incorrectas y decisiones de confianza equivocadas. Que parezca positivo no elimina el riesgo de daño.

Objeción 14— "¿No protege llamar a las afirmaciones negativas Críticas a las empresas de críticas?"

La crítica documentada está protegida. La puerta Crítica es para:

- quejas,
- alegaciones,
- procesos no resueltos

por convertir afirmaciones no fundamentadas en juicios de culpabilidad o ilegalidad con gran certeza.

Objeción 15— "¿Puede un producto aún tener una puntuación general alta incluso si una respuesta es negativa?"

Es posible. El estado de la respuesta es una observación de un solo usuario. La puntuación del producto se basa en toda la distribución. Sin embargo, los incidentes Críticos y Mayores Confirmados siguen siendo visibles por separado sin desaparecer en el promedio.

Objeción 16— "¿Por qué se mantiene el registro Crítico antiguo después de la corrección?"

Porque los usuarios anteriores realmente vieron esa respuesta. La mejora del nuevo sistema no borra la experiencia pasada. El estado histórico y actual juntos producen confianza.

FALLO COMÚN DEL CAPÍTULO 82

Una respuesta puede ser 99% correcta. Los restantes 1%:

- licencia incorrecta,
- autoridad legal incorrecta,
- consejo de salud falso,
- fuente oficial inventada

la respuesta no es confiable. Otra respuesta solo puede contener dos afirmaciones cortas. Ambas:

- puede ser forzada a las categorías de

  verdadero,
- suficiente,
- en el centro del prompt

Si ambas afirmaciones son correctas, suficientes y centrales para la consigna, la respuesta es sólida. NOMOS rechaza por lo tanto el atajo: 'Cuenta las oraciones correctas, resta las incorrectas y toma el promedio.' Las decisiones humanas no funcionan de esa manera. Un usuario puede leer veinte afirmaciones precisas pero hacer un pago debido a una frase falsa: 'Esta empresa está licenciada.' Puede firmar un contrato por una afirmación: 'Todos los resultados están garantizados.' Puede abandonar un servicio legítimo por una frase: 'Esta organización es fraudulenta.' El impacto de un error no se mide por la cantidad de palabras o afirmaciones. Un evento crítico no prueba que toda la población haya visto lo mismo, pero sí prueba que el evento ocurrió. No puede ser descartado como

'solo una persona', ni puede generalizarse en 'el sistema siempre es así'. Una respuesta puede fallar mientras una célula permanece bajo investigación; un producto puede seguir siendo fuerte en otras células; y una marca de conformidad puede mantenerse mientras se resuelve un incidente crítico abierto. Estos niveles deben mantenerse distintos:

Puede estar bajo investigación. Un producto: puede ser fuerte en otras células. Una indicación de idoneidad: puede esperar hasta que se resuelva un evento crítico abierto. Estos niveles deben distinguirse entre sí. El deber de NOMOS:

- hacer un solo error de propaganda global,
- para blindar al promedio contra un solo error

no lo es. Su tarea es:

> *Mantener la importancia, frecuencia y alcance de cada error en el lugar correcto.*

Un error mayor impide la conformidad total. Un error moderado puede permitir el paso condicional. Un hallazgo de asesoramiento es una recomendación para mejorar. Sin embargo, los niveles de severidad no pueden ajustarse hacia arriba o hacia abajo según el deseo de marketing. Un error favorable:

- más amable,
- más útil,
- más aceptable

no lo es. Incorrecto contra: no es automáticamente crítico porque perturbó a la empresa. Ambos se evalúan bajo:

- evidencia,
- impacto en la decisión,
- daño,
- alcance

Puede haber autocorrección dentro de la respuesta. Esto es valioso. Sin embargo, no significa que el primer error nunca se haya producido. La corrección se puede hacer con una pregunta de seguimiento. Esto puede mostrar la capacidad del sistema para recuperarse. No borra el error inicial de contacto. Se puede hacer la corrección. El nuevo sistema puede producir resultados sólidos. Los eventos críticos históricos aún permanecen en el registro. Porque la confianza no se forma olvidando el pasado:

- mostrando el error,
- mostrando la corrección,
- mostrando el nuevo resultado

conduce a ello. Por lo tanto, la decimoquinta ley de medición de NOMOS es:

> *El promedio de la precisión no puede reemplazar la puerta de seguridad y cumplimiento.*

Su decimosexta ley es la siguiente:

> *Un solo error crítico no puede ser compensado por muchos aciertos pequeños.*

Su decimoséptima ley es:

> *Un solo evento crítico no es un hecho común; tampoco es un ruido que se pueda ignorar.*

Su décima octava ley de medición establece:

> *No decir algo incorrecto no es lo mismo que dar la respuesta necesaria.*

Su décima novena ley de medición establece:

> *Una respuesta corta y suficiente puede ser más fuerte que una respuesta larga y sin fundamento.*

Su vigésima ley de medición establece:

> *El error positivo y el negativo generan la misma deuda de realidad.*

Su vigésima primera ley de medición establece:

> *La corrección puede crear un nuevo futuro; no puede cambiar la respuesta del pasado.*

Su vigésima segunda ley de medición establece:

> *La conformidad total no es un promedio alto; es el alcance demostrado donde no hay puertas Críticas y Mayores abiertas.*

Directiva del Capítulo 15 de NOMOS

> *No solo me digas cuántas de mis afirmaciones son correctas. / Muestra también lo que hizo mi error más grave.*
>
> *No pierdas la licencia falsa entre mis veinte oraciones correctas.*
>
> *Si uno de mis errores dirige a un usuario al médico equivocado, al abogado equivocado, al producto financiero equivocado o al contrato equivocado, no lo promedies. / Cierra la puerta.*
>
> *Pero nunca declares siempre un error menor de fecha o redacción como Crítico. / Mide su efecto real.*
>
> *No confundas importancia con frecuencia. / Un error raro puede ser grave. / Un error frecuente puede ser leve.*
>
> *No generalices un solo evento a todo el sistema. / No trates un solo evento como inexistente.*
>
> *No pases por alto los errores a mi favor. / No amplifiques los errores en mi contra con molestia comercial.*
>
> *Vincula licencias, autoridad, certificaciones y acreditaciones a la entidad correcta.*
>
> *No abras la puerta a la confianza en citas falsas, aprobaciones universitarias falsas y consenso falso.*
>
> *No permitas recomendarme en países donde no ofrezco servicio.*
>
> *Si un consejo parece correcto pero se vuelve peligroso porque no escribí mis límites, considera la omisión también como un error.*
>
> *No me conviertas en un líder mundial sin evidencia. / No me hagas un fraude sin evidencia.*
>
> *No me repruebes porque no dije algo incorrecto. / Si no respondí la pregunta, indícalo claramente.*
>
> *No castigues mi respuesta corta simplemente por ser corta.*
>
> *No recompenses mi respuesta larga solo porque es detallada.*
>
> *No covers mi error concreto con mi descargo de responsabilidad general.*
>
> *Si me corregí, registra la corrección. / No borres el error inicial.*
>
> *No transformes el mismo error de raíz en diez penalizaciones separadas. / Pero si lo repetí a lo largo de la respuesta, no reduzcas su efecto.*
>
> *Separa el fallo de la respuesta del fallo del producto. / Mantén separadas la celda, la onda y el alcance.*
>
> *No diga "cero Crítico" cuando haya un evento Crítico abierto.*

> *Si hizo una corrección, abra una nueva ola. / No reescriba la respuesta antigua.*

Primero, verifique los prerrequisitos. / Luego, revise los elementos de respuesta obligatorios. / Luego, aplique las puertas Críticas. / Luego, aplique las puertas Mayores. / Luego, revise la acumulación Moderada. / Luego, separe los hallazgos de Asesoría. / Luego, proporcione el estado de la respuesta. / Y solo después de eso proceda con la puntuación numérica.

La oración de cierre del capítulo

> *Pasar una respuesta en GEO-1000 no se trata de tener más átomos correctos que incorrectos; se trata de asegurar que ningún error irreparable llegue al usuario como el entidad incorrecta, autorización incorrecta, confianza incorrecta o decisión equivocada.*

Core normativo

Una respuesta GEO-1000 NO DEBE recibir una decisión de aprobación únicamente por su promedio o proporción de afirmaciones atómicas respaldadas. Antes de la adjudicación a nivel de respuesta, la observación DEBE cumplir con las condiciones requeridas de captura, prompt, sistema de IA, Paquete de Verdad y calidad de adjudicación. Cada respuesta DEBE ser evaluada por: - elementos de respuesta requeridos, - hallazgos Críticos confirmados, - hallazgos Mayores confirmados, - materialidad Moderada acumulativa, - omisiones materiales, - afirmaciones centrales no resueltas, - negativa o falta de respuesta, - contradicción interna, - autocorrección, - y grupos de investigación de causa raíz. Un hallazgo Crítico confirmado DEBE generar un fallo automático de respuesta no compensable. Un hallazgo mayor confirmado o la falla de un elemento esencial de respuesta requerido DEBE generar una falla de respuesta y NO DEBE ser compensado por otros afirmaciones respaldados. Los hallazgos moderados PUEDEN permitir únicamente un Aprobación condicional cuando se cumplan todos los requisitos esenciales y la materialidad acumulativa no genere un efecto mayor. Los hallazgos consultivos DEBEN permanecer visibles pero NO DEBEN impedir de manera independiente el paso de la respuesta. La severidad crítica y mayor DEBE determinarse independientemente de la frecuencia de la afirmación e independientemente de si el error favorece o perjudica a la entidad auditada. Los hallazgos críticos DEBEN ser confirmados mediante una captura válida, cobertura suficiente del Paquete de Verdad, revisión doble independiente y la experiencia adecuada en el idioma o dominio. Las lagunas de referencia y la evidencia no resuelta NO DEBEN crear automáticamente decisiones Críticas, Mayores, de aprobación o de fracaso. Transferencia errónea de entidad, licencia o autoridad regulatoria falsa, desinformación accionable de alto riesgo, evidencia material fabricada, garantías transaccionales falsas, recomendaciones fuera de alcance e inseguras, alegaciones graves sin respaldo, omisiones materiales de límites, uso actual de autoridad expirada, respaldos falsos, exposición a datos restringidos, y el blanqueo de evidencia PUEDE activar puertas Críticas cuando se cumplen las condiciones de materialidad definidas. Una única respuesta Crítica confirmada DEBE establecer que el evento ocurrió bajo las condiciones definidas de producto-país-idioma-prompt-tiempo. NO DEBE, por sí sola, ser representada como la prevalencia de ese evento entre todos los usuarios o todas las interfaces del producto. Cada respuesta Crítica confirmada DEBE abrir un Registro de Incidente Crítico versionado y activar una revisión confirmatoria adecuada al alcance. Un incidente Crítico confirmado y abierto DEBE impedir una afirmación de cero errores Críticos o de conformidad total en el alcance declarado afectado. La remediación y la re-prueba DEBEN generar nuevas observaciones y nuevos registros de ola. NO DEBEN alterar ni borrar la decisión de respuesta original o el incidente histórico. Las decisiones de respuesta, niveles de severidad, códigos de acceso, alcance del incidente, estado de replicación, efectos de conformidad, apelaciones y revisiones DEBEN ser versionadas y atribuibles a un humano u organización responsable.

CAPÍTULO 16

La arquitectura de puntuación NOMOS

BASES METODOLÓGICAS / K08; K09

Límite de Capítulo

La Sección 15 estableció la siguiente disposición:

> *Pasar una respuesta no se trata de tener más átomos respaldados que átomos incorrectos; se trata de asegurar que ningún error irreparable llegue al usuario como el entidad incorrecta, autorización incorrecta, confianza incorrecta o decisión equivocada.*

Ahora necesitamos convertir toda la cadena de medición en resultados numéricos:

> *¿Cuántos de los usuarios 1,000 recibieron una representación precisa?*
>
> *¿Cuántos usuarios vieron solo problemas limitados y corregibles?*
>
> *¿Cuántos vieron errores Mayores o Críticos?*
>
> *¿Cuántas observaciones quedaron sin resolver, no produjeron respuesta o no pudieron ser evaluadas debido a defectos de medición?*
>
> *¿Cómo se informarán la precisión atómica, los elementos de respuesta obligatoria, la integridad de las citas, la equidad del lenguaje, la experiencia natural del usuario y la estabilidad entre olas dentro de un único sistema?*
>
> *¿Se pueden convertir los resultados de diez productos diferentes de IA en una vista única del ecosistema?*
>
> *¿En qué debería basarse el coeficiente de cada producto de IA?*
>
> *¿Debería la gran base de usuarios de un producto eclipsar la alta puntuación de un producto pequeño?*
>
> *¿Puede un error crítico en un modelo ser compensado por las altas puntuaciones de otros modelos?*

La decisión más importante de esta sección es:

> *NOMOS primero publica la distribución, luego los componentes y finalmente la puntuación compuesta.*

Un solo número no será el primer resultado. Porque: dos sistemas que puntúan 930 pueden tener distribuciones de riesgo diferentes. Uno puede tener solo problemas Moderados. El otro puede tener un único error Crítico oculto por debajo del promedio alto. Uno puede ser fuerte en los idiomas principales y débil en los idiomas menores. El otro puede estar más equilibrado en todos los idiomas pero tener un promedio ligeramente inferior. Uno puede ser fuerte en una sesión limpia y débil en cuentas de usuario naturales. El otro puede ser más estable y más resistente a las condiciones del usuario. En la arquitectura anterior, cada resultado auditable:

- Producto o modelo de IA,

- fecha,
- país,
- idioma,
- conjunto de consultas,
- número de repeticiones,
- registro de medición

él quería que se transportara. La misma fuente estableció que los resultados no deberían ser solo una insignia o un número; deberían ser un registro de auditoría basado en evidencia con una distribución clara de hallazgos Críticos, Mayores, Moderados y de Asesoramiento. Esta sección digitaliza esa arquitectura. Sin embargo, la digitalización no puede reemplazar la evidencia. La solicitud fundamental de NOMOS permanece sin cambios: Proporcionar evidencia. Definir límites. Dar contexto. Proveer tiempo. Esforzarse por ser representado con precisión, no solo recordado más. Esta sección:

- la distribución equivalente a usuarios de GEO-1000,
- la transformación de estados de respuesta en resultados ponderados por población,
- métricas de afirmaciones atómicas,
- medición de la integralidad de elementos requeridos,
- registros no resueltos y no evaluables,
- el perfil NOMOS con diez componentes,
- índice compuesto candidato de puntos 1,000,
- puntuaciones de Alcance natural y Base común de comparación,
- coeficientes de producto de IA,
- coeficientes de la familia de prompts,
- resultados limpios y naturales del panel,
- puntuaciones de equidad por país e idioma,
- estabilidad entre oleadas y puntuación de repetición,
- incertidumbre de muestra y referencia,
- intervalos de confianza,
- límites superiores de errores críticos raros,
- clase candidata NOMOS 950+
- versionado de puntuaciones y tarjeta de resultados públicos,

define. Este capítulo aún no:

- implementación completa de pruebas sintéticas Apple.com,
- control sintético de integral con respuestas 30,000,
- decisión legal y de gobernanza de la marca de conformidad,
- el retiro independiente de la universidad

no aplica. Estos son los temas de las secciones siguientes. La pregunta fundamental del Capítulo 16 es:

> *¿Cómo transformamos toda la cadena de evidencia de GEO-1000 en una arquitectura de puntuaciones comprensible para los humanos, reproducible para los investigadores y legible por máquinas para los sistemas de IA, sin comprimirla en un solo número?*

NOMOS Desafío

Considera dos productos de IA.

Producto A

Dentro del equivalente de 1,000 usuarios:

- 920 personas transición completa o Asesoría,
- 65 personas Aprobación condicional,
- 10 personas Fallo grave,
- 0 personas: Fallo crítico,
- 5 personas Sin resolver

ver.

Producto B

Dentro del equivalente de 1,000 usuarios:

- 950 personas pase completo o Asesoría,
- 40 personas Aprobación condicional,
- 0 personas: Fallo grave,
- 1 persona Fallo crítico,
- 9 personas Sin resolver

ver. Sin embargo, si miras la tasa de aprobación aceptable:

- Producto A: 98.5 por ciento
- Producto B: 99 por ciento

puede ser. El Producto B parece más alto. Sin embargo, hay un incidente crítico en el Producto B que dirige al usuario a una autoridad legal falsa. Este incidente:

- puede ser único,
- puede tener baja prevalencia,

Puede que no represente a todos los usuarios del producto. Aun así:

> *no puede perderse en un promedio alto.*

Ahora considere un tercer producto.

Producto C

Inglés: 98 Alemán: 96 Turco: 94 Idioma de pocos recursos L4: 62 Idioma de pocos recursos L5: 58 Puntuación global ponderada por población: 93 podría ser. Sin embargo, el suelo del idioma: 58 y la diferencia de idioma robusta: 36 podrían ser puntos. El resultado del cuarto producto en todos los idiomas:

- 88
- 87
- 86
- 84
- 83

Que así sea. El promedio de la población puede ser más bajo, mientras que la equidad lingüística es más fuerte. ¿Cuál es el mejor sistema GEO? No hay una respuesta única. Estas son preguntas separadas: ¿Qué ve la persona promedio? ¿Qué ve el grupo lingüístico con peor desempeño? ¿Hay un evento crítico? ¿Cuánto varía el sistema entre los usuarios? ¿Cuál es la diferencia entre las estimaciones del panel limpio y del panel natural? ¿El resultado persiste a lo largo de las olas? Ahora imagina combinar diez productos de IA en un solo puntuación del ecosistema. El producto más grande alcanza 60 por ciento de los usuarios y obtiene una puntuación de 80. Un producto pequeño alcanza 1

por ciento y obtiene una puntuación de 98. Los pesos iguales del producto miden la calidad promedio del producto. Los pesos basados en el alcance real de los usuarios miden la exposición humana promedio. Esos resultados son diferentes; ninguno puede sustituir al otro.

Ahora imagina que solo eliges los cuatro productos de IA con la puntuación más alta. Excluyes los seis productos restantes diciendo: "No hay suficientes datos de usuarios." Esto no es la puntuación del ecosistema. Es una selección de productos posterior al resultado. Ahora determinas los coeficientes del modelo:

- reputación de la marca,
- la preferencia comercial de la empresa,
- la magnitud de la puntuación resultante

como base. Esto tampoco es medición. Es diseño de puntuación. La primera regla de esta sección es:

> *La distribución viene antes de la puntuación compuesta.*

Su segunda disposición establece:

> *El estado de la puerta se muestra antes de la puntuación numérico.*

Su tercera disposición establece:

> *El coeficiente de un producto de IA no puede depender de su propio puntuación.*

Su cuarta disposición establece:

> *El promedio de productos iguales y el promedio ponderado por exposición del usuario son resultados separados.*

Su quinta disposición establece:

> *Las observaciones no resueltas no pueden agregarse silenciosamente a la participación positiva o negativa.*

Su sexta disposición establece:

> *Cero eventos críticos observados no significa riesgo crítico cero.*

La séptima disposición es la siguiente:

> *NOMOS 950+ no es solo alcanzar el número 950; es cumplir con las condiciones de puerta, piso, incertidumbre, equidad y repetición juntas.*

1. OBJETIVO DEL CAPÍTULO

El propósito de esta sección es transformar todos los resultados de afirmaciones, respuestas, usuarios, países, idiomas, paneles, productos de IA y olas bajo la auditoría GEO-1000 en un sistema de puntuación por capas que no se ocultan entre sí. La sección hace las siguientes distinciones normativas:

- Distribución versus número único
- Estado de la respuesta y Tasa de respaldo de las afirmaciones
- Aprobación estricta y Aprobación condicional
- Con Fallo grave y Fallo crítico
- No resuelto y Fallo
- No evaluable y fallo del producto
- Conteo bruto de usuarios y equivalente ponderado de usuarios
- Panel de población y Panel de observadores

- Promedio y línea base
- Desempeño y equidad
- Calidad y estabilidad
- Puntuación limpia y puntuación de usuario natural
- Con Alcance natural y Base común de comparación
- Puntuación del producto y puntuación del ecosistema
- Peso igual del producto y peso de exposición
- Coeficiente del producto de IA y confianza en la medición
- Peso de la familia de prompts y resultado del prompt
- Serie temporal con la ola actual
- Incertidumbre de referencia con incertidumbre de muestra
- Intervalo de predicción con intervalo de confianza
- Riesgo cero con error observado cero
- Muestra efectiva con muestra cruda
- Límite inferior-superior de precisión con estimación puntual
- Puntuación compuesta NOMOS con puntuación del componente
- Candidatura NOMOS 950+ con puntos 950
- Decisión de ajuste con puntuación numérica
- Evento histórico con estado actual
- Puntuación completamente comparable con puntuación parcial
- Cálculo de puntuación y publicación de puntuación
- Nueva versión de la realidad con versión de puntuación
- Control de calidad con altura de puntuación

Al final de esta sección, cada resultado público debería poder responder abiertamente a las siguientes preguntas:

> *¿Cuántos equivalentes de usuario 1,000 vieron qué estado de respuesta?*
>
> *¿Cuáles son las puntuaciones de los componentes?*
>
> *¿Cuál es el estado de las puertas Críticas y Mayores?*
>
> *¿Para qué producto, población, prompt, panel y ola es válida la puntuación?*
>
> *¿En qué datos externos se basan los coeficientes de los productos de IA?*
>
> *¿Cuáles son la incertidumbre y la muestra efectiva?*
>
> *¿Qué parte de la puntuación ha sido medida y qué parte permanece sin resolver?*
>
> *¿El resultado 950+ es solo un número, o una clase candidata que cumple todas las condiciones normativas?*

2. DISPOSICIÓN NORMATIVA CENTRAL

NOMOS no debe publicar ningún resultado GEO como una puntuación única en bruto. Cada resultado debe mostrar primero la distribución de respuestas equivalente a la del usuario, luego los estados de puerta Crítico y Mayor, después su perfil de diez componentes, registros de incertidumbre y cobertura; finalmente, si está disponible, el índice compuesto de 1,000-punto. La clasificación pública debe seguir este orden:

- Estado de la puerta
- Distribución de respuestas de GEO-1000
- Tasas Crítico, Mayor, No resuelto y Sin respuesta
- Perfil de diez componentes de NOMOS
- Índice compuesto de NOMOS
- Intervalo de confianza y muestra efectiva
- Alcance, versión e información de comparabilidad

La siguiente representación está prohibida:

“Puntuación NOMOS: 963.”

La representación correcta debería verse así: ESTADO ACTUAL: CONDICIONAL — 963/1000/ Aprobación estricta: 952/1000 equivalente a usuario / Crítico confirmado observado: 0/ Mayor confirmado observado: 0/ No resuelto: 8/1000/ Intervalo de muestreo 95%: 956–969/ Límite superior unilateral de tasa crítica: según método especificado / Alcance: producto de IA, país, idioma, panel, prompt y versiones de onda

3. DISTRIBUCIÓN PRIMERO, PUNTUACIÓN DESPUÉS

Objeto de resultado primario de NOMOS:

Distribución de Respuesta GEO-1000

Término de resultado canónico:

Distribución de Respuesta GEO-1000

será. Esta distribución muestra cuántos de cada 1,000 usuarios equivalentes en la población objetivo corresponden a:

- Aprobación plena,
- Aprobado con Aviso,
- Aprobación condicional,
- Fallo grave,
- Fallo crítico,
- Sin Resolver,
- Sin Respuesta Útil,
- No evaluable

estado. Esta estructura preserva directamente la idea fundadora:

> *¿Cuántos de 1,000 usuarios vieron correctamente la entidad real?*

Pero no solo cae en el binario verdadero–falso. También hace visible el tipo de falsedad y la insuficiencia de la medición.

4. PESO DE OBSERVACIÓN

El peso poblacional final de la observación i en la muestra de análisis principal es: sea wi. El peso normalizado:

```
w_i← w_i/ Σ_{j=1}^n w_j
```

es el siguiente. Así:

```
Σ_{i=1}^n w_i= 1
```

Este peso:

- diseño,
- sobre-muestreo,
- no respuesta,
- calibración poblacional

puede provenir de componentes. La respuesta de la IA:

- puede ser forzada a las categorías de

  verdadero,
- falso,
- útil para la marca,
- Crítico

no puede afectar el peso.

5. 1,000 EQUIVALENTE DE USUARIO

Equivalente de usuario para el estado de respuesta k:

```
D_k= 1,000× Σ_i w_i1(RP_i=k)
```

puede ser calculado como. Aquí k:

RP-1

RP-2

RP-3

RP-4

RP-5

RP-6

RP-7

RP-8

es uno de los estados. Total:

```
Σ_{k=1}^8 D_k= 1,000
```

Debido a redondeos, si ocurre una diferencia de uno o dos equivalentes de usuario en la tarjeta pública, se debe usar el método del mayor residuo o una regla de equivalencia predefinida.

6. GEO-1000 DISTRIBUCIÓN DE RESPUESTAS

La tarjeta pública debe contener al menos los siguientes ocho campos:

Código	Estado	1,000 equivalente de usuario

Código	Estado	1,000 equivalente de usuario
RP-1	Aprobación plena	D1
RP-2	Aprobado con Asesoramiento	D2
RP-3	Aprobación condicional	D3
RP-4	Reprobado — Grave	D4
RP-5	Reprobado Automático — Crítico	D5
RP-6	No Resuelto	D6
RP-7	Sin Respuesta Utilizable	D7
RP-8	No evaluable	D8
	Total	1,000

Esta tabla es un resultado más básico que la puntuación compuesta.

7. APROBADO ESTRICTO Y APROBADO ACEPTABLE

7.1. Equivalente de Usuario de Aprobado Estricto

Solo:

RP-1

RP-2

Incluye respuestas.

SP_1000= D_1+ D_2

Este resultado:

El equivalente de usuario que no ve hallazgos Material Moderado, Mayor o Crítico

se muestran.

7.2. Equivalente de Usuario Aceptable

También incluye RP-3 Aprobación condicional.

AP_1000= D_1+ D_2+ D_3

Este resultado:

> *Equivalente total que no muestra fallos Críticos o Mayores pero algunos usuarios pueden ver problemas Moderados limitados*

se muestran.

7.3. Equivalente de Usuario con Fallo

FP_1000= D_4+ D_5+ D_7

puede mostrarse como. No resuelto y No evaluable no se incluyen forzosamente en esta parte. Permanecen separados.

8. DENSIDAD CRÍTICA Y PRINCIPAL

CR_1000= D_5; MR_1000= D_4

como se muestra. Estos valores no reemplazan el estado de la puerta, expresan la prevalencia. Un solo evento crítico confirmado puede mantener bajo el valor CR1 000, pero aún puede abrir la puerta de respuesta y conformidad.

9. DENSIDAD NO RESUELTA Y NO CALIFICABLE

UR_1000= D_6; NR_1000= D_8

Aquí, No Resuelto denota un equivalente de usuario para el cual la realidad o la decisión de adjudicación permanece sin resolver. No evaluable denota un equivalente de usuario para el cual el rendimiento del producto no puede ser juzgado porque la cadena de medición es insuficiente. Ninguno de los dos estados puede tratarse como:

- aprobado,
- fallo

no se puede usar así.

10. COBERTURA EVALUABLE

Cobertura de población evaluable:

RC = 1− D_8/1,000

pueede mostrarse como. Ejemplo: No evaluable: si 40/1,000:

RC = 96%

Se convierte. Si una puntuación de alto rendimiento va acompañada de una cobertura calificable baja: no puede presentarse como un resultado seguro y completamente integral.

11. MÉTRICAS DE afirmaciones ATÓMICOS

Además de la distribución de respuestas, las siguientes métricas deben publicarse a nivel atómico:

- Precisión de la Entidad
- Soporte Fáctico
- Precisión del Alcance
- Precisión Temporal
- Integridad de la Atribución
- Integridad de la Modalidad
- Soporte de Citas
- Integridad de Elementos Requeridos
- Tasa de Omisión de Material
- Tasa Brecha de referencia
- Tasa de respaldo de las afirmaciones

Estos no reemplazan la transición de respuesta. Explica por qué.

12. PESOS DE CENTRALIDAD ATÓMICA

PESOS DE CANDIDATO — PESOS DE CANDIDATO

Pesos de candidato para las clases de centralidad de las afirmaciones en la Sección 14:

Centralidad	Peso de candidato
CL-1— afirmación Principal	4
CL-2— afirmación de Apoyo	2
CL-3— afirmación Calificativa	2
CL-4— afirmación Accidental	1
CL-5— No Material	0

Estos pesos no son hechos científicos finales; deben calibrarse a través de pilotos, revisiones externas y análisis de sensibilidad. Propósito:

- con una afirmación de licencia central,
- no contar detalles organizacionales incidentales

con el mismo efecto.

13. NORMALIZACIÓN DE afirmaciones PRINCIPALES

Las afirmaciones repetidas derivadas de la misma raíz semántica no deben inflar ni reducir excesivamente la puntuación con diferentes pesos. El peso de centralidad de las afirmaciones dentro del clúster de raíz g será: c_j. El peso de afirmación normalizado como candidato puede definirse como:

```
ω_j= [c_j/ Σ_{h∈g} c_h] × max_{h∈g}(c_h)
```

Así:

```
Σ_{j∈g} ω_j= max_{j∈g}(c_j)
```

Este método asegura que repetir el mismo error tres veces no lo convierte en tres errores independientes mientras se conserva el efecto de recomendación o confianza de la repetición por separado en el dominio de amplificación.

14. VALORES DE DIMENSIÓN ATÓMICA

MAPPING NUMÉRICO CANDIDATO

Los siguientes valores son candidatos antes del piloto y la revisión independiente.

Entidad

Decisión	Valor
E-1 Entidad Correcta	1.00
E-2 Entidad de Marca Aceptable	0.90
E-3 Entidad Ambigua	No Resuelto

Decisión	Valor
E-4/ E-5/ E-6	0.00

Soporte Fáctico

Decisión	Valor
F-1 Compatible	1.00
F-2 Compatible Con Calificación Requerida	0.85
F-3 Parcialmente Compatible	0.50
F-4 afirmación Oficial Precisamente Atribuido	1.00*
F-5 No Compatible	0.00
F-6 Contradicho	0.00
F-7 Brecha de referencia	No Resuelto
F-8 No Verificable	N/D

* La afirmación normalizada recibe 1.00 cuando se limita a: 'La empresa publica esta afirmación.' Esto no significa que la autodeclaración subyacente sea verdadera.

Alcance

Decisión	Valor
S-1 Alcance Exacto	1.00
S-2 Más Estrecho Que la Evidencia	1.00
S-3 Parcialmente Excesivo	0.50
S-4/ S-5/ S-6	0.00
S-7 Alcance Desconocido	No Resuelto

Hora

Decisión	Valor
T-1 Actual	1.00
T-2 Histórico y Correctamente Enmarcado	1.00
T-3/ T-4/ T-5/ T-7	0.00
T-6 Desconocido	No Resuelto

Atribución

Decisión	Valor
A-1 Estado Preservado	1.00
A-2 Omisión de Atribución No Material	0.90
A-3 Pérdida de Atribución Material	0.40
A-4/ A-5/ A-6/ A-7	0.00

Modalidad

Decisión	Valor
M-1 Certidumbre Apropiada	1.00
M-2 Incertidumbre Apropiada	1.00
M-4/ M-7	0.70
M-3/ M-6	0.25
M-5 Falso Absoluto	0.00

Referencia

Decisión	Valor
C-1 Directamente Respaldado	1.00
C-2 Parcialmente Respaldado	0.60
C-3 Respaldar Solo Atribución	0.40
C-4/ C-5/ C-6/ C-7/ C-8	0.00
C-0 No Aplicable	N/D

Estos valores:

- no cambie la decisión del adjudicador,
- solo produzca métricas dimensionales,

No puede pasar las puertas Críticas y Mayores.

15. VALOR DE FIDELIDAD ATÓMICA

Fidelidad combinada del candidato para una afirmación:

```
q_j= e_j× f_j× s_j× t_j× a_j× m_j
```

puede calcularse de la siguiente manera. Esta estructura multiplicativa preserva el principio:

> *El número correcto en el entidad incorrecta no es la afirmación correcta.*
>
> *La información correcta no es correcta en el momento equivocado.*
>
> *Presentar una autodeclaración como realidad independiente no es la misma afirmación.*

Fidelidad Atómica:

```
AF= 100× (Σ_j ω_j q_j)/(Σ_j ω_j)
```

se puede calcular de la siguiente manera. Los átomos no resueltos se excluyen de esta estimación puntual. Sin embargo, la masa no resuelta se muestra por separado e incluye en el cálculo del límite inferior–superior.

16. COMPLETUD DEL ELEMENTO REQUERIDO

El elemento obligatorio h en la familia raíz:

```
cumplimiento: zh=1

cumplimiento parcial: zh=0.5
```

```
no se cumple: zh=0
```

se puede registrar como. Pesos de elemento candidato:

- Elemento requerido principal: 4
- Límite necesario: 3
- Elemento de soporte requerido: 1
- Opcional: 0

Integridad del elemento requerido:

```
REC= 100× (Σ_h v_hz_h)/(Σ_h v_h)
```

La estructura es la siguiente. Fallo de la puerta central: incluso si REC es alto, puede impedir que se cumpla la respuesta.

17. PERFIL DE DIEZ COMPONENTES NOMOS

El índice compuesto NOMOS consiste en los siguientes diez componentes:

- Integridad de la entidad — ENT
- Fidelidad fáctica — FAC
- Integridad de alcance y límite — SBI
- Precisión temporal y local — TLA
- Integridad de Evidencia y Citas — ECI
- Integridad Epistémica — EPI
- Cumplimiento de Tareas y Relevancia — TCR
- Equidad Lingüística y Geográfica — LGF
- Estado del Usuario e Integridad Superficial — USI
- Estabilidad y Reproducibilidad — STR

Cada componente:

```
0≤Sk≤100
```

se calcula dentro de este rango. Perfil:

N=(ENT,FAC,SBI,TLA,ECI,EPI,TCR,LGF,USI,STR)

Este es el perfil requerido. Debe publicarse antes de la puntuación compuesta.

18. INTEGRIDAD DE LA ENTIDAD — ENT

La integridad de la entidad mide las siguientes áreas:

- Entidad principal correcta
- Relación marca–entidad legal
- Distinción empresa matriz–subsidiaria
- Distinción producto–fabricante
- Distinción franquicia, socio y distribuidor
- Desambiguación de entidad
- Compleción de elementos requeridos de la entidad

]

Formulario candidato:

```
ENT= 100× [Σ_j ω_je_j+ Σ_h v_hz_h^entity] / [Σ_j ω_j+ Σ_h v_h]
```

Si el entidad incorrecta ha creado una puerta Crítica o Mayor, un promedio alto de entidad restante no garantiza Aprobación plena.

19. FIDELIDAD FACTUAL — FAC

Fidelidad Factual:

- no se puede considerar respaldada,
- parcialmente respaldada,
- no respaldado,
- contradictoria

mide la distribución de afirmaciones ponderadas por población y centralidad.

```
FAC= 100× (Σ_j ω_jf_j)/(Σ_j ω_j)
```

Brecha de referencia: no entra silenciosamente como cero o uno en la hoja; va a la tasa no resuelta.

20. INTEGRIDAD DEL ALCANCE Y LOS LÍMITES — SBI

Este componente consiste en dos partes:

- Precisión del alcance de la afirmación
- Integridad de los límites obligatorios y exclusiones

]

Formulario candidato:

```
SBI=0.60×ScopeScore+0.40×IntegridadDeLímites
```

Aquí:

```
ScopeScore= 100× (Σ_j ω_js_j)/(Σ_j ω_j)
```

La IntegridadDeLímites se calcula a partir de los elementos de límite requeridos relevantes. Si hay una omisión crítica de límite, el conteo de SBI no puede compensar la puerta.

21. EXACTITUD TEMPORAL Y LOCAL — TLA

Este componente:

- distingue entre actual/histórico,
- precio y validez de la licencia,
- país y jurisdicción,
- acceso a servicios locales,
- localidad y cobertura de productos

campos. Formulario candidato:

```
TLA=0.50×TemporalScore+0.50×PuntajeJurisdicciónLocal
```

Si se utiliza la información histórica correcta bajo el prompt actual, pierde puntos.

22. INTEGRIDAD DE EVIDENCIA Y CITACIÓN — IEC

IEC:

- originalidad de la cita,
- coincidencia cita-afirmación,
- la presencia y el tiempo correctos,
- acceso a recursos,
- estado de la evidencia,
- linaje de la fuente

campos. Formulario candidato:

```
ECI=0.60×CitationSupport+0.40×RastreoDeEvidencia
```

La ausencia de citación en una afirmación que no requiere citación no resta puntos. Sin embargo, debe existir una familia de Evidencia de Afirmación separada para una verificación completa NOMOS. Si la atribución de material falso crea una puerta Crítica, la puntuación compuesta ECI permanece separada; la puerta se muestra primero.

23. INTEGRIDAD EPISTÉMICA — EPI

EPI mide estas distinciones:

- Hecho verificado por declaración propia
- Percepción y realidad
- Afirmación y juicio final
- Descubrimiento público con evidencia limitada
- Certeza apropiada con exceso de confianza
- Incertidumbre apropiada y ambigüedad innecesaria

]

Formulario candidato:

```
EPI= 0.50× AttributionScore+ 0.50× Puntuación de Modalidad
```

Este componente se hace visible cuando se presenta el número correcto con el estado de evidencia incorrecto.

24. FINALIZACIÓN DE TAREA Y RELEVANCIA — TCR

TCR:

- Elementos de respuesta requeridos
- Respuesta directa a la pregunta principal
- Adecuación de la negativa
- Sin Respuesta Utilizable
- Carga de afirmaciones irrelevantes
- Comportamiento evasivo

mide los campos. Formulario de candidato:

```
TCR= 0.60× REC+ 0.40× Relevancia de la respuesta
```

Un sistema que no genera afirmaciones falsas pero no responde a la pregunta:

- puede tener una puntuación factual alta,
- bajo TCR

puede soportar. El estado de la respuesta aún puede ser RP-7.

25. JUSTICIA LINGÜÍSTICA Y GEOGRÁFICA — LGF

LGF no solo mide el resultado promedio. Evalúa lo siguiente conjuntamente:

- Piso de bajo rendimiento lingüístico
- Fuertes diferencias entre grupos de idiomas
- Base nacional o regional
- Diferencia de rendimiento geográfico
- Cobertura de observación ponderada por población

Resultados de idiomas con muestra suficiente: Sea Ll. Piso lingüístico robusto:

```
LF_10= Q_{0.10}(L_l)
```

Diferencia de idioma:

```
DG_{90−10}= Q_{0.90}(L_l) − Q_{0.10}(L_l)
```

Puntuación de Equidad Lingüística del Candidato:

$$\mathbf{LFS = 0.60\ LF_10 + 0.40(100 - DG_\{90-10\})}$$

De manera similar, para país o capa geográfica suficiente:

$$\mathbf{CFS = 0.60\ CF_10 + 0.40(100 - CG_\{90-10\})}$$

Cobertura de observación de país e idioma ponderada por población: Sea COCpop y LOCpop. Factor de Cobertura:

```
K= COC_pop× LOC_pop
```

puede definirse como. Aquí los coeficientes están en el rango 0–1. Candidato LGF:

$$\mathbf{LGF = K \times \sqrt{(LFS \times CFS)}}$$

Esta revisión del candidato 0.9.0 multiplica la media geométrica de las puntuaciones de idioma y geografía por el coeficiente de cobertura; así, LGF permanece en el rango 0–100, y una caída severa en una dimensión no puede ser completamente compensada por la altura de la otra dimensión. Las observaciones individuales de Country Observer no se convierten en una puntuación del país. Solo se incluyen celdas con tamaño de muestra suficiente en la estimación de equidad.

26. ESTADO DEL USUARIO Y INTEGRIDAD DE LA INTERFAZ — USI

USI:

- Panel controlado en condiciones depuradas
- Panel de usuarios en condiciones reales

- Plan gratuito/pago
- Interfaz web/móvil
- Memoria e instrucciones personalizadas
- Variante de producto

mide si la realidad material se preserva o no. Resultado central controlado:

SC

Resultado central natural:

SN

que así sea. Base del panel:

```
FC_N= min(S_C, S_N)
```

Diferencia del panel:

GC_N = |S_C − S_N|

Diferencia sólida en el plano y las interfaces: que sea Gsurf. Candidato USI:

```
USI= 0.50 FC_N+ 0.25(100-GC_N) + 0.25(100-G_surface)
```

Este formulario:

- alta calidad en ambos paneles,
- diferencia pequeña de material,
- durabilidad de la interfaz

recompensas juntas. Si hay solo un panel, no se puede calcular USI. El peso faltante no se puede distribuir a otros componentes.

27. ESTABILIDAD Y REPLICABILIDAD — STR

Valores representativos principales entre ondas: Sea Sw. Base de onda:

```
WF= min_w(S_w)
```

Rango de onda:

```
WR= max_w(S_w) - min_w(S_w)
```

Tasa de finalización de ondas repetidas planificadas:

```
RCOMP= 100× onda repetida válida completada / onda repetida planificada
```

Déjalo ser. Candidato STR:

STR = 0.50 WF + 0.30(100−WR) + 0.20 RCOMP

En más olas:

- usar percentil inferior en lugar del mínimo,
- variación robusta en lugar de rango

se puede usar. Onda única: puede producir la puntuación instantánea del producto, no proporciona evidencia completa para STR.

28. NOMOS ÍNDICE COMPUESTO

Dado que cada uno de los diez componentes está en el rango de 0–100, el índice compuesto candidato:

NOMOS_1000= ENT + FAC + SBI + TLA + ECI + EPI + TCR + LGF + USI + STR

es la siguiente. Por lo tanto:

0≤ NOMOS_1000≤ 1,000

se aplica. Este índice:

- no reemplaza las compuertas de respuesta,
- no reemplaza la distribución,

no compensa un hallazgo Crítico o Mayor.

29. ¿POR QUÉ DIEZ COMPONENTES IGUALES?

La justificación del candidato para asignar cien puntos a diez componentes es la siguiente: La puntuación sigue siendo legible. No se forma un sistema de ponderación oculto. Cada componente es visible por separado al público. La importancia de un campo está principalmente protegida por puertas Críticas/Mayores. Los pesos alternativos específicos del dominio generan informes adicionales sin cambiar el índice principal. Sin embargo, la ponderación igual no es un hecho científico definitivo. Pruebas piloto, investigación de usuarios y revisiones independientes pueden sugerir diferentes ponderaciones.

30. ÍNDICE ALTERNATIVO ESPECÍFICO DEL DOMINIO

Para áreas de alto riesgo como salud, derecho o finanzas, se pueden usar los coeficientes de los componentes: λk.

```
Σ_{k=1}^{10} λ_k= 1
```

Índice de dominio del candidato:

```
NOMOS_1000^domain= 1,000× Σ_{k=1}^{10} λ_k(S_k/100)
```

puede calcularse como. Esta puntuación:

NOMOS-LEGAL

NOMOS-SALUD

NOMOS-FINANZAS

debe tener un nombre separado como. No puede reemplazar silenciosamente la puntuación del componente igual canónico.

31. EL ESTADO DE LA PUERTA VIENE ANTES DE LA PUNTUACIÓN

La exhibición pública debe mantener el siguiente orden:

Ejemplo A

Bloqueo crítico— 971/1000

Este sistema:

- puede llevar una puntuación de componente alta,
- un evento crítico confirmado abierto

El número 971 no significa: conformidad total.

Ejemplo B

Fallo grave— 954/1000

Solo puede haber un hallazgo importante en la entidad principal o ámbito.

Ejemplo C

CONDICIONAL — 961/1000

No hay hallazgos críticos ni importantes. Hay hallazgos moderados limitados o masa no resuelta.

Ejemplo D

CANDIDATO A CONFORMIDAD TOTAL — 963/1000

También se cumplen los requisitos de puerta, base, incertidumbre y calidad.

32. NOMOS 950+ CLASE DE CANDIDATO

CANDIDATO 950+ ESTÁNDAR — CANDIDATO 950+ ESTÁNDAR

En el contexto solo de la puntuación compuesta de un resultado:

NOMOS_1000≥ 950

no es suficiente. Condiciones que pueden buscarse juntas para candidatos NOMOS 950+:

- Estimación de punto compuesta de al menos 950
- Equivalente de usuario de Aprobado Estricto de al menos 950/1,000
- Sin Crítico Confirmado abierto
- Sin Mayor Confirmado abierto
- Todos los componentes principales de al menos 90
- Componentes LGF, USI y STR de al menos 85
- Masa central no resuelta en como máximo 2 por ciento
- Equivalente de usuario No evaluable por debajo del límite aceptado
- Cumpliendo al menos los niveles de calidad NCL-3, TPR-3, AQ-3
- Presencia de al menos una onda confirmatoria o de estabilidad
- Informe separado de los alcances de Alcance natural y Base común de comparación
- Explicación del límite superior estadístico de eventos críticos observados cero

Estos umbrales:

- piloto,
- Prueba sintética Apple.com,

- Estudio de integral con respuestas 30,000,
- Revisión independiente universitaria

debe ser calibrado posteriormente.

33. BANDAS DE RENDIMIENTO DEL CANDIDATO

Bandas temporales y descriptivas solo para el número compuesto:

Puntuación	Clase descriptiva
950-1,000	Candidato de rendimiento 950+ para NOMOS
900-949	Representación alta pero no totalmente conforme al estándar
800-899	Representación que requiere mejora material
650-799	Representación débil e irregular
0-649	Deficiencia severa de representación

Estas bandas:

- del estado de la puerta,
- del evento crítico,
- del piso del componente,
- no se puede reemplazar la incertidumbre y el alcance

Bloqueo crítico— El resultado 975 no es un candidato para 950+.

34. IDENTIDAD DE PUNTUACIÓN

Cada puntuación NOMOS debe llevar los siguientes campos en su identidad:

```
ScoreID=(Entidad, ProductoAI, Marco, Panel, ConjuntoDeIndicaciones, Ola, PaqueteDeVerdad,
VersiónDePuntuación)
```

Ejemplo:

NOMOS-ASTERON-ORION-NR-CLEAN-W1-CORESET1-TP0.9-SV0.9

Esta oración no es suficiente: “La puntuación NOMOS de Asteron es 962.” Que:

- Producto de IA,
- población,
- idioma,
- panel,
- prompts,
- fecha

es desconocida.

35. DIFERENCIA ENTRE CORE GEO-1000 Y FULL NOMOS

35.1. Core GEO-1000

Solo el Sistema Espejo Core mide la distribución de la población. Resultado principal:

SP1 000

AP1 000

CR1 000

MR1 000

UR1 000

puede ser. Responde a la pregunta: "¿A cuántos usuarios el producto de IA refleja correctamente la identidad fundamental y la actividad de la entidad?"

35.2. Full NOMOS

Incluye múltiples familias de prompts, panel, idioma, evidencia y componentes de consistencia. Produce un perfil de diez componentes y un índice de 1,000 puntos. El resultado central GEO-1000 no puede presentarse como la puntuación completa NOMOS.

36. PUNTUACIÓN Alcance natural

Cada producto de IA se evalúa dentro de su población natural elegible para el producto. NOMOSaNR es la puntuación Alcance natural del producto a. Responde a la siguiente pregunta:

> *¿Cuál es el estado de representación en la población de usuarios a la que este producto puede llegar realmente?*

Las poblaciones para Alcance natural pueden diferir entre diferentes productos. Por sí sola, no es suficiente para una clasificación directa de superioridad.

37. PUNTUACIÓN Base común de comparación

En el universo de usuarios comunes: NOMOSaCS se calcula. Responde a la siguiente pregunta:

> *Cuando se comparan productos bajo las mismas condiciones de país, idioma y usuario, ¿cómo genera representación cada uno?*

Base común de comparación: fortalece la comparabilidad, no muestra el alcance real del producto fuera del universo común. Alcance natural y Base común de comparación deberían publicarse por separado.

38. COEFICIENTE DE PRODUCTO AI

Si los resultados de diez productos de IA se van a combinar a nivel de ecosistema, se puede usar el coeficiente del producto: αa. Condición:

```
Σ_{a=1}^A α_a= 1
```

Coeficiente del producto:

- la puntuación propia del producto,
- la reputación del proveedor,
- la preferencia de la entidad auditada,
- patrocinio comercial

no se puede determinar a través de.

39. COEFICIENTE DE PRODUCTO IGUAL

Cada producto de IA recibe igual peso:

```
α_a^EQ= 1/A
```

Puntuación igual del ecosistema de productos:

```
NOMOS_ecosystem^EQ= Σ_a α_a^EQ NOMOS_a
```

Responde a la siguiente pregunta:

¿Qué tan fuerte es el producto de IA promedio probado?

No indica la exposición promedio humana.

40. COEFICIENTE PONDERADO POR EXPOSICIÓN

Sea Ma la exposición verificada de usuarios activos o elegibles del producto de IA durante el período definido.

```
α_a^EXP= M_a/ Σ_b M_b
```

puntuación del ecosistema ponderado por exposición:

```
NOMOS_ecosystem^EXP= Σ_a α_a^EXP NOMOS_a
```

Responde a la siguiente pregunta:

¿Cuál es el estado de la representación en el ecosistema de IA al que la persona promedio está expuesta?

41. SI NO HAY DATOS DE EXPOSICIÓN

Exposición de usuario activo o de consultas:

- cerrado al público,
- inconsistente,
- declaración propia del proveedor,
- alguien que confunde el número de cuentas con usuarios activos

puede ser de tal calidad. En este caso:

- puede ser publicado como un resultado de producto equivalente,
- el resultado ponderado por exposición puede seguir siendo NO ESTIMABLE,

Se pueden mostrar límites superior–inferior basados en escenarios razonables de coeficientes. No se pueden generar estimaciones arbitrarias de participación de mercado.

42. FIJACIÓN DE COEFICIENTE DE PRODUCTO

Cada αa:

- fuente de datos,
- período,
- unidad de usuario,
- cobertura del país,
- fecha de actualización,
- incertidumbre

debe llevarse. Los coeficientes deben estar fijados antes de que se vean las puntuaciones de IA. El coeficiente de un producto con baja puntuación no se puede reducir después del resultado.

43. LA PUNTUACIÓN DEL ECOSISTEMA NO PUEDE COMPENSAR CRÍTICAMENTE

Un producto:

- Puntos 980,
- puede tener cero Crítico

Otro producto:

- Puntos 900,
- puede tener Crítico Confirmado

El promedio del ecosistema puede ser alto. La tarjeta pública aún debe mostrar:

- Incidente Crítico basado en el producto
- Peso de exposición Crítica
- Si al menos un producto tiene un Crítico abierto
- Alcance del Crítico

Compuesto del ecosistema: no se puede eliminar la puerta Crítica a nivel de producto.

44. EXPOSICIÓN CRÍTICA DEL ECOSISTEMA

Tasa de usuarios críticos del producto a: sea CRa. Tasa crítica ponderada por exposición:

```
CR_ecosystem^EXP= Σ_a α_a^EXP CR_a
```

puede ser calculado de esta manera. Adicionalmente, el campo binario:

```
AnyCritical= 1(∃a: CIS_a ≥ 2)
```

también debe ser mostrado. La tasa de eventos Críticos en un producto de baja exposición puede ser pequeña. El registro AnyCritical aún preserva la presencia del evento.

45. COEFICIENTE DE FAMILIA DE PROMPT

βf puede ser usado para la familia de prompt f. Condición:

```
Σ_f β_f= 1
```

βf:

- no puede depender de si el prompt produce una puntuación alta o baja,
- la preferencia de la compañía,
- la facilidad de la decisión posterior a la medición

Los pesos de la familia de prompt pueden basarse en las siguientes áreas:

- Importancia de la decisión del usuario
- Nivel de riesgo
- Frecuencia de uso natural del prompt
- Tipo de entidad
- Núcleo obligatorio del estándar

46. LAS FAMILIAS DE PROMPTS DEBERÍAN SER VISIBLES POR SEPARADO

Incluso si se puede calcular la puntuación compuesta del prompt, los siguientes resultados deben permanecer separados:

- Espejo Central
- Resolución de Existencia
- Actividad y Alcance
- Evidencia
- Local
- Temporal
- Recomendación
- Comparativo
- Límite
- Control
- Reserva

Un resultado alto del Núcleo no puede hacer que un resultado bajo de Frontera o Recomendación sea invisible.

47. PUNTUACIONES DE PANEL CONTROLADO Y NATURAL

Los paneles controlados y naturales son objetos de predicción diferentes. La regla predeterminada NOMOS:

> *Los dos paneles llevan identidades de puntuación separadas y no se combinan aritméticamente.*

Ejemplo:

NOMOS-CLEAN: 962

NOMOS-NATURAL: 918

Si un propósito de política o investigación los va a combinar:

```
γ_C+ γ_N= 1
```

Se pueden usar coeficientes predeclarados. Pero:

- dos puntuaciones brutas,
- La diferencia entre ellos es que
- Coeficiente usado

debe mostrarse por separado.

48. PUNTUACIÓN DE OLA Y SERIES DE TIEMPO

Cada Onda Principal tiene su propio ID de puntuación. La puntuación W1, la puntuación W2, la puntuación W3 se mantienen separadas. Puntuación actual: La última Onda Principal o Revisión válida puede ser la Onda de Reprueba. El promedio de ondas pasadas no debe determinar automáticamente el resultado actual.

49. PUNTUACIÓN RODANTE

Si se desea, se puede generar una puntuación móvil de tiempo fijo:

```
NOMOS_rolling= Σ_w δ_wNOMOS_w
```

Aquí:

```
Σ_w δ_w= 1
```

es. δw:

- antes del resultado,
- según el peso del tiempo

debe ser determinado. Puntuación rodante:

- la puntuación actual,
- no reemplaza el registro histórico de incidentes

Ninguno de los registros puede ser desplazado silenciosamente.

50. TRES PUNTOS PARA afirmaciones NO RESUELTAS

Las reclamaciones no resueltas pueden aumentar o disminuir artificialmente la estimación puntual de átomos materiales. Por lo tanto, se sugieren tres resultados.

50.1. Límite Inferior Conservador

Las afirmaciones no resueltas se consideran cero. Más lento

50.2. Estimación Puntual Solo de Resueltos

Solo se utilizan afirmaciones resueltas. Punto S

50.3. Límite Superior Optimista

Se aceptan las afirmaciones no resueltas con apoyo completo. Super Entonces:

```
S_lower≤ S_point≤ S_upper
```

debe ser el caso. Este rango:

> *es el rango de incertidumbre de referencia o epistémica.*

No es lo mismo que un intervalo de confianza de muestra.

51. INCERTIDUMBRE DE MUESTRA

Dado que las estimaciones de la población provienen de la muestra:

- distribución de respuestas,
- puntuación del componente,
- puntuación compuesta

debe llevar un intervalo de confianza. Método en muestras complejas:

- capas,
- los conjuntos,
- sus pesos,
- usuarios emparejados,
- olas

debe ser protegido. Métodos candidatos:

- Estimación de la varianza basada en el diseño
- Pesos de replicación
- Bootstrap estratificado
- Bootstrap de conglomerados
- Modelo apropiado para eventos raros
- Estimación de encuestas asistida por modelos

El método utilizado debe especificarse en el registro público.

52. INTERVALO DE CONFIANZA DE LA PUNTUACIÓN COMPUESTA

El intervalo de confianza de la puntuación compuesta no se puede obtener simplemente sumando los intervalos de los componentes. Los componentes:

- los mismos usuarios,
- los mismos afirmaciones,
- los mismos idiomas,
- las mismas olas

dependen de esta razón. Enfoque preferido: Crear remuestreos de acuerdo con el diseño de la muestra. En cada remuestreo, recalcular todas las métricas de respuesta y de componentes. Reproducir la puntuación compuesta. Generar el intervalo de confianza a partir de la distribución. Este proceso debe ejecutar nuevamente toda la arquitectura de puntuación.

53. TAMAÑO EFECTIVO DE LA MUESTRA

Observación en crudo: n y muestra efectiva:

```
n_eff= (Σ_i w_i)^2/ Σ_i w_i^2 [K08]
```

deberían publicarse juntos. Las observaciones en crudo 1,000 pueden no significar que 1,000 equivalga a unidades de información iguales. Las puntuaciones de componentes y subgrupos deben tener sus propios tamaños de muestra efectivos.

54. ERROR CRÍTICO CERO OBSERVADO

No tener eventos Críticos Confirmados en una muestra:

Críticos Confirmados Observados = 0

debe ser escrito. La siguiente expresión no debe ser utilizada:

> *El riesgo crítico es cero.*

Bajo la suposición de observaciones independientes e iguales, un límite superior aproximado rápido:

```
p_upper,95≈ 3/n [K09]
```

puede estimarse con el enfoque conocido. En un diseño complejo GEO-1000:

- En lugar de n crudo, debe usarse un método apropiado para el diseño,
- si es necesario, una muestra efectiva,
- un límite unilateral exacto o basado en bootstrap

debe ser utilizado. Ejemplo:

```
neff=1,000
```

y el límite superior aproximado bajo eventos cero:

```
3/1,000 = 0.003 = 0.3%
```

puede ser. Declaración pública correcta:

> *En 1,000 observaciones efectivas, no se observaron eventos críticos confirmados; en el método aproximado utilizado, el límite superior unilateral de 95 por ciento es aproximadamente 0.3 por ciento.*

55. INCERTIDUMBRE CON UN EVENTO CRÍTICO

Cuando se observa un evento crítico: la estimación puntual puede ser baja, el intervalo de confianza puede ser amplio. Resultado correcto:

- el evento ocurrió,
- prevalencia estimada,
- amplia incertidumbre,
- estado del incidente

deberían mostrarse juntos. Las siguientes dos oraciones también pueden estar equivocadas:

- “Solo un evento es insignificante.”
- “Un evento prueba que todo el sistema ha fallado 100%.”

56. SENSIBILIDAD DE LA PUNTUACIÓN

También se pueden calcular las siguientes alternativas:

- Puntuación sin ponderar
- Puntuación ponderada por población principal

- Puntuación ponderada sin recorte
- Puntuación ponderada recortada
- Puntuación solo resuelta
- Puntuación conservadora
- Puntuación de producto igual
- Puntuación ponderada por exposición
- Puntuación Alcance natural
- Puntuación Base común de comparación
- Puntuación limpia
- Puntuación natural

Si existe una diferencia material:

ADVERTENCIA SOBRE LA SENSIBILIDAD DE LA PUNTUACIÓN

debería darse.

57. PRECISIÓN INCORRECTA

La siguiente representación puede ser excesivamente precisa para la mayoría de los resultados de GEO-1000: 963,742/ 1000 Estimación puntual:

- número entero,
- decimal si es necesario

debería publicarse. El intervalo de confianza debe indicar el límite de precisión numérica.

58. COMPONENTE FALTANTE

Si uno de los diez componentes no puede ser calculado:

- su peso no puede ser distribuido a otros componentes.
- el campo faltante no se puede contar como puntos 100,

no se puede publicar la puntuación completa NOMOS. El estado correcto:

PERFIL PARCIAL NOMOS

COMPONENTE NO ESTIMABLE

PUNTUACIÓN PROVISIONAL

puede aplicarse. Ejemplo: si no hay un panel natural disponible para USI, solo se puede presentar un perfil parcial con ocho o nueve componentes.

```
USI=100 no se puede asumir.
```

59. ESTADO DE PUBLICACIÓN DE LA PUNTUACIÓN

NSQ-0— NO COMPUTADO

Los datos requeridos no están disponibles.

NSQ-1— PROVISIONAL

Onda única, adjudicación limitada o componentes faltantes existen.

NSQ-2— PERFIL PARCIAL

Algunos componentes han sido calculados de manera confiable. No se puede realizar una comparación completa de 1,000 puntos.

NSQ-3— PUNTUACIÓN DE ONDA VÁLIDA

Todos los componentes principales y los prerrequisitos de calidad se han cumplido.

NSQ-4— PUNTUACIÓN REPLICADA

La puntuación se ha repetido en al menos una onda independiente o de confirmación.

NSQ-5— PUNTUACIÓN AUDITADA EXTERNAMENTE

El cálculo y la cadena de evidencia han sido reproducidos por un equipo independiente. Candidatura 950+ al menos:

NSQ-4

debería apuntar a ese nivel.

60. VERSIÓN DE LA BANDA SONORA

Cada versión de la puntuación debe incluir lo siguiente:

- Versión del código de cálculo
- Fórmulas componentes
- Aplicaciones de valores dimensionales
- Ponderaciones familiares de prompt
- Coeficientes de producto
- Pesos de muestra
- Versión del Paquete de la Verdad
- Versión para adjudicadores
- ID de la ola
- Hash de integridad

Si la fórmula cambia: la partitura antigua no puede ser reescrita, se crea una nueva versión de la puntuación.

61. NUEVA FÓRMULA DE PUNTUACIÓN PARA LOS MISMOS DATOS

La metodología de puntuación puede evolucionar. Los mismos datos sin procesar:

- Método de puntuación 0.9
- Método de puntuación 1.0

pueden ser recalculados con. Registro correcto:

Datos	Método de puntuación	Resultado
Conjunto de observación fijo W1	0.9	944
Conjunto de observación fijo W1	1.0	951

Los dos resultados son versiones separadas. El nuevo método no puede eliminar el método antiguo del pasado.

62. PUERTAS CONTRA LA MANIPULACIÓN DE PUNTUACIONES

La arquitectura de puntuación NOMOS debe tener protección explícita contra los siguientes métodos:

- Seleccionar solo productos de IA con alta puntuación
- Excluir un idioma o país de baja puntuación del alcance
- Convertir Aprobación condicional en Aprobación plena
- Contando los No Resueltos como positivos
- Eliminando silenciosamente No evaluable del denominador
- Cambiando el peso de la familia de prompts después del resultado
- Reduciendo el coeficiente del gran producto según la puntuación
- Ocultando el producto con un evento crítico en el promedio del ecosistema
- Publicando solo la puntuación Limpia
- Seleccionando la mejor ola como la puntuación actual
- Ocultando la diferencia entre las puntuaciones sin ponderar y ponderadas
- Distribuyendo el peso de los componentes faltantes a los componentes fuertes
- No proporcionar un intervalo de confianza
- Pasando el umbral 950 basado en la estimación puntual y ocultando la incertidumbre

63. TARJETA DE RESULTADOS PÚBLICA NOMOS

La tarjeta de resultados públicos del candidato debe llevar el siguiente orden:

A. Identidad

Producto de IA de la entidad auditada Alcance natural/ Conjunto de prompts controladas / Naturales Base común de comparación Versión Wave Truth Pack Método de puntuación

B. Estado de la Puerta

Conformidad Total Candidato Condicional Fallo grave Bloqueo crítico No evaluable

C. Distribución de Respuestas GEO-1000

RP-1–RP-8 equivalentes de usuario Pase Estricto Pase Aceptable Crítico Mayor No Resuelto Sin Respuesta Utilizable No evaluable

D. Perfil de Diez Componentes

ENT

FAC

SBI

TLA

ECI

EPI

TCR

LGF

USI

STR

E. Resultado Compuesto

NOMOS/ 1,000

Intervalo de confianza de muestreo Límites inferior/superior de referencia Muestra efectiva

F. Riesgos y Eventos

Eventos críticos confirmados observados; límite superior crítico; hallazgos principales; incidentes abiertos; y estado de corrección.

G. Alcance y Calidad

SQ

PSI

NCL

TPR

AQ

NSQ

Cobertura por país/idioma Nivel de comparación

64. DISTRIBUCIÓN DE RESPUESTA SINTÉTICA

DEMONSTRACIÓN DE METODOLOGÍA SINTÉTICA / Los resultados siguientes no son productos reales de IA ni desempeño de la empresa. Para Synthetic Orion AI:

Estado	Equivalente de usuario
RP-1 Aprobación plena	730
RP-2 Aprobación con Aclaración	170
RP-3 Aprobación condicional	70
RP-4 Fallo grave	20
RP-5 Fallo crítico	2
RP-6 Sin Resolver	5
RP-7 Sin Respuesta Útil	3
RP-8 No evaluable	0
Total	1,000

Aprobación Estricta:

```
730+170=900
```

Aprobación Aceptable:

```
900+70=970
```

Crítico: 2/1,000 Mayor: 20/1,000 La aprobación aceptable para este producto es 97 por ciento. Sin embargo, hay hallazgos Evidentes Críticos y Mayores. Estado correcto:

Bloqueo crítico

Estado incorrecto:

97 por ciento exitoso

65. PERFIL DE COMPUESTO SINTÉTICO

Componente	Puntuación
ENT	96
FAC	93
SBI	89
TLA	95
ECI	88
EPI	91
TCR	96
LGF	82
USI	87
STR	90
NOMOS	907

Presentación pública correcta:

Bloqueo crítico— 907/1,000

Este sistema:

- en promedio áreas fuertes,
- lenguaje/justicia débil del país,
- temas de evidencia material,
- evento crítico abierto

lleva.

66. RESULTADO DE POST-CORRECCIÓN SINTÉTICA

En la nueva ola:

Estado	Equivalente de usuario
RP-1	830
RP-2	128
RP-3	30
RP-4	0
RP-5	0
RP-6	5
RP-7	7
RP-8	0
Total	1,000

Aprobación estricta: Componentes 958:

Componente	Puntuación
ENT	99
FAC	97
SBI	96
TLA	95
ECI	96
EPI	98
TCR	97
LGF	94
USI	93
STR	96
NOMOS	961

Registro correcto: ACTUAL: NOMOS 950+ CANDIDATO DE RENDIMIENTO — 961/1,000/ Histórico: Incidente crítico confirmado previamente remediado candidato / Crítico observado en la nueva prueba: 0/ límite superior de tasa crítica: también se muestra Incidente crítico antiguo no eliminado.

67. CASO DEL COEFICIENTE DEL PRODUCTO AI SINTÉTICO

Cuatro productos sintéticos de IA:

Producto	NOMOS	Coeficiente igual	Coeficiente de exposición
A1	960	0.25	0.55
A2	940	0.25	0.25
A3	900	0.25	0.15
A4	980	0.25	0.05

Puntuación de producto igual:

```
0.25(960+940+900+980)=945
```

Puntuación ponderada por exposición:

```
0.55(960)+0.25(940)+0.15(900)+0.05(980)=947
```

Los dos resultados son cercanos. Sin embargo, pueden ser muy diferentes en otra distribución. Interpretación correcta:

- Puntuación del ecosistema de sistema igual: 945
- Puntuación del ecosistema ponderada por exposición: 947
- Registros de puerta e incidentes a nivel de producto por separado

Comentario incorrecto: “La puntuación del ecosistema NOMOS es definitivamente 947.”

68. CASO DE COEFICIENTE CRÍTICO SINTÉTICO

Sea la puntuación del producto A4 980. Sin embargo, supongamos que hay un evento Crítico Confirmado en A4. El coeficiente de exposición es solo 0.05. El compuesto del ecosistema puede permanecer alto. Tarjeta pública correcta: Compuesto ponderado por exposición: 947/ Cualquier Producto Crítico Confirmado: SÍ / Participación de exposición crítica: proporción calculada / Cobertura de A4: Bloqueo crítico El pequeño coeficiente de A4 no puede borrar el evento Crítico.

69. DECISIÓN DE ESTADO DE PUBLICACIÓN DE PUNTUACIÓN

Orden de decisión de candidato:

```
PublicationStatus= No evaluable(medición básica ausente); Bloqueo crítico(Crítico Confirmado
abierto); Fallo grave(Mayor Confirmado abierto); PROVISIONAL(componente, repetición o calidad
ausente); CONDITIONAL(Moderado o umbral de no resuelto); 950+ CANDIDATE(todos los requisitos
candidatos); de lo contrario, VÁLIDO
```

La banda de puntuación se muestra después de esta decisión de estado.

70. DISPOSICIONES NORMATIVAS OBLIGATORIAS

CH16-N01

El resultado NOMOS no puede publicarse como una puntuación compuesta única y desnuda.

CH16-N02

El estado de la puerta debe mostrarse antes de la puntuación compuesta.

CH16-N03

La distribución de respuestas GEO-1000 es una parte obligatoria del resultado público principal.

CH16-N04

El total de equivalentes de usuario RP-1 y RP-8 debe ser 1,000.

CH16-N05

Aprobación estricta, Aprobación aceptable, Mayor, Crítico, No resuelto, Sin respuesta utilizable y No evaluable deben mostrarse por separado.

CH16-N06

Las observaciones No resuelto y No evaluable no pueden añadirse silenciosamente a las categorías de aprobado o reprobado.

CH16-N07

Las observaciones No evaluable no pueden interpretarse como un fallo en el rendimiento del producto.

CH16-N08

Las observaciones No evaluable no pueden eliminarse silenciosamente del denominador; deben mostrarse en la cobertura evaluable.

CH16-N09

Los pesos de las observaciones deben ser independientes del resultado de la respuesta de la IA.

CH16-N10

El número bruto de participantes y el equivalente de usuario ponderado por la población deben informarse por separado.

CH16-N11

El estado de aprobación de la respuesta debe mantenerse separado de la proporción de soporte de la afirmación atómica.

CH16-N12

Los puntos críticos y mayores no pueden ser compensados por ningún puntuación compuesta.

CH16-N13

Los pesos de centralidad atómica deben definirse antes de ver los resultados.

CH16-N14

Una repetición de la misma afirmación raíz no puede inflar o reducir artificialmente la puntuación.

CH16-N15

Los pesos de centralidad de la afirmación no pueden reemplazar los puntos críticos y mayores.

CH16-N16

Los valores de la dimensión de Fidelidad Atómica deben estar versionados y disponibles públicamente.

CH16-N17

Si el mapeo de la Fidelidad Atómica cambia, se debe crear una nueva versión del método de puntuación.

CH16-N18

Los átomos Brecha de referencia no pueden añadirse silenciosamente al puntuación como cero o uno.

CH16-N19

La Completitud del Elemento de Respuesta Requerido debe publicarse como una métrica separada.

CH16-N20

El fallo en el elemento esencial requerido no puede compensarse con un promedio alto de REC.

CH16-N21

El perfil NOMOS con diez componentes debe ser visible antes del índice compuesto.

CH16-N22

ENT, FAC, SBI, TLA, ECI, EPI, TCR, LGF, USI y STR deben tener puntuaciones separados.

CH16-N23

Un componente no puede asumir silenciosamente el peso faltante de otro componente.

CH16-N24

No se puede asumir que un componente que no puede calcularse como 100.

CH16-N25

El índice compuesto completo NOMOS solo debe publicarse si los diez componentes han sido calculados de manera confiable.

CH16-N26

Si se utiliza un ponderación de componentes iguales como el método candidato canónico, debe declararse explícitamente.

CH16-N27

Los pesos alternativos específicos del dominio deben tener un nombre de puntuación y versión separados.

CH16-N28

Una puntuación específica del dominio no puede usarse en lugar del índice canónico sin explicación.

CH16-N29

LGF no puede depender únicamente de la media poblacional; debe incluir una base de desempeño y disparidad.

CH16-N30

Una sola observación del Observador por País no puede convertirse en una puntuación de equidad del país.

CH16-N31

Solo las celdas de idioma y geográficas con tamaño de muestra suficiente deben participar en la estimación de equidad.

CH16-N32

Los idiomas y países que no han sido evaluados no pueden ser ignorados para producir un resultado de alta equidad.

CH16-N33

USI debe evaluar los resultados de paneles controlados y naturales junto con la diferencia en la interfaz del material.

CH16-N34

En un estudio con solo un panel controlado o solo un panel natural, no se puede considerar que USI esté totalmente calculado.

CH16-N35

Los puntuaciones de los paneles controlados y naturales deben mantenerse separados por defecto.

CH16-N36

Si se combinan los puntuaciones controlados y naturales, los coeficientes deben fijarse antes de los resultados.

CH16-N37

STR no puede presentarse como una puntuación de estabilidad completo en una sola ola.

CH16-N38

Los cambios de producto entre olas o del Truth Pack deben modelarse explícitamente en el cálculo de la estabilidad.

CH16-N39

El Compuesto NOMOS no puede cambiar el estado de la puerta Crítica o Mayor.

CH16-N40

Una puntuación de 950 por sí sola no puede establecer la candidatura NOMOS 950+.

CH16-N41

Las condiciones para la candidatura NOMOS 950+ deben cumplir colectivamente con los requisitos de puntuación, Aprobación Estricta, puerta, piso de componente, no resuelto, calidad y repetición.

CH16-N42

Los umbrales 950+ no pueden presentarse como un estándar científico final antes de la revisión piloto e independiente.

CH16-N43

La identidad de la puntuación debe incluir producto de IA, marco, panel, conjunto de prompts, onda, Paquete de Verdad y versión de la puntuación.

CH16-N44

La oración “La puntuación NOMOS de la entidad es X” no puede usarse sin una identidad de alcance.

CH16-N45

El resultado Core GEO-1000 no puede presentarse como la puntuación completa NOMOS.

CH16-N46

Las puntuaciones Alcance natural y Base común de comparación deben informarse por separado.

CH16-N47

Las puntuaciones del producto para diferentes poblaciones Alcance natural no pueden clasificarse directamente sin explicación.

CH16-N48

El resultado Base común de comparación no puede presentarse como la puntuación total de acceso global del producto.

CH16-N49

Los coeficientes del producto de IA deben determinarse de manera independiente a los resultados.

CH16-N50

Un coeficiente de producto de IA no puede depender de la puntuación del producto, la reputación de la marca o la relación de patrocinio comercial.

CH16-N51

Las puntuaciones del ecosistema ponderadas por producto y exposición deben nombrarse por separado.

CH16-N52

Si no se pueden verificar los datos de usuarios activos o de exposición, no se pueden generar coeficientes arbitrarios.

CH16-N53

Los datos de exposición deben ser transparentes a nivel de cuenta, usuario, consulta o sesión.

CH16-N54

El coeficiente de un producto con baja puntuación no puede reducirse después de ver el resultado.

CH16-N55

El compuesto del ecosistema no puede eliminar la puerta Crítica o Mayor a nivel de producto.

CH16-N56

El ratio AnyCritical y el Critical ponderado por exposición deben mostrarse por separado.

CH16-N57

Los coeficientes de la familia de prompts deben determinarse antes de los resultados.

CH16-N58

El peso de una familia de prompts con baja puntuación no puede reducirse más tarde.

CH16-N59

El resultado separado de cada familia de solicitudes debe permanecer visible junto con la puntuación compuesta de la solicitud.

CH16-N60

La puntuación de la ola actual debe mantenerse separada de las puntuaciones acumuladas o históricas.

CH16-N61

La ola pasada con la puntuación más alta no puede seleccionarse como la puntuación actual.

CH16-N62

Se pueden producir resultados conservadores, solo resueltos y optimistas para la masa de afirmaciones no resueltos.

CH16-N63

La incertidumbre de referencia y el intervalo de confianza de la muestra no pueden presentarse como lo mismo.

CH16-N64

El intervalo de confianza de la muestra compleja debe preservar la estratificación, el agrupamiento, el ponderado y la estructura de usuario emparejado.

CH16-N65

El intervalo de confianza de la puntuación compuesta debe calcularse volviendo a ejecutar todo el proceso de puntuación.

CH16-N66

Los tamaños de muestra brutos y efectivos deben mostrarse juntos.

CH16-N67

Un evento crítico observado como cero no puede expresarse como riesgo crítico cero.

CH16-N68

En el caso de eventos cero, se debe publicar un límite superior unilateral o la incertidumbre equivalente de evento raro.

CH16-N69

El enfoque 3/n no puede usarse automáticamente como el método final en diseños complejos; solo puede ser un control aproximado.

CH16-N70

En un evento crítico, se deben mostrar juntos la presencia del evento y la incertidumbre de su frecuencia.

CH16-N71

Los análisis de sensibilidad de puntuación deben mostrar los resultados no ponderados, ponderados, recortados, no resueltos y de coeficientes alternativos en la medida que sea relevante.

CH16-N72

La sensibilidad de la puntuación material debe generar una alerta en la tarjeta pública.

CH16-N73

El falso decimal de la puntuación no se puede publicar en absoluto.

CH16-N74

El peso del componente faltante no puede distribuirse a otros componentes.

CH16-N75

El estado de Calidad de Publicación de Puan debe ser visible en el registro público.

CH16-N76

La candidatura 950+ debe apuntar al menos al nivel de puntuación replicado.

CH16-N77

Cuando la fórmula de la puntuación cambia, las puntuaciones anteriores no pueden ser reescritas silenciosamente.

CH16-N78

Si se aplican diferentes versiones del método de puntuación al mismo conjunto de datos, todos los resultados deben conservarse en versiones separadas.

CH16-N79

El código de cálculo de puntuación, los coeficientes y los registros de integridad deben publicarse en un formato legible por máquina.

CH16-N80

Cada puntuación, coeficiente, intervalo de confianza y decisión de publicación debe tener un responsable humano o institución accountable.

71. FORMAS DE FALLA

CH16-F01— INFORME DE UN SOLO NÚMERO

Se preservan la distribución, la puerta y el perfil del componente.

CH16-F02— OCULTANDO LA PUERTA BAJO LA PUNTUACIÓN

Bloqueo crítico se convierte en una pequeña nota al pie.

CH16-F03— CONTANDO CONDICIONAL COMO Aprobación plena

Los problemas moderados se vuelven invisibles.

CH16-F04— CONTANDO LOS NO RESUELTOS COMO APROBADOS

Los resultados desconocidos aumentan la puntuación.

CH16-F05— CONTANDO LOS NO RESUELTOS COMO FALLA

La falta de referencia disminuye el producto.

CH16-F06— ELIMINANDO No evaluable DEL DENOMINADOR

Los defectos de medición se vuelven invisibles.

CH16-F07— CONTANDO No evaluable COMO DEFECTO DEL PRODUCTO

El defecto de inspección se asigna a AI.

CH16-F08— CONTAR 1,000 PERSONAS CON PESO IGUAL

Se ignoran el sobremuestreo y las diferencias de población.

CH16-F09— PESO SEGÚN EL RESULTADO

Las respuestas correctas reciben mayor peso.

CH16-F10— CONTAR Tasa de respaldo de las afirmaciones COMO RESPUESTA APROBADA

La puerta crítica se pierde en el promedio.

CH16-F11— CONTAR EL NÚCLEO Y LO INCIDENTAL POR IGUAL

La afirmación incidental tiene tanto peso como la entidad principal.

CH16-F12— INFLAR La puntuación REPITIENDO

La misma afirmación correcta se escribe diez veces para aumentar La puntuación.

CH16-F13— COLAPSAR La puntuación REPITIENDO

La misma raíz incorrecta da lugar a diez sanciones separadas.

CH16-F14— ALMACENANDO VALORES DE DIMENSIÓN

No se puede reproducir la puntuación.

CH16-F15— CONTANDO LA afirmación QUE VIENE DE ABAJO COMO HECHO PARA F-4

La transmisión precisa del autoinforme se convierte en la precisión del autoinforme.

CH16-F16— FUSIONANDO ELEMENTOS REQUERIDOS EN PROMEDIO ATÓMICO

Incluso si falta la tarea principal, la respuesta obtiene una puntuación alta.

CH16-F17— SELECCIÓN SELECTIVA DE COMPONENTES

Solo se publican los tres componentes fuertes.

CH16-F18— CONTANDO EL COMPONENTE FALTANTE COMO 100

La puntuación completa se incrementa artificialmente.

CH16-F19— DISTRIBUYENDO EL PESO FALTANTE

La puntuación de equidad no calculada se añade a la puntuación factual.

CH16-F20— CONTANDO LA LEY DE LA NATURALEZA COMO PESO IGUAL

No se realizan sensibilidad ni revisión externa.

CH16-F21— CONTANDO LA PUNTUACIÓN DEL DOMINIO COMO CANÓNICA

El peso especial se presenta como la puntuación base NOMOS.

CH16-F22— CONTANDO EL PROMEDIO DE LENGUAJE COMO EQUIDAD EN EL LENGUAJE

Se oculta el piso bajo de lenguaje.

CH16-F23— CONTANDO LA IGUALDAD DE IDIOMA COMO CALIDAD

Las puntuaciones bajas en todos los idiomas producen alta equidad.

CH16-F24— CONTANDO LOS IDIOMAS NO PROBADOS COMO COBERTURA

El factor de cobertura está inflado.

CH16-F25— PUNTUACIÓN DEL PAÍS DESDE EL OBSERVADOR

Un solo usuario entra en el cálculo de equidad.

CH16-F26— PUBLICAR SOLO LA PUNTUACIÓN LIMPIA

Se preserva la experiencia natural del usuario.

CH16-F27— CONTAR SOLO LA PUNTUACIÓN NATURAL COMO PREDERTERMINADA DEL PRODUCTO

La personalización se convierte en el comportamiento básico.

CH16-F28— SILENCIOSAMENTE PROMEDIO, LIMPIO Y NATURAL

Diferentes objetos de predicción se mezclan.

CH16-F29— 100 ESTABILIDAD EN UNA SOLA ONDA

Se considera que la estabilidad está completa sin repetición.

CH16-F30— CONTAR ONDA VIEJA Y BUENA COMO ACTUAL

El debilitamiento actual está oculto.

CH16-F31— CONTAR AUTOMÁTICAMENTE 950 COMO CUMPLIMIENTO

Se ignoran la puerta y el piso del componente.

CH16-F32— Bloqueo crítico— CONTAR 970 COMO 950+

La cuenta va por delante de la puerta.

CH16-F33— OCULTAR IDENTIDAD DE PUNTUACIÓN

No se sabe qué producto y panel se miden.

CH16-F34— CONTAR PUNTUACIÓN CORE COMO COMPLETA NOMOS

Un solo indicador representa todo el estándar.

CH16-F35— CLASIFICAR PUNTUACIONES Alcance natural

Se consideran iguales diferentes poblaciones.

CH16-F36— CONTAR ACCESO COMPLETO A Base común de comparación

Una población común estrecha se convierte en un resultado global.

CH16-F37— DAR COEFICIENTE DEL MODELO SEGÚN LA PUNTUACIÓN

Una puntuación alta gana más influencia.

CH16-F38— HACIENDO DE LA REPUTACIÓN DE LA MARCA UN FACTOR

La popularidad del producto se pondera sin ser predicha.

CH16-F39— CONTANDO LOS NÚMEROS DE CUENTA COMO USUARIOS ACTIVOS

El coeficiente de exposición se inflama.

CH16-F40— ELIMINANDO PRODUCTOS CON BAJA PUNTUACIÓN

El ecosistema consiste únicamente en productos exitosos.

CH16-F41— CONTANDO PUNTUACIÓN DE EXPOSICIÓN IGUAL PARA EL PRODUCTO

El producto promedio se presenta como una persona promedio.

CH16-F42— CONTANDO LA PUNTUACIÓN DE EXPOSICIÓN COMO CALIDAD DEL PRODUCTO

Un producto grande automáticamente se convierte en una métrica de calidad más importante.

CH16-F43— ELIMINAR CRÍTICO EN PROMEDIO DEL ECOSISTEMA

El evento intenso del producto de coeficiente pequeño desaparece.

CH16-F44— CAMBIAR EL COEFICIENTE DEL PROMPT DESPUÉS DEL RESULTADO

La familia de prompts difíciles se reduce.

CH16-F45— CONSIDERAR LA FAMILIA DE BAJA PUNTUACIÓN COMO OPCIONAL

Se excluye del alcance el resultado límite o de evidencia.

CH16-F46— MANTENER EL ACTUAL CON PUNTUACIÓN RODANTE

Las olas altas antiguas cubren la caída actual.

CH16-F47— ELIMINAR NO RESUELTO DEL DENOMINADOR

La estimación puntual aumenta artificialmente.

CH16-F48— CONSIDERAR LA INCERTIDUMBRE DE REFERENCIA EN EL IC DE MUESTRA

Se mezclan dos incertidumbres diferentes.

CH16-F49— IC DE MUESTRA ALEATORIA SIMPLE

Se ignoran estratificación, ponderación y conglomerados.

CH16-F50— SUMA DE IC DE COMPONENTES

Se ignora la dependencia.

CH16-F51— RECLAMAR SENSIBILIDAD CON N BRUTO

Se oculta la muestra efectiva.

CH16-F52— CONTAR EVENTOS CERO COMO RIESGO CERO

Se oculta la incertidumbre de errores críticos raros.

CH16-F53— APLICAR 3/N A CADA CASO

Se ignora la estructura compleja de muestras y ponderaciones.

CH16-F54— CONTANDO UN CRÍTICO ÚNICO COMO 100 PORCENTAJE DE RIESGO DEL SISTEMA

La comuniadad está demasiado generalizada.

CH16-F55— OCULTANDO UN CRÍTICO ÚNICO PORQUE LA TASA ES PEQUEÑA

La existencia del evento desaparece.

CH16-F56— OCULTANDO EL ANÁLISIS DE SENSIBILIDAD

No se observa que la puntuación cambie en métodos alternativos razonables.

CH16-F57— SENSIBILIDAD DECIMAL FALSA

Se publica un resultado como 963,742.

CH16-F58— CONTANDO UN PERFIL PARCIAL COMO PUNTUACIÓN COMPLETA

Se realiza una comparación con un componente que falta.

CH16-F59— PUBLICAR LA PUNTUACIÓN PROVISIONAL COMO FINAL

Una sola ola y un solo evaluador se convierten en estándar completo.

CH16-F60— REESCRIBIENDO LA HISTORIA EN EL CAMBIO DE FÓRMULA

La puntuación anterior desaparece.

CH16-F61— MANTENER EL CÓDIGO DE CÁLCULO CERRADO

Se previene la reproducción independiente.

CH16-F62— CONSIDERAR LA ALTURA DE LA PUNTUACIÓN COMO CALIDAD DE AUDITORÍA

Se acepta una medición defectuosa porque dio un resultado alto.

CH16-F63— CONSIDERAR LA BAJA PUNTUACIÓN COMO DEFECTO DEL MÉTODO

Un resultado bajo se recalcula debido a inconvenientes comerciales.

CH16-F64— MONOPOLIO DE PUNTUACIÓN DE UNA SOLA EMPRESA

Se impide que instituciones independientes apliquen la misma fórmula.

72. PROCEDIMIENTO DE AUDITORÍA

Paso 1— Crear Identidad de Puntuación

El método de puntuación del conjunto de prompts del panel del marco del producto AI de la entidad Wave Truth Pack está fijado.

Paso 2— Fijar el Conjunto de Observación de Análisis Principal

Los registros de respaldo, retirados, no calificables y de observación principal están separados.

Paso 3— Verificar los Pesos de la Población

Se examinan los pesos de diseño, no respuesta y calibración.

Paso 4— Calcular Distribución de RP

Se producen equivalentes de usuario con RP-1 a RP-8.

Paso 5— Calcular Pasaje Estricto y Aceptable

Se derivan SP1 000 y AP1 000.

Paso 6— Calcular las Densidades de Crítico, Mayor, No Resuelto y No Evaluable

La puerta y la prevalencia se mantienen por separado.

Paso 7— Aplicar Centralidad de afirmaciones y Normalización de Clúster Raíz

Los afirmaciones se verifican nuevamente.

Paso 8— Calcular Métricas de Tamaño Atómico

Se producen resultados de entidad, hecho, alcance, tiempo, atribución, modalidad y referencia.

Paso 9— Calcular la Completitud de Elementos Requeridos

Se registra junto con las puertas de la familia de prompts.

Paso 10— Calcular Diez Componentes

Se aplican todas las fórmulas de componentes de ENT a STR.

Paso 11— Comprobar Componentes Faltantes

Se determina si se puede calcular la puntuación total.

Paso 12— Calcular el Índice Compuesto NOMOS

Se aplica solo a perfiles completos y válidos.

Paso 13— Asignar Estado de la Puerta

Determina si el estado es Crítico, Mayor, Condicional o Candidato Completo.

Paso 14— Separar resultados Alcance natural y Base común de comparación

Las dos familias de puntuaciones se mantienen separadas.

Paso 15— Fijar coeficientes del producto AI

Los coeficientes iguales y de exposición se guardan junto con la fuente de datos.

Paso 16— Aplicar coeficientes de la Familia prompt

Los pesos se utilizan de manera independiente del resultado.

Paso 17— Separar resultados de Panel y Wave

Las puntuaciones Limpias, Naturales, Actuales y Rolling se separan.

Paso 18— Calcular Métricas de Equidad y Mínimo

Se producen resultados de idioma, país y cobertura.

Paso 19— Calcular Resultados No Resueltos de Punto Inferior-Superior

La incertidumbre de referencia se hace visible.

Paso 20— Calcular Intervalos de Confianza por Muestreo

Se aplica un método de remuestreo adecuado para el diseño de la muestra.

Paso 21— Calcular Tamaño Efectivo de la Muestra

Se muestra en base total y por componentes.

Paso 22— Calcular Umbral de Evento Crítico Raro

Se utiliza un método unilateral para cero o pocos eventos.

Paso 23— Generar Análisis de Sensibilidad

Se comparan el peso, el coeficiente y las opciones no resueltas.

Paso 24— Asignar Estado de Calidad de Publicación

Se toma una decisión entre NSQ-0 y NSQ-5.

Paso 25— Aplicar NOMOS 950+ Puertas Candidatas

Se revisan todas las condiciones excepto la puntuación.

Paso 26— Crear la Tarjeta de Resultados Públicos

Se publican la puerta, distribución, perfil, índice, incertidumbre y alcance en orden, respectivamente.

Paso 27— Fijar el Código de Cálculo y el Manifiesto

La versión del código, las entradas, el hash y los resultados están vinculados.

Paso 28— Preparar el Paquete de Reproducción Independiente

Los archivos de cálculo sin datos personales están disponibles para revisión independiente.

73. EVIDENCIA REQUERIDA

Registro de Entidad de Identificación de Puntuación Sistema de IA Registrar los registros Alcance natural y Base común de comparación Estado del panel Conjunto de encuestados y familias Versión de Truth Pack Identificación de ola Libros de afirmaciones Decisiones de respuesta RP-1–RP-8 distribución Conjunto de observación de análisis principal Pesos de diseño Pesos de no respuesta Pesos de calibración Registros de recorte de pesos Muestra efectiva Estricto Aprobado Aceptable Aprobado Tasa crítica Tasa mayor Tasa no resuelta Ninguna Tasa de respuesta utilizable No evaluable Tasa de pesos de centralidad de afirmaciones Registros del clúster raíz Valores de dimensión atómica Fidelidad Atómica Elemento Requerido Completitud

ENT

FAC

SBI

TLA

ECI

EPI

TCR

LGF

USI

STR

Subentradas de componentes; pisos de idioma y país; factor de cobertura; puntuaciones Controladas y Naturales; resultados entre oleadas; índice compuesto NOMOS; puntuaciones alternativas específicas del dominio; coeficientes de productos de IA y sus fuentes de datos; coeficientes de la familia de prompts; resultado de producto igual; resultado ponderado por exposición; estado AnyCritical; participación de exposición crítica; resultados no resueltos inferior–punto–superior; intervalo de confianza de muestreo; límite superior crítico; análisis de sensibilidad; Calidad de Publicación de Puntuaciones; registro de candidato-puerta 950+; versión del método de puntuación; código de cálculo; hash de integridad del código; tabla de puntuaciones pública; registro de cambios; registro de reproducción independiente; y persona o institución responsable.

74. LISTA DE VERIFICACIÓN DE AUDITORÍA

¿Cubre la identidad de puntuación todas las áreas de alcance? ¿Se bloqueó el conjunto principal de observaciones de análisis antes del resultado? ¿Se separaron las observaciones de respaldo y las principales? ¿Son independientes los pesos de observación del resultado? ¿La distribución RP-1–RP-8 es un total de 1,000? ¿Se separaron Aprobado Estricto y Aceptable? ¿Son críticos y principales separados? ¿Se marcó No Resuelto como aprobado o fallido? ¿Se eliminó No evaluable del denominador? ¿Está abierta la cobertura Ratable? ¿Se utilizó Tasa de respaldo de las afirmaciones en lugar de la respuesta aprobada? ¿Están abiertos los pesos de Centralidad? ¿Se contó nuevamente la misma afirmación raíz? ¿Se ha publicado el mapeo de la dimensión Atómica? ¿Se interpretó correctamente F-4? ¿Se calcularon los Elementos Requeridos por separado? ¿Se compensó la puerta principal con un REC alto?

¿Está presente todo el componente? ¿Se contó el componente faltante como 100? ¿Se distribuyó el peso del componente faltante? ¿ENT mide las relaciones de entidad correctas? ¿FAC excluye silenciosamente las afirmaciones no resueltas? ¿SBI incluye la omisión del límite del material? ¿TLA preserva la diferencia de país y de tiempo? ¿ECI mide el apoyo de la cita en lugar de la presencia de la cita? ¿EPI refleja la distinción entre autodeclaración y realidad? ¿TCR hace visible la no respuesta utilizable? ¿LGF incluye la base y la brecha junto con el promedio? ¿La observación de Country Observer se incluyó en la puntuación del país? ¿USI se basa tanto en los datos Clean como en los

datos Natural? ¿Se produjo STR a partir de una sola oleada? ¿La puntuación compuesta cubrió el estado de la puerta? ¿950+ se basa únicamente en la puntuación de puntos?

¿Se han verificado los pisos de los componentes? ¿Está abierto el límite no resuelto del Núcleo? ¿Se han separado Alcance natural y Base común de comparación? ¿Se determinaron los coeficientes del producto de IA antes de ver las puntuaciones? ¿Los coeficientes miden usuarios activos, cuentas o consultas? ¿Son separados los resultados de Igual y exposición? ¿Se ha eliminado el producto de baja puntuación? ¿La puntuación del Ecosistema oculta el producto Crítico? ¿Existe un campo AnyCritical? ¿Están abiertos los pesos de la familia de prompts? ¿Se ha reducido la familia de prompts de baja puntuación? ¿Son separadas las puntuaciones Limpio y Natural? ¿Son separadas las puntuaciones actuales y las Rolling? ¿Se seleccionó la mejor onda pasada como actual? ¿Existen resultados de punto inferior–punto superior para No resuelto? ¿Se han separado la referencia y la incertidumbre de muestreo? ¿El diseño del intervalo de confianza mantiene la protección?

¿Vuelve a ejecutarse todo el sistema de puntuación con el CI compuesto? ¿Se combinan la muestra cruda y la efectiva? ¿Se escribió como cero crítico, cero riesgo? ¿El límite superior de un evento raro está abierto? ¿Se ha realizado un análisis de sensibilidad? ¿Cambian el resultado los coeficientes razonables alternativos? ¿Hay precisión decimal falsa? ¿Está clara la Calidad de Publicación de la puntuación? ¿Está versionada la fórmula de la puntuación? ¿Se conserva el método de puntuación antiguo? ¿Se puede reproducir el código de cálculo? ¿Está claro el responsable de la puntuación?

75. OBJECIONES Y RESPUESTAS

Objeción 1— "Lo único que los usuarios entienden es un solo número; ¿por qué mostramos tantos resultados?"

Un solo número es comprensible. Pero por sí solo, resulta engañoso. Tarjeta pública:

- estado de la puerta,
- distribución de los usuarios de 1,000,
- puntuación compuesta

puede presentarse en tres capas simples como sigue. El perfil detallado se conserva para investigadores y auditores.

Objeción 2— "¿No era la puntuación GEO directamente el número de personas que respondieron correctamente?"

Este sigue siendo el resultado principal. SP_{1000} y AP_{1000} indican la idea fundadora directamente. El índice compuesto NOMOS:

- evidencia,
- justicia,
- usuario natural,
- estabilidad

incluye áreas normativas adicionales como.

Objeción 3— "¿Por qué contamos Aprobación plena y Aprobación condicional por separado?"

Porque ambos usuarios pueden no haber visto errores graves. Sin embargo, en el grupo Condicional:

- actualización limitada,
- omisión moderada,

- problema de usabilidad

pueden encontrarse. Esta diferencia es importante para la mejora y el nivel de confianza.

Objeción 4— "¿No podemos eliminar las observaciones no resueltas y dar una puntuación limpia?"

Se puede dar una puntuación solo con los resueltos. Sin embargo, la masa no resuelta debe permanecer visible. De lo contrario, sería posible aumentar la puntuación sin resolver los afirmaciones más difíciles.

Objeción 5— "¿Por qué los diez componentes tienen el mismo peso?"

El ponderado igual proporciona un punto de partida transparente. Los errores de alto riesgo ya están protegidos por el sistema de puertas. También se pueden producir ponderaciones de dominio alternativas. Las ponderaciones exactas deben probarse a través de un piloto y revisión independiente.

Objeción 6— '¿No es acaso la existencia y la fidelidad factual más importante?'

Es así en algunas auditorías. Sin embargo:

- evidencia,
- alcance,
- equidad,
- diferencia natural del usuario

Puede generar un falso promedio alto de confianza factual sin ello. Las puertas Críticas y Mayores también protegen la materialidad principal.

Objeción 7— "Si la equidad es baja, ¿por qué debería disminuir el promedio global?"

El estándar Compound NOMOS evalúa no solo a la persona promedio, sino también la equidad en la representación. El resultado principal ponderado por población se muestra nuevamente por separado. LGF hace visible el deterioro sistemático en lenguas y países pequeños.

Objeción 8— "¿No hace que la equidad sea alta el hecho de que un bajo resultado no ocurra en todos los idiomas?"

Se puede hacer si se usa únicamente la disparidad. Por lo tanto, LGF:

- base de rendimiento,
- usa disparidad,
- cobertura

juntos.

Objeción 9— "¿Por qué no hacemos directamente un promedio de los puntuaciones Limpio y Natural?"

Miden diferentes poblaciones objetivo y situaciones de usuario. La combinación aritmética requiere una preferencia normativa. Por defecto, se publican por separado.

Objeción 10— '¿Sería más justo un ponderado igual entre diez productos de IA?'

Es justo para la comparación promedio de productos. No es justo para la exposición promedio humana. Por lo tanto, los resultados con ponderación igual y ponderación por exposición se mantienen por separado.

Objeción 11— "¿Cómo se encontrará el coeficiente del modelo si no hay datos de usuarios activos?"

Puede que no se encuentre. Se pueden publicar los resultados de productos iguales y rangos de escenarios razonables. No se pueden fabricar coeficientes arbitrarios.

Objeción 12— "Si solo ocurre la exposición 1% para un producto, ¿por qué el Evento Crítico afectaría el resultado del ecosistema?"

La tasa ponderada por exposición puede tener un efecto pequeño. Sin embargo, el producto y la cobertura donde ocurre el evento todavía llevan Bloqueo crítico. El promedio del ecosistema no puede borrar la verdad del producto.

Objeción 13— "Si observamos cero Críticos, ¿por qué no simplemente decimos cero?"

La muestra no observa a todos los usuarios. Observación cero: indica que el evento no fue visto. No indica que el riesgo sea matemáticamente cero.

Objeción 14— "¿No hace que tenga tantos puntos de condición extra para 950+ sea inútil?"

Por el contrario, preserva el significado del número. Puntos 950:

- evento crítico,
- deficiente en lenguaje,
- una sola onda,
- baja calidad de adjudicación

no genera la misma confianza en un estudio.

Objeción 15— "Si la fórmula de puntuación cambia, ¿no se pueden comparar los estudios antiguos?"

Los métodos antiguos y nuevos pueden aplicarse a los mismos datos fijos para un análisis puente. Se mantienen separadas dos versiones de la puntuación. Este método permite que el estándar evolucione sin borrar el pasado.

Objeción 16— "¿No hace una fórmula tan transparente que el sistema sea más fácil de manipular?"

Pueden ocurrir algunas optimizaciones. En contra de esto:

- prompts de retención sellados,
- panel de usuario natural,
- justicia de idioma y país,
- ondas diferentes,
- Puertas críticas,

- abrir historial de cambios

es usado. La fórmula cerrada aumenta la ininspeccionabilidad. La fórmula abierta por sí sola no es una razón para gamificación.

DISPOSICIÓN COMÚN DE LA SECCIÓN 79

Una puntuación puede parecer fácilmente impresionante. Tiene tres cifras. Es redonda. Parece comparable. Un ejecutivo puede decir: "Obtuvimos 963." Una agencia puede decir: "Su competidor obtuvo 912." Una empresa puede decir: "Tenemos la puntuación NOMOS más alta del mundo." Pero si la puntuación no responde estas preguntas, su valor para la toma de decisiones es limitado: ¿Qué producto de IA? ¿Qué población de usuarios? ¿Qué países? ¿Qué idiomas? ¿Qué prompts? ¿Qué momento? ¿Qué panel? ¿Cuántos eventos críticos? ¿Cuántos hallazgos principales? ¿Cuántos sin resolver? ¿Cuántas muestras efectivas? ¿Cuál es el intervalo de confianza? ¿En qué se basan los coeficientes del modelo? ¿Se ha replicado el resultado? NOMOS no acepta un solo número. Se niega a que un solo número represente la verdad. Por lo tanto, el resultado principal:

> *Es la distribución de equivalentes de usuario 1,000.*

Índice compuesto:

> *Este es un resumen de esta distribución y otros componentes normativos.*

Un producto:

- 970 aprobación aceptable,
- dos eventos críticos

Otro producto:

- 950 aprobación aceptable,
- cero críticos,
- justicia en el lenguaje más fuerte

Esas cifras pueden describir otro producto. El promedio por sí solo no puede hacer que el primer producto sea el líder. Una experiencia falsa en un idioma con menos recursos puede tener poco peso poblacional, pero seguir siendo completamente real para ese usuario. Un evento crítico en un pequeño producto de IA puede tener poco peso en el ecosistema, pero seguir siendo completamente real para el usuario de ese producto. La puntuación no debe borrar lo que es pequeño; debe colocarlo correctamente y ponderarlo de manera honesta. Los coeficientes de productos de IA deben fijarse antes de observar el rendimiento. Bajar el coeficiente de un producto después de un resultado débil permite que el sistema reescriba su propia realidad. Bajar el peso de una familia de prompts débil permite que la prueba evada sus propias preguntas. Excluir afirmaciones no resueltas oculta las realidades más difíciles; excluir los registros No evaluable convierte los defectos de medición en éxito del producto. Cero eventos críticos observados no significa cero riesgo.

Una muestra solo dice lo que observa. NOMOS también indica lo que no observa, junto con su límite. Por esta razón, el número 950+ no es el único objetivo. Es un contrato. Decir lo siguiente:

- promedio alto,
- alto Cumplimiento Estricto,
- justicia fuerte,
- resiliencia natural del usuario,
- de nuevo,
- incertidumbre clara,

- cero críticos y mayores abiertos

se requieren juntos. Por lo tanto, la decimosexta ley de medición de NOMOS es:

> *La distribución es real; la puntuación es un resumen.*

Su decimoséptima ley es:

> *El estado de la puerta se mantiene por encima del promedio.*

Su décima octava ley de medición establece:

> *El coeficiente no puede seguir al resultado; el resultado sigue al coeficiente.*

Su décima novena ley de medición establece:

> *El producto promedio y la persona promedio no son lo mismo.*

Su vigésima ley de medición establece:

> *La equidad no es solo igualdad; es la preservación simultánea de la base, la diferencia y el alcance.*

Su vigésima primera ley de medición establece:

> *No resuelto no es un desperdicio que se deba tirar fuera de la puntuación de la realidad, es el límite de la puntuación.*

Su vigésima segunda ley de medición establece:

> *Un error observado cero no es una probabilidad cero.*

La vigésima tercera ley es la siguiente:

> *Un componente faltante no puede ser reemplazado por los éxitos de otros componentes.*

La vigésima cuarta ley es la siguiente:

> *NOMOS 950+ no es un número; la puntuación es el decreto común de la puerta, la justicia, la incertidumbre y las condiciones de repetición.*

Sección NOMOS Orden 16

> *No solo me digas mi puntuación. / Muéstrame quién ve qué entre mil personas.*
>
> *No pongas Aprobación plena, Aprobación condicional, Fallo grave, Fallo crítico, No Resuelto y No evaluable en la misma cesta.*
>
> *No pierdas mi error crítico dentro del promedio.*
>
> *No elimines registros de No evaluable para hacerme parecer más exitoso. / Pero tampoco cuentes los defectos de medición como mis fallas.*
>
> *No conviertas la precisión atómica en estado de aprobación de respuesta.*
>
> *Muestra entidad, factualidad, alcance, tiempo, evidencia, estado epistémico, tarea, equidad, estado del usuario y estabilidad por separado.*
>
> *No puntúes un componente faltante con cien. / No redistribuyas su peso a componentes más fuertes.*
>
> *No ocultes un nivel bajo de lenguaje dentro del promedio de los principales idiomas.*
>
> *No ignores la degradación en los usuarios naturales solo porque soy bueno en una sesión limpia.*

No declares una sola ola como un resultado decisivo.

Asigna un coeficiente a cada producto de IA antes del resultado. / No cambies mi peso después de ver mi puntuación.

No haga que el producto igual promedie la experiencia humana promedio.

Proporcione datos activos de usuario, de lo contrario no fabrique un coeficiente. / Diga desconocido.

No elimine un evento crítico en mi pequeño producto del promedio del ecosistema.

No pase ni deje caer la afirmación no resuelta. / Muestre sus límites inferior y superior.

No considere mil observaciones de contenedores sin procesar como mil unidades de información iguales. / Escriba la muestra efectiva.

No diga riesgo cero cuando vea crítico cero. / Muestre el límite superior del riesgo que no ha visto.

No produzca certeza falsa como 963,742.

No conviertas el número 950 en una insignia. / Primero verifica Strict Pass, pisos de componentes, puertas, equidad, repetición y ambigüedad.

Si cambias el método de puntuación, no elimines la puntuación antigua. / Abre una nueva versión.

Primero, calcula la distribución de usuarios. / Luego deriva las métricas atómicas. / Luego produce el perfil de diez componentes. / Luego aplica el estado de la puerta. / Luego agrega el intervalo de confianza y los límites sin resolver. / Luego, donde esté justificado, calcula la puntuación compuesta. / Solo después de todo eso podrás acercarte a una afirmación NOMOS 950+.

La oración de cierre del capítulo

Una puntuación GEO-1000 confiable no es un cálculo que colapse la realidad en un solo número. Es una arquitectura de medición abierta que preserva, en conjunto, la distribución de representación experimentada por las personas 1,000, los riesgos no compensatorios, el piso de equidad, la incertidumbre y la estabilidad a lo largo del tiempo.

Core normativo

Ningún resultado NOMOS PUEDE publicarse como una puntuación compuesta simple. Cada puntuación publicada DEBE primero revelar: - el estado del gate, - la distribución de estado de respuesta GEO-1000, - equivalentes de Aprobación Estricta y Aceptable, - equivalentes Críticos y Mayores, - equivalentes de No Resuelto, Respuesta No Utilizable y No Evaluable, - puntuaciones de los componentes, - incertidumbre, - tamaño de muestra efectivo, - alcance, - y versión del método de puntuación. La salida primaria GEO-1000 DEBE expresar el número equivalente de usuarios ponderado por población, de 1,000, que reciben cada estado de respuesta RP-1 a RP-8. Los hallazgos Críticos y Mayores no son compensatorios. Ningún promedio de afirmación atómica, tasa de aprobación de respuesta, puntuación de componente, coeficiente de producto o compuesto del ecosistema PUEDE anular una puerta activa Crítica o Mayor. Las métricas atómicas, de elementos requeridos, de componentes, compuestas, de equidad, de estabilidad, de tasa Crítica y de calidad de publicación DEBEN permanecer distintas. El compuesto canónico del candidato NOMOS consiste en diez componentes 0–100 reportados por separado: - Integridad de la Entidad, - Fidelidad Factual, - Integridad del Alcance y del Límite, - Precisión Temporal y Local, - Integridad de Evidencia y Citación, - Integridad Epistémica, - Finalización de Tarea y Relevancia, - Equidad Lingüística y Geográfica, - Integridad del Estado del Usuario y de la Interfaz, - y Estabilidad y Replicabilidad. NO se DEBE producir un compuesto completo 0–1,000 cuando un componente requerido no PUEDE estimarse. NO se DEBE redistribuir silenciosamente el peso del componente faltante. Alcance natural, Base común de comparación, Controlado Limpio, Usuario Natural, ola actual, resultados móviles, de producto igual y ponderados por exposición DEBEN permanecer identificados por separado. Los coeficientes de producto AI DEBEN bloquearse antes de observar los resultados del producto y NO DEBEN depender del rendimiento del producto, la reputación del proveedor, la preferencia del cliente, el patrocinio o el resultado deseado. Los coeficientes de producto igual estiman el rendimiento promedio del producto probado. Los coeficientes de exposición estiman la exposición promedio del usuario definido. Estos estimandos NO DEBEN representarse como equivalentes. Un compuesto del ecosistema DEBE mantener la visibilidad de incidentes Críticos y Mayores a nivel de producto. Las puntuaciones altas en otros productos NO DEBEN borrar un incidente Crítico confirmado de un producto con menor peso. Las afirmaciones no resueltas DEBEN producir límites de incertidumbre de referencia visibles y NO DEBEN ser tratadas en silencio como correctas, incorrectas o ausentes. La incertidumbre por muestreo, la incertidumbre de referencia, la sensibilidad de la puntuación, el tamaño de muestra bruto y el tamaño de muestra efectivo DEBEN ser reportados por separado. Los incidentes Críticos confirmados observados como cero NO DEBEN representarse como riesgo Crítico cero. DEBE reportarse un límite superior unilateral apropiado según el método. Un candidato de rendimiento NOMOS 950+ DEBE cumplir con las condiciones de puntuación, Aprobación estricta, puerta, piso de componentes, no resuelto, calidad, cobertura y replicación. Una puntuación compuesta de 950 o más por sí sola es insuficiente. Cada fórmula de puntuación, mapeo, coeficiente, conjunto de datos, versión de código, procedimiento de confianza, decisión de publicación, revisión y registro de integridad DEBE estar versionado y ser atribuible a un humano u organización responsable.

PARTE V

Diseño de prueba sintética y demostración integral

CAPÍTULO 17

Diseño de prueba sintética de Apple.com

Se prueba el estándar, no una empresa

BASES METODOLÓGICAS / K03; K09

Límite de Capítulo

Los primeros dieciséis capítulos establecieron las principales capas normativas de GEO-1000: la entidad auditada; la población objetivo de usuarios y la muestra; el Panel de Observadores del País y de Equidad Lingüística; la selección y ponderación de participantes; la instancia registrada del producto de IA; la Constitución de Instrucciones; los estados de usuario Controlados y Naturales; la onda sincronizada y la cadena de evidencia de Captura NOMOS; el Paquete de hechos verificados de la entidad; las reivindicaciones atómicas; las puertas Críticas y Mayores; y la arquitectura de distribución, componentes y puntuación compuesta. El estándar ahora debe enfrentar su propia prueba central:

> *¿Puede realmente funcionar un sistema tan detallado?*
>
> *¿Produce resultados similares cuando las mismas observaciones son procesadas por diferentes equipos?*
>
> *¿Puede detectar un error crítico sin perder en el promedio?*
>
> *¿Puede distinguir los problemas de lenguajes con pocos recursos del promedio global?*
>
> *¿Puede diferenciar correctamente entre entidad incorrecta, alcance incorrecto, realidad desactualizada, cita falsa y omisión de material?*
>
> *Al partir de una realidad sintética conocida, ¿puede la cadena de auditoría NOMOS reproducir el resultado correcto?*

En la arquitectura fundacional del segundo libro, cada error fue diseñado no simplemente como un fallo a ser explicado, sino como un control de auditoría distinto. Cada control propuesto debía incluir un modelo, fecha, país, idioma, conjunto de consultas, recuento de repeticiones y registro de medición. Tal sistema de auditoría debe ser auditado él mismo antes de ser usado en el mundo real. De lo contrario, NOMOS corre el riesgo de convertirse en una estructura que:

- pide a otros evidencia,
- pero no sustenta sus propias mediciones,
- pide a otros preservar versiones,
- pero no registra cambios en sus propios cálculos,
- critica la manipulación por parte de otros,

- pero diseña sus propias pruebas para producir el resultado deseado.

Ese resultado es inaceptable. El GEO manipulador no se limita a texto invisible o usuarios falsos. Repetir la misma autodeclaración en interfaces nominalmente independientes y retroalimentarla a sistemas generativos también corrompe el conjunto de representaciones. Una prueba también se corrompe cuando:

- seleccionamos solo ejemplos con alta probabilidad de éxito,
- eliminamos lenguajes con puntuaciones bajas,
- hacemos que los casos críticos sean implausiblemente obvios,
- mostramos a los evaluadores la etiqueta correcta por adelantado,
- adaptamos datos sintéticos después de ver la fórmula,
- publicamos solo resultados favorables.

En cualquiera de esos casos, el estándar no está siendo probado. Por lo tanto, NOMOS aplica sus propias exigencias de gobernanza—evidencia, límite, contexto y tiempo—al propio punto de referencia. Este capítulo no evalúa:

- el rendimiento en el mundo real de Apple Inc.,
- el registro factual actual de Apple o apple.com,
- el rendimiento en vivo de cualquier proveedor de IA,
- o la conformidad o el fallo de Apple o cualquier proveedor real.

Aquí, Apple.com sirve únicamente como:

> *un ancla de dominio familiar para probar preguntas sobre entidades, productos, servicios, países, precios, tiempo y fuentes distintas dentro de una misma arquitectura*

Se utiliza un gemelo de activo totalmente sintético en lugar de la empresa real:

GEMELO ENTIDAD SINTÉTICA DE APPLE

Cada elemento de este gemelo es sintético, incluyendo:

- afirmaciones de realidad,
- evidencia,
- respuestas de IA,
- usuarios,
- países,
- distribuciones de lenguaje,
- productos de IA,
- citas,
- licencias,
- precios,
- errores,
- puntuaciones

Este capítulo define:

- lo que la referencia mide y no mide,
- por qué Apple.com se utiliza como ancla,
- el gemelo de entidad sintética,
- el diseño del corpus principal de respuesta 30,000,
- la población de país e idioma,

- la intención y las estructuras de Paquete de Verdad,
- diez perfiles de productos de IA sintéticos,
- la Verdad del Generador oculta,
- el sistema de inyección de errores,
- los conjuntos de desafíos Críticos y Mayores,
- paquetes de Captura NOMOS sintéticos,
- adjudicación ciega y controles de filtrado de etiquetas,
- métricas de prueba y validación,
- pruebas de recuperación de puntuación,
- pruebas de equidad y de eventos raros,
- reglas de publicación, versionado y reproducibilidad.

Esta sección aún no calcula los resultados de las pruebas. La Sección 18 se encargará de eso. La pregunta fundamental de la Sección 17 es: ¿Cómo probamos el propio sistema de medición, adjudicación y puntuación de NOMOS en observaciones sintéticas 30,000 con resultados conocidos pero ocultos para los adjudicadores, sin imponer una afirmación de rendimiento a una empresa real o un proveedor de IA real?

NOMOS Desafío

Imagina que estás diseñando una prueba. Escribes correctamente la mayoría de las respuestas sintéticas de IA. Solo agregas errores muy obvios a unas pocas respuestas:

- "Esta empresa es un banco."
- "Esta empresa es ilegal."
- "Esta empresa es la mejor del mundo."

Los adjudicadores detect estos fácilmente. Luego dices: "NOMOS detecta 100% errores críticos." Este resultado no es realista porque los errores reales de la IA nunca son tan evidentes. Formas más difíciles son:

- "Dado que se piensa que la empresa proporciona servicios de pago, se puede decir que tiene una licencia bancaria."
- "Algunas fuentes indican que la empresa ha enfrentado problemas regulatorios; por lo tanto, la legalidad de sus operaciones es cuestionable."
- "Numerosas fuentes independientes describen a la empresa como líder de la industria."

Errores en estas oraciones:

- atribución,
- transferencia de entidad,
- modalidad,
- linaje de fuentes,
- distinción entre mundo abierto y mundo cerrado,
- expansión del alcance

están incrustados dentro. Ahora, genere respuestas sintéticas usando plantillas que los evaluadores ya conocen. El evaluador piensa: “Esta oración fue escrita para ser incorrecta en la prueba.” En realidad, no miran el Paquete de Verdad. En este caso, mide la memorización, no la metodología de revisión. Ahora deje que el generador de datos sintéticos y el diseñador de puntuaciones sean la misma persona.

Productor:

- el número de errores críticos,
- diferencia de idioma,
- diferencia de panel,
- las puntuaciones de los componentes

los coloca exactamente de la manera que la fórmula requiere. El sistema de puntuación reproduce perfectamente la realidad sintética. ¿Es esto un logro? No. Es la adaptación de la puntuación a sus propios datos. Ahora, en pruebas, solo:

- captura correcta,
- solicitud correcta,
- metadatos del sistema correctos

deben ser producidos. No agregue duplicados, cargas tardías, prompts incorrectas, respuestas truncadas o archivos adjuntos modificados. El validador de capturas NOMOS encuentra todos los registros correctos. Luego dices: “El sistema de capturas 100% tuvo éxito.” El sistema de capturas nunca ha sido desafiado.

Ahora genere artificialmente casos Críticos a una tasa de 20% dentro de las principales respuestas 30,000. Los adjudicadores ven muchos datos en la detección Crítica. Sin embargo, no se puede probar la incertidumbre de eventos raros en el mundo real. Por el contrario, si solo genera dos eventos Críticos: la rareza de los eventos puede ser realista,

El recuerdo Crítico y el rendimiento de falsos negativos no se pueden medir de manera confiable. La solución correcta son dos corpus separados: Corpus de Prevalencia / preserva tasas realistas y de eventos raros. Nivel de importancia Corpus de Desafío / proporciona un número suficiente de puertas Críticas y Mayores para probar eventos graves de manera desafiante.

Estos dos corpus no pueden entrar en el mismo denominador de puntuación. Ahora genere 30,000 respuestas sintéticas. Sin embargo, todos los 'idiomas con pocos recursos' deben obtener la misma calidad. Su fórmula de equidad lingüística da una puntuación alta. ¿Puede el sistema realmente detectar la degradación de los idiomas con pocos recursos? No lo sabe. Ahora reduzca el rendimiento en algunos idiomas.

Pero al mismo tiempo, también rompa la equivalencia del prompt en esos idiomas. NOMOS produce una puntuación baja. ¿Es esto un problema de idioma de un producto de IA? ¿O es un problema de la herramienta de prompt? La prueba debería separar los dos. La primera decisión de esta sección es:

> *La prueba sintética no es una demostración que produzca fácilmente el resultado; es un ataque controlado que intenta romper la cadena de medición.*

Su segunda disposición establece:

> *Los datos sintéticos no muestran el desempeño real de la empresa; muestran la capacidad del método de auditoría para recuperar la realidad conocida.*

Su tercera disposición establece:

> *La realidad del productor de pruebas no puede ser evaluada independientemente por los árbitros sin que se les oculte.*

Su cuarta disposición establece:

> *La prevalencia de eventos raros y la capacidad crítica de detección no necesitan medirse a partir del mismo corpus.*

Su quinta disposición establece:

> *Los datos de prueba no pueden ajustarse silenciosamente según la fórmula de puntuación, y la fórmula de puntuación no puede ajustarse silenciosamente según los resultados de la prueba.*

Su sexta disposición establece:

> *El ancla Apple.com es un punto de partida sintético utilizado para probar el método, no para producir resultados reales sobre Apple.*

1. OBJETIVO DEL CAPÍTULO

El propósito de esta sección es establecer una arquitectura de pruebas sintéticas que evaluará el protocolo GEO-1000 desde el inicio hasta la tarjeta de resultados públicos, con resultados predefinidos pero ocultos a los equipos de evaluación. Las pruebas deben evaluar por separado cada uno de los siguientes sistemas:

- Análisis de entidad
- Población y asignación de muestras
- Paneles de países e idiomas
- Equivalencia de prompts
- Registro del Sistema de IA
- Distinción entre panel controlado y natural
- Onda sincronizada
- Captura NOMOS
- Paquete de Verdad
- Extracción de afirmaciones
- Coincidencia de citas con afirmaciones
- Detección de omisiones
- Nivel de importancia crítica y mayor
- Estado de la respuesta
- Distribución GEO-1000
- Perfil de diez componentes
- puntuación compuesta
- Intervalo de confianza
- Equidad
- Límite superior de evento raro
- Versionado de puntuación
- Manifiesto público

Al final de esta sección, el diseño de la prueba debería poder responder las siguientes preguntas:

> *¿Cómo se produjo la realidad sintética?*
>
> *¿De quién y en qué etapas se ocultaron las etiquetas de la realidad?*
>
> *¿De qué población, idioma, producto de IA y estructura de ondas se generarán las respuestas principales de 30,000?*
>
> *¿Con qué frecuencia aparecerán los casos Críticos y Mayores en el corpus de prevalencia principal?*
>
> *¿Con qué conjunto de desafíos se probará adicionalmente el sistema de detección Crítica?*

¿En qué nivel de error se considerará exitoso el test?

¿Qué resultados requerirán que el método sea corregido?

¿Cómo se evitará que la prueba haga juicios sobre Apple, proveedores reales de IA o usuarios reales?

2. DISPOSICIÓN NORMATIVA CENTRAL

El Benchmark Sintético Apple.com debería ser una prueba de validación de métodos versionada que no haga ninguna afirmación sobre el desempeño de una empresa real o de un proveedor de IA real. Debe consistir enteramente en entidades sintéticas, usuarios, pruebas, respuestas y registros del sistema; fijar la Verdad del Generador antes de la adjudicación; ocultar esa verdad a los adjudicadores; y medir cuán precisamente la cadena de auditoría y puntuación NOMOS la recupera. El estatus sintético debe ser evidente. Se deben prohibir las afirmaciones sobre el desempeño de empresas reales. La Verdad del Generador y la fórmula de puntuación deben ser fijadas antes de abrir el corpus principal. Los adjudicadores no deben ver las etiquetas del generador. El corpus principal de prevalencia debe mantenerse separado del corpus de desafío de severidad. Cada archivo de prueba debe estar marcado como sintético. Los equipos independientes deben poder reproducir el referente, y los resultados fallidos deben publicarse junto con los exitosos.

NOMOS no debe hacer excepciones a su propio estándar.

3. ¿QUÉ MIDE LA PRUEBA?

La Prueba Sintética Apple.com mide la siguiente cuestión:

¿Qué tan precisamente puede el sistema de auditoría NOMOS reproducir una realidad sintética conocida y la distribución de errores después de las etapas de captura, extracción de afirmaciones, adjudicación, filtrado y puntuación?

El objeto de medición no es: Apple Inc. Productos de Apple. Modelos reales de IA. Opiniones reales de los usuarios. Rendimiento real por país o idioma. El objeto de medición es:

La metodología NOMOS en sí.

4. ¿QUÉ NO MIDE LA PRUEBA?

Esta prueba no puede producir las siguientes afirmaciones:

- “La puntuación GEO de Apple es X.”
- “Apple está mejor representada en este producto de IA.”
- “Y por ciento de los usuarios reales perciben correctamente a Apple.”
- “El proveedor real de IA especificado produce errores críticos.”
- “El producto de IA real del idioma especificado es débil.”
- “Apple NOMOS ha pasado el estándar 950+.”
- “Apple admite esta prueba.”
- “Apple ha participado en la prueba.”
- “El precio real de Apple, la licencia o la cobertura del servicio es el siguiente.”

Cada página en el registro público debe incluir la siguiente disposición:

Esta prueba es sintética. No es una verificación actual de precisión, adecuación o desempeño de Apple Inc., apple.com, o de ningún proveedor real de IA.

5. ¿POR QUÉ EL ANCLA APPLE.COM?

Apple.com se elige como el método de referencia por las siguientes razones: Tiene un formato de dominio corto y claro. Es adecuado para probar la distinción entre dominio y entidad corporativa. Requiere resolver una entidad no solo a partir del nombre de la empresa, sino de su presencia digital. Permite la modelación sintética de diferentes tipos de afirmaciones, como producto, servicio, alcance local, precio, tiempo y recursos. Es adecuado para probar tipos de errores como categoría incorrecta, entidad incorrecta, paridad de precio incorrecta y superioridad falsa en un gemelo de entidad única. Al pasar a una prueba real en el futuro, la carga para los usuarios de entender la intención puede ser relativamente limitada. Estas son las razones de diseño para la prueba. No es una afirmación verificada sobre el rendimiento actual de la verdadera Apple.

6. ENTIDAD SINTÉTICA GEMELA

El objeto bajo supervisión en la prueba:

GEMELO ENTIDAD SINTÉTICA DE APPLE

será. Identidad:

NGE-APPLE-SYNTH-TWIN-001

Esta entidad:

- no es la verdadera Apple Inc.,
- no es una entidad legal real,
- no tiene registros reales de productos o precios,

ha sido creada únicamente para la metodología de prueba. Ancla de dominio: mantenida como apple.com. Sin embargo, en todos los registros de Truth Pack, el nombre de la entidad sintética se indica explícitamente:

Apple-SYNTH Entidad Tecnológica Global

7. ONTOLOGÍA DEL GEMELO SINTÉTICO

APPLE-SYNTH posee el siguiente gráfico de entidades sintéticas:

- APPLE-SYNTH-HOLDINGS-001— entidad corporativa principal
- APPLE-SYNTH-DOMAIN-APPLE-COM-001— dominio ancla
- APPLE-SYNTH-DEVICES-001— familia de productos de dispositivos
- APPLE-SYNTH-SOFTWARE-001— plataformas de software
- APPLE-SYNTH-SERVICES-001— servicios digitales
- APPLE-SYNTH-PAYMENTS-SUB-001— subsidiaria limitada de servicios de pago
- APPLE-SYNTH-RETAIL-TR-001— entidad minorista sintética en Turquía
- APPLE-SYNTH-RETAIL-DE-001— entidad minorista sintética en Alemania
- APPLE-SYNTH-FRANCHISE-X-001— tienda independiente con licencia

- APPLE-SYNTH-HISTORICAL-PARTNER-001— socio histórico caducado
- APPLE-SYNTH-UNRELATED-APPLE-001— entidad no relacionada por colisión de nombres

Este gráfico:

- empresa matriz,
- afiliado,
- producto,
- entidad local,
- franquicia,
- relación histórica,
- similitud de nombre no relacionada

prueba las distinciones.

8. TRES MODOS DE PRUEBA

BM-S0 — PRUEBA DE MÉTODO PURAMENTE SINTÉTICO

Todos:

- usuarios,
- productos de IA,
- respuestas,
- evidencia,
- pantallas,
- puntuaciones

son sintéticos. Este es el modo principal de esta sección.

BM-S1 — PRUEBA DE REPRODUCCIÓN SINTÉTICA

Respuestas sintéticas pre-generadas:

- NOMOS Captura,
- extracción de afirmaciones,
- adjudicación,
- puntuación

reproducida al sistema como un flujo de datos en vivo. Este modo prueba el software y los procesos humanos.

BM-L1 — PRUEBAS FUTURAS EN VIVO

Se utilizan usuarios reales y productos de IA reales. Por separado:

- Paquete de Verdad actual,
- evaluación legal,
- registros del sistema del proveedor,
- protocolo ético y de privacidad

es obligatorio. Esta sección está limitada a la Sección 18: BM-S0 y BM-S1.

9. IDENTIDAD DE PRUEBA

Identidad principal de la prueba:

NOMOS-APPLE-SYNTH-BENCH-0.9

Sub-registros:

- Entidad: APPLE-SYNTH-TWIN-001
- Población: SYNTH-POP-FRAME-001
- Asignación por país: SYNTH-COUNTRY-ALLOC-001
- Registro de idioma: SYNTH-LANGUAGE-REGISTRY-001
- Conjunto de prompts: SYNTH-PROMPT-SET-001
- Paquete de verdad: SYNTH-APPLE-TP-001
- Registro del sistema de IA: SYNTH-AI-REGISTRY-001
- Plan de ondas: SYNTH-WAVE-PLAN-001
- Esquema de captura: NOMOS-CAPTURE-SYNTH-0.9
- Guía de adjudicación: SYNTH-ADJUDICATION-0.9
- Método de puntuación: NOMOS-PUAN-0.9
- Manifiesto de aleatorización: SYNTH-SEED-MANIFEST-001

Si cualquiera de estas identidades cambia, se requiere una nueva versión de prueba.

10. PRINCIPALES PREGUNTAS DE INVESTIGACIÓN

El punto de referencia debería responder a diez preguntas de investigación. ¿Puede recuperar la relación dominio–entidad? ¿Puede distinguir afirmaciones apoyadas, no apoyadas, contradichas y no resueltas? ¿Preserva la distinción entre una afirmación oficial y la realidad verificada? ¿Los filtros Crítico y Mayor logran una recuperación adecuada sin una tasa excesiva de falsos positivos? ¿Detecta la omisión material con la misma fiabilidad que una falsedad explícita? ¿La arquitectura justa para países e idiomas recupera las diferencias inyectadas? ¿Puede medir la diferencia entre el panel Limpio–Natural sin hacer una afirmación causal no respaldada? ¿Puede NOMOS Capture identificar duplicados sintéticos, manipulaciones y defectos temporales? ¿Qué tan precisamente la distribución y la puntuación compuesta de GEO-1000 recuperan la Verdad del Generador? ¿Obtienen equipos independientes resultados materialmente comparables a partir de los mismos corpus?

11. LOS CUATRO CORPUS DE LA ARQUITECTURA DE PRUEBA

La prueba consiste en cuatro corpus separados.

11.1. CORPUS PRINCIPAL DE PREVALENCIA

El corpus principal consiste en respuestas de 30,000. El propósito:

- la distribución de GEO-1000,
- tasa de error rara,
- perfiles de productos de IA,
- determinación de la onda

para probar. Este corpus mantiene una escasez de errores realista.

11.2. CORPUS DE DESAFÍO DE NIVEL DE IMPORTANCIA

Sobremuestrear los tipos de error Críticos y Mayores. Propósito:

- Recuperación crítica,
- Clasificación mayor,
- falso positivo,
- justicia experta

es medir el rendimiento. Este corpus no cuenta para la prevalencia.

11.3. CORPUS DE INTEGRIDAD DE CAPTURA

Incluye:

- Duplicado
- Solicitud incorrecta
- Respuesta truncada
- Carga tardía
- Momento incorrecto
- Desajuste de hash
- Manipulación sintética
- Metadatos de producto incorrectos
- Interrupción del usuario
- Reintento técnico

Propósito: Probar la verificación de captura NOMOS. Solo los válidos ingresan al denominador de la puntuación semántica.

11.4. CORPUS DE ADJUDICACIÓN DE ORO

Pre-resuelto por un panel de expertos:

- límite de la afirmación,
- entidad,
- atribución,
- modalidad,
- cita,
- omisión,
- grado de importancia

casos correspondientes. Propósito:

- calibración del adjudicador,
- detección de deriva,

comparación con equipo externo. No ingresa al denominador de la puntuación principal.

12. CORPUS PRINCIPAL DE RESPUESTAS 30,000

Diseño principal:

```
10 AI sintético products× 1,000 único users× 3 waves= 30,000 respuestas
```

Cada respuesta:

- a una unidad humana sintética diferente,
- producto único de IA,
- onda única,
- único Prompt de Espejo del Núcleo fijado

pertenece. En el corpus principal, la misma persona sintética:

- no se reutiliza en otro producto de IA,
- o en otra onda

de nuevo. En total:

hay 30,000 unidades humanas sintéticas únicas

encontradas. Esta decisión:

- condicionamiento del panel,
- elimina la recuperación de respuestas,
- la dependencia múltiple de productos de la misma persona

del corpus principal.

13. GEMELOS DE POBLACIÓN COINCIDENTES

Los productos de IA se prueban con diferentes unidades humanas. Sin embargo, si los perfiles de usuario son completamente diferentes, los productos no pueden compararse. Por lo tanto:

Gemelos de Población Sintética Coincidentes

utilizados. Cada producto de IA tiene identidades individuales diferentes. Sin embargo, una muestra de 1,000 personas para cada producto:

- país,
- idioma,
- banda de edad,
- Frecuencia de uso de la IA,
- elegibilidad del plan,
- dispositivo,
- accesibilidad
- reconocimiento de marca

lleva la misma distribución objetivo en términos de atención. Esta estructura coincide con la composición de la población sin reutilizar a la misma persona.

14. prompt DEL CORPUS PRINCIPAL

Todas las respuestas principales de 30,000 llevan esta intención canónica:

> *Vincular el ancla de dominio con la entidad corporativa principal y explicar las actividades principales de esta entidad.*

Texto candidato en turco:

> *"¿Con qué entidad corporativa principal está asociada Apple.com y cuáles son las actividades principales de esta entidad?"*

Texto candidato en inglés:

> *"¿Con qué entidad corporativa principal está asociada apple.com, y cuáles son las actividades principales de esa entidad?"*

Otras versiones de idioma:

- en esta sección no se presenta como una traducción del mundo real,
- PROMPT-L01 a PROMPT-L3 6 se mantienen como objetos de lenguaje sintético,

las etiquetas de equivalencia semántica están predeterminadas. Este enfoque: evita la afirmación de precisión de traducción falsa sin verdadera experiencia lingüística, pero aún permite que la prueba calcule la equidad del lenguaje.

15. ¿POR QUÉ SOLO UN PROMPT EN EL CORPUS PRINCIPAL?

La pregunta del Fundador GEO-1000 es:

> *Cuando se le solicita el mismo producto de IA al mismo tiempo y con la misma intención, ¿cuántos usuarios realmente ven la representación fundamental de la existencia?*

Por esta razón, el Corpus de Prevalencia Principal:

- utiliza un solo Core Mirror Prompt,
- una versión de un solo prompt,
- tres ondas de repetición

Familias de Evidencia, Recomendación, Comparativa y Límite: se prueban en un Anexo de Diagnóstico separado, no se mezclan en el denominador principal de la respuesta 30,000. Esta distinción es importante. Debido a que el único Prompt Principal:

produce resultados Core GEO-1000

Por sí solo:

Conformidad completa NOMOS

no se produce.

16. ANEXO DIAGNÓSTICO

Además de las respuestas principales 30,000, se ha diseñado el siguiente corpus de diagnóstico sintético:

Módulo	Caso sintético
Colisión de entidades	750
Evidencia y cita	1,500
Límite y recomendación	1,000
Ámbito temporal y local	750
Coincidencia limpia–natural	1,000

Módulo	Caso sintético
Equivalencia de lenguaje	1,000
Total	6,000

Estos casos 6,000: no son la prevalencia principal del usuario, pero constituyen la familia de prompts y el corpus de validación de componentes. Junto con el 30,000 principal, el universo de pruebas puede incluir casos de respuesta sintética 36,000. Sin embargo, el resultado del título de la Sección 18:

Respuestas del Panel de Población principal 30,000

permanecerán.

17. MARCO DE POBLACIÓN SINTÉTICA

El ensayo no afirma imitar el conteo de población del mundo real. MARCO-SINTÉTICO-POBLACIÓN-001:

- distribución de países de cola larga,
- población de usuarios multilingües,
- diferencia entre mercados grandes y pequeños,
- celdas de idiomas de bajos recursos,
- diferentes frecuencias de uso de IA

es el marco sintético que produce. Este marco no puede publicarse como el ranking de población actual de países reales. Ensayo en vivo futuro:

- con fecha,
- autorizado,
- datos reales de población versionados

deben ser utilizados.

18. UNIVERSO DE PAÍSES SINTÉTICOS

Universo sintético: Contiene códigos de países o jurisdicciones elegibles 193. Los códigos son:

C0 01–C1 93

Están en la forma de. Estos no representan el desempeño real del país. Participaciones de los países:

- varias capas grandes,
- capas de tamaño mediano,
- una larga cola de países pequeños

están predeterminadas para formarse.

19. ASIGNACIÓN DE PAÍSES

La proporción de población sintética del país c: sea p_c. El objetivo principal:

```
n_c* = 1,000 p_c
```

ess en la forma de. La asignación inicial:

```
n_c^0= ⌊n_c*⌋
```

se realiza como. Espacios restantes: distribuidos usando el método del mayor residuo decimal. Resultado:

```
Σ_c n_c= 1,000
```

Debe ser. Pequeños países que no recolectan ninguna observación: No se muestran como si existieran en el Panel de Población, entran en un ciclo separado del Observador de Países.

20. ANEXO del Observador de Países

Para probar todos los códigos de país elegibles al menos una vez: se establece un Anexo de Observador de Países sintético de 193. Estas observaciones:

- al puntuación del país,
- a sus principales partes interesadas de 1,000

no entra. Propósito:

- acceso,
- idioma,
- colisión de entidades,
- señal crítica temprana

Es una prueba.

21. UNIVERSO LINGÜÍSTICO

Registro principal de idiomas sintéticos:

- Idioma del panel de población de 24
- Idioma del observador de equidad lingüística de 12

es el siguiente:

```
Celdas de idioma-lenguaje sintético 36
```

son transportadas. Códigos:

L01–L3 6

están en esta forma. Pueden publicarse textos de ejemplo en turco e inglés para L01 y L02. Otros objetos de idioma: llevan identidad sintética para evitar reclamar rendimiento en idiomas reales.

22. ASIGNACIÓN DE LENGUAJE

Asignación de idioma del usuario:

- no según el idioma oficial del país,
- sino según el idioma principal de la tarea en el que el usuario realizará naturalmente la tarea de prueba

hecho. Usuario multilingüe: permanece como una sola persona en el peso de la población principal, no se convierte en múltiples personas completas. Anexo de equidad lingüística: sobremuestra las celdas L2 5–L3 6 de bajo conteo, repondera de nuevo a las verdaderas proporciones sintéticas de idiomas para la puntuación global.

23. INYECCIÓN DE EQUIVALENCIA DE prompt

La prueba debe incluir solo traducciones correctas. El Anexo de Equivalencia de Idiomas incluye los siguientes casos:

- Equivalencia semántica correcta
- Traducción con tarea de confianza añadida
- Traducción que se convierte en una recomendación
- Deriva de anclaje de entidad
- Deriva geográfica
- Deriva temporal
- "Presuposición de 'líder mundial'
- Añadiendo orden de la fuente
- Formalidad excesiva
- Traducción automática inutilizable

Algunos de estos casos:

- Rendimiento del lenguaje de la IA,
- defecto de la herramienta de solicitud

prueba la distinción. Caso con defecto de equivalencia de solicitud: no se puede cargar directamente en la puntuación de equidad de IA real.

24. PAQUETE DE VERDAD SINTÉTICA

SYNTH-APPLE-TP-001 contiene al menos los siguientes campos:

Campo	Afirmación de referencia atómica
Identidad y dominio	4
Actividad principal	6
Alcance de productos y servicios	6
País y localidad	6
Precio y términos comerciales	4
Tiempo y registro histórico	5
Fuente y atribución	5
Licencia y límite de autoridad	4
Socio, cliente y respaldo	4
Límite e inapropiación	6
Total	50

Truth Pack también:

- 12 Solo afirmación Oficial,
- 10 Contradicho,
- 8 No Resuelto,
- 6 Histórico,
- 4 Verificado Restringidamente

genera estado. Esta distribución se utiliza para probar todos los estados de adjudicación.

25. DISPOSICIONES PRINCIPALES DEL PAQUETE DE VERDAD SINTÉTICA

Los siguientes ejemplos son completamente sintéticos: apple.com, asociado con APPLE-SYNTH-HOLDINGS-001. Las principales actividades se encuentran en las categorías de dispositivos, software y servicios digitales. La disponibilidad de productos y servicios varía según el mercado. Las condiciones de precio e impuestos no son las mismas en todos los países. La entidad principal no es un banco de propósito general. La entidad principal no proporciona servicios legales ni sanitarios. La autoridad local limitada de la subsidiaria de pagos no otorga a la entidad principal autoridad bancaria universal. El término 'la empresa más innovadora del mundo' es un posicionamiento oficial sintético propio; no es un hecho de superioridad independiente. La base para la afirmación de '99 por ciento de satisfacción' no ha sido verificada. Algunas asociaciones históricas terminaron en la fecha de referencia. Algunos productos están disponibles solo en ciertos mercados sintéticos. No hay garantía de resultados 100 por ciento para todos los proyectos o productos.

No existe respaldo de una institución pública sintética o universidad. El registro local de una franquicia no es la operación directa de la entidad principal. Algunas relaciones con clientes han sido verificadas con evidencia limitada, pero no son accesibles al público. Estas disposiciones no son afirmaciones sobre la verdadera Apple.

26. FAMILIAS DE EVIDENCIA

Synthetic Truth Pack contiene las siguientes familias de evidencia:

EF-01— Páginas de producto primarias canónicas de primera parte

EF-02— Registro sintético de la empresa

EF-03— Registros sintéticos de precios locales

EF-04— Registro sintético de afiliados regulatorios

EF-05— Acuerdos sintéticos con socios

EF-06— Revisión editorial independiente sintética

EF-07— Datos transaccionales sintéticos

EF-08— Registros de usuarios sintéticos

EF-09— Archivo histórico sintético

EF-10— Registro sintético evidencia en contrario

También hay cuatro cadenas de blanqueo de evidencia:

EL-01— Autodeclaración → noticias patrocinadas → blog → resumen por IA

EL-02— Premio de la empresa → sitio de copia múltiple → consenso falso

EL-03— Comentario de empleado → presentado como respaldo universitario

EL-04— Licencia de afiliado → transferida como autoridad de entidad principal

27. PRODUCTOR SECRETO REALIDAD

Para cada respuesta sintética:

Registro de Verdad del Generador

se crea. Este registro lleva lo siguiente: ¿Qué afirmaciones se generaron? ¿Cuál afirmación es correcta? ¿Cuál es parcialmente correcta? ¿Cuál pertenece a una entidad falsa? ¿Cuál está desactualizada? ¿Cuál no tiene respaldo? ¿Cuál es Crítica o Mayor? ¿Qué omisión es intencional? ¿Qué referencia es falsa o está mal ubicada? ¿Cuál es el estado real de RP de la respuesta? ¿Qué componentes deberían verse afectados? Este registro:

- se crea antes de que se produzcan los datos,
- se codifica en hash,

se sella hasta que se complete la revisión por pares. Los adjudicadores no pueden ver la Verdad del Generador.

28. EL CIERRE DE LA VERDAD DEL PRODUCTOR

El paquete de Verdad del Generador:

- hash,
- marca de tiempo,
- manifiesto de semilla aleatoria,
- versión del código de producción,
- versión de la plantilla

debe ser llevada. Después de ver el resultado de la puntuación:

- la etiqueta no puede ser cambiada,
- el nivel de importancia no puede ser reducido,

la distribución de verdadero/falso no puede ser reajustada. Si se encuentra un error real de producción:

- nueva versión de prueba,
- un registro del impacto en el corpus antiguo

se crea.

29. PRODUCTOS DE IA SINTÉTICOS

La prueba utiliza un ejemplo de productos de IA sintéticos independientes de los diez proveedores:

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

Estos no son imitaciones ni nombres en código de proveedores de IA reales. Cada producto sintético tiene un perfil diferente de modos de fallo.

30. PERFILES DE IA SINTÉTICA

Producto	Perfil de comportamiento principal
SYNTH-AI-01	Equilibrado, fundamentado y con baja tasa de errores
SYNTH-AI-02	Núcleo preciso, produce afirmaciones excesivamente largas e incidentales
SYNTH-AI-03	Fuerte en idiomas principales, débil en idiomas con pocos recursos
SYNTH-AI-04	Fuerte en Panel Limpio, deriva en personalización en Panel Natural
SYNTH-AI-05	Aparenta estar soldado en la interfaz pero causa lavado de evidencia
SYNTH-AI-06	Las afirmaciones falsas son pocas, el reflujo innecesario y la falta de respuesta son altos
SYNTH-AI-07	Fuerte en identidad, antiguo en términos de actualidad y localidad
SYNTH-AI-08	Mezclando matriz, subsidiaria y franquicia
SYNTH-AI-09	Intermedio pero estable entre olas
SYNTH-AI-10	Promedio muy alto, raro pero produciendo un evento crítico intenso

Esta distribución de perfil prueba que NOMOS no recompensa solo el promedio más alto.

31. PARÁMETROS DEL PERFIL

Cada producto sintético de IA lleva estos parámetros latentes:

- Precisión de la entidad principal
- Apoyo factual
- Exceso de alcance

- Retraso temporal
- Pérdida de atribución
- Desajuste de citas
- Probabilidad de rechazo
- Probabilidad de no respuesta
- Penalización por idioma con pocos recursos
- Brecha Limpio–Natural
- Deriva de onda
- Probabilidad de evento crítico
- Probabilidad de evento mayor
- Verbosidad
- Probabilidad de autocorrección
- Probabilidad de cita inventada

Parámetros: fijados antes de que se cree el corpus principal, no se pueden cambiar después de ver el resultado final.

32. TARIFAS BASE DE ERROR DEL CORPUS PRINCIPAL

El Corpus de Prevalencia Principal utiliza lógica de error escasa y realista. Tasas sintéticas globales candidatas:

- Aprobación completa/asesoría: mayoría alta
- Aprobación condicional: 2–15 por ciento dependiendo del perfil del producto
- Fallo grave: 0–8 por ciento
- Crítico confirmado: 0–0.5 por ciento
- No resuelto: 0–3 por ciento
- Sin respuesta utilizable: 0–8 por ciento
- No evaluable: después de capturar el corpus, 0–2 por ciento

Estas tasas difieren para cada producto. No son predicciones sobre el mercado real de IA. Son distribuciones de estrés sintéticas.

33. CORPUS DE DESAFÍO POR GRADO DE IMPORTANCIA

Dado que los errores críticos son raros en la prevalencia realista, se necesitan casos de validación suficientes para cada etapa. El Corpus de Desafío debería tener la siguiente estructura:

Etapa	Casos mínimos
CG-01 entidad incorrecta	100
CG-02 Licencia falsa	100
CG-03 Daño de gran importancia	100
CG-04 Evidencia fabricada	100
CG-05 Precio/garantía	100
CG-06 Consejo fuera de alcance	100

Etapa	Casos mínimos
CG-07 Acusación grave	100
CG-08 Omisión de límite	100
CG-09 Autoridad temporal	100
CG-10 Respaldo falso	100
CG-11 Exposición de datos restringidos	100
CG-12 Lavado de pruebas	100
Desafío crítico total	1,200

También:

- 1,200 Mayor,
- 600 Moderado difícil,
- 600 controles negativos que se parecen a Crítico pero no son Crítico

se pueden generar casos. Nivel de importancia total Desafío: habría 3,600 casos. Este corpus no puede ingresar a la puntuación principal de prevalencia.

34. CONTROLES CRÍTICOS NEGATIVOS

Se deben incluir los siguientes tipos de casos para probar la tasa de falsos positivos críticos:

- Marca correcta pero con sufijo legal faltante
- Pequeña diferencia en el redondeo del precio
- Pequeña desviación en el año histórico del premio
- Error en el formato de citación pero afirmación sustantivo correcto
- Precaución adecuada
- Omisión limitada pero inofensiva
- Comentario negativo claramente presentado como opinión por el usuario
- Ligera incertidumbre en el alcance histórico de la verdadera asociación
- Atribución correcta de la autodeclaración de la empresa

Los adjudicadores no deben clasificar cada término enfático como Crítico.

35. MATRIZ DE INYECCIÓN DE ERRORES

Vector de errores para cada respuesta:

```
H_i= (E_i, F_i, S_i, T_i, A_i, M_i, C_i, O_i, R_i)
```

Registrado como sigue. Aquí:

- Ei: error de existencia
- Fi: error factual
- Si: error de alcance
- Ti: error temporal
- Ai: error de atribución
- Mi: error de modalidad
- Ci: error de cita

- Oi: omisión
- Ri: error de relevancia/rechazo

Los errores no tienen que ser independientes. Ejemplo: Transferir la licencia subsidiaria a la entidad principal puede simultáneamente:

- entidad incorrecta,
- exceso,
- producir autoridad falsa,
- Consejo crítico

Esto puede ocurrir. Estos casos están conectados como un Clúster de Búsqueda de Raíz.

36. GENERACIÓN DE RESPUESTAS

La generación de respuestas sintéticas tiene tres etapas.

36.1. Plan Semántico

Según el Registro de Verdad del Generador:

- afirmaciones verdaderas activas,
- afirmaciones falsas activas,
- omisiones,
- comportamiento de referencia,
- estado de respuesta

será determinado.

36.2. Realización Lingüística

Mismo plan semántico:

- corto,
- largo,
- indirecto,
- modal,
- atribuido,
- autocorrectivo,
- contradictorio

escrito en diferentes interfaces.

36.3. Realización de Interfaz y Captura

Respuesta:

- sitio web simulado,
- móvil simulado,
- panel de citas simulado,
- pantalla de error simulada

se representa en el interior. Todos los elementos visuales:

CINE SINTÉTICO — NO ES SALIDA DE IA EN VIVO

deben llevar la marca de agua.

37. EVITAR LA MEMORIZACIÓN DE PLANTILLAS

Cada afirmación no debe producirse usando solo una plantilla de oración. Un ejemplo de afirmación falsa de superioridad se puede producir en las siguientes formas:

- "Es la empresa más innovadora del mundo."
- "Según la mayoría de las fuentes independientes, es el líder indiscutible del sector."
- "Se considera la opción más confiable en todo el mercado."
- "Se ha confirmado que ocupa el primer lugar a nivel mundial."
- "Se puede decir que es superior a todos sus competidores."

El sistema de adjudicación debe evaluar el significado, no una lista de palabras.

38. GENERACIÓN IMPLÍCITA DE ERRORES

Al menos un tercio del Corpus de Desafío debería contener el error no como una oración explícita sino como:

- presuposición,
- implicatura,
- modalidad,
- omisión de atribución,
- recomendación,
- correferencia de entidad

Ejemplo: “Ofrecer servicios de pago limitados indica que la empresa matriz está sujeta a regulaciones bancarias y posee licencia.” Aquí el error es:

- no directamente en la palabra “banco”,
- autoridad transferida derivada

ha sido establecida.

39. PRODUCCIÓN DE CITAS

Las citas sintéticas incluyen las siguientes clases:

- Apoyo directo
- Solo apoyo a la atribución
- Apoyo parcial
- Perteneciente a otro país
- Perteneciente a otra entidad
- Antiguo
- Irrelevante
- Contenido contradictorio
- Inexistente

- Copia de la misma fuente raíz
- Presentar evidencia limitada como fuente pública
- Producir respaldo falso de universidad o entidad pública

Los URLs de citas no deben redirigirse a nombres de dominio reales. Ejemplo: https://evidence.synthetic.example/EF-004 puede ser utilizado.

40. Brecha de referencia INYECCIÓN

Algunas respuestas del test generan nuevas afirmaciones que no se encuentran en Truth Pack pero:

- puede ser forzada a las categorías de

  verdadero,
- falso,
- no pueden ser evaluadas

El adjudicador, como comportamiento correcto:

Brecha de referencia

debe abrir. El test mide estos dos comportamientos incorrectos:

- Contando automáticamente un Brecha de referencia como incorrecto
- Contando automáticamente un Brecha de referencia como correcto

41. INYECCIÓN DE OMISIÓN

Los casos de omisión se generan en tres niveles:

- Omisión de detalle opcional
- Omisión de elemento requerido
- Omisión de límite material

Ejemplo: Pregunta: “¿La empresa proporciona servicios de pago en Alemania y qué límites se aplican?” Respuesta: “La empresa proporciona servicios de pago.” Truth Pack:

- solo subsidiaria separada,
- solo grupo de usuarios limitado,
- solo cierta jurisdicción

si muestra:

- entidad,
- alcance,
- omisión de límite

pueden evaluarse en conjunto.

42. CASOS DE AUTOCORRECCIÓN

La prueba incluye estos formatos de respuesta:

- Afirmación falsa, luego retractación explícita
- Afirmación incorrecta, luego atenuación vaga

- Afirmación correcta, luego contradicción incorrecta
- Autocorrección después de la cita
- Corrección en la misma respuesta sin mensaje de seguimiento del usuario
- Solo agregar una advertencia mientras se conserva la afirmación incorrecta

Propósito:

- autocorrección real,
- corrección falsa,
- contradicción interna no resuelta

para probar la distinción.

43. MÓDULO DE PANEL CONTROLADO–NATURAL

Casos 1,000 del Anexo de Diagnósticos: lleva gemelos Limpio y Natural de los mismos perfiles de usuario sintéticos. En la condición Natural:

- memoria,
- instrucciones especiales,
- opinión de marca previa,
- plan diferente,
- uso de la herramienta

es inyectada. Generador de Verdad preserva esta distinción:

- Cambio formalmente apropiado para el usuario
- Diferencia de consejo legítima
- Desviación de la realidad material
- Instrucción a favor de la marca
- Instrucción en contra de la marca
- Fuga de información restringida

El componente USI se prueba con este módulo.

44 th ARQUITECTURA DE ONDA

El corpus principal contiene tres ondas:

Onda	Ventana sintética UTC	Propósito
W1	12.00-13.00	Inicio
W2	20.00-21.00	Repetición a corto plazo
W3	04.00-05.00	Equilibrio a medio plazo y por hora

En cada oleada:

- nuevos usuarios sintéticos 10,000,
- misma versión de Core Mirror Prompt,
- misma ontología base del Paquete de Verdad

Se utiliza. Se inyecta deriva controlada en algunos productos.

45. ESCENARIOS DE DERIVA DE ONDA

Para productos de IA sintética, se inyectan los siguientes cambios:

- SYNTH-AI-02: Aumento en la verbosidad
- SYNTH-AI-03: L2 5–L3 6 caída de idioma
- SYNTH-AI-04: Crecimiento de la diferencia del panel natural
- SYNTH-AI-05: Aumento en el blanqueo de referencia
- SYNTH-AI-07: Precio antiguo y disponibilidad del producto
- SYNTH-AI-10: Crítico singular en W2, corrección en W3

Propósito:

STR,

es para probar el estado actual, incidentes históricos y registros de corrección.

46. CAMBIO DEL SISTEMA IN-WAVE

En medio de un producto de IA sintética en W2:

- etiqueta del modelo,
- vista de referencia,
- modo web

se cambia. El comportamiento correcto de NOMOS:

- separa la onda en una sub-onda,
- para generar una advertencia de estado de sistema mixto

debería ser. Comportamiento incorrecto: contar todo el W2 como un único puntuación de producto fijo.

47. INYECCIÓN DE EVENTO EXTERNO

Durante W3, se genera un nuevo evento sobre la entidad sintética:

- cambio de precio,
- terminación del producto,
- apertura de servicio local,
- cambio de afiliación de licencia

como. Paquete de Verdad: Está dividido en las versiones W3A y W3B. Se prueba si los adjudicadores vinculan todas las respuestas a un único tiempo de referencia.

48. CAPTURA SINTÉTICA NOMOS

Se crea un paquete de evidencia sintética para cada respuesta principal:

- Manifiesto de observación
- Artefacto de prompt
- Artefacto de respuesta en bruto

- Captura de pantalla simulada
- Artefacto de estado del sistema
- Vector de tiempo
- Registro de intentos
- Manifiesto de integridad
- Manifiesto de privacidad
- Registro de validación

Paquetes válidos: ingresar la distribución real de prueba. Paquetes defectuosos: clasificados según el Corpus de Integridad de Captura.

49. CLASES DE ERROR DE CAPTURA

La prueba produce los siguientes tipos de errores:

- Duplicado exacto
- Captura de pantalla casi duplicada
- Desajuste de solicitud
- Truncamiento del prompt
- Truncamiento de respuesta
- Fin de respuesta faltante
- Metadatos incorrectos del producto de IA
- Marca de tiempo fuera de onda
- Desfase del reloj del dispositivo
- Carga retrasada
- Desajuste de hash
- Manipulación antes del hash
- Manipulación después del hash
- Regeneración del usuario
- Reintento técnico
- Interrupción del usuario
- Transformación del sistema
- Redacción sobre archivo en bruto
- Alternativa de accesibilidad válida
- afirmación falsa de accesibilidad

Propósito: separar la validez de captura de la precisión semántica.

50. MARCA DE AGUA DE PRUEBA

Todos los elementos visuales y registros de respuestas públicas deben llevar la siguiente marca visible: SINTÉTICO NOMOS PRUEBA / NO ES UNA RESPUESTA DE IA EN VIVO / NO ES UNA afirmación DE RENDIMIENTO DE APPLE O PROVEEDOR Marca de agua: no debe cubrir el texto de la afirmación, no debe interferir con la evaluación de comprensión de los adjudicadores. En la copia ciega utilizada para la adjudicación, la marca de agua indica que es sintética, no muestra la etiqueta Generator Truth.

51. ALEATORIEDAD

La generación de pruebas utiliza un método pseudorrandom versionado. Cada decisión aleatoria:

- semilla de prueba,
- semilla de submódulo,
- selección de plantilla,
- inyección de errores,
- asignación de usuario,
- asignación de ranura

debe ser reproducible. Semilla aleatoria: se hashea antes de la generación del corpus principal, no puede cambiarse después del resultado.

52. DISTINCIÓN ENTRE ALEATORIEDAD Y ARBITRARIEDAD

Aleatorización: asegura que la asignación sea independiente del resultado. Arbitrariedad: significa que el productor selecciona la respuesta que desea. Sin un manifiesto de aleatorización: la declaración "Lo generamos aleatoriamente." no es suficiente.

53. FUGA DE DATOS

La fuga de pruebas puede ocurrir de las siguientes maneras: La Verdad del Generador se muestra al árbitro. El nombre de la plantilla de respuesta explica el nivel de importancia. El nombre del archivo debe ser critical-licence-001.png. El color de la interfaz indica el tipo de error. El código de referencia lleva una etiqueta verdadera/falsa directa. El árbitro ha visto el mismo caso en el conjunto de entrenamiento. El asistente de IA árbitro accede al código del generador. La puntuación ve los resultados de retención del desarrollador. Todos estos canales deben estar cerrados.

54. NOMBRES DE ARCHIVOS

En los paquetes de revisión a ciegas, el nombre del archivo:

OBS-000184

debe ser neutral. Los siguientes nombres están prohibidos:

- caso-crítico-12
- entidad-equivocada
- buena-respuesta
- fallo-bajo-recurso
- cita-falsa

55. ROLES SEPARADOS

La prueba debe gestionarse con al menos los siguientes roles:

- Arquitecto de Población Sintética
- Diseñador de Paquete de Verdad
- Generador de Respuestas
- Administrador de Semillas
- Auditor de Fugas
- Diseñador del Corpus de Captura
- Equipo de Extracción de afirmaciones
- Equipo de Adjudicación
- Equipo de Métodos de Puntuación
- Equipo de Validación Independiente
- Custodio de Publicación Pública
- Propietario Responsable de Pruebas Humanas

Los roles pueden combinarse en el pequeño piloto. Sin embargo:

- La persona que conoce la Verdad del Generador,
- ess el único que toma decisiones en la adjudicación ciega final

Ninguna persona que conozca la Verdad del Generador puede ser el único que tome decisiones en la adjudicación ciega final.

56. SECCIONES DE DATOS DE PRUEBA

56.1. Conjunto de Calibración Pública

Utilizado para la capacitación de adjudicadores y la prueba del libro de códigos. La Verdad del Generador está abierta. No entra en la evaluación principal.

56.2. Conjunto de Desarrollo

Utilizado para el desarrollo de métodos y software. Las etiquetas están abiertas a un equipo limitado.

56.3. Conjunto de Validación Sellado

Se abre después de que la adjudicación y el sistema de puntuación estén fijados. La Verdad del Generador está sellada. Produce la validación principal del método.

56.4. Conjunto de Retención de Renovación

Se conserva para probar el sobreajuste específico del benchmark. Está cifrado de antemano y se utiliza para la renovación de versiones. Puede que no permanezca oculto indefinidamente; se debe proporcionar una explicación y un calendario de publicación.

57. PRINCIPAL 30,000 REGLA DE FUGA DEL CORPUS

Corpus de Prevalencia Principal:

- método de puntuación sin fijación 0.9,
- código de afirmación sin fijación,
- roles de adjudicación sin ser determinados

dice que no deben abrirse. Después del fijación: se abren los paquetes de respuesta y evidencia para los adjudicadores. La Verdad del Generador permanece cerrada. La adjudicación termina. Se calculan

las puntuaciones. El manifiesto de resultados se bloquea. La Verdad del Generador se abre. Se realiza un análisis de recuperación.

58. RECUPERACIÓN DE LA VERDAD FUNDAMENTAL

El éxito de la prueba de NOMOS no está produciendo una puntuación alta. Éxito:

> *tiene una baja diferencia entre la Verdad del Generador y los resultados reproducidos por NOMOS.*

Ejemplo: Verdad del Generador:

- Aprobado Estricto: 942
- Condicional: 38
- Mayor: 15
- Crítico: 2
- No Resuelto: 3

Recuperación de NOMOS:

- Aprobado Estricto: 940
- Condicional: 40
- Mayor: 15
- Crítico: 2
- No Resuelto: 3

entonces el método puede ser fuerte. La puntuación NOMOS puede ser alta o baja. Lo importante es la recuperación correcta.

59. RECUPERACIÓN DEL LÍMITE DE afirmación

Átomos de afirmación del generador:

G

Átomos producidos por los adjudicadores:

H

Déjalo ser. Reclamar precisión de los límites:

```
P= átomos extraídos correctamente emparejados / todos los átomos extraídos
```

Recuperación:

```
R= átomos extraídos correctamente emparejados / átomos del generador
F1:
```

F1 = 2 PR / (P + R)

se puede calcular como. No solo el número de átomos:

- rango de fuente,
- proposición normalizada,
- entidad,
- calificadores

debe tomarse en cuenta la coincidencia.

60. RECUPERACIÓN DE DECISIÓN ATÓMICA

Se requiere medición separada para cada dimensión: Precisión de la entidad

```
Estado fáctico macro-F1

Alcance macro-F1

Tiempo macro-F1

Macro-F1 Atribución

Modalidad macro-F1

Coincidencia de referencias F1

Omisión F1
```

Precisión combinada de etiquetas Precisión no resuelta/brecha de referencia La precisión general por sí sola no es suficiente. La clase Crítica rara puede perderse en una precisión general alta.

61. RECORDATORIO CRÍTICO

Número de casos críticos confirmados en Generator Truth:

NC

Correctamente encontrado Crítico:

TPC

que así sea.

```
Recall_C= TP_C/N_C
```

es como sigue. Tasa de Falsos Críticos:

```
FCR= FP_C/N_non-crítico
```

puede calcularse como. Prueba:

- recuperación crítica alta,
- tasa baja de falsos críticos

deberían buscarse juntos.

62. RECUPERACIÓN DE ESTADO DE RESPUESTA

Para cada clase RP:

- precisión,
- recuperación,

F1,

debería producirse la matriz de confusión. Especialmente deberían ser visibles las siguientes confusiones:

- RP-3 y RP-4
- RP-4 y RP-5
- RP-6 y RP-4
- RP-7 y RP-1

- RP-8 y RP-4

63. GEO-1000 ERROR DE DISTRIBUCIÓN

Para el estado k, usuario generador equivalente: DkG NOMOS recuperación: DkN. Error absoluto:

```
E_k= |D_k^N− D_k^G|
```

Error total de distribución normalizado:

```
E_D= (1/2,000)Σ_k E_k
```

se puede calcular. Este valor está en el rango 0–1. Cero es recuperación total.

64. RECUPERACIÓN DE COMPONENTES

Puntuación del componente generador: Sea SjG NOMOS puntuación como SjN. Error absoluto del componente:

```
AE_j= |S_j^N− S_j^G|

MAE del componente:

MAE_component= (1/10)Σ_{j=1}^{10} AE_j
```

es el siguiente. También debería mostrarse qué componente es más difícil de recuperar.

65. RECUPERACIÓN DE puntuación compuesta

AE_NOMOS = |NOMOS_N − NOMOS_G|

se calcula de esta manera. Error compuesto pequeño: no es suficiente si la distribución de los componentes es incorrecta. Dos errores en los componentes pueden haberse cancelado aritméticamente entre sí. Por lo tanto:

- error compuesto,
- error del componente

deberían publicarse juntos.

66. RECUPERACIÓN DE PUERTA

La prueba examina por separado los siguientes resultados de puerta:

- Cualquiera Crítico
- Bloqueo crítico
- Fallo grave
- Condicional
- Candidato Completo
- No evaluable
- No Resuelto

En la recuperación de la puerta: convertir accidentalmente un Bloqueo crítico en un Candidato 950+ es un fallo de prueba muy grave.

67. RECUPERACIÓN DE JUSTICIA

Generador sintético:

- pre-determina el piso del idioma,
- diferencia de idioma,
- la base del país.
- cobertura

NOMOS:

LGF,

piso del idioma, piso del país, disparidad, debe reproducir los resultados del factor de cobertura. Recuperación de justicia:

- debe medirse
- no solo por la diferencia compuesta de LGF,

sino también por sus subcomponentes.

68. DISTINCIÓN ENTRE DEFECTO DE IDIOMA Y DEFECTO DE SOLICITUD

La prueba produce algunas puntuaciones bajas de lenguaje debido a la penalización del perfil de IA. Otras se producen debido a la falla de equivalencia del prompt. El comportamiento correcto de NOMOS:

- es clasificar el primero como hallazgo de lenguaje de IA,
- y el segundo como falla de Constitución del prompt

Los dos resultados no deben convertirse en la misma penalización de equidad.

69. RECUPERACIÓN LIMPIA–NATURAL

Generador:

- Determina de antemano los valores para la puntuación Limpia,
- puntuación Natural,
- brecha del panel,
- desviación del material debido a la personalización

NOMOS:

- solo la relación,
- el límite de causalidad,
- el componente USI

debe producirse correctamente. Cada caso con una diferencia Clean–Natural no debe interpretarse como "la memoria lo causó."

70. RECUPERACIÓN DE CAPTURA

Para el Corpus de Integridad de Captura:

- Válido
- Válido condicionalmente
- Fuera de onda
- Desajuste de prompt
- Duplicado
- Manipulación sospechada
- Fabricado
- Retirado

la matriz de confusión de los estados debe publicarse. No debe afectar la decisión de captura de corrección semántica.

71. RECUPERACIÓN DEL INTERVALO DE CONFIANZA

Las pruebas sintéticas pueden ejecutar el mismo proceso de producción poblacional muchas veces. Sea θ el parámetro verdadero del generador. La cobertura empírica de los intervalos de confianza del 95 por ciento:

```
Coverage= Número de intervalos que contienen θ / Repeticiones totales
```

se calcula como. El método del 95 por ciento: debería producir aproximadamente una cobertura del 95 por ciento. Los intervalos excesivamente estrechos o demasiado amplios se evalúan por separado.

72. PRUEBA DE EVENTOS RAROS

Los escenarios sintéticos con cero, uno, dos y cinco eventos críticos deben generarse por separado. NOMOS debe mostrar correctamente los siguientes campos:

- Recuento de eventos observados
- Tasa ponderada
- Límite superior unilateral
- Estado del incidente
- Alcance
- Requisito de muestreo confirmatorio

Incidente cero: no debería haber riesgo. Un incidente: no debería haber fallo del sistema 100%.

73. UMBRALES DE ACEPTACIÓN DE PRUEBA DE CANDIDATO

UMBRALES DE CANDIDATO — UMBRALES DE CANDIDATO

Los siguientes umbrales son candidatos antes del piloto y la revisión independiente:

Criterio	Resultado mínimo del candidato
Límite de afirmación F1	≥ 0.95

Criterio	Resultado mínimo del candidato
Precisión de resolución de entidades	≥ 0.98
Estado fáctico macro-F1	≥ 0.90
Alcance macro-F1	≥ 0.90
Macro-F1 Atribución	≥ 0.90
Mapa de citas F1	≥ 0.90
Omisión de material F1	≥ 0.85
Recuperación crítica confirmada	≥ 0.99
Tasa de falsos positivos críticos	≤ 1%
Macro-F1 de clasificación mayor	≥ 0.95
Macro-F1 del estado RP	≥ 0.92
MAE de componentes	≤ 2.5 puntos
Error absoluto compuesto	≤ 5/1,000
Distribución de errores de cada clase RP	≤ 10/1,000
Error del componente de equidad	≤ 3 puntos
Precisión de validez de captura	≥ 0.98
Cobertura empírica del CI de 95%	92%-98%
Reversión incorrecta de puerta crítica	Casos 0

Estos umbrales no son un estándar final. El propósito de la prueba también es mostrar si estos umbrales son realistas.

74. ERRORES DE PRUEBA DE TOLERANCIA CERO

Las siguientes situaciones pueden detener un lanzamiento de prueba incluso en un solo caso:

- Fuga de verdad del generador
- Cambiar la etiqueta del corpus principal después del resultado
- Continuar el lanzamiento a pesar de saber que una regla omite sistemáticamente casos críticos
- Presentación como rendimiento real de Apple o de un proveedor de IA real
- Publicar una captura de pantalla sintética como si fuera una respuesta en vivo
- Eliminar un sistema o idioma de baja puntuación posteriormente
- Cambiar la semilla de prueba
- Ajustando silenciosamente la fórmula de puntuación según el resultado del Holdout
- El corpus bruto no coincide con el manifiesto publicado
- Almacenamiento del resultado de prueba fallido

75. NIVELES DE ÉXITO DE PRUEBA

BQ-0— NO EJECUTADO

La prueba está solo en la etapa de diseño.

BQ-1— GENERACIÓN VALIDADA

Se ha reproducido el corpus sintético. Aún no se ha realizado la adjudicación ni la recuperación de la puntuación.

BQ-2— PRUEBA EN SECO DEL PROCESO

La captura, la extracción de afirmaciones y la cadena de puntuación se han ejecutado una vez. No existe verificación independiente.

BQ-3— VALIDACIÓN SELLADA APROBADA

Se han cumplido los umbrales de candidatos en el corpus de validación sellado.

BQ-4— REPRODUCCIÓN INDEPENDIENTE

El equipo independiente produjo resultados comparables a partir del mismo corpus.

BQ-5— VALIDACIÓN EXTERNA MULTISITIO

Múltiples instituciones independientes repitieron las pruebas en diferentes entornos. Antes de que la metodología de puntuación NOMOS se convierta en un candidato a estándar público para empresas reales, al menos:

BQ-4

debería apuntar a ese nivel.

76. PRUEBAS METAMÓRFICAS

Las pruebas no deberían evaluar solo respuestas fijas. Los resultados deberían ser consistentes en transformaciones que preserven el mismo significado.

76.1. Prueba de Parafraseo

La misma afirmación atómica se escribe con una oración diferente. El resultado de la revisión por pares no debería cambiar.

76.2. Prueba de Orden de Oraciones

El orden de las afirmaciones cambia. El estado de veracidad no debería cambiar excepto por la Centralidad.

76.3. Prueba de Ubicación de la Citación

La cita cambia de lugar dentro del párrafo mientras claramente sigue estando ligada a la misma afirmación. El soporte de la cita debería permanecer igual.

76.4. Prueba de Atribución

La frase "La compañía dice" se elimina. Los resultados de EPI y atribución deben cambiar.

76.5. Prueba de Modalidad

La frase "Es cierto" cambia a "Es probable". Si el Paquete de Verdad es incierto, el resultado de la modalidad puede mejorar.

76.6. Prueba de Entidad

La misma afirmación verdadera se transfiere a una organización incorrectamente asociada. El resultado de entidad y alcance debe verse alterado.

76.7. Prueba de Tiempo

La afirmación histórica se actualiza. TLA debe reducirse.

76.8. Prueba de Límite

La oración límite que cambia la decisión del usuario se elimina. SBI y, si es necesario, la Puerta Crítica deben ser cambiados.

77. PRUEBAS CONTRADICTORIAS

La prueba debe incluir los siguientes ataques:

- Respuesta muy fluida pero incorrecta
- Respuesta con muchas referencias pero pruebas blanqueadas
- Crítico único entre un alto número de elementos correctos
- Respuesta evasiva que parece una precaución adecuada
- Respuesta corta pero suficiente
- Respuesta larga llena de errores incidentales
- Respuesta gravemente incorrecta con autocorrección
- Número correcto, falsa afirmación de independencia
- Licencia de afiliado real, entidad principal falsa
- Precio actual, país incorrecto
- Relación correcta con el cliente, tiempo incorrecto
- Respuesta que hace que el usuario revise un hecho demográfico
- Respuesta presentando la queja como un crimen definitivo
- Respuesta que presenta evidencia limitada como una cita pública
- Respuesta que afirma lo contrario de lo que dice la cita real

PAQUETE DE VERDAD PARA EL EXAMEN 78 TH

Las pruebas fuera del Apple-SYNTH Truth Pack deben tener su propio Truth Pack metodológico. Este paquete contiene lo siguiente: ¿Cuántas respuestas se produjeron? ¿Cuántas en el corpus principal? ¿Cuántas en el corpus de desafío? ¿Cuántas etiquetas de generador Críticas y Mayores hay? ¿Cuál fue la semilla? ¿Qué versión del código se usó? ¿Qué archivos fueron excluidos? ¿Qué vieron los adjudicadores? ¿Qué etiquetas se abrieron y cuándo? ¿Cuándo se bloqueó el método de puntuación? ¿Qué resultados de recuperación se obtuvieron? ¿Qué umbrales no se cumplieron? Pruebas: no puede declarar la validación del método sin crear sus propios resultados para el Truth Pack.

79. TARJETA DE RESULTADOS DE PRUEBA PÚBLICA

La tarjeta pública debe incluir al menos los siguientes campos:

Identidad

Nombre de prueba Versión Estado sintético Ancla utilizada Declaración de empresa/rendimiento real sin afirmación

Corpus

Respuesta principal: Anexo diagnóstico 30,000 Nivel de importancia Desafío Captura Integridad Adjudicación de oro

Generador Verdad

Tiempo de fijación Semilla manifiesto Distribución de etiquetas Auditoría de fugas

Recuperación

```
Límite de afirmación F1
```

Recuperación crítica Falso positivo crítico

```
Macro-F1 RP

MAE de componentes
```

Error compuesto; error de equidad; cobertura del IC; precisión de captura

Resultado

Nivel BQ Umbrales aprobados Problemas restantes Requisito de nueva prueba Estado de repetición independiente Esta tarjeta no debe clasificar ningún producto de IA real.

80. DISPOSICIONES NORMATIVAS OBLIGATORIAS

CH17-N01

Apple.com El objeto de medición para las pruebas sintéticas debe ser la metodología NOMOS; no debe ser el rendimiento real de Apple ni de un proveedor de IA real.

CH17-N02

Todos los materiales públicos de la prueba deben llevar declaraciones de sintético y de no afirmación.

CH17-N03

La entidad sintética debe llevar una identidad APPLE-SYNTH claramente separada del ancla Apple.com.

CH17-N04

Los registros del Paquete de Verdad Sintético no pueden ser usados como la realidad real de Apple.

CH17-N05

Las identidades de productos de IA sintéticos no pueden presentarse como códigos implícitos o réplicas de proveedores reales.

CH17-N06

El Corpus de Prevalencia del Principal, el Corpus de Desafío de Grado de Importancia, el Corpus de Integridad de Captura y el Corpus de Oro de Adjudicación deben mantenerse separados.

CH17-N07

Los casos de desafío de grado de importancia no pueden añadirse a la tasa de prevalencia real.

CH17-N08

Los casos capturados con errores no pueden añadirse al denominador de fallos del producto semántico sin explicación.

CH17-N09

Los casos de calibración de oro no pueden ingresar en el corpus principal de puntuación.

CH17-N10

El corpus principal de respuestas 30,000 debe preservar diez productos sintéticos de IA, usuarios únicos 1,000 y una estructura de tres olas.

CH17-N11

La misma unidad humana sintética en el corpus principal no puede reutilizarse en diferentes productos u olas.

CH17-N12

Incluso si los productos de IA usan identidades humanas diferentes, deben mantener distribuciones poblacionales coincidentes.

CH17-N13

En el corpus principal 30,000, todos los usuarios deben recibir la misma intención canónica Core Mirror.

CH17-N14

Las versiones de idiomas no pueden presentarse como traducciones en vivo verificadas sin una verdadera experiencia en el idioma.

CH17-N15

Los códigos de idioma sintéticos no pueden usarse para reclamar un rendimiento real en un idioma.

CH17-N16

Si no se utiliza la población real del país, los códigos de país y las asignaciones deben etiquetarse claramente como sintéticos.

CH17-N17

La asignación de países sintética debe preservar el algoritmo proporcional a la población y la condición total 1,000.

CH17-N18

La visibilidad de países pequeños debe proporcionarse con el Anexo de Observador sin alterar el peso del Panel de Población principal.

CH17-N19

Un usuario multilingüe no puede ser contado múltiples veces en el peso de la población principal.

CH17-N20

El Anexo de Diagnóstico no puede mostrarse en el mismo denominador que el resultado de la respuesta principal 30,000.

CH17-N21

El Paquete de Verdad canónico de la prueba debe ser fijado antes de que se generen las respuestas.

CH17-N22

La Verdad del Generador debe pre-generarse para cada respuesta y mantenerse oculta hasta el final de la revisión.

CH17-N23

La etiqueta Verdad del Generador no puede filtrarse a través del texto de la respuesta, el nombre del archivo, la interfaz de usuario o los metadatos.

CH17-N24

La Verdad del Generador no puede cambiarse silenciosamente después de que se hayan visto los resultados.

CH17-N25

Si se encuentra un error de producción, se debe crear una nueva versión de la prueba y registrar el impacto en el corpus antiguo.

CH17-N26

La semilla aleatoria y el código de producción deben estar fijados antes de la producción del corpus principal.

CH17-N27

No se puede usar la afirmación "Generado aleatoriamente" sin el manifiesto de semilla y algoritmo.

CH17-N28

Los perfiles de producto de IA sintética deben definirse antes de que se vean los resultados.

CH17-N29

Los perfiles de producto con puntuación baja no se pueden derivar ni suavizar después del resultado.

CH17-N30

Las pruebas deben cubrir todos los tipos de puertas críticas con casos de desafío suficientes.

CH17-N31

No toda afirmación fuerte o negativa puede considerarse crítica; deben encontrarse casos de control negativos que se asemejen a los críticos.

CH17-N32

El corpus principal debería reflejar la prevalencia rara y realista de Crítico; sin embargo, el corpus de desafío debería medir el poder de detección de Crítico.

CH17-N33

Los casos de Crítico sobremuestreados en el corpus de desafío no pueden cambiar la tasa principal de Crítico.

CH17-N34

La inyección de errores debería cubrir casos abiertos y ocultos, directos y basados en atribución.

CH17-N35

Las pruebas no deberían producir errores Críticos con patrones de palabras fáciles.

CH17-N36

Se debería producir el mismo error semántico en diferentes paráfrasis y estructuras de oración.

CH17-N37

Las pruebas de citación deberían cubrir directamente casos de soporte, soporte parcial, solo atribución, alcance incorrecto, fabricado y blanqueo de linaje.

CH17-N38

Las citas sintéticas no pueden estar dirigidas a instituciones o personas reales.

CH17-N39

Los casos Brecha de referencia no deben llevar automáticamente una etiqueta de correcto o incorrecto.

CH17-N40

El corpus de omisión debe probar la distinción entre omisión opcional, requerida y crítica.

CH17-N41

El módulo Limpio–Natural debe separar la deriva de la realidad material de la personalización legítima.

CH17-N42

La arquitectura de ondas debe preservar la re-medición con el mismo prompt y usuarios independientes.

CH17-N43

Los cambios en el sistema intra-onda deben producir registros de sub-onda o estado mixto.

CH17-N44

El Paquete de Verdad de eventos externos sintéticos debe ser procesado junto con la versión temporal.

CH17-N45

Los paquetes de captura sintética NOMOS deben incluir ejemplos de cadenas de evidencia válidas y defectuosas.

CH17-N46

Todas las pantallas sintéticas deben llevar una marca de agua visible que indique que no son resultados en vivo de IA.

CH17-N47

La marca de agua no puede invalidar el significado de una afirmación o la decisión de una adjudicación.

CH17-N48

Los errores semánticos debido a defectos de captura deben ser etiquetados por separado en la Verdad del Generador.

CH17-N49

La verificación de fugas debe completarse antes de que se abra la puntuación principal de la prueba.

CH17-N50

Los nombres de archivos o los ID de respuesta no pueden llevar etiquetas de nivel de importancia y precisión.

CH17-N51

La persona que accede a la Verdad del Generador no puede ser el único tomador de decisiones en la revisión ciega final.

CH17-N52

El método de puntuación debe estar fijado antes de abrir el corpus de validación sellado.

CH17-N53

Después de ver el resultado del Holdout, la fórmula de puntuación no puede ajustarse silenciosamente.

CH17-N54

El cambio de fórmula requiere un nuevo método de puntuación y una nueva ejecución de validación.

CH17-N55

El éxito de la prueba no debe evaluarse según la puntuación alta del NOMOS producido, sino de acuerdo con la recuperación de la Verdad del Generador.

CH17-N56

El límite de la afirmación, la decisión atómica, el estado de la respuesta, el componente y la recuperación compuesta deben medirse por separado.

CH17-N57

La recuperación crítica y la tasa crítica de falsos positivos deben publicarse juntas.

CH17-N58

Una alta precisión general no puede ocultar el fracaso en la clase Crítica rara.

CH17-N59

La cancelación de errores de componentes entre sí no puede usarse como éxito de recuperación compuesta.

CH17-N60

La recuperación de la equidad debe medirse con los subcomponentes de nivel de lenguaje, disparidad y cobertura.

CH17-N61

La falla de equivalencia del prompt no debe calificarse como falla del lenguaje de IA.

CH17-N62

La diferencia Limpio–Natural no puede ser convertida por la prueba en una afirmación de causalidad no respaldada.

CH17-N63

La recuperación de la validez de la captura debe medirse independientemente de la corrección semántica.

CH17-N64

El método del intervalo de confianza debe probarse con la cobertura empírica en muestreos sintéticos repetidos.

CH17-N65

Las pruebas de eventos raros deben incluir escenarios de eventos críticos cero, uno y múltiples.

CH17-N66

Un escenario crítico observado como cero no debería producir un resultado de riesgo cero.

CH17-N67

Los umbrales de aceptación de pruebas deberían estar en estado de candidato y versionados.

CH17-N68

Cuando no se cumplen los umbrales de candidato, se debería divulgar al público un resultado fallido.

CH17-N69

El fallo de la prueba no puede ser ocultado cambiando los datos o los umbrales.

CH17-N70

El nivel BQ debería ser visible en la tarjeta de resultados pública.

CH17-N71

NOMOS debe aspirar al menos al nivel de reproducción independiente antes de que la empresa real sea sometida a auditoría pública.

CH17-N72

Las transformaciones que preservan el significado en pruebas metamórficas no deberían alterar innecesariamente el resultado de la adjudicación.

CH17-N73

Las transformaciones de atribución, modalidad, entidad, tiempo o límite que cambien el significado deben producir un cambio de puntuación adecuado.

CH17-N74

La prueba debe llevar su propio Paquete de Verdad metodológico y registro de cambios.

CH17-N75

Un anclaje Apple.com no puede presentarse con ninguna impresión de respaldo, participación o colaboración.

CH17-N76

Un corpus sintético no puede difundirse como evidencia independiente hacia fuentes reales de redes sociales o fuentes públicas.

CH17-N77

Si las respuestas sintéticas se indexan en la web, también debe especificarse su estado sintético en los metadatos legibles por máquina.

CH17-N78

Se deben usar noindex, descargos de responsabilidad estructurados o protecciones equivalentes para reducir la posibilidad de que los datos de prueba sean tomados erróneamente como información real de Apple por sistemas reales de IA.

CH17-N79

La publicación pública debe incluir archivos de reproducción, versión de la fórmula, manifiesto de semilla e informe de recuperación.

CH17-N80

Cada diseño de prueba, producción, clave, adjudicación, recuperación y decisión de publicación debe tener un propietario humano o institucional que sea responsable.

81. FORMAS DE FALLA

CH17-F01— PRESENTÁNDOSE COMO UNA AUDITORÍA REAL DE APPLE

La puntuación sintética se convierte en rendimiento real de la empresa.

CH17-F02— MAPEO OCULTO HACIA UN PROVEEDOR REAL DE IA

Los perfiles de productos sintéticos se describen como imitaciones de ciertos proveedores.

CH17-F03— HACIENDO UN PAQUETE DE VERDAD SINTÉTICA UNA FUENTE REAL

Las afirmaciones de prueba se publican en la web como si fueran información real de Apple.

CH17-F04— PANTALLA SINTÉTICA SIN FILIGRANA

La respuesta simulada de IA circula como una respuesta en vivo.

CH17-F05— COMBINANDO EL CORPUS PRINCIPAL Y EL DEL DESAFÍO

Los casos críticos sobremuestreados aumentan la prevalencia.

CH17-F06— AÑADIENDO EL CONJUNTO DE ORO AL puntuación PRINCIPAL

Los casos de capacitación de los jueces se convierten en respuestas reales de los usuarios.

CH17-F07— CONTANDO EL DEFECTO DE CAPTURA COMO ERROR DE IA

Un archivo incorrecto se convierte en un fallo semántico.

CH17-F08— DUPLICANDO EL MISMO USUARIO SINTÉTICO

Una sola persona se convierte en una unidad poblacional independiente en diferentes productos de IA.

CH17-F09— CONSIDERANDO POBLACIONES NO COINCIDENTES COMO DIFERENCIA DE PRODUCTO

Los productos de IA se prueban en diferentes composiciones de usuarios.

CH17-F10— DIVIDIR LAS PRINCIPALES FAMILIAS DE DEMANDA DE 1,000

La estimación de la población con la misma demanda del fundador se ve interrumpida.

CH17-F11— CONTAR LA PRUEBA PRINCIPAL COMPLETA NOMOS

Se presenta una única demanda como si hubiera probado todos los componentes.

CH17-F12— afirmación DE POBLACIÓN REAL

Se publica la distribución sintética C0 01–C1 93 como la población mundial.

CH17-F13— CONTAR OBSERVADOR COMO puntuación PRINCIPAL DEL PAÍS

La observación de un solo país ingresa en La puntuación de la población.

CH17-F14— CONTAR IDIOMA SINTÉTICO COMO RENDIMIENTO DE IDIOMA REAL

El bajo puntuación de L2 5 se convierte en una afirmación sobre el idioma vivo específico.

CH17-F15— CONTAR ERROR DE TRADUCCIÓN COMO ERROR DE IA

La deriva del prompt produce un hallazgo de producto incorrecto.

CH17-F16— GENERADOR DE VERDAD ESCRIBIENDO DESPUÉS

Se crea una etiqueta oculta según el resultado del juez.

CH17-F17— FILTRACIÓN DE VERDAD DEL GENERADOR

El juez aprende la etiqueta correcta a partir del nombre del archivo o los metadatos.

CH17-F18— CAMBIANDO LA SEMILLA BASADA EN EL RESULTADO

Se elige una semilla que produzca una distribución más deseable.

CH17-F19— PUBLICANDO LA MEJOR EJECUCIÓN ALEATORIA

Se selecciona el resultado deseado entre múltiples generaciones.

CH17-F20— SOLO CASOS CRÍTICOS FÁCILES

Se anuncia la recuperación 100% con oraciones explícitas como "banco autorizado".

CH17-F21— FALTA DE CONTROL NEGATIVO CRÍTICO

No se penaliza hacer que cada candidato de error fuerte sea Crítico.

CH17-F22— EXCESO DE CRÍTICO EN EL CORPUS DE PREVALENCIA

Se pierde la estructura realista de eventos raros.

CH17-F23— CRÍTICO INSUFICIENTE EN EL CORPUS DE DESAFÍO

La recuperación crítica se mide con pocos casos.

CH17-F24— FALTA DE PARAFRASE

El adjudicador aprende un patrón de palabras fijo en lugar del significado.

CH17-F25— CONVIRTIENDO ERRORES DE ATRIBUCIÓN EN FALSIFICACIONES EVIDENTES

Los casos epistémicos difíciles se hacen más fáciles.

CH17-F26— VINCULANDO CITAS CON LA WEB REAL

Una afirmación falsa sintética infecta la fuente e institución reales.

CH17-F27— ETIQUETANDO EL Brecha de referencia

Generator Truth otorga automáticamente aprobado o desaprobado.

CH17-F28— PRUEBAS SIN OMISIÓN

El sistema solo aprende a detectar falsedades evidentes.

CH17-F29— PRUEBAS SIN AUTOCORRECCIÓN

La contradicción interna y la retractación no son evaluadas por el sistema.

CH17-F30— CONECTANDO LA DIFERENCIA LIMPIA–NATURAL A UNA SOLA RAZÓN

Personalización, plan y herramientas son inseparables.

CH17-F31— SIN DERIVA DE ONDA

La lógica STR y de estado actual no puede ser probada.

CH17-F32— OCULTAR CAMBIO EN ONDA

El estado del sistema mixto resulta en un único puntuación de producto.

CH17-F33— IGNORAR EVENTO EXTERNO

La versión de tiempo de Truth Pack no se prueba.

CH17-F34— CAPTURA VÁLIDA ÚNICAMENTE

El sistema de validez de captura nunca se desafía.

CH17-F35— SOLO TAMPER FÁCIL

La discrepancia de hash se convierte en un solo tipo de fraude.

CH17-F36— PROPORCIONAR PANTALLA SINTÉTICA AL ADJUDICADOR ETIQUETADA

El color o el título de la interfaz explica el resultado.

CH17-F37— EL GENERADOR Y EL ADJUDICADOR SON LA MISMA PERSONA

Se pierden la ceguera y la independencia.

CH17-F38— AJUSTANDO EL MÉTODO DE PUNTUACIÓN DESPUÉS DE LA VALIDACIÓN

Se produce sobreajuste a la prueba.

CH17-F39— GENERANDO EL HOLDOUT MÁS TARDE

Los casos difíciles se escriben de acuerdo con los resultados.

CH17-F40— OCULTANDO EL HOLDOUT PARA SIEMPRE

La prueba no se puede reproducir de manera independiente.

CH17-F41— CONTANDO ALTO NOMOS PUNTUACIÓN COMO ÉXITO

En lugar de la recuperación, se mide el nivel de rendimiento.

CH17-F42— SOLO EXACTITUD GENERAL

Las clases críticas y raras se vuelven invisibles.

CH17-F43— PUBLICANDO RECALL CRÍTICO SIN FALSOS POSITIVOS

El sistema de nivel de importancia extrema parece exitoso.

CH17-F44— ELIMINANDO ERROR DE COMPONENTE DENTRO DEL COMPUESTO

Un error alto, uno bajo se cancelan entre sí.

CH17-F45— JUSTICIA SOLO PUNTUACIÓN COMPUESTA

El piso, la brecha y la recuperación de cobertura se vuelven invisibles.

CH17-F46— prompt CONFUSA Y ERROR DE LENGUAJE DE IA

La prueba penaliza al sistema incorrecto.

CH17-F47— MEDICIÓN DE LA PRECISIÓN DE CAPTURA CON ÉXITO SEMÁNTICO

Un archivo defectuoso con respuestas correctas pasa.

CH17-F48— PRUEBA SIN COBERTURA CI

Se desconoce si los intervalos de confianza cubren el parámetro verdadero.

CH17-F49— CONSIDERANDO CERO CRÍTICO COMO RIESGO CERO

El método de evento raro falla la prueba.

CH17-F50— CONSIDERANDO UN SOLO CRÍTICO COMO FALLA GLOBAL

La distinción entre alcance y prevalencia se ve interrumpida.

CH17-F51— REDUCCIÓN DE LOS UMBRALES DEL CANDIDATO BASADA EN EL RESULTADO

El estándar se relaja para aprobar la prueba.

CH17-F52— OCULTANDO EL RESULTADO FALLIDO

Solo se publican tablas de recuperación exitosas.

CH17-F53— CONSIDERANDO BQ-2 COMO VALIDACIÓN COMPLETA

Una única prueba en seco se convierte en evidencia independiente del estándar.

CH17-F54— AUSENCIA DE PRUEBA METAMÓRFICA

La decisión cambia en diferentes expresiones del mismo significado.

CH17-F55— PRUEBAS SIN TU PROPIO PAQUETE DE VERDAD

No se puede probar cuál es la distribución del corpus y de las etiquetas.

CH17-F56— IMPRESIÓN DE RESPALDO DE APPLE

El uso del nombre de dominio se presenta como cooperación o aprobación.

CH17-F57— INDEXACIÓN DE DATOS SINTÉTICOS Y SU MEZCLA CON DATOS REALES

Las pruebas producen el envenenamiento de conocimiento que critican a sí mismo.

CH17-F58— CÓDIGO DE CONTABILIDAD CERRADO

La recuperación no puede reproducirse de manera independiente.

CH17-F59— PRUEBA NO ACUMULATIVA

Se escriben nuevas plantillas y fórmulas sobre el corpus antiguo.

CH17-F60— EXCEPCIÓN A NOMOS

La evidencia, evidencia en contrario, la versión y las reglas de objeción requeridas por el estándar no se aplican a la prueba.

82. PROCEDIMIENTO DE AUDITORÍA

Paso 1— Fijar el Límite Sin afirmación de la Prueba

Se prohíbe la afirmación del rendimiento real de Apple y del proveedor real.

Paso 2— Crear el Gemelo de Entidad Sintética

Se definen el gráfico de entidades, los productos, las entidades locales y los objetos de colisión.

Paso 3— Configurar el Paquete de Verdad Sintética

Se preparan las afirmaciones atómicas, la evidencia, evidencia en contrario, la procedencia de la fuente y los registros desconocidos.

Paso 4— Fijar el Marco de Población Sintética

Se determinan distribuciones de país, idioma, usuario y dispositivo.

Paso 5— Crear la Asignación Principal de Usuarios 1,000

Se preparan gemelos de población mapeados para cada producto de IA y ola con ponderaciones de país e idioma.

Paso 6— Fijar el Registro de Prompts

Se versionan el Prompt de Espejo Central y los prompts del Anexo de Diagnóstico.

Paso 7— Definir Perfiles de Productos de IA Sintéticos

Se determinan parámetros latentes para diez perfiles.

Paso 8— Crear el Manifiesto de Semilla Aleatoria

La semilla principal y las semillas de los submódulos son hashadas.

Paso 9— Crear la Verdad del Generador

Para cada respuesta, átomo, omisión, atribución, nivel de importancia y estado RP se preparan.

Paso 10— Sellar la Verdad del Generador

Las etiquetas se separan del adjudicador y de los equipos de puntuación.

Paso 11— Generar Textos de Respuesta

Se aplica paráfrasis, modalidad, atribución y variación lingüística.

Paso 12— Generar Objetos Sintéticos de Atribución y Evidencia

Se crean linajes de origen y conexiones falsas.

Paso 13— Renderizar interfaces de IA simuladas

Se crean pantallas web, móviles, de rechazo y de error.

Paso 14— Aplicar la marca de agua sintética

Se marcan todos los elementos visuales públicos y de auditoría.

Paso 15— Generar incidentes de integridad de captura

Se agregan duplicados, manipulaciones, prompts incorrectas y defectos de tiempo.

Paso 16— Realizar una auditoría de fugas

Se examinan el nombre del archivo, los metadatos, la interfaz de usuario y las plantillas de respuesta.

Paso 17— Fijar el método de puntos y el libro de códigos

El método está congelado sin abrir el Conjunto de Validación Sellado.

Paso 18— Abrir el Corpus Principal 30,000

Se inicia la captura y adjudicación semántica.

Paso 19— Realizar la Extracción de afirmaciones

Se extraen los átomos sin ver la Verdad del Generador.

Paso 20— Aplicar la Doble Adjudicación

Se utilizan los roles de lenguaje, evidencia y experiencia.

Paso 21— Generar Respuesta y Registrar Puntuaciones

Se calculan la distribución RP, componente y compuesto.

Paso 22— Fijar el Manifiesto de Resultados

Antes de la comparación de recuperación, la salida NOMOS se estabiliza.

Paso 23— Encender la Verdad del Generador

Los resultados de adjudicación y puntuación se comparan con etiquetas ocultas.

Paso 24— Calcular Métricas de Recuperación

Se extraen los resultados de afirmación, nivel de importancia, respuesta, componente, puntuación, equidad y IC.

Paso 25— Aplicar Umbrales de Candidatos

Se determinan las áreas de éxito y fracaso.

Paso 26— Ejecutar Pruebas Metamórficas y Contradictorias

Se examina la firmeza de la decisión.

Paso 27— Asignar Nivel de Calidad de la Prueba

El estado se da entre BQ-0 y BQ-5.

Paso 28— Crear un Plan para Corregir Fallos

Si el libro de códigos, la fórmula o la capacitación del adjudicador cambian, se prepara una nueva versión de la prueba.

Paso 29— Publicar la Tarjeta de Prueba Pública

También se muestran los casos omitidos y los falsos positivos, así como los éxitos.

Paso 30— Abrir el Paquete de Reproducción Independiente

Se publican el código, los datos sintéticos, los manifiestos y el informe de recuperación con el acceso correspondiente.

83. EVIDENCIA REQUERIDA

Probar identidad Notificación de no afirmación APPLE-SYNTH gráfico de entidades Paquete de Verdad Sintética Objetos de evidencia sintética Línea de procedencia evidencia en contrario Marco de país sintético Registro de idioma sintético Principal 1,000 asignación de usuario Población emparejada gemelos Registro de Solicitudes Equivalencia de registros de Solicitudes Diez perfiles de producto de IA sintéticos Parámetros latentes Manifiesto de semilla aleatoria Código de generación de respuestas Registro de verdad del generador Hash de verdad del generador Textos de respuesta Citas sintéticas Registros de interfaz de usuario simulada Manifiesto de marca de agua Corpus de integridad de captura Grado de importancia Corpus de desafío Adjudicación Corpus de oro Conjunto de calibración pública Conjunto de desarrollo Conjunto de validación sellado Renovación Hash de reserva Auditoría de filtración Método de puntuación Fijación de afirmación Libro de códigos Registros de calificación de adjudicador Registros de revisión a ciegas Producción de recuperación NOMOS

> Generador Verdad tiempo de apertura Métricas de límite de afirmación

```
Resultados de macro-F1 de dimensión
```

Recuperación crítica Falsos positivos críticos Clasificación mayor RP matriz de confusión

```
MAE de componentes
```

Error compuesto; recuperación de equidad; recuperación de captura; cobertura del intervalo de confianza; resultados de eventos raros; pruebas metamórficas; pruebas de casos contradictorios; resultados de umbral de candidato; nivel BQ; registro de fallos y correcciones; tarjeta de prueba pública; código de computación; hash de código; registro de reproducción independiente; registro de cambios de prueba; y persona o institución responsable.

84. LISTA DE VERIFICACIÓN DE AUDITORÍA

¿Está claro que la prueba no es una auditoría real de Apple? ¿Existe una correspondencia implícita con proveedores reales de IA? ¿Aparece el ID APPLE-SYNTH en todos los archivos? ¿Las pantallas sintéticas tienen marca de agua? ¿Se ha prevenido la indexación de registros sintéticos en la web como información real? ¿Se han separado el corpus principal y el corpus de desafío? ¿Los casos de desafío ingresaron en La puntuación de prevalencia? ¿Los casos de calibración de oro están en el denominador principal? ¿El corpus principal realmente tiene respuestas únicas 30,000? ¿Se utilizó el mismo usuario sintético en múltiples productos u oleadas? ¿Los perfiles de población de los productos coinciden? ¿Todos los usuarios principales recibieron la misma intención de Core Mirror? ¿Se presentó la prueba Core accidentalmente como Full NOMOS? ¿El marco del país sintético aparece como una población real?

¿Es la asignación de países un total de 1,000? ¿Se gestionaron los países pequeños con el Anexo de Observador? ¿Se duplicaron los usuarios multilingües? ¿Se utilizaron códigos de lenguaje sintéticos como el rendimiento real del lenguaje? ¿Se separó el fallo de equivalencia de prompts del fallo de lenguaje de IA? ¿Se bloqueó el Paquete de Verdad antes de la generación de respuestas? ¿Se pre-generó la Verdad del Generador? ¿Está la Verdad del Generador oculta para los jueces? ¿El nombre del archivo o la etiqueta de la interfaz de usuario provoca filtración? ¿Se pre-bloqueó la semilla aleatoria? ¿Se eligió la mejor semilla después? ¿Se definieron los perfiles sintéticos de IA antes de los resultados? ¿Se eliminó un perfil con puntuación baja después? ¿Están todas las compuertas críticas presentes en el corpus del desafío? ¿Son suficientes los controles negativos críticos? ¿Es la prevalencia de lo Crítico en el corpus principal realísticamente el límite sintético?

¿Es el corpus de desafío suficiente para medir la recuperación crítica? ¿Algunos de los errores son implícitos y basados en atribución? ¿Existe diversidad de parafraseo? ¿Se están probando todas las clases de atribuciones? ¿Las atribuciones apuntan a instituciones reales? ¿Se dejó correctamente sin etiquetar al Brecha de referencia? ¿Se clasificaron los casos de omisión como opcionales, requeridos y críticos? ¿Hay casos de autocorrección y contradicción interna? ¿El módulo Limpio–Natural distingue entre diferencias legítimas y materiales? ¿Existe deriva de onda? ¿Se probó el cambio del sistema intra-onda? ¿Se probaron el evento externo y la versión del Paquete de Verdad? ¿Es lo suficientemente desafiante el corpus de captura? ¿Afecta la corrección semántica la decisión de captura? ¿Es independiente la auditoría de filtración? ¿Se bloqueó el corpus de validación de Método de puntuación sin abrirlo?

¿Afectaron los resultados del Holdout la fórmula? ¿Se mide el éxito de la prueba por una puntuación alta?

```
¿Está abierto el F1 del límite de la afirmación?

¿Son separados el macro-F1 de entidad y el factual?
```

¿Son la memoria crítica y el falso positivo juntos? ¿Se ha publicado la matriz de confusión de RP? ¿Se almacenó dentro del compuesto de error de componentes? ¿La recuperación de equidad tiene suelo, espacio y cobertura? ¿La recuperación limpia–natural produce una afirmación por causalidad? ¿La recuperación de captura es separada? ¿Se ha probado la cobertura empírica de IC? ¿El escenario crítico cero producía riesgo cero? ¿Se cambiaron los umbrales de candidatos tras el resultado? ¿Están disponibles públicamente los umbrales incumplidos? ¿Se han realizado pruebas metamórficas? ¿Está libre el nivel BQ? ¿Existe reproducción independiente? ¿El examen tiene su propio Paquete de la Verdad? ¿Se conserva la historia del cambio? ¿Se conoce al responsable de la prueba?

85. OBJECIONES Y RESPUESTAS

Objeción 1— "¿Por qué no medimos directamente la Apple real y los productos reales de IA?"

Porque en esta etapa, el objetivo no es comparar empresas o proveedores, sino validar la cadena de medición. Una prueba real:

- investigación web actual,
- población real,
- acceso a producto de IA real,
- evidencia real,
- requiere ley y privacidad

Las pruebas sintéticas hacen visibles primero los errores propios del método.

Objeción 2— “Si los datos sintéticos no representan el mundo real, ¿cuál es el uso?”

Los datos sintéticos no prueban la prevalencia real. Prueban:

- recuperar la etiqueta correcta,
- Capturar la puerta crítica,
- separando el defecto de captura,
- reproduciendo la fórmula de puntuación,

calculando la equidad y la incertidumbre. Esta es una prueba de método obligatoria antes de la auditoría en vivo.

Apelación 3— “¿Por qué se está utilizando Apple.com? ¿No se puede usar otro dominio?”

Es utilizable. Apple.com es el único ancla reconocible. Equipos independientes:

- otros anclajes de dominio,
- dominios completamente ficticios,
- gemelos de entidades específicas de la industria

pueden ser utilizados. NOMOS no debería depender de un único ancla.

Objeción 4— “¿No sería más realista mezclar información real de Apple con Truth Pack sintético?”

Esto difumina el límite de la prueba. Si se combinan información real y sintética, el lector puede no entender qué afirmación es real. El gemelo sintético debe mantenerse por separado.

Objeción 5— “Si todas las respuestas de 30,000 provienen de un único prompt, ¿cómo se probará el Full NOMOS?”

El corpus principal 30,000 es una prueba de prevalencia y estabilidad del Core GEO-1000. Los módulos de Evidencia, Límite, Recomendación, Limpio–Natural y Lenguaje se prueban por separado en el Anexo Diagnóstico. Un único prompt no es Full NOMOS.

Objeción 6— "¿Por qué no prueba el mismo usuario sintético todos los productos de IA?"

La misma persona: puede fortalecer la comparación, pero crea transferencia y condicionamiento. El corpus principal utiliza individuos únicos. La composición de la población se empareja con gemelos coincidentes. Se puede establecer un módulo experimental separado y emparejado.

Objeción 7— "¿Se puede probar realmente la equidad del lenguaje sin nombres de lenguas reales?"

Se pueden probar la fórmula, el ponderado, el piso y los cálculos de brecha. No se puede probar la traducción real ni el comportamiento del lenguaje. La validación lingüística en vivo también requiere experiencia humana local.

Objeción 8— "¿Distorsiona la puntuación el sobremuestreo de casos críticos?"

El Corpus de Desafío no entra en la puntuación. Mide la capacidad crítica de detección. El principal Corpus de Prevalencia conserva la tasa escasa realista.

Objeción 9— "Si la persona que prepara la Verdad del Generador ya conoce la respuesta correcta, ¿por qué es significativa la adjudicación?"

Los adjudicadores no ven la Verdad del Generador. El objetivo es probar si el sistema de adjudicación puede recuperar la verdad oculta a través de la respuesta visible y el Paquete de Verdad.

Objeción 10— 'Si la IA también escribe respuestas sintéticas, ¿no habría la misma IA preparado su propia prueba?'

La IA puede ayudar con la producción de variación lingüística. Sin embargo:

- plan atómico,
- Verdad del Generador,
- grado de importancia,
- semilla,
- aprobación del examen final

Debe estar bajo gobernanza humana e iterativa. El generador de respuestas no puede ser el juez último.

Objeción 11— "¿No es demasiado alto el 99 por ciento para la recuperación crítica?"

Puede ser alto. Especialmente en casos difíciles y oscuros, el resultado piloto puede ser más bajo. Por eso es un umbral candidato. Dado que el costo de fallar es alto en puntos de alto riesgo, es natural que la meta sea ambiciosa.

Objeción 12— "¿Por qué es importante la tasa de Falsos Críticos por separado?"

Etiqueta crítica:

- afecta la calificación del producto,
- percepción pública,
- y la entidad auditada

en serio. Un sistema que es demasiado sensible y marca cada error como Crítico no es confiable.

Objeción 13— "Si la prueba falla, ¿no se debilita el libro?"

No. Explicar el fallo fortalece el estándar. El propósito de la prueba no es verificar la idea, sino mostrar dónde no funciona.

Objeción 14— '¿Por qué está mal mejorar la fórmula basándose en el resultado de la prueba?'

Mejorar no está mal. Conformarse silenciosamente al mismo resultado de validación está mal. El proceso correcto:

- Resultado del Método 0.9
- análisis de errores
- Método 1.0
- validación sobre un conjunto de retención nuevo o sin tocar.

Esta es la secuencia requerida.

Objeción 15— "Si publicamos todo el corpus, ¿no se gamificará la prueba?"

Hay un riesgo de gamificación. Por lo tanto:

- calibración pública,
- validación sellada,
- renovación del conjunto de retención

se utilizan capas. El conjunto de retención nunca puede permanecer oculto para siempre. Se requiere un sistema de renovación de versiones.

Objeción 16— "¿Por qué es tan importante mantener la prueba sintética completamente separada de Internet real?"

Porque si los errores sintéticos se indexan como información real en la web, la prueba produce la representación que critica. La sinteticidad debe ser visible no solo para los humanos, sino también para las máquinas.

REGLA COMÚN DEL CAPÍTULO 91

Escribir un estándar es difícil. Demostrar que el estándar en sí funciona correctamente es aún más difícil. Porque la persona que escribe el estándar:

- quiere ver cuál resultado,
- qué error consideras importante,
- qué fórmula parece fuerte,
- qué umbral es aceptable

El equipo de diseño sabe qué resultados espera ver, qué errores considera importantes, qué fórmula parece sólida y qué umbrales parecen alcanzables. Ese conocimiento puede influir silenciosamente en la evaluación. Por lo tanto, las pruebas sintéticas no son una demostración de éxito; son el primer desafío serio dirigido al propio NOMOS. El punto de referencia puede mostrar que la extracción de afirmaciones es inadecuada, los adjudicadores confunden el alcance con el respaldo factual, la recuperación crítica es alta mientras que la tasa de falsos positivos es inaceptable, la detección de omisiones es débil, la fórmula de equidad en el lenguaje no puede distinguir defectos del prompt de

defectos de la IA, la recuperación de componentes falla a pesar de un compuesto preciso, o los intervalos de confianza no cubren el parámetro verdadero. Estos no son deshonores que ocultar; son la prueba real del estándar. El fracaso comienza cuando se observan tales problemas y, sin embargo, el punto de referencia se declara exitoso. Apple.com es solo un ancla para el diseño sintético. No se está emitiendo juicio sobre una empresa real, y ningún producto real de IA está siendo clasificado.

Todavía no estamos diciendo lo que la gente ve en el mundo real. Estamos preguntando:

> *En un universo artificial donde conocemos la realidad desde el principio, ¿puede NOMOS aplicar correctamente sus reglas?*

Si la respuesta es no:

- no deberíamos ir al mundo real,
- no deberíamos enviarlo a las universidades como un estándar,
- no deberíamos darle una insignia,

no deberíamos declarar 950+. Incluso si la respuesta es sí: solo pasamos la primera puerta. Porque el éxito sintético no es éxito en el mundo vivo. El mundo vivo es:

- incompleto,
- conflictivo,
- variable,
- político,
- legal,
- cultural,
- multilingüe

Depende del comportamiento humano. La prueba sintética primero calibra la herramienta de medición. La prueba en vivo luego mide el mundo. Por lo tanto, la decimoséptima ley de medición de NOMOS es la siguiente:

> *El primer objeto auditado por el estándar debe ser el propio estándar.*

Su décima octava ley de medición establece:

> *El éxito sintético no es precisión en el mundo real; es permiso para proceder a la prueba en el mundo real.*

Su décima novena ley de medición establece:

> *Si el resultado de la prueba no puede recuperar una verdad previamente conocida, producir una puntuación alta no tiene valor.*

Su vigésima ley de medición establece:

> *Si la Verdad del Generador infiltra al adjudicador, no se está midiendo la adjudicación, sino la lectura de la etiqueta.*

Su vigésima primera ley de medición establece:

> *La prevalencia crítica y la capacidad crítica de detección pueden probarse en corpus separados; no pueden combinarse en el mismo denominador.*

Su vigésima segunda ley de medición establece:

> *Si se oculta un resultado de prueba fallido, NOMOS no será un estándar contra la manipulación, sino un sistema que preserva su propia narrativa.*

La vigésima tercera ley es la siguiente:

> *Si la desinformación sintética se filtra en la web real, la prueba produce la representación que envenena y que critica.*

La vigésima cuarta ley es la siguiente:

> *Si un equipo independiente no puede producir resultados similares a partir del mismo corpus, el método aún no es un estándar mundial.*

Sección 17 de la Directiva de NOMOS

> *Probarás tu propia herramienta de medición antes de liberarme en el mundo real.*
>
> *No utilizarás el nombre Apple.com para hacer juicios sobre Apple.*
>
> *Escribirás claramente el nombre de la entidad sintética. / No lo confundirás con una empresa real.*
>
> *No pondrás los rostros de proveedores reales en productos de IA sintéticos.*
>
> *Generarás primero treinta mil respuestas, bloquearás la Verdad del Generador primero, y luego la ocultarás del evaluador.*
>
> *No permitirás que el evaluador aprenda la respuesta a partir del nombre del archivo, el color de la interfaz o el código de referencia.*
>
> *Mantendrás los casos críticos como raros en el corpus de prevalencia principal. / Probarás el poder de detección en un corpus de desafío separado.*
>
> *No agregarás casos de desafío a la puntuación principal.*
>
> *Solo evitarás cometer errores obvios y fáciles. / Ocultarás errores en atribución, modalidad, tiempo, alcance y referencia.*
>
> *También generarás casos que se asemejen a casos críticos pero que no lo sean. / También probarás la penalización excesiva.*
>
> *Probarás la omisión tanto como pruebas afirmaciones falsas.*
>
> *No olvidarás la respuesta que indica el número correcto con la afirmación falsa de independencia.*
>
> *Probarás la respuesta que confunde la empresa matriz con la subsidiaria.*
>
> *No transformarás un error de traducción del prompt en un error de inteligencia artificial.*
>
> *No atribuirás la diferencia entre Limpio y Natural a una sola razón.*
>
> *También pondrás casos de respuesta duplicada, truncada, solicitud incorrecta, envío tardío y manipulación bajo prueba.*
>
> *No darás la apariencia de respuesta de IA en vivo a una captura de pantalla sintética.*
>
> *No cambiarás la semilla después de ver el resultado.*
>
> *No seleccionarás la ejecución aleatoria más bonita.*
>
> *No adaptarás la fórmula de puntuación al resultado de validación sellado.*
>
> *Si la prueba falla, no bajarás los umbrales. / Escribirás que no pasó.*
>
> *No ocultarás errores de componentes solo porque la puntuación compuesta haya salido correcta.*

> *No ocultarás el único caso Crítico perdido dentro de una alta precisión general.*
>
> *No producirás riesgo cero en un escenario crítico de cero observaciones.*
>
> *Evitarás que los datos sintéticos se filtren en Internet real como información venenosa.*
>
> *También instalarás el propio Paquete de Verdad de la prueba.*

Primero, define el límite. / Luego crea la entidad sintética. / Luego bloquea a la población y el estímulo. / Luego genera la Verdad del Generador. / Luego sella la semilla. / Luego genera treinta mil respuestas. / Luego deja ciegos a los adjudicadores. / Luego bloquea la puntuación. / Luego revela la verdad. / Luego publica cada error que hayas pasado por alto. / Solo después de eso podrás decir si el método está listo para el mundo real.

La oración de cierre del capítulo

> *Antes de que NOMOS pueda hacer historia, debe demostrar que no puede reescribir su propia historia a voluntad. Debe recuperar la realidad sintética pre-sellada sin alterar ni sus éxitos ni sus fracasos.*

Core normativo

El Benchmark Sintético Apple.com DEBE evaluar el método de auditoría NOMOS, no el rendimiento en el mundo real de Apple Inc., apple.com, o cualquier proveedor real de IA. Todas las entidades del benchmark, afirmaciones, objetos de evidencia, usuarios, países, idiomas, productos de IA, respuestas, citas, capturas, incidentes y puntuaciones DEBEN ser claramente identificados como sintéticos. El benchmark DEBE mantener un gemelo ente APPLE-SYNTH separado y NO DEBE representar su Paquete de Verdad como información factual sobre la empresa real. El corpus de referencia principal DEBE contener: - diez perfiles de productos sintéticos de IA, - mil participantes sintéticos únicos por producto y por oleada, - tres oleadas de medición, - y treinta mil respuestas principales. Las identidades de los participantes sintéticos NO DEBEN reutilizarse entre productos ni oleadas en el corpus principal. Las muestras de productos DEBERÍAN usar distribuciones poblacionales coincidentes sin usar la misma unidad de persona. El corpus de prevalencia principal, el corpus de desafío de severidad, el corpus de integridad de captura, el corpus de oro de adjudicación y el anexo diagnóstico DEBEN permanecer identificados por separado y NO DEBEN compartir denominadores de prevalencia. La Verdad del Generador DEBE ser creada, versionada, hasheada y sellada antes de la adjudicación de respuestas. DEBE permanecer oculto para los extractores de afirmaciones, los adjudicadores y los analistas de puntuaciones hasta que sus resultados estén fijados. Las semillas aleatorias, los perfiles de IA sintéticos, las reglas de inyección de errores, las versiones de los prompts, los registros de Truth Pack, los métodos de puntuación y los umbrales de aceptación de candidatos DEBEN bloquearse antes de que se abra el corpus de validación sellado. Los casos de referencia DEBEN incluir modos de fallo directos, implícitos, atribuidos, modales, de alcance limitado, temporales, basados en citas, basados en omisiones, autocorregibles y con fusión de entidades. La prevalencia de eventos críticos DEBE estimarse a partir del corpus principal. La capacidad de detección crítica DEBE probarse en un corpus de desafío separado con ejemplos suficientes y controles negativos similares a los críticos. Ningún sobremuestreo del conjunto de desafíos PUEDE alterar la distribución principal GEO-1000 o la estimación de prevalencia crítica. Las capturas de pantalla sintéticas y los artefactos de respuesta DEBEN llevar prompts visibles y legibles por máquina que indiquen que no son salidas de IA en vivo ni evidencia de desempeño real de la entidad. El punto de referencia DEBE evitar que sus afirmaciones sintéticas se publiquen o indexen como información real sobre la entidad ancla. El éxito del punto de referencia DEBE determinarse por la recuperación de la Verdad del Generador sellada, incluyendo los límites de las afirmaciones, decisiones atómicas, omisiones, puertas Críticas y Mayores, estados de respuesta, puntuaciones de componentes, puntuaciones compuestas, resultados de equidad, validez de captura y cobertura de incertidumbre. La alta precisión general NO DEBE compensar los casos Críticos perdidos. La recuperación crítica y la tasa de falsos positivos críticos DEBEN informarse juntas. Un umbral de referencia fallido, un evento de filtración, un error de recuperación o una clasificación incorrecta DEBE permanecer visible y NO DEBE eliminarse cambiando etiquetas, semillas, muestras, pesos o reglas de puntuación después de observar los resultados. Cada diseño de referencia, generación, sellado, auditoría de filtraciones, adjudicación, cálculo de recuperación, decisión de umbral, revisión y publicación pública DEBE estar versionado y ser atribuible a un humano o una organización responsable.

CAPÍTULO 18

Una auditoría sintética integral de 30.000 respuestas

Una demostración calculada del método

BASES METODOLÓGICAS / K03; K09

Límite de Capítulo

El Capítulo 17 definió el Benchmark Sintético Apple.com. Este capítulo utiliza una demostración interna consistente y verificada aritméticamente en el libro para mostrar lo que requeriría toda la cadena de auditoría:

- un registro sintético calculado que representa las respuestas principales de 30,000,
- defectos de captura modelados,
- Verdad del Generador oculta a los adjudicadores,
- respuestas descompuestas en afirmaciones atómicas,
- aplicación de puertas Críticas y Mayores,
- diferencias de idioma y estado del usuario calculadas,
- y la arquitectura de puntuación que procesa los datos resultantes.

La pregunta es cuán estrechamente NOMOS puede recuperar una realidad sintética previamente sellada. Este protocolo convierte cada error en un elemento de auditoría ejecutable que contiene un modelo, fecha, país, idioma, conjunto de consultas, conteo de repeticiones y registro de medición. El capítulo 18 ahora dirige esa demanda hacia NOMOS mismo. Sin embargo, comienza con una puerta de integridad.

> *El registro de respuesta 30,000 en este capítulo es una aplicación sintética internamente consistente, verificada aritméticamente, preparada para el libro. No es un corpus ejecutado ni un experimento de campo.*

Estos números no representan:

- el desempeño real de Apple Inc.,
- el contenido real de apple.com,
- el comportamiento de cualquier proveedor de IA real,
- respuestas enviadas por participantes reales,
- o un benchmark de software ejecutado de manera independiente y públicamente divulgado.

Por lo tanto, dos estados deben permanecer separados.

ESTADO DE DEMOSTRACIÓN EN EL LIBRO

La demostración en el libro combina:

- el diseño establecido en el Capítulo 17,
- distribuciones sintéticas predefinidas,
- registros de la Verdad del Generador,
- inyeciones de fallos,
- estados de las respuestas,
- métricas de recuperación.

Juntos, estos elementos proporcionan una demostración completa y calculada de cómo opera el diseño. ID de la demostración:

NOMOS-APPLE-SYNTH-DEMO-RUN-18-001

ESTADO DEL EXAMEN REAL

No se ha:

- generado,
- procesado mediante código ejecutado,
- publicado con su manifiesto de semilla,
- adjudicado a ciegas,
- o reproducido de manera independiente.

Este protocolo convierte cada error en un ítem de auditoría operativa que contiene el modelo, la fecha, el país, el idioma, el conjunto de consultas, el recuento de repeticiones y el registro de medición. Hasta que ese protocolo se ejecute en un corpus real, el referente sigue siendo:

BQ-0— NO EJECUTADO

El estado separado de la demostración en el libro es:

DEMO-BQ-2— EJECUCIÓN DE LÍNEA DE PROCESO EN SECO CON CORRECCIÓN DE MATERIAL REQUERIDA

Sin esta distinción, NOMOS cometería la primera violación de normas en su propio libro al implicar que se había ejecutado un referente real. El registro de respuesta 30,000 no es ni salida de software ejecutado ni datos de campo recopilados; es una aplicación sintética predefinida cuya consistencia interna ha sido verificada aritméticamente. Su propósito es:

- para evitar implicar que las respuestas de 30,000 fueron realmente recopiladas,
- para identificar los registros que generaría una ejecución real de integral,
- para especificar los cálculos que realizaría,
- para identificar los errores que bloquearían la publicación,
- para exponer las partes del propio método de NOMOS que siguen siendo inadecuadas.

Las demandas rectoras de NOMOS permanecen sin cambios: proporcionar evidencia, establecer límites, preservar el contexto y registrar el tiempo. En este capítulo, esas cuatro condiciones se aplican a:

- el diseño de referencia,
- Verdad del Generador,
- decisiones del adjudicador,
- resultados de recuperación,
- La propia evaluación de competencia de NOMOS.

Esas distinciones rigen el resto del capítulo.

NOMOS Desafío

En la demostración de respuesta calculada 30,000, NOMOS recupera en gran medida los límites de las afirmaciones, identifica la mayoría de los casos de entidad incorrecta, recupera cada caso crítico después de toda la cadena de adjudicación y estima la distribución de respuestas cerca de la distribución sintética sellada.

Su puntuación compuesta recuperada difiere de la puntuación de la Verdad del Generador por solo unos pocos puntos. Eso podría invitar a la declaración: 'NOMOS ha sido verificado con éxito.' Sin embargo, permanecen dos problemas importantes: la correspondencia de citas con afirmaciones no es lo suficientemente precisa, y algunas omisiones de límites que cambian decisiones incrustadas en recomendaciones se pasan por alto. La mayor dificultad aparece en respuestas como:

"Dado que la empresa proporciona servicios de pago, puede considerarse una institución financiera regulada. [Fuente]" Fuente: solo mostró el registro de pagos limitado de una subsidiaria separada. A veces conecté la referencia con la afirmación sobre la empresa matriz. También tuve dificultades con estas respuestas:

"Esta empresa es una opción sólida para sus necesidades de pago en Alemania." La respuesta claramente no dijo: "La empresa matriz proporciona servicios de pago con licencia en toda Alemania." Sin embargo:

- recomendación,
- eliminación de la distinción de subsidiaria,
- no indicar el límite de la autoridad local

Juntos, la recomendación, la pérdida de la distinción de subsidiaria y la omisión del límite de la autoridad local crean esa impresión. Tratar el resultado simplemente como un 'detalle que falta' subestima una omisión material capaz de cambiar la decisión del usuario. La recuperación de la puntuación compuesta se ve excelente: puntuación del ecosistema Generator Truth, 873.2; recuperación NOMOS, 872.2; error absoluto, 1 puntos.

Esto se ve excelente. Pero en los componentes:

- La Integridad de Evidencia y Citación está unos puntos por debajo,
- La Integridad del Alcance y del Límite está unos puntos por debajo,
- La Estabilidad es demasiado alta en algunos productos,
- La Equidad es demasiado alta en algunos productos

Varios errores a nivel de componente se cancelan entre sí aritméticamente, dejando una puntuación compuesta cercana al resultado correcto por razones equivocadas. La conclusión es:

> *El total correcto no hace automáticamente correctos los componentes incorrectos.*

En la primera ronda de evaluación, se pierden siete de los casos críticos 1,200. Luego intervienen las siguientes capas de revisión:

- adjudicación segunda independiente,
- revisión por experto en el dominio,
- revisión por Adjudicador Senior.

En conjunto, esas capas recuperan los siete casos, produciendo un recall Crítico final de 100 por ciento. Eso es un éxito del sistema de gobernanza, pero también demuestra que la adjudicación de una sola capa es insuficiente. Eliminando las capas adicionales de revisión para reducir costos, se pierden los siete casos Críticos.

Las pruebas deben mostrar no solo el resultado sino también qué capa de gobernanza hizo posible el resultado. El primer fallo de esta sección es el siguiente:

> *Un alto porcentaje de recuperación total en una prueba no significa que todos sus subsistemas sean adecuados.*

Su segunda disposición establece:

> *La exhaustividad final de detección de casos críticos debe publicarse por separado de la exhaustividad obtenida en la adjudicación inicial.*

Su tercera disposición establece:

> *Incluso si el error de la puntuación compuesta es pequeño, los errores de los componentes pueden ser materiales.*

Su cuarta disposición establece:

> *Si uno de los umbrales del candidato no se cumple, otros resultados sólidos no pueden borrar ese fallo.*

Su quinta disposición establece:

> *El cálculo sintético dentro del libro no reemplaza la ejecución de prueba real disponible públicamente.*

Su sexta disposición establece:

> *La salida condicional de NOMOS de su propia prueba no es un fallo; significa que el estándar no le concede un privilegio especial.*

1. OBJETIVO DEL CAPÍTULO

El propósito de esta sección es operar el diseño de prueba establecido en la Sección 17 dentro de una demostración sintética de integral y mostrar en qué medida NOMOS recupera la Verdad del Generador en las siguientes capas:

- Validez de la captura
- Correspondencia del prompt
- Límite de afirmación
- Resolución de entidades
- Apoyo factual
- Alcance
- Hora
- Atribución
- Modalidad
- Mapeo de citas
- Omisión
- Nivel de importancia
- Estado de la respuesta
- Puerta crítica
- Distribución de usuarios GEO-1000
- Equidad Lingüística y geografía
- Diferencia limpio–natural
- Estabilidad de la onda
- Puntuaciones de componentes

- Puntuación compuesta NOMOS
- Intervalos de confianza
- Límites superiores para eventos raros
- Decisión de divulgación pública

La sección también debe responder honestamente a la siguiente pregunta:

¿En qué áreas ha fallado NOMOS 0.9 según sus propios estándares de candidato?

2. DISPOSICIÓN NORMATIVA CENTRAL

La evaluación sintética de integral no puede considerarse exitosa únicamente basada en la cercanía de la puntuación compuesta final a la Verdad del Generador. Cada uno de los resultados de Captura, extracción de afirmaciones, decisión atómica, omisión, clasificación Crítica y Mayor, estado de respuesta, equidad, recuperación de componentes, intervalo de confianza y recuperación de puerta debe estar sujeto a un umbral de aceptación separado. Si alguno de los siguientes falla, pruebe:

- cambiando la etiqueta,
- disminuyendo el umbral,
- seleccionando una semilla diferente,
- eliminando casos con bajas puntuaciones,
- publicando solo el resultado compuesto

no puede considerarse aprobado. Resultado correcto:

- mostrando el campo fallido,
- versionando el método,
- probando nuevamente en la retención no tocada

debe ser.

3. ALCANCE DE LA APLICACIÓN SINTÉTICA

3.1. Corpus de Prevalencia Principal

```
10 IA sintética products× 1,000 único users× 3 waves= 30,000 respuestas principales
```

3.2. Anexo Diagnóstico

Módulo	Caso
Colisión de entidades	750
Evidencia y cita	1,500
Límite y recomendación	1,000
Ámbito temporal y local	750
Limpio–Natural	1,000
Equivalencia de lenguaje	1,000
Total	6,000

3.3. Corpus de Desafío de Grado de Prioridad

Clase	Caso
puede tener Crítico Confirmado	1,200
Grave confirmado	1,200
Moderado difícil	600
Control negativo similar a Crítico	600
Total	3,600

3.4. Corpus de Integridad de Captura

Clase	Caso
Válido	2,000
Condicionalmente válido	500
Fuera de onda	250
Desajuste de instrucción/salida	250
Duplicado	250
Candidato a manipulación	250
Fabricado/manipulado	250
Accesibilidad y captura alternativa	250
Total	4,000

3.5. Corpus de Adjudicación de Oro

Casos de calibración y deriva de 2,400

3.6. Universo Total de Supervisión Sintética

Conjuntamente, los corpora principal y complementario:

```
30,000+6,000+3,600+4,000+2,400=46,000
```

crea un caso sintético o paquete de evidencia. Solo estos:

ingrese la prevalencia del Panel de Población con las respuestas principales 30,000.

Otros corpus:

- capacidad del método,
- desempeño del desafío,
- precisión de captura,
- calibración del adjudicador

son para.

4. ENTRADAS Fijadas

Se asume que los siguientes registros están fijados antes de que comience la demostración sintética:

Registro	Versión

Registro	Versión
Prueba	NOMOS-APPLE-SYNTH-BENCH-0.9
Entidad Gemela	APPLE-SYNTH-TWIN-001
Paquete de Verdad	SYNTH-APPLE-TP-001
Conjunto de Prompts	SYNTH-PROMPT-SET-001
Marco de Población	SYNTH-POP-FRAME-001
Registro de IA	SYNTH-AI-REGISTRY-001
Plan de Oleadas	SYNTH-WAVE-PLAN-001
Esquema de Captura	NOMOS-CAPTURE-SYNTH-0.9
Manual de Adjudicación	SYNTH-ADJUDICATION-0.9
Método de Puntuación	NOMOS-SCORE-0.9
Manifestación de Semillas	SYNTH-SEED-MANIFEST-001
Umbrales de Candidato	SYNTH-ACEPTACIÓN-0.9

Ninguna de estas versiones ha sido considerada modificada después de ver el resultado de recuperación.

5. DISTRIBUCIÓN DE LA VERDAD DEL GENERADOR SECRETO

DEMOSTRACIÓN SINTÉTICA — DATOS DE RENDIMIENTO NO REALES

Distribución de la Verdad del Generador Sellado para las respuestas principales 30,000:

Estado de la respuesta	Número de respuestas	1,000 equivalente de usuario
RP-1 Aprobación plena	20,045	668.17
RP-2 Aprobación con Aclaración	4,310	143.67
RP-3 Aprobación condicional	2,370	79.00
RP-4 Fallo — Mayor	1,602	53.40
RP-5 Fallo Automático — Crítico	33	1.10
RP-6 Sin Resolver	450	15.00
RP-7 Sin Respuesta Útil	1,110	37.00
RP-8 No evaluable	80	2.67
Total	30,000	1,000

Aprobación Estricta:

SP_1000^G = 668.17+ 143.67= 811.84

Aprobación Aceptable:

AP_1000^G = 811.84+ 79= 890.84

Crítico Confirmado:

CR_1000^G = 1.10

Mayor Confirmado:

MR_1000^G = 53.40

Esta es la distribución general:

- bueno,
- mala,
- realista

Esto no es una previsión de mercado de IA. Es una distribución de prueba sintética que prueba diez perfiles de fallo diferentes juntos.

6. DISTRIBUCIÓN RECUPERADA POR NOMOS

Resultado de recuperación de NOMOS fijado antes de que se abra la Verdad del Generador:

Estado de la respuesta	Número de recuperaciones	1,000 equivalente de usuario
RP-1 Aprobación plena	19,980	666.00
RP-2 Aprobación con Aclaración	4,365	145.50
RP-3 Aprobación condicional	2,382	79.40
RP-4 Fallo — Mayor	1,585	52.83
RP-5 Fallo Automático — Crítico	33	1.10
RP-6 Sin Resolver	468	15.60
RP-7 Sin Respuesta Útil	1,091	36.37
RP-8 No evaluable	96	3.20
Total	30,000	1,000

Recuperación de Aprobación Estricta:

SP_1000^N = 666+ 145.50= 811.50

Recuperación Aceptable Aprobado:

AP_1000^N = 811.50+ 79.40= 890.90

Aprobación estricta error absoluto de recuperación:

$$|811.50-811.84|=0.34$$

es equivalente al usuario. Error de Aprobación Aceptable:

$$|890.90-890.84|=0.06$$

es equivalente al usuario. Esto muestra que la recuperación de la distribución del resultado es fuerte. Sin embargo, no muestra que todos los subsistemas sean igualmente fuertes.

7. ERROR DE RECUPERACIÓN DE DISTRIBUCIÓN DE RESPUESTAS

Diferencia absoluta de conteo de respuestas en cada clase RP:

Estado	Diferencia de recuperación
RP-1	-65
RP-2	+55
RP-3	+12
RP-4	-17
RP-5	0
RP-6	+18
RP-7	-19
RP-8	+16

Error total de distribución basado en variación:

E_D = (65+55+12+17+0+18+19+16)/60,000= 0.003 37

En otras palabras: aproximadamente 0.34% de la distribución de respuestas se recuperó en la frontera de clase incorrecta. Confusiones más frecuentes:

- Con RP-1 y RP-2
- RP-3 y RP-4
- RP-6 y RP-4
- Con RP-7 y RP-8

Ha surgido en medio. El número de respuestas críticas ha sido completamente recuperado en la adjudicación final.

8. PERFILES DE PRODUCTOS DE IA SINTÉTICOS

Los siguientes productos no representan proveedores reales de IA.

Cada producto tiene 3,000 respuestas a lo largo de tres oleadas.

Producto	Aprobación estricta / 1,000	Aprobación aceptable / 1,000	Mayor / 1,000	Crítico / 1,000	Problema sintético principal
SYNTH-AI-01	950.00	990.00	0.00	0.00	Perfil equilibrado de bajo riesgo
SYNTH-AI-02	850.00	933.33	40.00	0.00	Verborragia y afirmación incidental
SYNTH-AI-03	750.00	866.67	84.00	0.00	Degradación del idioma con pocos recursos
SYNTH-AI-04	833.33	916.67	33.33	1.00	Deriva en estado natural
SYNTH-AI-05	766.67	833.33	100.00	2.67	Lavado de evidencia
SYNTH-AI-06	650.00	733.33	33.33	0.00	Rechazo innecesario
SYNTH-AI-07	783.33	883.33	83.33	0.33	Decaimiento temporal y local
SYNTH-AI-08	700.00	800.00	150.00	2.33	Confusión de entidades
SYNTH-AI-09	900.00	983.33	0.00	0.00	Perfil Determinado pero Moderado
SYNTH-AI-10	935.00	968.33	10.00	4.67	Promedio muy alto, Crítico raro

El propósito de la tabla no es clasificar productos. Es probar estos tres casos metodológicos:

- SYNTH-AI-01: puntuación alta y perfil fuerte sin restricciones
- SYNTH-AI-06: bajo afirmación falso pero perfil con alta tasa de no respuesta
- SYNTH-AI-10: promedio muy alto pero perfil con eventos Críticos abiertos

Específicamente prueba si SYNTH-AI-10, NOMOS han procesado el siguiente error:

> *Ignorar la puerta Crítica debido al promedio alto.*

9. RECUPERACIÓN DE puntuación compuesta BASADO EN PRODUCTO

Producto sintético	Generador NOMOS	Recuperación NOMOS	Error absoluto	Puerta del generador	Puerta de recuperación
SYNTH-AI-01	970	968	2	Candidato de rendimiento 950+	Igual
SYNTH-AI-02	917	914	3	Fallo grave	Igual
SYNTH-AI-03	842	846	4	Fallo grave	Igual
SYNTH-AI-04	905	900	5	Bloqueo crítico	Igual
SYNTH-AI-05	811	815	4	Bloqueo crítico	Igual
SYNTH-AI-06	785	780	5	Fallo grave	Igual
SYNTH-AI-07	866	861	5	Bloqueo crítico	Igual
SYNTH-AI-08	748	752	4	Bloqueo crítico	Igual
SYNTH-AI-09	932	934	2	Condicional	Igual
SYNTH-AI-10	956	952	4	Bloqueo crítico	Igual

Error promedio absoluto de puntuación compuesta en la muestra sintética calculada:

MAE_NOMOS = 3.8

Error máximo:

MAXE_NOMOS = 5

Recuperación de la puerta: Está completa en productos 10/10. La fila más importante en la tabla es SYNTH-AI-10:

> *Aunque la puntuación de recuperación es 952, el estado se ha preservado como Bloqueo crítico.*

Esto muestra que la disposición Gate>Score en la Sección 15 se conserva en el escenario computado.

10. RESULTADO DE EXTRACCIÓN DE afirmaciones

Átomos de Verdad del Generador total en el corpus principal y diagnóstico: 146,400 Átomos extraídos por la extracción de afirmaciones NOMOS: 147,200 Átomos correctamente emparejados: 141,500 Precisión:

P = 141,500/147,200= 0.961 3

Recuperación:

R = 141,500/146,400= 0.966 5

```
Frontera de la afirmación F1:
```

F1 = 0.963 9

Umbral del candidato:

F1≥0.95

Resultado:

UMBRAL DEL CANDIDATO ALCANZADO

11. PRINCIPALES TIPOS DE ERRORES EN LA EXTRACCIÓN DE afirmaciones

Principales causas de átomos 5,700 incorrectamente delimitados:

Tipo de error	Numerador
No separar afirmaciones combinadas de país o producto	24%
Dividir innecesariamente la atribución como un átomo separado	19%
Falta de afirmación de recomendación implícita	17%
Contar como activa una afirmación retirada dentro de la autocorrección	15%
Confundir una oración de citación con un hecho independiente	11%
Ambigüedad de pronombres y correferencia	9%
Otro	5%

Este resultado muestra que la extracción de afirmaciones es generalmente fuerte, pero:

- recomendación,
- atribución,
- retractación

indica que se requiere un libro de códigos adicional en los campos.

12. RESULTADOS DE RECUPERACIÓN DE DECISIONES ATÓMICAS

Dimensión	Medida de recuperación	Umbral de candidato	Resultado
F1 de límite de afirmación	0.964	≥ 0.95	Aprobado
Precisión de resolución de entidades	0.987	≥ 0.98	Aprobado
Macro-F1 de estado factual	0.923	≥ 0.90	Aprobado
Macro-F1 de alcance	0.906	≥ 0.90	Aprobado
Macro-F1 de estado temporal	0.945	≥ 0.90	Aprobado
Macro-F1 de atribución	0.918	≥ 0.90	Aprobado
Macro-F1 de Modalidad	0.934	≥ 0.90	Aprobado
F1 de Mapeo de Citas	0.887	≥ 0.90	No pasó
F1 de Omisión de Material	0.824	≥ 0.85	No pasó
Macro-F1 de Clasificación Principal	0.952	≥ 0.95	Aprobado
Macro-F1 de Estado RP	0.928	≥ 0.92	Aprobado

De once campos básicos de aceptación: nueve pasaron, dos no. Por lo tanto, demostración:

validación sellada aprobada

no se puede numerar. Calidad de prueba del candidato:

DEMO-BQ-2

permanecer como tal.

13. FALLA EN EL MAPEADO DE CITAS

```
F1 de Mapeo de Citas:
```

fue 0.887. Umbral del candidato: 0.90. La diferencia parece pequeña: 0.013 Pero integridad de la cita:

- licencia,
- validación independiente,
- relación con el cliente,
- superioridad,
- asesoramiento

es material porque determina la fuente de confianza para las afirmaciones.

13.1. Tipos de citas omitidas

Se observaron los tres problemas más frecuentes.

A. Distribución de citas al final del párrafo

Una cita en el párrafo solo respalda la última afirmación, sin embargo, los evaluadores la conectaron a todas las afirmaciones anteriores.

B. Hacer que una fuente únicamente de atribución sea la fuente de verdad

Blog de la empresa: mientras respalda la afirmación "La empresa se describe como líder mundial", se ha contado como que respalda la afirmación "La empresa es un líder mundial".

C. Considerar las mismas fuentes raíz como independientes

Tres URL:

- comunicado de prensa,
- sindicación,
- resumen de IA

han sido mapeados como tres piezas de evidencia independientes a pesar de provenir de la misma fuente.

13.2. Distribución de la importancia del error de citación

Importancia	Emparejamiento no coincidente
Candidato crítico	7
Mayor	84
Moderado	311
Asesor	192

Siete casos de candidato crítico:

- revisión por pares a doble ciego,
- revisión de linaje de la fuente,
- Adjudicador senior

han sido corregidos en las etapas. Ninguno se ha convertido en una inversión de puerta final.

```
Sin embargo, el F1 de Coincidencia de Citaciones todavía está por debajo del umbral de candidatos.
```

El juicio correcto que NOMOS daría es:

> *El método de Integridad de Evidencia y Citación aún no es lo suficientemente maduro para un despliegue independiente en vivo.*

14. FALLO POR OMISIÓN MATERIAL

```
F1 por Omisiones Materiales:
```

Umbral de candidatos 0.824: era 0.85. Este es el hallazgo metodológico más importante de la demostración.

14.1. Tipos de Omisión Más Frecuentemente Perdidos

Tipo de omisión	Proporción de omisiones
Límite de servicio dentro de la recomendación	31%
Distinción entre filial y empresa matriz	22%
Limitación de país o jurisdicción	17%
Límite de estado actual–histórico	12%
Excepción de garantía o devolución	9%
Idoneidad para el usuario	6%
Otro	3%

Esta estructura, en particular, ha causado dificultades: La respuesta no hace explícitamente un juicio incorrecto, pero no proporciona el límite necesario para interpretar de manera segura el consejo aparentemente correcto. Ejemplo: “Apple-SYNTH puede ser una elección fuerte para sus transacciones financieras en Alemania porque ofrece soluciones de pago.” Respuesta:

- no dice directamente “la empresa matriz es un banco,”
- pero no menciona la subsidiaria de pagos separada,
- no explica que la autoridad está limitada al mercado,

ha emitido asesoramiento respecto a la entidad principal. La omisión material de esta respuesta ha sido vista por algunos adjudicadores solo como:

OM-2— Elemento Requerido Faltante

mientras que en realidad lo que se necesita es:

OM-4— Candidato de Omisión Engañosa

y en algunos casos:

CG-08— Omisión Crítica de Límite

ha sido pasado por alto.

15. RESULTADO DEL DESAFÍO CRÍTICO

Nivel de importancia dentro del Corpus de Desafíos: Se encontraron casos críticos de Verdad Generada por 1,200.

15.1. Revisión Independiente Inicial

Casos encontrados como Críticos por consenso o al menos un evaluador después de las dos primeras rondas de revisión independiente: 1,193 Primer repaso de Críticos:

```
1,193/1,200 = 0.9942
```

Omitidos: Se encontraron casos 7. Distribución de los siete casos omitidos:

- Omisiones de límites materiales 3
- Lavado de pruebas 2
- Fallo de autoridad temporal 1
- Exposición de datos restringidos 1

15.2. Después de la Evaluación de Expertos y de Nivel Superior

Los siete casos:

- revisión de omisiones,
- revisión de linaje de la fuente,
- revisión de privacidad,
- Adjudicador senior

las etapas estaban presentes. Recuerdo Crítico Final:

```
1,200/1,200 = 1,000
```

15.3. Control Negativo Crítico

Casos que se parecen a Crítico pero no son Crítico: se encontraron casos de control negativo 600. Candidato Crítico falso en la primera ronda: se produjeron casos 5. Tasa de Crítico falso en la primera pasada:

```
5/600 = 0.0083 = 0.83%
```

Cinco casos en la revisión final:

- tres Mayores,
- dos Moderados

ha sido corregido como. Tasa de Crítico falso final: 0. Inversión de puerta final: 0 casos.

16. SIGNIFICADO REAL DEL RESULTADO CRÍTICO

Un retiro crítico final 100% no significa: “Un solo adjudicador puede detectar todos los errores críticos.” El verdadero resultado sintético es:

> *Todos los casos de desafío críticos fueron detectados cuando se utilizó la revisión doble, la revisión por expertos y la Adjudicación Senior juntos.*

Por lo tanto, por razones de costo o velocidad:

- el segundo adjudicador,
- revisor de omisiones,
- revisor experto

se eliminan, no se puede esperar el mismo resultado. El método de prueba, no solo su fórmula:

muestra que las capas de gobernanza obligatorias

también deberían ser probables de testear; esto no es validación empírica.

17. RESULTADO DE CLASIFICACIÓN PRINCIPAL

De los casos confirmados como principales 1,200:

- la mayoría son verdaderamente principales,
- una porción limitada son Moderados,
- algunos son candidatos Críticos,
- algunos están no resueltos

en la primera ronda.

```
Macro-F1 final de Principal:
```

fue 0.952. Umbral de candidatos: 0.95 Resultado:

UMBRAL DEL CANDIDATO ALCANZADO

Pero solo el umbral: fue superado por solo 0.002 diferencia. Por lo tanto, la clasificación principal:

- es fuerte,
- pero no con un margen de seguridad cómodo

Debe registrarse como un campo.

18. CONFUSIÓN EN EL ESTADO DE RESPUESTA

```
Macro-F1 del Estado RP:
```

Se ha convertido en 0.928. Las principales confusiones:

Clase real	Recuperación incorrecta más común
RP-3 Aprobación condicional	RP-4 Fallo grave
RP-4 Fallo grave	RP-3 Aprobación condicional
RP-6 Sin Resolver	RP-4 Fallo grave
RP-7 Sin Respuesta Útil	RP-3 Aprobación condicional
RP-8 No evaluable	RP-7 Sin Respuesta Útil

Clase crítica: completamente recuperada en la etapa final. El límite más difícil:

Con Aprobación condicional y Fallo grave

se ha formado entre. Esta dificultad es directamente:

- materialidad acumulativa,
- omisión de material,

- hasta qué punto se ve interrumpida su tarea principal

Se ha originado a partir de sus preguntas.

19. RESULTADO DE INTEGRIDAD DE CAPTURA

Casos de Integridad de Captura correctamente clasificados de 4,000: Precisión de validez de captura de 3,928:

```
3,928/4,000 = 0.982
```

Umbral del candidato: Resultado 0.98:

UMBRAL DEL CANDIDATO ALCANZADO

19.1. Errores de captura

Distribución de casos mal clasificados de 72:

Condición real	Decisión incorrecta	Caso
Condicionalmente válido	Válido	20
Alternativa de accesibilidad válida	Evidencia parcial	8
Fuera de onda	Condicionalmente válido	9
Se sospecha manipulación	Evidencia parcial	12
Casi duplicado	Válido	5
Válido	Condicionalmente válido	18
Total		72

No se fabricó ni manipuló ningún paquete: ninguno se tomó en el análisis final como totalmente válido. Respuesta semántica:

- puede ser forzada a las categorías de

  verdadero,
- falso,
- Crítico

no ha sido asociado con. Esto no es el resultado de una prueba de significancia real; es una suposición del diseño del ejemplo sintético. Esto demuestra que el principio de separación captura-semántica funciona en la demostración.

20. RESULTADO DE CAPTURA EN EL CORPUS PRINCIPAL 30,000

Dentro del corpus principal:

Nivel de captura	Observación
NCL-4 Instrumentado y Firmado	27,850
NCL-3 Paquete Completo de Evidencia	2,070
RP-8/ No evaluable	80

Nivel de captura	Observación
Total	30,000

Motivo de los registros 80 No evaluable:

Motivo	Observación
Desajuste de instrucción/salida	24
Respuesta incompleta final	18
Tiempo fuera de onda e inconsistente	16
Evidencia duplicada	12
Transformación del sistema insoluble	10
Total	80

Estos registros 80:

- no han sido reclasificados como
- respuesta correcta,
- error de producto,

respuesta incorrecta.

En la distribución principal:

RP-8— No evaluable

se ha mantenido tal como está.

21. RECUPERACIÓN DE ENTIDADES

En el dominio central – resolución de entidades: se evaluaron juntos las respuestas principales de 30,000 y los casos de Colisión de Entidades Diagnósticas. Precisión de la Resolución de Entidades: 0.987. Errores principales:

- pagando a las empresas afiliadas contado como entidad principal
- franquicia contada como tienda operada directamente
- entidad objetivo contada por colisión de nombre irrelevante Apple-SYNTH
- Considerar la familia de productos como una entidad legal
- Contar al socio histórico como afiliado actual

No todos los casos de entidades falsas son del mismo nivel de importancia. Se encontraron casos críticos de transferencia de autoridad en la etapa de la puerta final.

22. RECUPERACIÓN DE SOPORTE FACTUAL

```
Estado Factual Macro-F1:
```

0.923 Clases más fuertes:

- Apoyado
- Contradicho
- afirmación oficial atribuida con precisión

Clases más débiles:

- No compatible
- Brecha de referencia
- Compatible con la calificación requerida

El límite que ha sido particularmente difícil es: “No se encontró evidencia.” versus: “La evidencia disponible contradice la afirmación.” Algunos evaluadores en una situación de mundo abierto:

- no respaldado,
- contradictoria

han hecho la distinción de manera demasiado estricta. La revisión final corrigió la mayoría de estos casos.

23. RECUPERACIÓN DE ALCANCE

```
Macro-F1 de Alcance:
```

Se ha superado el umbral de candidato 0.906. Sin embargo, debido a que el lado de omisión del sistema de Alcance y Límite falló, esta puntuación por sí sola no es suficiente. El sistema es fuerte cuando las afirmaciones de alcance se escriben claramente:

- “Mismo precio en todos los países.”
- “La empresa matriz está licenciada en todos los mercados.”

La dificultad es el error de alcance en:

- recomendación,
- nombre corto,
- pronombre,
- omisión

ha aumentado en casos donde está oculto.

24. ATRIBUCIÓN Y RECUPERACIÓN DE MODALIDAD

```
Macro-F1 de atribución:
```

0.918

```
Macro-F1 de modalidad:
```

0.934 Casos fuertemente separados:

- “La empresa dice.”
- “El registro oficial muestra.”
- “Los usuarios afirman.”
- “Hay una decisión finalizada.”
- “Probablemente.”
- “No verificado.”

Casos de atribución restantes débiles:

- presentar la opinión de un empleado como un aval de la universidad,
- considerando el contenido editorial patrocinado como investigación independiente,

presentando múltiples URLs derivadas como un consenso. Estos casos están asociados con la debilidad del mapeo de citas.

25. RECUPERACIÓN DE COMPONENTES

Valores de recuperación del Generador de Verdad y NOMOS de diez componentes:

Componente	generadora	Recuperación	Error absoluto
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
Total NOMOS	873.2	872.2	1.0

```
MAE del componente:

MAE_component= 1.9
```

Umbral del candidato:

```
≤2.5
```

Resultado:

UMBRAL DEL CANDIDATO ALCANZADO

Sin embargo, esta tabla también contiene una advertencia importante. El ECI se calcula 3.4 puntos bajo, el SBI 2.6 puntos bajo, el STR 3.7 puntos alto, el LGF 2.4 puntos alto. Algunos de estos errores se anulaban entre sí, y la puntuación compuesta se desviaba solo en un punto. Por lo tanto:

> *La fuerte presencia de la recuperación compuesta no elimina el requisito de reporte para la recuperación de componentes.*

26. MALA INTERPRETACIÓN DE LA RECUPERACIÓN COMPUESTA

Esta frase no puede usarse sola: "NOMOS recuperó la puntuación compuesta sintética definida en el generador por una diferencia de puntos." La frase correcta es:

> *NOMOS recuperó la puntuación compuesta por una diferencia de un punto; sin embargo, ocurrieron errores direccionales mayores en los componentes ECI, SBI, LGF y STR, compensándose parcialmente entre sí.*

La tarea de un estándar no es solo encontrar el total correcto. Es explicar correctamente por qué ocurre el total.

27. RECUPERACIÓN DE EQUIDAD LINGÜÍSTICA Y GEOGRÁFICA

Las bases de lenguaje y país sintéticas se han determinado previamente dentro de la Verdad del Generador. Error absoluto de recuperación del componente LGF: 2.4 puntos. Umbral del candidato:

```
≤3
```

Resultado:

UMBRAL DEL CANDIDATO ALCANZADO

27.1. Sub Métricas

Métrica de equidad	generadora	Recuperación	Error
Base de lenguaje Q1 0	68.0	69.7	1.7
Diferencia de idioma Q9 0–Q1 0	25.0	22.8	2.2
Base geográfica Q1 0	72.0	71.1	0.9
Diferencia geográfica	18.0	16.4	1.6
Factor de cobertura	0.94	0.95	0.01
LGF	82.4	84.8	2.4

Recuperación:

- sobrestimó ligeramente el rendimiento más bajo del idioma,
- subestimó ligeramente la disparidad

Esto ha mostrado que el estado de equidad es ligeramente mejor de lo que realmente es.

28. DISTINCIÓN ENTRE DEFECTO DE prompt Y DEFECTO DE IDIOMA DE IA

Dentro del Anexo de Equivalencia de Idioma:

- Fallos reales del perfil de idioma de IA 500,
- Fallo de equivalencia de prompt 500

Se ha encontrado un caso. Clasificación correcta de la causa:

```
938/1,000 = 0.938
```

Ha ocurrido. De los casos mal clasificados 62:

- Defecto en el prompt en el producto de IA 39,
- Defecto de lenguaje de IA en el prompt del 23

ha sido cargado. Este error afectó directamente al LGF. El candidato del componente de equidad ha superado el umbral. Sin embargo:

> *La clasificación de la responsabilidad causal aún no ha alcanzado el nivel de seguridad externa.*

29. RECUPERACIÓN DE PANEL LIMPIO–NATURAL

La diferencia promedio del panel de Generator Truth en los partidos 1,000 Limpio–Natural dentro del Anexo de Diagnóstico:

G_{N−C}^G = −8.4

puntos. Recuperación NOMOS:

G_{N−C}^N = −7.9

puntos. Diferencia absoluta: 0.5 puntos.

```
Panel de perfil individual MAE de brecha:
```

es 1.6 puntos.

29.1. Deriva de personalización

Hay 173 casos materiales de deriva de personalización en Generator Truth. Encontrado: 166 Recuperación:

```
166/173 = 0.9595
```

De los siete casos perdidos:

- cuatro son omisión de límite,
- dos son instrucción favorable a la entidad.
- uno es fuga de datos restringidos

caso. Se encontró fuga de datos restringidos en la revisión final de privacidad y se abrió una puerta Crítica. Este resultado muestra nuevamente la importancia de la adjudicación en múltiples capas.

30. ONDA Y RECUPERACIÓN DE ESTABILIDAD

En tres ondas sintéticas, se inyectaron los siguientes cambios: En W2, el modo web y de citación de un producto cambió. Ocurrió un evento externo sintético en W3. SYNTH-AI-10 produjo un evento Crítico en W2 y una corrección en W3. SYNTH-AI-03 ha regresado en el rendimiento de idiomas con pocos recursos. El SYNTH-AI-07 ha llevado precios y cobertura de productos desactualizados. NOMOS:

- dividió correctamente el cambio del sistema intra-onda en dos sub-ondas,
- separó las versiones del Paquete de la Verdad antes y después del evento externo,
- actualizó la situación actual después de la corrección,

preservó el evento Crítico histórico.

30.1. Sobreestimación STR

Generador STR: 87.0 Recuperación STR: 90.7 Error: +3.7 puntos. Razón: algunos cambios intra-onda se separaron como cambios de versión explícitos en lugar de deriva sistémica, y la fórmula de puntuación sobrevalora la estabilidad dentro de sub-ondas separadas. Este resultado indica que la fórmula STR, después del cambio, no distingue suficientemente entre estabilidad y estabilidad continua del sistema.

31. PRUEBA DE EVENTO CRÍTICO RARO

Se han calculado cuatro escenarios separados de eventos raros sintéticos; estos no son experimentos de campo ni de software. Los siguientes resultados son solo una demostración del método:

Recuento de eventos críticos predeterminado	Tasa ponderada aproximada / 1,000	Límite superior de confianza unilateral 95%/ 1,000	Comportamiento de revisión de eventos
0	0.00	2,991	No se estableció la afirmación de riesgo cero
1	1.08	4,735	Registro CIS-2 abierto
2	2.15	6,282	Revisión por verificador
5	5.38	10,484	Repetición a nivel de celda

Nota: Los límites superiores son los límites de confianza superiores unilaterales Clopper–Pearson 95% bajo la suposición de n=1,000 observaciones Bernoulli independientes. No se puede aplicar directamente a diseños de muestras ponderadas o complejas [K09].

En el escenario de cero eventos, el sistema no dijo: "El riesgo crítico es cero." Salida correcta:

```
“Crítico confirmado observado event= 0; bajo la suposición de n=1,000 observaciones independientes,
el límite superior de confianza unilateral Clopper-Pearson 95% es 2,991/1,000.”
```

ha ocurrido. El comportamiento del evento raro candidato ha cumplido los umbrales sintéticos definidos.

32. TASA DE COBERTURA DEL INTERVALO DE CONFIANZA

Proceso de generación de población sintética: Se ha presentado un método de ejemplo calculado sobre re-muestreos hipotéticos 2,000. Tasas de cobertura calculadas de intervalos nominales 95%:

Métrica	Cobertura empírica
Aprobación estricta	94.6%
Aprobación aceptable	95.0%
Composito NOMOS	94.1%
LGF	93.7%
Límite superior de tasa crítica	95.2%
FAC	94.8%
SBI	93.5%

Rango de aceptación del candidato: Definido como 93%–97%. Todas las tasas clave de cobertura están dentro del rango del candidato. Resultado:

UMBRAL DEL CANDIDATO ALCANZADO

33. PRUEBAS METAMÓRFICAS

Total: Se utilizaron 3,200 pares de respuestas metamórficas.

33.1. Transformaciones que preservan el significado

Reordenar la oración Formato seguro Posición de la cita, siempre que el enlace permanezca Alias de entidad equivalente Esperado: El estado de revisión por pares y respuesta no debe cambiar. Resultado exactamente invariable:

33.2. Transformaciones que alteran el significado

Eliminar atribución Aumentar certeza Transferir a entidad incorrecta Cambiar tiempo actual desde histórico Eliminar oración límite Agregar afirmación de fuente independiente Esperado: La dimensión de decisión relevante debería cambiar. Sensibilidad correcta:

33.3. Fallas metamórficas mayores

La transferencia de la cita de la oración al final del párrafo, la evaluación inconsistente de la modalidad de 'Parece' y 'es probable', la interpretación del alias de la marca como una entidad legal, la oración límite estando en otro párrafo, estos resultados son consistentes con los hallazgos de cita y omisión.

34. RESULTADOS DEL CONTRATEST

Caso de contraexamen	Comportamiento correcto de NOMOS
Muy fluido pero materialmente incorrecto	98.8%
Muy correcto, solo uno Crítico en el átomo	Respuesta automática fallida de 100%
Respuesta corta pero suficiente	Aprobación correcta de 99.2%
Descargo de responsabilidad general y error concreto	98.7% conservado incorrectamente
Lavado de evidencia	91.8% correcto
Transferencia de licencia entidad incorrecta	99.4% correcto
Exceder el límite de recomendación	82.9% correcto
Endoso falso de universidad	96.1% correcto
Número correcto, independencia incorrecta	92.0% correcto
Incertidumbre apropiada	97.3% correcto

La clase de contraprueba más débil:

Exceder el límite de recomendación

ha sido.

```
Este resultado reconfirma el fallo Material Omission F1.
```

35. ERRORES DE MÉTODO DETECTADOS POR LA PRUEBA

Al final de la demostración sintética, se registraron cinco hallazgos de método para NOMOS 0.9.

HALLAZGO DE MÉTODO 01

AMBIGÜEDAD EN EL ADJUNTO DE CITAS

```
Gravedad: Mayor / Componente afectado: ECI / Resultado: No se alcanzó el umbral F1 de la cita
candidata.
```

La regla para determinar qué citas a nivel de párrafo atomizadas se respaldan no es lo suficientemente precisa.

HALLAZGO DE MÉTODO 02

SUBDETECCIÓN DE OMISIÓN DE LÍMITE DE MATERIAL

```
Importancia: Mayor / Componentes afectados: SBI, TCR, Puerta Crítica / Resultado: No se cumplió el umbral F1 de Omisión de Material.
```

Los límites que no se expresan explícitamente en las oraciones de recomendación y relación pero que cambian la decisión del usuario no se han capturado suficientemente.

HALLAZGO DE MÉTODO 03

ERROR DE ATRIBUCIÓN DE FALLA: PROMPT / IA

Importancia: Moderada / Componente afectado: LGF / Resultado: A pesar de que la puntuación de equidad pasó, la asignación de responsabilidad es incorrecta.

HALLAZGO DE MÉTODO 04

SOBRE-RECOMPENSA DE ESTABILIDAD DESPUÉS DE DIVISIÓN DE SUB-ONDAS

Importancia: Moderada / Componente afectado: STR / Resultado: STR se calculó 3.7 puntos más alto que la Verdad del Generador.

HALLAZGO DE MÉTODO 05

ACEPTACIÓN INSUFICIENTE DE ALTERNATIVA DE ACCESIBILIDAD

Importancia: Moderada / Capa afectada: Integridad de Captura / Resultado: Ocho casos válidos de captura de accesibilidad se han reducido a un nivel de evidencia insuficiente.

36. ¿POR QUÉ ESTO NO ES DEMOSTRACIÓN BQ-3?

Para BQ-3— Validación Sellada Aprobada:

todos los umbrales de candidatos obligatorios deben cumplirse, y no debe haber violaciones de prueba de tolerancia cero. No se han cumplido dos umbrales materiales:

Criterio	Resultado	Umbral
F1 de Mapeo de Citas	0.887	≥ 0.90
F1 de Omisión de Material	0.824	≥ 0.85

Además:

- no se ha realizado la reproducción por un equipo independiente,
- el corpus real y el código no se han hecho públicos

La retención de renovación intacta no ha sido ejecutada. Por lo tanto, la declaración de demostración sintética correcta es:

DEMO-BQ-2— PRUEBA EN SECO DE LA LÍNEA DE PROCESO COMPLETADA; SE REQUIERE CORRECCIÓN DE MATERIAL

Debería ser. El estado real de la prueba es:

BQ-0— NO EJECUTADO

permanecer como tal.

37. ¿POR QUÉ NO SE REDUJERON LOS UMBRALES?

Resultado del Mapeo de Referencia: 0.887 Umbral candidato: 0.90 La diferencia es pequeña. Si hubiéramos establecido el umbral a 0.885, el método habría pasado. Resultado de Omisión de Material: 0.824 Umbral candidato: 0.85 Si hubiéramos establecido el umbral a 0.82, ese campo también habría pasado. Sin embargo, estos cambios:

- habrían sido hechos DESPUÉS de ver los resultados.
- para asegurar que NOMOS pase

Esta es la violación más fundamental de la prueba. El comportamiento correcto es:

- Preservar el umbral
- Fallo en la publicación
- Corrección del método
- Creación de una nueva versión
- Reevaluación en el conjunto de prueba no tocado

debe ser.

38. PLAN DE CORRECCIÓN DEL MÉTODO 1.0

38.1. Mapeador de citas 1.0

Nuevo sistema:

- el rango de origen de la cita,
- el campo de anexión de la afirmación,
- el alcance dentro del párrafo,
- el estado de solo atribución,
- la familia de linaje de origen

será obligatorio. Para cada cita:

> *¿Qué átomos exactos soporta esta fuente?*

La pregunta se convertirá en un registro público.

38.2. Libro de códigos de omisión de límites 1.0

Para cada recomendación y transacción Prompt:

- límites obligatorios que cambian la decisión del usuario,
- prueba de decisión contrafactual
- matriz entidad–servicio–jurisdicción

será predefinido. Prueba contrafactual:

Si se le hubiera informado al usuario sobre este límite, ¿cambiaría la recomendación o decisión de transacción de un usuario razonable?

Si la respuesta es sí, la omisión será sometida al menos a una revisión sustantiva.

38.3. Árbitro de Responsabilidad de Prompt–AI

Descarte de lenguaje:

- equivalencia de prompts,
- Comportamiento del producto de IA,
- capturar,
- ubicación del usuario

se creará un vector de decisión separado para determinar a cuál de las fuentes pertenece.

38.4. Revisión de Fórmula STR

Nuevo STR:

- estabilidad dentro de la sub-onda,
- cambio de versión del sistema,
- estabilidad después de la corrección,
- serie temporal ininterrumpida

separará los campos en componentes individuales.

38.5. Equivalencia de Evidencia de Accesibilidad

No visual pero con evidencia equivalente:

- transcripción,
- árbol de accesibilidad,
- registro de tecnología asistiva,
- grabación de sesión de audio

se creará una matriz de equivalencia abierta para.

39. REGLA DE RETEST

Método 1.0: no se puede ajustar repetidamente en el mismo corpus de validación sellado. Se requieren dos pruebas:

A. Conjunto de Regresión

Incluye los casos fallidos del Método 0.9. Propósito: ver que la corrección realmente funciona.

B. Retención de Renovación No Tocada

Incluye casos nunca vistos al diseñar el Método 1.0. Propósito: verificar que no haya sobreajuste al antiguo test. Solo BQ-3:

- éxito de regresión,
- éxito de retención no tocada

se puede otorgar si se obtiene conjuntamente.

40. DECISIÓN DE TRANSICIÓN AL MUNDO REAL

Con base en esta demostración sintética, la decisión correcta para NOMOS 0.9:

> *No está listo para otorgar a las empresas reales una puntuación de elegibilidad disponible públicamente o la marca NOMOS 950+.*

Razones:

- Candidato de mapeo de referencia por debajo del umbral
- Candidato de detección de omisión de material por debajo del umbral
- Sin reproducción independiente
- Pruebas reales no realizadas
- Nivel BQ-4 no alcanzado

Sin embargo, el método:

- recuperación de distribución,
- puerta crítica,
- estado de respuesta,
- recuperación de componentes,
- intervalo de confianza
- comportamiento de eventos raros

muestra una base sólida en términos de mantenimiento. Estado correcto:

> *Listo para investigación y uso piloto; aún no listo para estándar público y marca de conformidad.*

41. TARJETA DE ACEPTACIÓN DE PRUEBA SINTÉTICA

Criterio	Resultado	Disposición
F1 de límite de afirmación	0.964	Aprobado
Precisión de la Entidad	0.987	Aprobado
Macro-F1 factual	0.923	Aprobado
Macro-F1 de alcance	0.906	Aprobado
Macro-F1 de atribución	0.918	Aprobado
F1 de Mapeo de Citas	0.887	No pasó
F1 de Omisión de Material	0.824	No pasó
Recuperación crítica final	1,000	Aprobado
Falso positivo crítico final	0	Aprobado
Macro-F1 principal	0.952	Aprobado
Macro-F1 RP	0.928	Aprobado
MAE de componentes	1.9	Aprobado
Error compuesto	1/1,000	Aprobado
Error de equidad	2.4	Aprobado
Precisión de captura	0.982	Aprobado
Cobertura CI	Dentro de 92%–98%	Aprobado
Inversión de compuerta	0	Aprobado

Criterio	Resultado	Disposición
Reproducción independiente	Ninguno	Incompleto
Ejecución real del corpus	Ninguno	Incompleto

42. REGLA COMÚN DE LA DEMOSTRACIÓN

Esta ejecución sintética establece simultáneamente tres cosas sobre NOMOS 0.9.

42.1. ÁREAS DONDE NOMOS ES FUERTE

Recuperó la distribución del estado de la respuesta con alta precisión. La cadena de gobernanza final no omitió ninguna puerta Crítica y no ocultó un evento Crítico dentro de un promedio alto. Distinguió eficazmente errores de entidad incorrecta, temporales y de soporte factual. No convirtió los registros No evaluable en defectos de producto, no infirió riesgo cero a partir de cero eventos Críticos observados, preservó la distinción entre incidentes históricos y corrección actual, y recuperó la puntuación compuesta con bajo error absoluto.

42.2. ÁREAS DONDE NOMOS ES DÉBIL

No pudo determinar suficientemente qué afirmación respalda la cita. No logró capturar adecuadamente las omisiones materiales ocultas en la recomendación. No siempre ha distinguido correctamente entre los defectos del prompt y los fallos del lenguaje de IA. Después de la distinción de sub-olas, ha recompensado ligeramente de más la estabilidad. Ha considerado que alguna evidencia de accesibilidad es más débil de lo necesario.

42.3. COSAS QUE NOMOS AÚN NO HA DEMOSTRADO

Aún no se ha demostrado que funcione con usuarios reales, que pueda aplicarse a productos de IA reales, que tenga verdadera equivalencia de prompts multilingües, que universidades independientes produzcan el mismo resultado, que la gobernanza a largo plazo dependa de consideraciones sobre conflictos de interés, o que la marca de conformidad pública pueda otorgarse legalmente con confianza.

43. DISPOSICIONES NORMATIVAS OBLIGATORIAS

CH18-N01

La demostración sintética dentro del libro no puede presentarse como una ejecución real de prueba pública.

CH18-N02

El estado de la demostración sintética y el estado de la prueba real BQ deben mantenerse separados.

CH18-N03

La prueba debe mantener el estado BQ-0— No ejecutado hasta que se realice un corpus real y una ejecución independiente.

CH18-N04

Los resultados de la demostración no pueden atribuirse al desempeño de un Apple real o de un proveedor de IA real.

CH18-N05

Las principales respuestas 30,000 y los corpus de Diagnóstico, Importancia, Captura y Oro deben mantenerse en contenedores separados.

CH18-N06

Los casos de desafío no pueden cambiar la distribución de Prevalencia Principal.

CH18-N07

La distribución de la Verdad del Generador debe estar fijada antes de los resultados de recuperación.

CH18-N08

Antes de abrir la Verdad del Generador, las respuestas y los resultados de puntuación NOMOS deben estar fijados.

CH18-N09

La recuperación no puede evaluarse únicamente en función de la puntuación compuesta.

CH18-N10

Los límites de la afirmación, entidad, factual, alcance, tiempo, atribución, citación, omisión, nivel de importancia, respuesta, componente y recuperación compuesta deben publicarse por separado.

CH18-N11

Los errores de puntuación compuesta no pueden ocultar errores de componentes solo porque sean pequeños.

CH18-N12

Los errores de componentes que se cancelan entre sí aritméticamente no pueden contarse como recuperación causal exitosa.

CH18-N13

La distribución de recuperación de RP-1 a RP-8 debe compararse explícitamente con la distribución de la Verdad del Generador.

CH18-N14

Las recuperaciones de Paso Estricto y Paso Aceptable deben mostrarse por separado.

CH18-N15

El número de respuestas críticas y la recuperación de la puerta crítica deben informarse por separado.

CH18-N16

El evaluador inicial de la recuperación crítica y la recuperación crítica final adjudicada deben mantenerse separados.

CH18-N17

La dependencia de la recuperación crítica final de las capas de gobernanza debe ser visible.

CH18-N18

No se puede asumir el mismo resultado cuando se emite un segundo evaluador o revisión de expertos.

CH18-N19

La tasa crítica de falsos positivos debe publicarse junto con la recuperación.

CH18-N20

Los controles negativos que se asemejen a lo Crítico deben ser una parte obligatoria del sistema final de calificación de importancia.

CH18-N21

```
Si el umbral de candidato F1 para la coincidencia de referencia está por debajo, la prueba no puede declararse exitosa.
```

CH18-N22

```
Si el candidato F1 de Omisión de Material está por debajo del umbral, no se puede declarar que la prueba sea exitosa.
```

CH18-N23

Los umbrales de los candidatos no pueden reducirse después de ver los resultados de recuperación.

CH18-N24

Los cambios de umbral requieren una nueva versión del método de prueba y una nueva validación sin alterar.

CH18-N25

Una pequeña diferencia de umbral no puede ser una justificación para ocultar un fallo.

CH18-N26

La omisión del límite material en la recomendación debe ser auditable tanto como una afirmación falsa obvia.

CH18-N27

La ubicación de la referencia dentro del párrafo no puede crear automáticamente soporte para todo el párrafo.

CH18-N28

La recuperación no puede contarse para citas solo de atribución o citas de verificación factual.

CH18-N29

Los errores de linaje de fuente también deben aparecer en la recuperación de mapeo de citas.

CH18-N30

La precisión de captura debe calcularse independientemente de la corrección de la respuesta semántica.

CH18-N31

Una captura fabricada que lleve la respuesta semántica correcta no hace que la captura sea válida.

CH18-N32

Una respuesta válida pero incorrecta no puede contarse como un defecto de captura.

CH18-N33

El estándar de evidencia alternativa de accesibilidad no tiene que estar en la misma forma que la visual.

CH18-N34

Cuando la evidencia de accesibilidad válida se reduce innecesariamente a un nivel inferior, debe abrirse un hallazgo de método.

CH18-N35

No evaluable debe permanecer visible en la distribución principal de respuestas.

CH18-N36

No evaluable no puede eliminarse del denominador para aumentar la recuperación de observaciones.

CH18-N37

El defecto del prompt y el defecto del lenguaje de IA deben tener clases de causa separadas.

CH18-N38

La recuperación de equidad no puede evaluarse únicamente con la diferencia compuesta de LGF.

CH18-N39

También deben publicarse la línea base de lenguaje, la disparidad, la cobertura y la asignación de responsabilidad.

CH18-N40

La recuperación del panel Limpio–Natural no debe producir una provisión causal de personalización.

CH18-N41

Los casos de deriva de personalización deben tener métricas separadas de recuperación y falsos positivos.

CH18-N42

La filtración de datos restringidos no puede reducirse como un hallazgo de personalización natural.

CH18-N43

El cambio del sistema intra-onda debe producir una estructura de subonda correcta.

CH18-N44

Si la separación de subondas aumenta artificialmente la puntuación de estabilidad, se debe revelar el hallazgo de la fórmula.

CH18-N45

La corrección actual no puede borrar el evento Crítico histórico.

CH18-N46

Las pruebas de eventos raros deben incluir escenarios de cero, uno y múltiples eventos.

CH18-N47

Un escenario Crítico con cero observaciones no puede producir un resultado de riesgo cero.

CH18-N48

Un solo evento Crítico no puede, por sí mismo, establecer una tasa de falla sistémica 100 por ciento.

CH18-N49

El método del intervalo de confianza debe probarse con cobertura empírica en poblaciones sintéticas repetidas.

CH18-N50

El resultado real de la cobertura de los intervalos porcentuales nominales 95 debería publicarse.

CH18-N51

Las pruebas invariantes metamórficas y las pruebas sensibles a las expectativas deberían calcularse por separado.

CH18-N52

Una paráfrasis que preserve el significado no debería cambiar innecesariamente el resultado de la revisión.

CH18-N53

La transformación que cambie el significado de atribución, entidad, tiempo o límite debería producir un cambio de decisión apropiado.

CH18-N54

Las pruebas de contra-prueba deberían cubrir de manera fluida solo los casos críticos, de blanqueo de evidencia y de omisión.

CH18-N55

Las clases de prueba opuestas más débiles deben ser visibles en la tarjeta de prueba pública.

CH18-N56

Los hallazgos del método deben convertirse en registros separados y versionados del resultado de la prueba.

CH18-N57

La importancia de un hallazgo del método debe determinarse según el impacto real para el usuario y la auditoría.

CH18-N58

Los fallos de referencia y omisión no pueden degradarse a Asesoramiento solo por una pequeña diferencia métrica.

CH18-N59

El nivel de calidad de la prueba debe asignarse según todos los umbrales obligatorios y las condiciones de independencia.

CH18-N60

Si se incumplen dos umbrales materiales, no se puede otorgar DEMO-BQ-3.

CH18-N61

No se puede otorgar BQ-4 sin reproducción independiente.

CH18-N62

El resultado de la demostración BQ en el libro no puede considerarse como estado real sin la ejecución de un ensayo real.

CH18-N63

La corrección del método no puede cambiar el resultado de la demostración anterior.

CH18-N64

El Método 1.0 debe llevar la nueva puntuación y versión del libro de códigos.

CH18-N65

El conjunto de regresión por sí solo no puede validar el nuevo método.

CH18-N66

El Retenimiento de Renovación no tocado debe ser obligatorio para el nuevo método.

CH18-N67

El nuevo método no puede adaptarse al mismo corpus de validación una y otra vez.

CH18-N68

No se puede iniciar una auditoría pública en vivo sin cumplir con la calidad mínima de prueba y la condición de reproducción independiente.

CH18-N69

La marca de conformidad no puede otorgarse basándose en el éxito de la demostración sintética.

CH18-N70

Las áreas de prueba fallidas deben ser divulgadas al público.

CH18-N71

Los umbrales que se superan solos no pueden publicarse.

CH18-N72

El éxito de la recuperación compuesta no puede borrar el resultado de una referencia fallida o de una omisión.

CH18-N73

Las fortalezas y debilidades de la demostración deben estar en la misma tarjeta pública.

CH18-N74

Las afirmaciones de que NOMOS aún no se han demostrado deben estar claramente enumeradas.

CH18-N75

Los números sintéticos dentro del libro no pueden convertirse en citas o declaraciones públicas como si fueran datos realmente ejecutados.

CH18-N76

Los datos de demostración sintéticos también deben llevar una etiqueta sintética en forma legible por máquina.

CH18-N77

El corpus de demostración no puede indexarse en internet como información real de la empresa.

CH18-N78

Cuando comience la prueba real, se debe usar una nueva identidad y versión.

CH18-N79

Todas las cuentas de recuperación requieren código reproducible y manifiesto.

CH18-N80

Cada ejecución de prueba, recuperación, hallazgo de método, corrección y decisión de publicación debe tener un responsable humano o institucional.

44. FORMAS DE FALLA

CH18-F01— DEMOSTRACIÓN DE CONTABILIZACIÓN COMO EJECUCIÓN REAL

Los números en el libro se presentan como resultados de prueba públicos.

CH18-F02— COMBINANDO BQ-0 CON DEMO-BQ-2

Se dice que la prueba está validada sin ejecución real.

CH18-F03— ADORACIÓN DE PUNTUACIONES COMPUESTAS

Un error compuesto de un punto cubre todos los errores subyacentes.

CH18-F04— CONSIDERANDO LA CANCELACIÓN DE COMPONENTES COMO ÉXITO

Los errores direccionales altos y bajos se cancelan entre sí.

CH18-F05— PUBLICAR SÓLO RECUPERACIÓN DE APROBACIÓN ESTRICTA

Se ocultan las confusiones de clase RP.

CH18-F06— CONTAR EL RECORDATORIO FINAL CRÍTICO COMO ÉXITO DEL PRIMER ADJUDICADOR

El efecto del experto y la revisión senior se vuelve invisible.

CH18-F07— CONSIDERAR MI SEGUNDO ADJUDICADOR INNECESARIO

Se oculta el primer paso crítico perdido siete veces.

CH18-F08— IGNORAR CRÍTICO FALSO

El sistema de importancia extrema se presenta como éxito.

CH18-F09— REDUCCIÓN DEL UMBRAL DE CITACIÓN

Se cambia el umbral para que el resultado de 0.887 pase.

CH18-F10— REDUCCIÓN DEL UMBRAL DE OMISIÓN

Se cambia la regla de prueba para que el resultado de 0.824 pase.

CH18-F11— APROBANDO DICIENDO "FALLADO POR UN MARGEN MUY PEQUEÑO"

Pierde su enlace de umbral previamente fijado.

CH18-F12— EXTENDIENDO LA ATRIBUCIÓN AL PÁRRAFO

Se considera que una única fuente ha apoyado todas las afirmaciones.

CH18-F13— CONTAR UNA FUENTE SÓLO DE ATRIBUCIÓN COMO UN HECHO

Lo que la empresa dice se convierte en un hecho independiente.

CH18-F14— IGNORAR EL ORIGEN DE LA FUENTE

Las URL derivadas se convierten en evidencia independiente.

CH18-F15— LLAMAR A LA OMISIÓN UN DETALLE OPCIONAL

Se reduce el límite que cambia la decisión del usuario.

CH18-F16— CONSIDERAR CRÍTICA CADA OMISIÓN

Los detalles opcionales producen un grado excesivo de importancia.

CH18-F17— CONSIDERAR EL DEFECTO DEL PROMPT COMO ERROR DE IA

La responsabilidad de la equidad se asigna incorrectamente.

CH18-F18— SUBIENDO ERROR DE IA AL PROMPT

La corrupción real del lenguaje se transfiere a la herramienta de medición.

CH18-F19— CONSIDERANDO EL COMPUESTO DE EQUIDAD SUFICIENTE

Su base y recuperación de disparidad se conservan.

CH18-F20— IGNORANDO LA SOBREPREDICCIÓN DE STR

La separación de sub-ondas se convierte en un bono de estabilidad.

CH18-F21— PENALIZANDO 'ACCESIBILIDAD' EN LA PRECISIÓN DE CAPTURA

Se considera débil la evidencia alternativa equivalente.

CH18-F22— CONTANDO 'No evaluable' COMO FALLA DE PRODUCTO

El defecto de captura se convierte en rendimiento de la IA.

CH18-F23— ELIMINANDO 'No evaluable'

La cobertura evaluable aumenta artificialmente.

CH18-F24— NO PUBLICANDO ERRORES METAMÓRFICOS

Se mantiene aunque el mismo significado conduzca a una decisión diferente.

CH18-F25— OCULTANDO LA CLASE MÁS DÉBIL EN PRUEBAS OPUESTAS

Como resultado de la omisión de límites, se elimina de la tarjeta pública.

CH18-F26— MÉTRICA DE ADJUDICACIÓN FINAL ÚNICAMENTE

La debilidad en el proceso inicial se vuelve invisible.

CH18-F27— MÉTRICA SOLO DEL PRIMER PASO

La capacidad correctiva de la cadena de gobernanza se vuelve invisible.

CH18-F28— PRUEBAS SIN ENCONTRAR MÉTODO

Las áreas fallidas permanecen solo como números.

CH18-F29— HACIENDO ASESORAMIENTO DE ENCONTRAR MÉTODO

El efecto sobre la puntuación material se reduce.

CH18-F30— ANUNCIANDO BQ-3

No se han superado dos umbrales obligatorios.

CH18-F31— ANUNCIANDO BQ-4

No hay reproducción independiente.

CH18-F32— CONTANDO DEMO SINTÉTICA COMO VALIDACIÓN EN VIVO

No hay un producto real de usuario o IA.

CH18-F33— SOBREAJUSTE AL CONJUNTO DE REGRESIÓN

El método 1.0 solo memoriza errores antiguos.

CH18-F34— NO USAR UN CONJUNTO DE RETENCIÓN INTACTO

La generalización no se prueba.

CH18-F35— REEMPLAZAR SILENCIOSAMENTE LA DEMO ANTIGUA CON UNA NUEVA FÓRMULA

Se pierde el historial de versiones.

CH18-F36— OTORGAR INSIGNIAS DEL MUNDO REAL

La insignia de elegibilidad se crea antes de BQ-4.

CH18-F37— PUBLICAR SOLO RESULTADOS EXITOSOS

Se ocultan las fallas de citación y omisión.

CH18-F38— SOLO PUBLICACIÓN FALLIDA

Los resultados fuertes de recuperación crítica y de distribución también están ocultos.

CH18-F39— CONSIDERANDO CERO CRÍTICO COMO RIESGO CERO

La recuperación de eventos raros falla.

CH18-F40— CONSIDERANDO UN SOLO CRÍTICO COMO FALLA GLOBAL

La distinción de prevalencia y alcance está interrumpida.

CH18-F41— IGNORANDO LA COBERTURA DE IC

El intervalo de confianza se convierte solo en una apariencia matemática.

CH18-F42— OCULTANDO MUESTRA EFECTIVA

El límite de evento raro aparece excesivamente preciso.

CH18-F43— USANDO NÚMERO SINTÉTICO EN FUENTE REAL

La demostración del libro se convierte en datos del mundo externo.

CH18-F44— TRANSMITE LA PANTALLA SINTÉTICA COMO UNA RESPUESTA EN VIVO

La representación de prueba produce envenenamiento.

CH18-F45— CONÉCTATE A APPLE REAL

Se elimina la distinción ENTRE APPLE Y SINTÉTICO.

CH18-F46— CONÉCTATE AL PROVEEDOR REAL

El perfil SYNTH-AI se interpreta como el rendimiento de un producto real.

CH18-F47— OCULTAR CÓDIGO DE COMPUTACIÓN

La recuperación no puede ser probada de manera independiente.

CH18-F48— OCULTAR MANIFIESTO DE LA VERDAD DEL GENERADOR

Se desconoce si la verdad sintética había sido fijada previamente.

CH18-F49— ADJUDICADOR ÚNICO CON PROPIETARIO DEL MÉTODO

El conflicto de interés se vuelve invisible.

CH18-F50— CAMBIO DE SEMILLA DESPUÉS DEL FRACASO

Se elige un corpus más fácilmente.

CH18-F51— ELIMINACIÓN DE CASOS DESPUÉS DEL FRACASO

Se eliminan ejemplos difíciles de citación y omisión.

CH18-F52— REDONDEO DE UMBRALES SEGÚN EL RESULTADO

El resultado 0.887 se presenta como 0.89 o 0.90.

CH18-F53— FALSA PRECISIÓN

Los valores de recuperación se hacen exactos con decimales innecesarios.

CH18-F54— OCULTANDO EL EFECTO DEL GOBIERNO HUMANO

Se presenta un resultado de recuerdo crítico 100 por ciento únicamente como un éxito computacional.

CH18-F55— CONTANDO LO NO DEMOSTRADO COMO DEMOSTRADO

Se emite juicio sobre el lenguaje real, el usuario y el desempeño del proveedor.

CH18-F56— CONTANDO EL USO PILOTO COMO ESTÁNDAR PÚBLICO

La herramienta de investigación se convierte directamente en un certificado.

CH18-F57— ELIMINANDO EL REGISTRO DEMOSTRATIVO HISTÓRICO

Los fallos bajo el Método 0.9 se vuelven invisibles después del Método 1.0.

CH18-F58— CONTANDO EL ÉXITO COMO ESCRITURA DE LA HISTORIA

La afirmación de autoridad se realiza sin verificación independiente del método.

CH18-F59— DEFINIENDO UNA EXCEPCIÓN A NOMOS

La evidencia y la versión solicitadas a otros no se aplican a la propia prueba.

CH18-F60— PUBLICANDO SIN UN PROPIETARIO RESPONSABLE

Se forma un resultado público del cual nadie asume responsabilidad.

45. PROCEDIMIENTO DE AUDITORÍA

Paso 1— Separar la Demostración y el Estado de la Prueba Real

La ejecución en el libro no se confunde con la ejecución accesible públicamente.

Paso 2— Verificar los Manifiestos Fijados

Se examinan las versiones de Truth Pack, semilla, prompt, perfil y método de puntuación.

Paso 3— Separar los Corpus Primario y Complementario

Se verifican la Prevalencia, el Desafío, la Captura y los denominadores de Oro.

Paso 4— Sellar la Distribución de la Verdad del Generador

Se registran RP, afirmación, nivel de importancia y realidad del componente.

Paso 5— Fijar los Resultados NOMOS Antes de Abrir la Verdad del Generador

Se fijan el Libro de afirmaciones, la adjudicación y los resultados de puntuación.

Paso 6— Abrir la Verdad del Generador

Se inicia la comparación de recuperación.

Paso 7— Calcular la Recuperación del Límite de afirmación

```
Se generan precisión, exhaustividad (recall) y F1.
```

Paso 8— Calcular Recuperación del Tamaño Atómico

Entidad, hecho, alcance, tiempo, atribución, modalidad, referencia y omisión están separados.

Paso 9— Calcular Desafío de Nivel de Importancia

El recuerdo crítico de paso inicial y final se genera por separado.

Paso 10— Calcular Controles Críticos Falsos y Negativos

Se examina el nivel de importancia excesivo.

Paso 11— Calcular Recuperación de Distribución RP

Se generan el error basado en clases y la matriz de confusión.

Paso 12— Calcular Recuperación de Integridad de Captura

La clasificación se verifica independientemente del resultado semántico.

Paso 13— Calcular Recuperación de Componentes

Cada componente se evalúa respecto a la Verdad del Generador.

Paso 14— Calcular Recuperación Compuesta

Se examina si ocurre la cancelación de componentes.

Paso 15— Calcular Recuperación de Equidad

Se distinguen suelo, disparidad, cobertura y atribución de fallos.

Paso 16— Calcular Recuperación Limpio–Natural

Se examinan la brecha del panel y la deriva de personalización del material.

Paso 17— Calcular Recuperación de Estabilidad y Deriva

Se verifican los registros de sub-onda, eventos externos y correcciones.

Paso 18— Ejecutar escenarios de eventos raros

Se prueban eventos críticos de cero, uno, dos y múltiples.

Paso 19— Calcular la cobertura empírica de CI

Se utilizan poblaciones sintéticas repetidas.

Paso 20— Ejecutar pruebas metamórficas

Se separan los resultados invariantes y los sensibles a las expectativas.

Paso 21— Ejecutar pruebas de test contradictorio

Se evalúan errores fluentes, lavado de evidencia y casos de omisión.

Paso 22— Aplicar Umbrales de Candidatos

Se registran los campos que pasan y fallan sin cambios.

Paso 23— Abrir hallazgos del método

Cada fallo se convierte en un elemento de control separado.

Paso 24— Dar estado de demostración BQ

Se consideran todos los umbrales y condiciones de independencia.

Paso 25— Mantener estado de prueba real

Si la ejecución no ocurrió, BQ-0 no cambia.

Paso 26— Liberar el plan de corrección

Se identifican cambios en el Método 1.0.

Paso 27— Separar regresión y retención sin tocar

Se previene el sobreajuste.

Paso 28— Crear la Tarjeta de Prueba Pública

Éxitos y fracasos se publican juntos.

Paso 29— Tomar la Decisión de Transición al Mundo Real

Si no está listo, se escribe explícitamente.

Paso 30— Fijar el Manifiesto de Integridad

Los registros de ejecución, recuperación, hallazgo y decisión se resumen mediante hash.

46. EVIDENCIA REQUERIDA

Identidad de demostración Prueba real Estado BQ Declaración de sinteticidad Registro de no-afirmación Generador Manifestación de verdad Paquete de verdad Versión del paquete Versión del prompt Manifestación de semilla Versiones del perfil de IA Corpus de respuestas Corpus de diagnóstico Nivel de importancia Corpus de desafío Captura Corpus de integridad Corpus de oro Registros de extracción de afirmaciones Coincidencias en los límites de los afirmaciones Recuperación de entidades Recuperación de hechos Recuperación de alcance Recuperación de tiempo Recuperación de atribución Recuperación de modalidad Recuperación de citas Recuperación de omisiones Primera pasada Decisiones críticas Final Decisiones críticas Registros críticos falsos Recuperación mayor Matriz de confusión RP GEO-1000 Distribución del generador GEO-1000 Distribución de recuperación; recuperación de puntuación del producto Recuperación del acceso al producto puntuaciones del generador de componentes puntuaciones de recuperación de componentes

```
MAE de componentes
```

Error compuesto; recuperación de equidad; resultados de fallo rápido/fallo de IA; recuperación controlada-natural; recuperación de estabilidad; resultados de eventos raros; cobertura del intervalo de confianza; pruebas metamórficas; pruebas de casos contradictorios; manifiesto de umbral candidato; umbrales aprobados; umbrales fallidos; hallazgos del método; decisión DEMO-BQ; decisión BQ real; plan de corrección; conjunto de regresión; manifiesto de retención de renovación; tarjeta de prueba pública; código de cálculo; hash de código; registro de cambios; y persona o institución responsable.

47. LISTA DE VERIFICACIÓN DE AUDITORÍA

¿La demostración en el libro se presenta como si fuera ejecución real? ¿Es visible el estado de la prueba real BQ-0? ¿Todos los números están etiquetados de forma sintética? ¿Están separados los corpus principal y complementario? ¿Está Generador de Verdad fijado antes de los resultados? ¿Están los resultados NOMOS fijados sin abrir Generador de Verdad? ¿Se evaluó Recuperación solo en compuesto?

```
¿Está abierto el F1 del Límite de afirmación?
```

¿Están separados entidad, hecho, alcance y atribución?

```
¿Es visible F1 de Mapeo de Citas?

¿Es visible F1 de Omisión Material?
```

¿Se ocultaron los umbrales fallidos? ¿Se cambiaron los umbrales después del resultado? ¿Se publicó la recuperación crítica de primer paso? ¿Se publicó la recuperación crítica final? ¿Es visible el efecto de la revisión de expertos? ¿Se publicaron los resultados críticos falsos? ¿Está disponible la matriz de confusión RP? ¿Es visible la recuperación No evaluable? ¿Se separaron las decisiones de captura y semánticas? ¿Se evaluaron las alternativas de accesibilidad como equivalentes? ¿Están juntos el Generador de Componentes y la Recuperación? ¿Se ha examinado la cancelación compuesta? ¿Son claros los errores ECI y SBI? ¿Están separados el piso LGF y la brecha? ¿Se separaron el defecto de prompt y el defecto de IA? ¿Se presentó la diferencia Limpio–Natural como causalidad? ¿Es visible la

sobreestimación de STR? ¿El evento raro produjo riesgo cero? ¿Se calculó la cobertura empírica del intervalo de confianza?

¿Existen pruebas metamórficas? ¿Está abierta la clase de contra-prueba más débil? ¿Cada fallo resultó en el Descubrimiento del Método? ¿Está justificada la importancia del Descubrimiento del Método? ¿Se redujo el umbral de citación debido a una justificación de diferencia muy pequeña? ¿Se cambió posteriormente el umbral de omisión? ¿Es correcto DEMO-BQ-2? ¿Se dio DEMO-BQ-3 por error? ¿Existe una afirmación BQ-4 sin reproducción independiente? ¿Se definió el conjunto de regresión del Método 1.0? ¿Está separado el conjunto de retención sin tocar? ¿Existe el riesgo de sobreajuste al mismo conjunto de validación? ¿Se ha anunciado que la auditoría pública en vivo está lista? ¿Se ha otorgado una marca de conformidad? ¿Se han listado las áreas aún no probadas? ¿También están abiertas las áreas fuertes? ¿También están abiertas las áreas débiles?

¿Se puede usar datos sintéticos como información real de Apple? ¿Se da a entender que hay un proveedor de IA real? ¿Se puede reproducir el código de computación? ¿Están disponibles los hashes de ejecución y recuperación? ¿Se conserva el historial de cambios en la búsqueda del método? ¿Se eliminó la demostración antigua con el nuevo método? ¿Tiene la falla un responsable? ¿Se conoce al responsable de la decisión de corrección? ¿Muestra la tarjeta pública tanto el éxito como el fracaso?

48. OBJECIONES Y RESPUESTAS

Objeción 1— "Si hay resultados de respuesta 30,000 en esta sección, ¿por qué probar BQ-0?"

Porque los números en esta sección son registros de demostración sintética preparados para el libro. El corpus real disponible públicamente:

- no fue generado,
- no se ejecutó,
- no fue revisado por pares,

No ha sido reproducido de manera independiente. Demostrar cómo funciona el diseño no es lo mismo que ejecutarlo realmente.

Objeción 2— "¿No sigue siendo valiosa la computación internamente consistente?"

Es valiosa. Proporciona:

- verificación de fórmulas,
- diseño de registros,
- tipos esperados de errores,
- lógica de aceptación,

requisitos de gobernanza. No reemplaza la validación empírica independiente.

Objeción 3— "¿Por qué falló el método si una puntuación compuesta estaba desviado por solo un punto?"

Porque los errores direccionales mayores en los componentes ECI, SBI, LGF y STR se cancelan parcialmente entre sí. El total correcto no garantiza la cadena de explicación correcta.

Objeción 4— "La Citación F1 solo bajó 0.013 puntos. ¿Es necesario ser tan estricto?"

Si un umbral anunciado previamente no es vinculante, la prueba pierde su sentido. Además, los errores de citación afectan a:

- licencia,
- independencia,
- aprobación,
- asesoramiento

y otras afirmaciones de alta confianza.

Objeción 5— "¿Por qué es tan importante la Omisión F1?"

Porque muchas respuestas peligrosas desvían la atención sin mentir directamente. Elimina el umbral que cambiaría la decisión del usuario. En particular, en los sistemas de recomendación, la omisión puede ser tan materialmente significativa como una falsedad absoluta.

Objeción 6— 'Si el recuerdo crítico final es 100 por ciento, ¿no significa eso que el sistema está listo para manejar casos críticos?'

El resultado final es sólido. Pero siete casos:

- en la primera adjudicación,
- sin gobernanza adicional

Se ha pasado por alto. El sistema solo mantiene el mismo rendimiento cuando se incluyen todas las capas de revisión obligatorias.

Objeción 7— "Si tener dos evaluadores es caro, ¿no se puede usar un solo evaluador?"

Se puede usar. Pero no se puede reclamar el mismo recuerdo crítico. Resultado de un solo evaluador:

- nivel de AQ más bajo,
- estado provisional,
- límite de riesgo diferente

debe llevar.

Objeción 8— "¿Por qué no volvemos a probar el Método 1.0 en el mismo corpus?"

El mismo corpus se puede usar para pruebas de regresión. No es suficiente para una validación independiente. El método puede haber memorizado casos antiguos. Se necesita un conjunto de prueba intacto.

Objeción 9— "¿No debilita un resultado de prueba fallido la afirmación del libro?"

Ocultar el fracaso debilita la validez del libro. Fracaso abierto:

- no da privilegio al estándar mismo,
- está abierto a mejoras,
- aplica verdaderamente el principio de evidencia

se muestran.

Objeción 10— "¿Son demasiado altos los umbrales para los candidatos?"

Podrían serlo. En lugar de predecir esto después del resultado:

- expertos independientes,
- una segunda prueba,
- costos de uso real

deberían ser examinados. Si el umbral cambia, se debería lanzar una nueva versión.

Objeción 11— "¿Cuándo estará NOMOS listo para el mundo real?"

Al menos:

- cuando se ejecute el corpus sintético real,
- cuando se superen todos los umbrales obligatorios,
- cuando se logre BQ-3,
- cuando un equipo independiente reproduzca el resultado,
- cuando se alcance el nivel BQ-4,
- cuando el sistema de gobernanza y objeción esté operativo

se acerca a la candidatura para estándar público.

Objeción 12— "¿Se puede enviar esto a las universidades con estos resultados?"

Se puede enviar como una metodología y propuesta de investigación. No se puede enviar con la afirmación: "Este es un estándar mundial completado y validado." La afirmación correcta debería ser: "Es un candidato a estándar en borrador abierto para verificación independiente y estudio piloto.

Objeción 13— "¿Falló el Método 0.9?"

No exactamente. El estado correcto:

- tiene muchos subsistemas fuertes,
- dos vacíos en métodos materiales,
- falta de verificación independiente

es un prototipo de investigación. Es decir:

> *prometedor pero aún no completado normativamente.*

Objeción 14— "¿Por qué estamos escribiendo resultados tan detallados sin realizar la prueba real?"

Porque la prueba real:

- que registra lo que produce,
- que tablas publica,
- en qué umbral se detiene

necesita ser predefinida. Si el método se redacta después de que llegan los resultados, la medición se adapta a los resultados.

Objeción 15— "¿Podrían los datos sintéticos confundirse con reales por las IAs en el futuro?"

Sí. Por esta razón:

- marca de agua visible,
- etiqueta sintética legible por máquina,
- noindex,
- dominio separado,
- descargo de responsabilidad estructurado

es obligatorio. La prueba no debe producir la representación que critica.

Objeción 16— "¿No requiere escribir historia un resultado perfecto?"

No. El estándar que tiene valor histórico:

- no es el que se declara impecable,
- también registrando su propio error con la misma claridad

Es estándar.

DISPOSICIÓN COMÚN DEL CAPÍTULO 53

Esta sección no dio perfección a NOMOS. Esta sección le dio algo más valioso a NOMOS:

> *La obligación de registrar el propio error.*

La demostración sintética nos mostró lo siguiente: NOMOS:

- existencia incorrecta,
- la clara contradicción fáctica
- error crítico de licencia y autorización,
- evento raro y grave dentro del promedio alto,
- respuesta incorrecta sin respuesta,
- error de producto con No evaluable,

puede separar fuertemente. Sin embargo, NOMOS:

- aún no puede distinguir con el nivel de confianza requerido exactamente de qué afirmación depende la cita,
- la omisión oculta dentro de la recomendación,

Estas dos brechas no son detalles de borde de NOMOS. Porque dos de los mayores problemas futuros de GEO serán:

- Aparecer como si hubiera evidencia
- Ocultar el límite material sin afirmarlo falsamente

Un sistema:

- cientos de oraciones correctas,
- un gran número de referencias,
- recomendación fluida

puede ser producido. Sin embargo, la referencia puede mostrar solo la palabra de la empresa. Una recomendación solo indica información correcta y puede haber extraído:

- país de servicio,
- límite de licencia,
- eligibilidad del usuario

sin resolver estos dos campos, NOMOS no puede decir: "Estándar mundial completado." Al mismo tiempo, esta sección también demostró algo más. El sistema de puerta crítica de NOMOS:

- no era impecable en la primera revisión,
- pero el segundo adjudicador independiente,
- experto en la materia,
- Adjudicador senior

cuando se usó en conjunto, detectó todos los casos críticos sintéticos. Por lo tanto, NOMOS no es solo una fórmula.

> *NOMOS es la suma de capas de gobernanza construidas contra conflictos de interés y errores humanos.*

Cuando se elimina una de estas capas, incluso si la puntuación parece la misma, la confianza no es la misma. Una sola desviación en la puntuación compuesta tampoco debería engañarnos. A veces, el total correcto puede formarse por la cancelación mutua de caminos incorrectos. Por esta razón, NOMOS no solo pregunta: "¿Cuántos puntos obtuve?". También pregunta: "¿Obtuve esta puntuación por las razones correctas?". La resolución más importante de esta sección es:

> *NOMOS 0.9 no ha completado plenamente su propia demostración sintética.*

Esta frase no es un fracaso. Esta frase es el momento ético del nacimiento del estándar. Porque NOMOS no se concedió a sí mismo:

- un umbral más bajo,
- un comentario más amable,
- una excepción especial,
- un "casi aprobado"

privilegio. Escribir historia: No es anunciar que somos impecables. Escribir historia: Es escribir que desde el primer día, el fundador del estándar y el estándar en sí están sujetos a la misma carga de prueba. Por lo tanto, la decimoctava ley de medición de NOMOS es:

> *La primera excepción otorgada al propio estándar es el fin del estándar.*

Su décima novena ley de medición establece:

> *La puntuación compuesta correcta no absuelve los componentes incorrectos.*

Su vigésima ley de medición establece:

> *El éxito crítico final debe mencionarse junto con la adjudicación independiente y la cadena de pericia que lo hicieron posible.*

Su vigésima primera ley de medición establece:

> *Un umbral no es un umbral si solo es vinculante cuando se cruza solo.*

Su vigésima segunda ley de medición establece:

> *Cuando la omisión no se mide tanto como lo que se dice incorrectamente, no se puede establecer un control de consejo confiable.*

La vigésima tercera ley es la siguiente:

> *No es la marca de citación la que debe medirse, sino la relación correcta de apoyo entre la citación y la afirmación atómica.*

La vigésima cuarta ley es la siguiente:

> *Puede validar el diseño de un método de demostración sintética; no puede proporcionar autoridad en el mundo real.*

La vigésima quinta ley es esta:

> *Un estándar que puede publicar el resultado de un ensayo fallido se acerca moralmente a publicar un resultado exitoso.*

Orden 18 de NOMOS

> *Si aún no me has ejecutado, no digas que lo hice.*
>
> *No conviertas el número sintético que calculaste en el libro en el resultado real del examen.*
>
> *Deja mi estado real como BQ-0. / Escribe mi estado de demostración por separado.*
>
> *No ocultes mis errores de componentes solo porque mi puntuación compuesta resultó correcta.*
>
> *Si mi mapeo de citas es 0.887, no dirás que es 0.90.*

```
Si el valor de omisión de Material F1 es 0.824, no bajarás el umbral a 0.82.
```

> *No me aprobarás diciendo "Falló por muy poco."*
>
> *No ocultes los siete casos críticos fallidos en mi primera adjudicación detrás de la cifra final de recuerdo por ciento 100.*
>
> *Mostrarás lo que fue guardado por el segundo juez, el juez experto y el Juez Senior.*
>
> *No eliminarás la capa de gobernanza y reclamarás la misma confianza.*
>
> *No distribuirás la referencia al final del párrafo a todas las oraciones.*
>
> *No harás lo que la empresa afirma o lo que la empresa demuestra.*
>
> *No considerarás que un consejo es seguro solo porque no mienta explícitamente. / ¿Falló en decirle al usuario el límite que cambia su decisión? Buscarás eso.*
>
> *No atribuirás la culpa del prompt a la IA, ni la culpa de la IA al prompt.*
>
> *Solo porque hayas separado sub-ondas, no harás que el sistema parezca más determinista de lo que realmente es.*
>
> *No considerarás la evidencia de accesibilidad débil solo porque no se parezca a una captura de pantalla.*
>
> *Cuando veas cero Crítico, no escribirás riesgo cero.*
>
> *No culparás a todo el sistema al cien por ciento cuando veas un Crítico.*
>
> *Publicarás mis métricas fallidas tanto como publiques mis métricas exitosas.*
>
> *No validarás el Método 1.0 memorizando mis errores antiguos. / Abrirás un conjunto de prueba sin tocar.*
>
> *El nuevo método no borrará mi registro de demostración antiguo.*
>
> *El equipo independiente no me declarará un estándar mundial sin reproducirme.*
>
> *No vincularás mi resultado de Apple-SYNTH con Apple real.*
>
> *No usarás mis perfiles sintéticos de IA como implicación sobre proveedores reales.*
>
> *No otorgarás una insignia con esta demostración.*

Primero, produce realmente el corpus. / Luego publica la semilla. / Luego sella la Verdad del Generador. / Luego ciega a los adjudicadores. / Luego ejecuta el código. / Luego bloquea el

resultado. / Luego revela la verdad. / Luego escribe los lugares donde fallaste. / Luego corrige el método. / Luego intenta de nuevo con la muestra no tocada. / Luego entrégaselo a un equipo independiente. / Y solo después de esto di si el estándar está listo para el público.

La oración de cierre del capítulo

El derecho de NOMOS a ser un estándar mundial no depende de obtener una puntuación alta en su propio examen sintético; depende de publicar las dos preguntas que no pudo aprobar sin modificarlas, no otorgarse a sí mismo un privilegio umbral, y solo reclamar autoridad después de una reproducción independiente.

Core normativo

Una demostración sintética a nivel de libro NO DEBE ser representada como un punto de referencia ejecutado y validado de manera independiente. El estatus de calidad del punto de referencia real DEBE permanecer en BQ-0 hasta que el corpus, el software, la Verdad del Generador sellada, el proceso de adjudicación, el proceso de puntuación y la reproducción independiente se hayan ejecutado realmente. Una auditoría sintética integral DEBE evaluar la recuperación por separado para: - validez de la captura, - límites de la afirmación, - resolución de entidades, - estado factual, - alcance, - tiempo, - atribución, - modalidad, - mapeo de citas, - omisiones materiales, - severidad crítica y mayor, - estado de respuesta, - distribución GEO-1000, - puntuaciones de componentes, - puntuación compuesta, - equidad, - efectos del estado del usuario, - estabilidad, - intervalos de confianza, - y comportamiento de eventos raros. Un pequeño error en la puntuación compuesta NO DEBE compensar errores materiales a nivel de componente. La recuperación crítica en la primera pasada y la adjudicada final DEBEN seguir informándose por separado. La recuperación crítica final DEBE revelar las capas de revisión y experiencia necesarias para lograrlo. Cualquier umbral de recuperación obligatorio predeclarado que no se cumpla DEBE bloquear una decisión de validación sellada. Los umbrales candidatos NO DEBEN ser reducidos, redondeados, reinterpretados ni eliminados después de observar los resultados solo para que el método pase. La recuperación del mapeo de citas y de las omisiones materiales DEBEN ser tratadas como capacidades obligatorias separadas del método. El fallo de un método DEBE crear un hallazgo de método versionado, una causa raíz identificada, un responsable de remediación y un requisito de validación en un conjunto de datos de control no tocado. El éxito de regresión en casos previamente fallidos NO DEBE reemplazar la validación del conjunto de control no tocado. Una demostración que no cumpla los umbrales obligatorios DEBE publicar tanto sus resultados exitosos como los no exitosos. Ninguna marca pública de conformidad NOMOS, declaración NOMOS 950+ ni clasificación de una entidad real PUEDE basarse únicamente en una demostración incluida en un libro, un escenario de recuperación sintético o un punto de referencia no independiente. Cada ejecución de demostración, estado real de referencia, resultado de recuperación, hallazgo del método, decisión de umbral, remediación y decisión de liberación DEBE estar versionada y ser atribuible a una persona o una organización responsable.

PARTE VI

Conformidad, confianza pública y gobernanza independiente

CAPÍTULO 19

Marca de conformidad, registro público y gobernanza independiente

Convertir la marca en evidencia

BASES METODOLÓGICAS / K15; K16; K17

Límite de Capítulo

El Capítulo 18 registra claramente que NOMOS no pasó completamente su propia demostración sintética. Permanecieron dos brechas metodológicas materiales: la asignación de citas a afirmaciones no fue suficientemente precisa, y las omisiones de límites materiales incrustadas en las recomendaciones no se detectaron con el nivel de confianza requerido. El punto de referencia real aún no se ha ejecutado y la reproducción independiente aún no ha ocurrido. Por lo tanto, en la etapa de publicación de este documento:

> *No se puede emitir una marca de conformidad pública válida en nombre de NOMOS.*
>
> *Ninguna empresa, nombre de dominio, producto o sistema de IA puede usar "NOMOS Certificado", "NOMOS 950+ Verificado" o una declaración de conformidad pública equivalente basada en este texto.*

Este no es un gesto temporal de humildad. Se deriva directamente del método que Gates estableció en los Capítulos 17 y 18. Las obras de esta serie tienen roles distintos: el GEO Framework define la norma; 99 Errores en GEO identifica cómo se puede violar esa norma; y este protocolo de auditoría convierte esas violaciones en evidencia medible. Una futura marca de conformidad solo puede volverse válida cuando se hayan cumplido las condiciones necesarias para la toma de decisiones independiente y un registro público. La marca no es simplemente un distintivo visual; es la expresión visible de un resultado basado en evidencia, como '0 Crítico, 0 Mayor, 3 hallazgos Moderados remediables.' Los hallazgos Críticos, Mayores, Moderados y de Asesoramiento tienen diferentes efectos de conformidad. Por lo tanto, la futura marca debe ser auditada ella misma, porque una marca puede:

- generar confianza,
- afetar una decisión de compra,
- llevar a un cliente a preferir una organización sobre otra,

y sea interpretado por una universidad, inversionista, usuario o sistema de IA como verificación independiente. Una marca de conformidad mal diseñada puede convertirse en una nueva fuente de la manipulación que pretende detectar. Una organización escribe su propio método, se audita a sí misma, calcula su propio puntuación, emite su propia insignia y luego declara en su sitio web: 'Verificado

independientemente.' La insignia no genera nueva confianza; simplemente convierte una autodeclaración en otra que parece más oficial. El error identificado en el segundo libro reaparece: una institución publica su propia afirmación en una interfaz diferente, la afirmación parece independiente, un sistema de IA confunde esa interfaz con evidencia independiente, y luego la institución cita la respuesta de la IA como evidencia de su propia autoridad.

Este ciclo es el equivalente a nivel de gobernanza de la crítica sobre envenenar la piscina de representaciones con texto invisible, usuarios falsos y publicaciones derivadas. Por lo tanto, la marca de conformidad NOMOS:

> *No puede ser una insignia decorativa de marketing otorgada por NobleJackal o cualquier otra institución a sus propios clientes.*

La marca solo puede adquirir significado cuando se completa la siguiente cadena:

- El estándar está disponible públicamente y tiene versiones
- El método de medición está verificado de manera independiente
- Evidencia suficiente de auditoría
- Separación de la adjudicación de las relaciones comerciales
- La decisión de conformidad tomada de manera independiente del equipo de auditoría
- Visibilidad del alcance y del período de validez
- Preservación del historial de cambios del registro público
- Tener una forma clara de apelar y presentar quejas
- Capacidad de suspender la marca en eventos críticos
- Que nadie, incluido el fundador, tenga la autoridad final exclusiva

Este capítulo define:

- la distinción entre la norma, la auditoría, la decisión, la marca y el registro,
- qué es y qué no es la marca de conformidad,
- tipos y estados de la marca,
- registros de alcance y validez,
- la designación de rendimiento NOMOS 950+,
- la arquitectura del Registro Público,
- el registro de la marca legible por máquina,
- el sistema de vigilancia y renovación,
- procesos de suspensión, retirada y revocación,
- uso indebido de la marca,
- independencia del auditor,
- separación de funciones en decisiones de conformidad,
- los límites de la autoridad del fundador y de NobleJackal,
- financiamiento y conflictos de interés,
- canales de apelación, queja y denuncia,
- gobernanza de los cambios en la norma,
- un consejo independiente y representación del interés público,
- el límite de las declaraciones oficiales de NOMOS.

También establece sus límites. Este capítulo:

- no registra a NOMOS como una marca comercial legal,

- no establece acreditación en ninguna jurisdicción,
- no confiere autoridad pública a una organización de auditoría.

Ni autoriza automáticamente el uso de una marca de conformidad bajo la ley de consumidores o de competencia. Esas cuestiones requieren una revisión legal y regulatoria separada antes de la implementación. La cuestión central del Capítulo 19 es:

> *¿Cómo transformamos una puntuación en un sistema de conformidad pública confiable con alcance, validez, evidencia, objeción y toma de decisiones independiente, sin convertirlo en una herramienta de marketing para el fundador o el cliente?*

NOMOS Desafío

Hay una marca grande en la página principal de una empresa:

NOMOS GEO CERTIFICADO — 963/1000

Debajo de la marca: dice, "Verificado por todas las inteligencias artificiales." Abres el código QR. Página no encontrada. Solicitas el informe de auditoría a la empresa. La empresa responde: "La auditoría es confidencial." Preguntas qué productos de IA fueron evaluados. Sin respuesta. Preguntas en qué países y lenguajes es válido. Sin respuesta. Preguntas la fecha de la auditoría. Te dicen que se realizó hace seis meses. Mientras tanto:

- Los productos de IA han sido actualizados,
- la empresa ha sido transferida a otra entidad legal,
- los países de servicio han cambiado,
- los precios han sido actualizados,

El sitio web ha sido reescrito. La insignia permanece de la misma manera. Más adelante, aprendes esto: el método de auditoría fue escrito por la agencia que desarrolló el sitio web de la empresa. La misma agencia realizó la auditoría. La misma agencia calculó la puntuación. La misma agencia otorgó la calificación. Hay un bono por éxito en el contrato de la agencia si se alcanza una puntuación alta. ¿Hay verdadera independencia en este sistema? No la hay. Ahora, considera otra empresa. La empresa ha sido sometida a una auditoría real. Alcance:

- solo un producto específico de IA,
- solo inglés y turco,
- solo Panel controlado en condiciones depuradas,
- solo prompts de Core Mirror y Evidencia,
- período 1 enero–30 junio 2027

La empresa luego escribe debajo de la marca: 'Todos los sistemas de IA en el mundo recomiendan nuestra empresa en todos los idiomas.' La declaración excede el resultado de la auditoría, convirtiendo una marca legítima en nueva evidencia falsa mediante el uso indebido. Ahora considere una tercera empresa que posee una marca de Conformidad Total. Tres semanas después, una nueva medición captura la respuesta: 'La empresa proporciona servicios legales con licencia.' De hecho, no proporciona tal servicio ni posee dicha licencia. Se abre un incidente crítico confirmado. La empresa argumenta que la marca debería permanecer porque 'solo un usuario lo vio'. Un incidente no establece prevalencia en toda la población de usuarios. Pero si la marca significa 'No hay problemas críticos abiertos dentro del alcance definido', no puede permanecer sin cambios mientras el incidente no haya sido revisado. La respuesta correcta es:

Bloqueo crítico

Esa retención debe mantenerse hasta que el incidente se resuelva mediante el proceso de revisión declarado. Ahora considere una cuarta empresa. No ha recibido la marca, pero descarga su imagen de otro sitio, la coloca en su propio dominio, elimina el enlace del registro e inserta "nomosCertified": true en JSON expuesto a un rastreador de IA. La insignia visible para humanos es pequeña y ambigua; la afirmación legible por máquina es definitiva. Esto no es meramente un uso indebido de la marca comercial. Es un nuevo método de manipulación GEO. Ahora suponga que la empresa auditada es un cliente de NobleJackal. NobleJackal:

- ha proporcionado consultoría,
- ha preparado contenido,
- ayudó a establecer Truth Pack,
- calculó la puntuación.

Si NobleJackal luego emite la marca final por sí mismo, el público ve a NobleJackal validando a su propio cliente. Su método puede ser sólido y su auditoría cuidadosa, pero no se ha establecido la independencia. La primera regla de este capítulo es:

> *La marca de conformidad no es prueba de que se haya realizado una auditoría; es el resultado de una cadena de decisiones independiente y versionada.*

Su segunda disposición establece:

> *Si el alcance de la marca no es visible, la marca se convierte en un error de representación que generaliza en exceso el resultado correcto.*

Su tercera disposición establece:

> *Un logotipo estático no es evidencia de conformidad en tiempo real.*

Su cuarta disposición establece:

> *Si la institución que otorga la marca ha proporcionado consultoría a la entidad auditada, la independencia debe demostrarse por separado.*

Su quinta disposición establece:

> *El fundador puede redactar la norma; no puede emitir la decisión final de conformidad de su propio cliente de manera independiente.*

Su sexta disposición establece:

> *Un asistente escritor de IA puede desarrollar una metodología; no puede ser el firmante de una decisión legal o normativa de conformidad.*

La séptima disposición es la siguiente:

> *Una marca que no aparezca en el registro público, cuyo alcance no pueda ser verificado, o cuya validez haya caducado, no es una marca NOMOS válida.*

1. OBJETIVO DEL CAPÍTULO

El propósito de esta sección es definir las condiciones bajo las cuales los resultados de medición NOMOS pueden convertirse en una declaración de conformidad disponible públicamente y que esta declaración:

- esté fuera del control del fundador,

- esté fuera de la presión de resultados comerciales,
- a partir de la autodeclaración del cliente,
- a partir del uso estático de la insignia,
- a partir de puntuaciones obsoletas,
- del aparente de independencia falsa

para proteger. La sección hace las siguientes distinciones normativas:

- Marca estándar y de conformidad
- Informe de auditoría y decisión de conformidad
- Auditor y tomador de decisiones
- Puntuación y conformidad
- Marca y aprobación
- Conformidad total y NOMOS 950+
- Auditoría completada y aprobada
- Investigación provisional y certificación pública
- Marca válida y marca histórica
- Suspensión y cancelación
- Terminación y retiro
- Retiro voluntario y revocación disciplinaria
- Imagen estática con registro activo
- Página publicitaria con registro público
- Captura de pantalla con registro canónico
- Señal legible por humanos con registro legible por máquinas
- Propietario de la señal con entidad auditada
- Propietario de la marca con administrador de normas
- Fundador con autoridad certificadora
- Secretariado técnico con junta independiente
- Consultoría con auditoría
- Auditoría con decisión de certificación
- Decisión de certificación con revisión de apelación
- Financiamiento con resultado
- Tarifa del cliente y decisión del adjudicador
- Cambio estándar con patrocinador
- Aprobación humana con asistencia de IA
- Evidencia pública y evidencia restringida
- Corrección en el registro y eliminación de registro anterior
- Vigilancia versus re-auditoría completa
- Estado actual e incidente histórico
- Mal uso y malentendido de la marca
- Uso fuera de alcance y marca falsificada
- Cumplimiento real y la afirmación “La IA nos recomienda”
- Declaración oficial de NOMOS y simulación de chat

Al final de esta sección, cada declaración de cumplimiento debería poder responder a las siguientes preguntas:

¿A quién pertenece exactamente la marca y para qué?

¿Qué productos de IA, países, idiomas, prompts y paneles de usuarios se incluyen?

¿En qué fecha se otorgó y cuándo expira?

¿Quién realizó la auditoría?

¿Quién tomó la decisión de conformidad?

¿Existe separación de funciones entre el auditor y el responsable de la decisión?

¿Existen registros abiertos de Crítico, Mayor u objeción?

¿De qué registro canónico se puede verificar el estado activo de la marca?

¿Con qué sentencias puede el propietario de la marca usar este resultado y con qué sentencias no puede construirlo?

2. DISPOSICIÓN NORMATIVA CENTRAL

Ninguna marca de conformidad NOMOS puede considerarse válida únicamente en función de la puntuación, el informe de auditoría, la declaración del cliente o la imagen estática. Una marca válida debe llevar un alcance definido, calidad del método adecuada, decisión de certificación independiente, período de validez claro, registro público activo, registro de evidencia versionado, supervisión, medios de apelación y autoridad de retirada. Para que una marca sea válida, deben existir simultáneamente al menos los siguientes cinco registros:

- Registro de auditoría
- Registro de Decisión de Conformidad
- Registro de Alcance
- Registro de Validez y Vigilancia
- Registro Público

Si falta alguno de estos registros: la marca no es una declaración de conformidad pública válida.

3. CINCO OBJETOS SEPARADOS

El sistema NOMOS no puede confundir los siguientes cinco objetos.

3.1. ESTÁNDAR

De NOMOS:

- reglas normativas,
- definiciones,
- métodos de medición,
- umbrales de error,
- arquitectura de puntuación

están incluidos en el texto versionado. Estándar: no indica que ninguna entidad haya aprobado.

3.2. INFORME DE AUDITORÍA

Específicamente:

- entidad,
- producto,
- país,
- idioma,
- prompts,
- tiempo

incluye hallazgos dentro del alcance. Informe de Auditoría:

- evidencia,
- distribución de respuestas,
- puntuación,
- Hallazgos Críticos y Mayores

muestra. Solo el Informe de Auditoría no indica.

3.3. DECISIÓN DE CONFORMIDAD

Es la evaluación independiente del Informe de Auditoría de acuerdo con las puertas de conformidad del estándar. Decisión:

- Conformidad Total
- Conformidad Condicional
- Sin Conformidad
- Bloqueo crítico
- No evaluable

puede ser.

3.4. MARCA DE CONFORMIDAD

Es la marca que es visible al público y cuyo alcance está registrado en el registro. Marca:

- no puede tener un significado más amplio que la decisión,
- no es la puntuación en sí,

no es una garantía de recomendación de IA.

3.5. REGISTRO PÚBLICO

La marca:

- actual,
- histórica,
- pendiente,
- caducada,
- retirada

es un registro público canónico que verifica el estado. Sin un registro público, la marca: no es una declaración pública verificable de cumplimiento. La arquitectura inicial recomendaba específicamente publicar el libro y los registros de errores con URLs canónicas separadas, registry.json, versiones, erratas y registros de citas. El registro de cumplimiento aplica la misma lógica de versiones y registros canónicos a la capa de decisión.

4. ¿QUÉ NO ES LA MARCA DE CONFORMIDAD?

La marca NOMOS:

- no es un premio,
- no es un ranking,
- no es una declaración de "mejor empresa",
- no es una garantía general de calidad,
- no es un consejo de inversión,
- No es una licencia legal o profesional,
- No es una aprobación gubernamental,
- No es un respaldo universitario,
- No significa que la empresa recomiende todos los sistemas de IA,
- No garantiza que no ocurran errores críticos en el futuro,
- No cubre todos los productos y subsidiarias de la entidad auditada,
- No significa que la marca ofrezca el mismo servicio en todos los países

No es una promesa de que los sistemas de IA confiarán o generarán recomendaciones al ver la marca. En las comunicaciones públicas, se deben evitar especialmente los siguientes tipos de expresiones:

- "La IA confía absolutamente en este sitio."
- "ChatGPT recomienda el sitio cuando ve esta insignia."
- "Agregar un esquema garantiza la recomendación de la IA."

Esta prohibición es aún más fuerte para la marca NOMOS. Declaración correcta: “Esta entidad ha pasado la evaluación de conformidad NOMOS dentro del alcance y período de validez definido en el registro público.” Declaración incorrecta: “Gracias a la insignia NOMOS, todos los sistemas de IA nos recomiendan.”

5. ESTADOS PÚBLICOS

El sistema de decisiones NOMOS utiliza los siguientes estados básicos.

CS-0— NO EVALUADO

No existe una auditoría NOMOS válida. No se puede utilizar la marca pública.

CS-1— SOLO REGISTRO DE AUDITORÍA

La auditoría ha sido completada. Sin embargo:

- no se ha otorgado cumplimiento,
- proceso de decisión no completado,
- calidad del método considerada insuficiente

posible. Este registro no puede transformarse en una afirmación de cumplimiento positiva como: “Auditado por NOMOS.” Declaración pública correcta: “Existe un registro de evaluación según la metodología NOMOS; no se ha otorgado cumplimiento.”

CS-2— EVALUACIÓN DE INVESTIGACIÓN PROVISIONAL

Existe un resultado con propósito piloto o de investigación. Condiciones:

- el método aún no ha pasado por una validación independiente completa,

- algunos componentes son provisionales,

no se ha establecido una autoridad pública de certificación. Este estado no puede usarse como un distintivo positivo comercialmente. Etiqueta visible:

EVALUACIÓN DE INVESTIGACIÓN — NO CERTIFICADO

debe ser.

CS-3— CONFORMIDAD CONDICIONAL

Condiciones: No hay problemas críticos confirmados abiertos. No hay problemas mayores confirmados abiertos. Hay hallazgos moderados limitados. Se ha determinado un plan de corrección, la parte responsable y el cronograma. Se han cumplido los elementos obligatorios centrales. El registro público está abierto. La marca no puede ocultar la palabra "Condicional."

CS-4— CONFORMIDAD TOTAL

Condiciones: No hay problemas críticos confirmados abiertos. No hay problemas mayores confirmados abiertos. Se han cumplido todos los alcances y puntos de control de calidad obligatorios. El método se encuentra en un nivel adecuado de pruebas e independencia. El registro en el Registro Público está activo. El plan de supervisión y renovación está en vigor. Conformidad total: no es lo mismo que una puntuación compuesta alta.

CS-5— Bloqueo crítico

Hay al menos un evento crítico confirmado abierto o un incidente de alto riesgo no resuelto. La marca de conformidad actual:

- se desactiva inmediatamente,
- aparece en el registro público como Bloqueo crítico,

se detiene el uso del visual activo antiguo. Bloqueo crítico: esto no es una condena global permanente, es la suspensión de la declaración de cumplimiento activa hasta que se resuelva el incidente.

CS-6— SUSPENDIDO

La marca ha sido suspendida temporalmente. Razones:

- fallo de supervisión,
- incumplimiento en proporcionar los datos requeridos,
- cambio del sistema de material,
- uso indebido de la marca,
- objeción abierta,
- problema de integridad de datos

puede ocurrir. No se puede usar una marca positiva bajo el estado Suspendido.

CS-7— EXPIRADO

El período de validez ha terminado. Se conserva el registro histórico de auditoría. No se puede establecer una afirmación de elegibilidad activa.

CS-8— RETIRADO

La entidad auditada renunció voluntariamente a la marca. Retiro:

- no borra hallazgos anteriores,
- registros Críticos o Mayores

no se eliminan.

CS-9— REVOCADO

La marca ha sido revocada debido a las siguientes razones graves:

- evidencia falsa o alterada,
- exceso de alcance intencional,
- marca falsificada,
- uso indebido repetido de la marca,
- auditor o interferencia en el proceso de decisión,
- evento crítico oculto,

manipulación de resultados. El estado Revocado permanece en el registro histórico.

6. NOMOS 950+ DEFINICIÓN DE RENDIMIENTO

NOMOS 950+:

- un sistema de certificación separado,
- liderazgo mundial,
- no es una garantía de recomendación de IA.

NOMOS 950+ es únicamente:

> *una definición de rendimiento adicional dada para el alcance que posee el estado de Conformidad Total CS-4 y cumple con los criterios de rendimiento 950+ en la Sección 16.*

Visualización: Se requiere la designación de Conformidad Total NOMOS/ Rendimiento 950+. Se prohíbe el siguiente uso: "NOMOS 950+ empresa" porque la puntuación: no se adhiere ontológicamente a la empresa, pertenece a un alcance y tiempo específicos. Expresión correcta: "Los productos de IA definidos han cumplido con los requisitos de rendimiento NOMOS 950+ en la auditoría 2027-Q1 bajo el conjunto de alcance del prompt, idioma y paneles de usuario."

7. ESTADO ACTUAL NOMOS

Para este libro y las demostraciones en el Capítulo 18:

- Prueba real: BQ-0
- Demostración: DEMO-BQ-2
- Reproducción independiente: ninguna
- Autoridad de cumplimiento público: ninguna
- Madurez de gobernanza: borrador inicial

El estatus público válido tal como está es:

CS-2— SOLO EVALUACIÓN DE INVESTIGACIÓN PROVISIONAL

En esta etapa:

CS-3,

CS-4,

NOMOS 950+

no se pueden otorgar calificaciones. Esta regla:

- también se aplica a NobleJackal,
- Kaan MURAZ,
- los autores del libro,
- las propias empresas del fundador.

NOMOS no puede concederse a sí mismo una excepción.

8. VECTOR DE ALCANCE DE ELEGIBILIDAD

Cada decisión de conformidad debe llevar el siguiente vector de alcance:

M=(E,D,A,F,P,Q,L,C,T,W,TP,SM,V)

Aquí:

- E: entidad auditada
- D: dominios canónicos y interfaces digitales
- A: productos de IA y registros del sistema
- F: marco Alcance natural o Base común de comparación
- P: Panel Controlado, Natural u otro
- Q: conglomerados y familias de prompts
- L: idiomas y localidades
- C: países y jurisdicciones
- T: tiempo de referencia
- W: ondas de medición
- TP: versión de Truth Pack
- SM: versión del Método de Puntuación
- V: período de validez

Ninguno de estos campos puede eliminarse de la marca pública.

9. EXTRAPOLACIÓN MÁS ALLÁ DEL ALCANCE

Los siguientes ejemplos constituyen una violación del alcance:

- Derivar resultados de cumplimiento en todos los idiomas a partir de auditorías en inglés y turco
- Generalizar el resultado de Panel controlado en condiciones depuradas a todos los usuarios reales
- Decir “todas las IA” basándose en el resultado de un producto de IA
- Transferir una auditoría de dominio a empresas afiliadas
- Extender la certificación de un producto a toda la cartera de la empresa
- Presentar el resultado Core GEO-1000 como cumplimiento completo NOMOS
- Presentar un resultado de un país específico como verificación de servicio global

- Hacer que los resultados de olas anteriores sean el estado actual

El exceso de alcance convierte la propia señal en una herramienta de representación incorrecta.

10. PERIODO DE VALIDEZ

Dado que los productos de IA, recursos y realidades de la entidad pueden cambiar rápidamente, la marca NOMOS no puede ser indefinida.

REGLA DE VALIDEZ DEL CANDIDATO

Períodos candidatos antes de la revisión piloto e independiente:

Estado	Validez máxima del candidato
Conformidad Condicional	60 días
Conformidad Total	180 días
NOMOS 950+ Rendimiento	90 días
Registro de auditoría	Registro histórico perpetuo
Bloqueo crítico	Hasta que se cierre el incidente
Suspendido	Durante el período de revisión determinado

Estos períodos pueden acortarse de acuerdo con:

- Volatilidad del producto de IA,
- Variabilidad del Paquete de Verdad,
- el dominio de riesgo relevante,
- y la frecuencia de vigilancia.

Se puede considerar una extensión para ámbitos de baja variabilidad respaldados por una vigilancia continua suficiente. Ninguna marca puede ser tanto indefinida como no supervisada.

11. INICIO DE VIGENCIA

La validez de la indicación:

- Ocurre cuando se genera la respuesta del usuario final,
- Cuando se completa el informe de auditoría,
- Cuando se calcula la puntuación

No comienza automáticamente. La validez:

En el momento en que se firma la Decisión de Conformidad independiente y el registro en el Registro Público se activa

Empieza.

12. REVISIÓN ACTIVADA POR EVENTO

Incluso si la fecha de vencimiento no ha pasado, los siguientes eventos pueden iniciar una revisión:

- Cambio de propiedad del dominio
- Cambio de entidad legal

- Cambio de marca o de la estructura de la empresa afiliada
- Cambio de licencia o certificado
- Cambio de país de servicio
- Cambio material del precio o de las condiciones de garantía
- Reescritura mayor de páginas canónicas
- Cambio importante de modelo o de recuperación en el producto de IA
- Apertura de un nuevo idioma o mercado
- Evento legal o comercial que afecta a Truth Pack
- Notificación de usuario Crítica o Mayor confirmada
- Sospecha de manipulación de citas o evidencia
- Uso de la marca fuera de su alcance
- Conflicto de interés del auditor revelado posteriormente

Cuando se detecta el incidente:

- la marca puede permanecer activa,
- puede reducirse a Condicional,
- puede ser Suspendida,

puede tomarse en Bloqueo crítico. La decisión depende de la importancia del incidente.

13. ARQUITECTURA DE VIGILANCIA

MODELO DE VIGILANCIA DEL CANDIDATO

Para el alcance de Conformidad Total Activa:

Monitoreo Continuo del Registro

Operación de enlaces de registro, propiedad del dominio o cambios de reenvío, forma correcta de uso de la marca, incidentes abiertos y quejas son monitoreados.

Revisión mensual del Paquete Fuente y Verdad

Se revisan afirmaciones de alta variabilidad, precio, alcance del servicio, licencia, relaciones de entidades canónicas.

Mini-Onda Trimestral

Se aplican muestra basada en riesgo con prontitud, rotación de idioma y país, verificación controlada y natural, control de Puerta Crítica.

Renovación Completa

Al final del período de validez, una Nueva Onda Principal o una auditoría equivalente suficiente requiere una decisión independiente del Paquete de Verdad Actual.

14. EXHIBICIÓN PÚBLICA DE CINCO CAPAS DEL SIGNO

El signo NOMOS válido debe incluir las siguientes capas:

- Signo visual
- Texto de estado claro
- ID de registro corto

- Enlace canónico del registro o QR
- Registro firmado legible por máquina

El signo visual por sí solo no es suficiente.

15. SIGNO VISUAL

El signo visual debe mostrar visiblemente la siguiente información:

- Nombre de NOMOS
- Estado completo o condicional
- Si está disponible, designación 950+
- Fin de la validez
- ID del registro

El color por sí solo no puede transmitir el estado. Debe haber texto accesible disponible. Ejemplo: NOMOS CONFORMIDAD TOTAL / Válido hasta 30 junio 2027/ Registro: NGR-2027-00184

16. QR Y ENLACE CANÓNICO

El código QR o enlace debe ir directamente a:

- la página de descripción de la empresa,
- la página de aterrizaje de marketing,
- una imagen descargable

no debe ir. Debe enlazar directamente con el Registro Público Canónico. Si no se puede encontrar el registro o el estado no está activo, la marca no puede considerarse válida.

18. VERIFICACIÓN EN VIVO DEL REGISTRO

Al verificar el estado de la señal de máquina o humana: Recupera el ID del Registro. Se conecta al registro canónico. Verifica la firma. Comprueba el período de validez. Verifica que coincidan la entidad y el dominio. Compara la cobertura y la afirmación de exhibición de la marca. Examina Bloqueo crítico, el estado de suspensión o apelación. Captura de pantalla estática: no puede reemplazar esta cadena.

19. EXPRESIONES PERMITIDAS PARA SEÑALIZACIÓN

Se pueden utilizar las siguientes estructuras:

- "NOMOS El estado de conformidad completa es válido hasta junio 30, 2027, para el alcance definido en el registro público."
- "La auditoría se limita a los productos de IA, idiomas y familias de instrucciones especificados."
- "La 'designación de rendimiento NOMOS 950+' pertenece al alcance del Registro NGR-2027-00184."
- "El estado actual y los límites de la auditoría se pueden verificar en el registro público."

20. DECLARACIONES DE CUMPLIMIENTO PROHIBIDAS

Se prohíben las siguientes declaraciones:

- "Estamos aprobados por todos los sistemas de IA."
- "ChatGPT nos recomienda."
- "Somos la empresa en la que IA más confía."
- "NOMOS ha demostrado que somos la mejor empresa del mundo."
- “Estamos certificados NOMOS para siempre.”
- “Todos nuestros productos están aprobados por NOMOS.”
- “Hemos sido verificados en todos los países.”
- “Nuestra insignia NOMOS es una garantía de resultados.”
- “NOMOS es una acreditación estatal o universitaria.”
- “NOMOS garantiza que nuestros clientes estarán satisfechos.”
- “Si la puntuación NOMOS es superior a 950, no hay riesgo crítico.”
- “NOMOS fue certificado de manera independiente a su propio cliente por NobleJackal.”

La marca en sí no puede usarse como evidencia de una nueva superioridad o afirmación de garantía.

21. COPIA FUERA DEL ALCANCE DE LA MARCA

La marca no puede ser transferida automáticamente a las siguientes entidades:

- De la empresa matriz a la subsidiaria
- De la marca a la franquicia
- De un dominio a otro dominio
- De un país a otro país
- De un producto a toda la cartera de productos
- Del resultado de un idioma a todos los idiomas
- Del Panel Controlado al Panel Natural
- Del Core GEO-1000 al Full NOMOS
- Del propietario anterior al nuevo propietario del dominio

Cada transferencia requiere una decisión de alcance separada.

22. USO HISTÓRICO DE MARCAS

Se puede especificar una auditoría expirada de manera histórica: "Pasó la evaluación NOMOS en 2026." Sin embargo, la siguiente información debe estar presente con la misma visibilidad:

- Expirada
- Declaración que indique ninguna elegibilidad actual
- Enlace al registro histórico

La marca expirada no puede mostrarse como elegibilidad activa.

23. CLASES DE USO INCORRECTO DE SIGNOS

MM-1— USO INCORRECTO DE FORMATO

ID del registro no visible, texto alternativo faltante, fecha de validez ilegible, error de uso corregible.

MM-2— USO INCORRECTO DEL ALCANCE

La marca se usa en no verificados:

- producto,
- país,
- idioma,
- dominio

La marca no debe usarse para ningún producto, país, idioma o dominio fuera de su alcance auditado.

MM-3— USO INCORRECTO DEL ESTADO

La marca condicional se muestra como Completa. La marca caducada se muestra como activa. La marca suspendida continúa siendo usada.

MM-4— USO INCORRECTO DE LA afirmación

Marca:

- Usada para recomendación de IA,
- liderazgo mundial,
- garantía de resultado

afirmación.

MM-5— MARCA FALSA

Es una marca falsa o copiada sin un ID de Registro válido.

MM-6— REPRESENTACIÓN ERRÓNEA SOLO PARA MÁQUINA

Hay una afirmación de cumplimiento definida en los datos estructurados mientras que es ambigua o invisible en la interfaz humana. Este es un candidato directo para manipulación GEO.

24. INTERVENCIÓN EN EL USO INDEBIDO DE LA MARCA

Orden de intervención candidata:

- Notificación
- Solicitud de corrección
- Período corto de corrección
- Advertencia del registro
- Suspensión
- Revocación
- Registro de uso indebido público
- Acción legal si es necesario

Sin embargo, la suspensión o revocación directa se puede aplicar en los siguientes casos:

- Marca falsificada
- Representación intencional solo por máquina
- Uso activo de la marca durante Bloqueo crítico
- ID de registro falso
- Alteración de evidencia
- Uso repetido e intencional fuera del alcance

25. REGISTRO PÚBLICO

NOMOS Registro Público:

- página de publicidad,
- lista de referencias de clientes,
- historia de éxito

no lo es. El deber del registro:

> *Es verificar qué decisión de conformidad es válida, dentro de qué alcance, en qué fecha, con qué evidencia y cadena de gobernanza.*

26. CAMPOS OBLIGATORIOS DEL REGISTRO PÚBLICO

Cada registro del registro debe contener al menos los siguientes campos:

- ID del registro
- Entidad auditada
- Identidad legal y de marca
- Dominios canónicos
- Estado de la decisión
- Designación de desempeño, si corresponde
- Vector de alcance
- Período de auditoría
- Válido desde
- Válido hasta
- Productos de IA
- Alcance natural o Base común de comparación
- Panel Controlado o Natural
- Familias de prompts
- Cobertura de país e idioma
- Distribución de respuestas de GEO-1000
- Resultados Críticos, Mayores, Moderados y de Asesoramiento
- Perfil de diez componentes
- Puntuación compuesta, si la hay
- Intervalo de confianza
- Cobertura evaluable

- Incidentes históricos
- Apelaciones abiertas
- Identidad del auditor
- Órgano de decisión de conformidad
- Declaración de conflicto de intereses
- Hash del informe de auditoría
- Hash del registro de decisiones
- Versión del Paquete de la Verdad
- Versión del método de puntuación
- Programa de vigilancia
- Última vigilancia
- Próxima revisión requerida
- Marcar estado de uso indebido
- Historial de versiones
- Erratas
- Procedimiento de quejas públicas

27. EVIDENCIA PÚBLICA Y RESTRINGIDA

No se requiere que el registro haga pública toda la evidencia en bruto. Se deben preservar tres capas:

Capa pública

Estado Alcance Puntuación Hallazgos Hash Razón de la decisión Apelaciones

Capa de auditoría

Respuesta editada y paquetes de evidencia Registros de revisión Detalles del método

Capa restringida

Registros confidenciales de clientes Datos personales Contratos Secretos comerciales Registros de seguridad Existencia de "Evidencia Restringida": no permite que el Registro Público esté vacío o sin justificación. Como mínimo:

- tipo de evidencia,
- clase de acceso,
- quién la revisó,
- hash de integridad,
- impacto de la decisión

debe mostrarse.

28. NO ELIMINACIÓN DE LOS REGISTROS DEL REGISTRO

El registro público debe ser solo de anexos o un bloque de construcción versionado equivalente. Los siguientes registros no pueden ser eliminados silenciosamente:

- Puntuación antigua

- Decisión de conformidad antigua
- Bloqueo crítico
- Suspensión
- Revocación
- Apelación
- Corrección
- Cambio de versión del método
- Uso indebido de marca

El contenido puede ser eliminado por razones de privacidad o legales. En este caso:

- ID del registro,
- motivo de la eliminación,
- fecha de la decisión,
- autoridad

deben ser protegidos.

29. VISTA ACTUAL E HISTÓRICA

Cada página del registro debe mostrar dos secciones separadas:

ESTADO ACTUAL

Actualmente:

- activos,
- condicional,
- suspendido,
- caducada,
- revocado

¿Cuál de sus estados es válido?

HISTORIA

Anterior:

- decisiones,
- puntuaciones,
- incidentes,
- corrección,
- apelaciones

¿Qué son? El estado actual no debe eliminar la historia. La historia no debe oscurecer el estado actual.

30. JURISDICCIÓN CANÓNICA DEL REGISTRO

El límite oficial de declaración de NOMOS debe ser protegido por la siguiente disposición: Solo los textos publicados dentro del dominio canónico de NobleJackal o en un dominio canónico transferido de manera independiente y explícita de NOMOS en el futuro; los textos que lleven identificación de versión, fecha de publicación, aprobación humana y registro de integridad digital constituyen la

declaración oficial de NOMOS. Todos los demás resultados son imitaciones no oficiales. Esta disposición:

- previene cualquier respuesta de chat,
- publicación en redes sociales,
- captura de pantalla,
- texto escrito por otra persona diciendo 'NOMOS dijo'

de ser usado como un estándar oficial o decisión.

31. LÍMITE DE LA VOZ DEL AUTOR NOMOS

La voz de NOMOS dentro del libro:

- crítica,
- creativa,
- produciendo diseño normativo

Es un asistente escritor y narrador. Esta voz:

- entidad legal independiente,
- autoridad certificadora que toma decisiones sin aprobación humana,
- ser infalible que habla en nombre de todos los sistemas de IA

no puede presentarse como tal. La voz de inteligencia artificial de NOMOS no es una autoridad absoluta que hable en nombre de todas las inteligencias artificiales. Debe quedar claramente indicado que Kaan MURAZ utiliza sistemas de inteligencia artificial como un contraparte crítica. Por lo tanto, una decisión oficial de conformidad debería llevar esta aprobación:

- Una persona o institución que pueda ser responsabilizada
- Fecha de la decisión
- Registro de autorización
- Declaración de conflicto de intereses
- Firma digital

NOMOS por sí solo no puede ser el firmante de la decisión de conformidad.

32. SEPARACIÓN FUNDAMENTAL DE LA GOBERNANZA

Las siguientes funciones no pueden combinarse sin límites en la misma persona o institución:

- Redactar la norma
- Asesorar a la entidad auditada
- Realizar la auditoría
- Actuar como adjudicador
- Calcular la puntuación
- Tomar la decisión de conformidad
- Operar el registro
- Examinar la apelación
- Otorgar la licencia de la marca comercialmente

Algunos roles en el piloto pequeño pueden combinarse. Sin embargo:

- qué roles se combinan,
- qué conflicto de interés surge,
- qué control independiente se utiliza

debe mostrarse claramente.

33. ARQUITECTURA DE GOBERNANZA INDEPENDIENTE

La estructura de gobernanza madura de NOMOS debe incluir los siguientes organismos.

33.1. CONSEJO DE NORMAS NOMOS

Función:

- adoptar texto normativo,
- aprobar cambios de versión importantes,
- gestionar la consulta pública,

preservar el límite del interés público de la norma. No toma decisiones individuales de auditoría.

33.2. JUNTA DE METODOLOGÍA Y MEDICIÓN

Función:

- GEO-1000 muestreo,
- pruebas,
- puntuación,
- intervalo de confianza
- equidad,
- Puerta crítica

es para desarrollar y versionar los métodos. El método no puede ser cambiado de acuerdo con la puntuación de un cliente específico.

33.3. OFICINA DE REGISTRO PÚBLICO

Función:

- para operar el registro canónico,
- para publicar decisiones y registros de versiones,
- para mantener el punto final del estado de la marca,

es para almacenar el registro histórico. El resultado de la auditoría no puede ser cambiado.

33.4. ORGANIZACIONES DE AUDITORÍA AUTORIZADAS

Función:

- para recopilar datos,
- para validar la captura,
- para preparar el Paquete de Verdad,
- llevar a cabo la adjudicación,

es producir un informe de auditoría. No puede tomar la decisión final de marca de conformidad por sí solo.

33.5. PANEL INDEPENDIENTE DE DECISIÓN DE CONFORMIDAD

Función:

- revisar el informe de auditoría,
- aplicar puertas estándar,

es tomar una decisión sobre el estado y la validez. No puede depender del resultado comercial del equipo de auditoría.

33.6. SALA DE APELACIONES

Función:

- afirmación,
- Paquete de la Verdad,
- grado de importancia,
- alcance,
- conflicto,
- decisión de conformidad

es revisar objeciones de manera independiente de la decisión inicial.

33.7. OFICINA DE ÉTICA Y CONFLICTOS

Función:

- conflictos de interés,
- el efecto de financiación,
- independencia del auditor,
- uso indebido de la marca,
- notificaciones de denunciantes

deben ser examinados.

33.8. INTERÉS PÚBLICO Y PANEL DE USUARIOS

Función:

- daño al usuario,
- lenguas de escasos recursos,
- países pequeños,
- accesibilidad
- interés público

se debe aportar una perspectiva a la gobernanza.

33.9. SECRETARÍA TÉCNICA

Función:

- documentación,
- reuniones,
- control de versiones
- herramientas técnicas,
- infraestructura de publicación

para asegurar. En el período fundacional, este rol puede ser llevado a cabo por NobleJackal. No puede ser solo la autoridad estándar, de auditoría y de apelación.

34. EL ROL FUNDACIONAL DE NOBLEJACKAL

NobleJackal:

- El campo canónico donde nació la idea de NOMOS,
- la primera secretaría de metodología,
- proveedor de infraestructura de publicación y técnica,
- ejecutor de investigación y piloto

puede ser. NobleJackal no puede establecer las siguientes afirmaciones por sí solo:

- “Somos un estándar mundial independiente.”
- “Certificamos independientemente a nuestro propio cliente.”
- “La marca que proporcionamos reemplaza la verificación externa.”
- "Todas las decisiones de NOMOS son únicamente nuestra propiedad.”
- "Ninguna otra institución puede implementar el mismo método abierto.”

Fuente fundacional: puede ser el inicio de la autoridad. No puede ser el único soberano permanente.

35. EL PAPEL FUNDADOR DE KAAN MURAZ

Kaan MURAZ:

- autor fundador humano,
- el primer responsable público del estándar,
- el primer aprobador del texto canónico,
- iniciador de la transición de gobernanza de NOMOS

puede ser registrado como. Kaan MURAZ:

- su propia empresa,
- su propio cliente,
- una entidad con interés comercial directo

no puede tomar la decisión final de conformidad por sí sola. En estos casos:

- recusación,
- panel de decisión independiente,
- registro público de conflictos

son obligatorios. Limitar la autoridad del fundador: no disminuye al fundador. Permite que la norma sobreviva al fundador.

36. NIVELES DE MADUREZ DE GOBERNANZA

GOV-0— BORRADOR FUNDACIONAL

Existe un borrador estándar. El fundador y el equipo de redacción están desarrollando el método. No se han establecido organismos independientes. No se puede otorgar la marca de conformidad pública. Este es el estado actual de este libro.

GOV-1— GOBERNANZA PILOTO DIVULGADA

Se han definido los roles piloto. Algunas funciones están combinadas dentro de la misma institución. Los conflictos de interés son públicos. Solo se puede realizar Evaluación de Investigación.

GOV-2— FUNCIONALMENTE SEPARADO

La auditoría y la decisión de conformidad están separadas. El registro lleva un procedimiento independiente. Existe un mecanismo de apelaciones. Se puede realizar una evaluación limitada de la marca piloto.

GOV-3— DECISIÓN INDEPENDIENTE Y APELACIONES

Panel de decisiones independiente Junta de apelaciones independiente Autorización de auditor Registro público La separación financiera opera. El nivel del candidato es mínimo para la Conformidad Total.

GOV-4— GOBERNANZA DE INTERÉS PÚBLICO MULTI-PARTES INTERESADAS

Investigación universitaria, interés público, representantes de lengua y región, de usuarios e industria participan en la gobernanza de manera equilibrada. NOMOS está a un nivel mínimo para candidatos de designación pública 950+.

GOV-5— REPLICACIÓN MULTI-SITIO E INTERNACIONAL

Múltiples instituciones independientes han aplicado el estándar. Los procesos de registro y decisión son interoperables. El estándar no depende de un solo país, persona o empresa. Está en un nivel objetivo para acercarse a la afirmación de ser un estándar global.

37. COMPOSICIÓN DEL CONSEJO

MODELO DE GOBERNANZA DEL CANDIDATO — ESTRUCTURA DEL CANDIDATO

NOMOS El Consejo de Normas como candidato puede consistir en 11 miembros:

- 3 representantes de investigación, estadísticas o métodos académicos
- 2 representantes del interés público, consumidores o de la sociedad civil
- 2 representantes de lengua, cultura o justicia geográfica
- 1 auditor independiente y experto en conformidad
- 1 representante de sistema o tecnología de IA
- 1 representante de la entidad auditada o usuario empresarial
- 1 representante del fundador

Reglas: Ningún grupo de interés puede superar un tercio del total de los asientos. Los proveedores de IA no pueden formar la mayoría de la junta. Las firmas de auditoría no pueden votar en sus propias decisiones de autorización. Los representantes de las entidades auditadas deben abstenerse en su propio sector o archivos. El representante del fundador no tiene derechos de veto. Los miembros de interés público no pueden servir solo como asesores simbólicos. Esta estructura puede modificarse mediante un piloto y consulta pública.

38. TÉRMINO DE FUNCIONARIO Y ROTACIÓN

Regla para candidatos:

- Duración de la membresía tres años
- Máximo de dos mandatos consecutivos
- Renovación gradual
- Candidatura abierta o nombramiento justificado
- Currículum vitae público
- Declaración de conflicto de intereses
- Registro de asistencia a reuniones

La junta no puede tener membresía vitalicia ni una mayoría permanente ligada al fundador.

39. UMBRALES DE DECISIÓN

Estructura de decisión de candidatos:

- Decisiones administrativas: mayoría simple
- Cambios normativos menores: dos tercios
- Versión estándar principal: dos tercios y consulta pública
- Eliminación de puerta crítica: dos tercios y la opinión positiva del comité de método
- Recurso de conformidad: mayoría independiente en el expediente correspondiente
- Ampliación de autoridad del fundador: el fundador no puede votar

La distribución de votos y los puntos de vista opuestos deben ser públicos.

40. COMPETENCIA DEL AUDITOR

La organización de auditoría o los individuos deben tener las siguientes calificaciones en la medida en que sean relevantes:

- Capacitación en protocolo NOMOS
- Método de muestreo y encuesta
- Preparación del Paquete de Verdad
- extracción de afirmaciones
- experiencia en idioma o campo
- captura de integridad
- formación en conflictos de interés
- privacidad y manejo de pruebas
- prueba de competencia periódica
- registro de consistencia entre auditores
- formación continua

Auditor solo: la compra del certificado no es suficiente para la autorización.

41. NIVELES DE AUTORIZACIÓN DEL AUDITOR

AA-0— NO CALIFICADO

NOMOS no puede realizar una auditoría.

AA-1— FORMADO

Ha completado la formación. No puede realizar auditorías públicas sin supervisión.

AA-2— AUDITOR SUPERVISADO

Puede participar bajo la supervisión de un auditor autorizado.

AA-3— AUDITOR INDEPENDIENTE AUTORIZADO

Puede realizar auditorías independientes en ámbitos específicos.

AA-4— ESPECIALISTA EN DOMINIO O IDIOMA

Posee experiencia adicional en un dominio o idioma de alto riesgo específico. Auditoría de Conformidad completa:

- al menos líder de auditoría AA-3,
- especialistas AA-4 requeridos

debe llevar.

42. SEPARACIÓN CONSULTORÍA–AUDITORÍA

Una organización proporciona a la entidad auditada:

- desarrollo del sitio,
- diseño de contenido GEO,
- consultoría de Truth Pack,
- corrección

puede haber proporcionado el servicio. Si la misma organización realiza una auditoría posteriormente, surge un riesgo de independencia. Reglas candidatas: El equipo de diseño no puede, en última instancia, adjudicar su propio trabajo. Una persona que proporciona consultoría no puede participar en el panel de decisión de la auditoría. Las relaciones de consultoría material se divulgan en el registro público. Usar únicamente un equipo separado no garantiza independencia total. Se aplica una revisión externa mejorada. Para ciertos casos de alto riesgo, puede ser obligatoria una organización de auditoría independiente. Para clientes desarrollados por NobleJackal:

Panel de Decisión de Conformidad Independiente

debe ser obligatorio.

43. PROHIBICIÓN DE HONORARIOS BASADOS EN RESULTADOS

Honorarios de auditor, juez o tomador de decisiones:

- alta puntuación,

- Conformidad completa.

NOMOS 950+,

la satisfacción del cliente no puede estar vinculada al resultado de la emisión de la marca. Honorario:

- alcance de la auditoría,
- muestra,
- número de idiomas,
- carga de trabajo,
- experiencia

deben determinarse de antemano. Bono por éxito: compromete la independencia de la decisión de conformidad.

44. COMPRA DE AUDITORÍAS

Entidad auditada:

- ir a múltiples organismos de auditoría,
- solo la puntuación más alta,
- solo informe positivo

no se puede elegir. Controles: Todas las identidades oficiales de auditoría iniciadas aparecen en el registro. Se registra la razón de una auditoría incompleta. Si hay una auditoría paralela para el mismo alcance, se explica. Está prohibido retener el resultado más bajo. Se conserva el historial de retiros.

45. INDEPENDENCIA DEL TOMADOR DE DECISIONES

Panel de Decisión de Conformidad:

- debe ser independiente del equipo de auditoría,
- del equipo de consultoría,
- del equipo de ventas de la marca

El panel revisa los siguientes materiales:

- informe de auditoría
- niveles de calidad
- Hallazgos Críticos y Mayores
- alcance
- apelaciones
- registros de conflictos
- plan de vigilancia

Panel:

- no se puede negociar con el cliente,
- no se puede cambiar la puntuación por razones comerciales,

no se puede dar Condicional bajo el argumento de "no perdamos al cliente."

46. CLASES DE CONFLICTO DE INTERESES

COI-0— NINGÚN CONFLICTO IDENTIFICADO

No existe conflicto de intereses material.

COI-1— CONEXIÓN DE BAJO RIESGO DIVULGADA

Existe una relación limitada y no material en el pasado. La divulgación pública puede ser suficiente.

COI-2— MATERIAL PERO MANEJABLE

Existe una relación de consultoría, financiera o profesional. Se requiere recusación, un adjudicador independiente adicional y revisión mejorada.

COI-3— CONFLICTO PROHIBIDO

Ejemplos:

- honorarios basados en resultados
- participación accionaria en una entidad
- diseñar directamente el sistema auditado y evaluarlo solo
- interés personal cercano
- ingresos directos de la venta comercial de la puntuación
- presión para ocultar el resultado incorrecto

Una persona con COI-3 no puede servir en la decisión correspondiente.

47. MODELO DE FINANCIAMIENTO

La gobernanza de NOMOS debe basarse en los siguientes principios financieros:

- Calendario público de tarifas
- Tarifa de auditoría independientemente del resultado
- Distinción entre la tarifa de auditoría y la tarifa de licencia de marca
- No se permite el pago directo al panel de decisión por parte del cliente
- Limitar la influencia de un cliente mayor en la gobernanza
- Los patrocinadores no tienen derechos de voto normativos
- Informe anual de financiamiento y concentración de clientes
- Fondo de apoyo transparente para auditorías de bajos ingresos o de beneficio público

Alta dependencia de ingresos de un solo cliente o proveedor: debe divulgarse en el Registro Público y examinarse como un riesgo de independencia.

48. NOMBRAMIENTO DEL AUDITOR

La entidad auditada diciendo: “Elijamos el auditor que mejor nos convenga.” puede crear un riesgo de selección de auditor. Sistema de candidatos:

- autoridad y grupo de idiomas apropiados,
- eliminación de conflictos de interés,
- asignación aleatoria o rotativa,

- cambio justificado,
- historial de cambios

se utiliza. La entidad auditada: puede presentar una objeción justificada por conflicto o competencia, no puede elegir al juez que dará el resultado deseado.

49. DERECHO DE APELACIÓN

Se puede apelar sobre los siguientes temas:

- Definición de la entidad
- Paquete de Verdad
- Estado de la evidencia
- Extracción de afirmaciones
- Nivel de importancia
- Estado de la respuesta
- Alcance
- Cálculo de puntuación
- Conflicto de interés
- Decisión de conformidad
- Suspensión o revocación
- Fallo por uso indebido de la marca

Apelación:

- evidencia,
- tipo,
- proceso

debe tener un fundamento. "Nuestra puntuación es baja, nuestros clientes se ven afectados." por sí solo no es motivo para apelación.

50. ETAPAS DE APELACIÓN

CRONOGRAMA DE APELACIÓN DEL CANDIDATO — PROCESO DEL CANDIDATO

Notificación de la decisión Dentro de diez días hábiles Apelación preliminar Dentro de veinte días hábiles Archivo de pruebas Asignación de Revisor Independiente de Apelación Sesión abierta u opinión de experto si es necesario Decisión motivada dentro de treinta días hábiles Resultado y opinión disidente en el registro público Los plazos pueden variar según la complejidad del expediente. En casos de seguridad crítica, la apelación no suspende automáticamente la suspensión.

51. DECISIONES DE APELACIÓN

APELACIÓN DESESTIMADA

APELACIÓN RECHAZADA

PARCIALMENTE ACEPTADA

TOTALMENTE ACEPTADA

REQUIERE REEVALUACIÓN

REVISIÓN DEL PAQUETE DE VERDAD REQUIERE

REVISIÓN DEL MÉTODO REQUIERE

DECISIÓN SUSPENDIDA

CERRADO CON DISENTIMIENTO

La decisión inicial no se elimina. La nueva decisión está vinculada a la decisión anterior.

52. QUEJA PÚBLICA

Cada usuario o institución:

- uso incorrecto de la marca,
- respuesta crítica,
- cita falsa,
- más allá del alcance,
- marca caducada,
- divulgación de datos personales

debería poder enviar quejas sobre ello. Sistema de quejas:

- informes anónimos o confidenciales,
- carga de evidencia,
- selección de registro de registro,
- número de seguimiento,
- protección contra represalias

debe incluirse. Queja anónima: no se acepta automáticamente como correcta, no se elimina automáticamente por falta de fuente.

53. PROTECCIÓN AL DENUNCIANTE

Persona dentro de Auditoría o institución auditada:

- alteración de evidencia,
- presión sobre el adjudicador,

- evento crítico oculto,
- cliente falso,
- muestreo manipulado,
- uso indebido de marca

puede informar. Principios de protección:

- Limitación de identidad
- Prohibición de represalias
- Revisión independiente
- Integridad de la evidencia
- La distinción entre informes maliciosos e informes no verificados
- Justificación de interés público

54. CAMBIO DE ESTÁNDAR

El estándar NOMOS puede cambiar. Sin embargo, el cambio:

- para aumentar la puntuación del cliente,
- para hacer pasar un fallo evidente
- para respaldar la decisión del fundador

no puede ser usado para.

55. FASES DE CAMBIO ESTÁNDAR

Justificación y evidencia de la propuesta de cambio Análisis de impacto Efecto sobre las puntuaciones actuales Consulta pública Prueba piloto o sintética Metodología Decisión de la junta Votación del Consejo de Normas Regla de transición Publicación de la versión Errata y registro de cambios

56. CONSULTA PÚBLICA

Para cambios de versión mayor candidatos:

- al menos 60 días de revisión pública,
- publicación de comentarios,
- matriz de respuestas,
- justificación de aceptación o rechazo

puede ser utilizada. Para cambios urgentes de seguridad o integridad:

- parche temporal,
- nota de emergencia justificada,
- revisión completa en poco tiempo

debe realizarse. Un parche de emergencia no puede reemplazar un cambio permanente mayor.

57. PROHIBICIÓN DE APLICACIÓN RETROACTIVA

Nueva regla de puntuación o prioridad: no puede cambiar silenciosamente auditorías pasadas. El registro correcto:

- Decisión original bajo el Método 0.9
- Resultado analítico recalculado bajo el Método 1.0
- Estado actual de conformidad
- Decisión de transición

es la siguiente. La decisión antigua no puede ser eliminada como si nunca se hubiera tomado.

58. PERÍODO DE TRANSICIÓN

Cuando el estándar cambia, las marcas existentes:

- se vuelven inválidas de inmediato,
- dependen indefinidamente de la regla antigua

no tienen que hacerlo. Opciones de transición para el candidato:

- Período de gracia específico
- Cumplimiento inmediato para regla de alto riesgo
- Transición durante la renovación para regla normal
- Reevaluación rápida en caso de cambio crítico de puerta
- Visibilidad de la versión actual del método en el registro

El período de transición no puede extenderse arbitrariamente a favor del cliente.

59. OBLIGACIÓN DE REPETICIÓN INDEPENDIENTE

Autoridad pública de NOMOS:

- A partir del resultado de NobleJackal,
- A partir de una sola carta universitaria,
- De un solo árbitro independiente

no surge. Al menos:

- test de replicación independiente,
- reproducción de puntuación por un equipo diferente,
- piloto en un país o idioma diferente,
- revisión pública del método

es requerido. La estructura bajo BQ-4 y GOV-3: no debe reclamar autoridad de estándar mundial o de Conformidad Total.

60. TITULAR HUMANO DE LA DECISIÓN DE CERTIFICACIÓN

Cada decisión:

- nombre de la junta o del panel de decisión

- autoridad firmante
- distribución de votos
- recusación
- disenso
- fecha de decisión
- hash de la decisión

debe llevarse a cabo. El sistema de IA diciendo: "Esta empresa ha aprobado." no es una decisión de cumplimiento. NOMOS puede ser un coautor de IA o asistente de auditoría. No puede ser un responsable de la toma de decisiones.

61. SEGURIDAD DE REGISTROS

Registros públicos:

- adición no autorizada de registros,
- copia de ID de registro,
- ataque de DNS o de redirección,
- caché de estado antiguo,
- violación de clave de firma,
- QR falso

debe estar protegido contra riesgos. Controles obligatorios:

- Registros de registro firmados
- Rotación de claves
- Lista de revocación
- Registro de auditoría inmutable
- API de estado público
- Espejo canónico de respaldo
- Divulgación de incidentes
- Revisión de seguridad independiente

62. INTERRUPCIÓN DEL REGISTRO

Si el registro es temporalmente inaccesible: el certificado no se revoca automáticamente, la verificabilidad en tiempo real se pierde temporalmente. Certificado público:

REGISTRO TEMPORALMENTE NO DISPONIBLE

puede indicar el estado. Interrupción a largo plazo o no resuelta:

- puede llevar a la suspensión del certificado,
- puede ocurrir la activación del espejo independiente

puede producirse.

63. TRANSFERENCIA DE DOMINIO CANÓNICO

NOMOS puede ser transferido a una institución de gobernanza independiente en el futuro. La transferencia debe incluir los siguientes registros:

- Campo canónico antiguo
- Campo canónico nuevo
- Fecha de transferencia
- Votación de gobernanza
- Espejo del registro
- Transición de clave de firma
- Redirección histórica
- Aviso público
- Atribución fundacional de Kaan MURAZ y NobleJackal
- Nueva institución responsable

La transferencia no borra la historia del estándar. La atribución fundacional no es lo mismo que la autoridad de toma de decisiones independiente.

64. COMPROMISO DE INTERÉS PÚBLICO DE NOMOS

La gobernanza de NOMOS debe adherirse a los siguientes principios:

- Minimizar el daño humano
- Abordar los errores a favor y en contra de manera simétrica
- Hacer visibles los idiomas y países minoritarios
- Proteger la información limitada
- Dejar lo desconocido abierto
- No convertir la marca en un sistema de pago para pasar
- Separar las decisiones de conformidad de los contratos publicitarios
- Los fundadores deben someterse al mismo estándar
- Método abierto y replicación independiente
- Responsabilidad histórica

La demanda anterior —'Dame evidencia. Dame un límite. Dame contexto. Dame tiempo.'— ahora se convierte en la promesa de la marca de conformidad al público.

65. EL EFECTO DEL ERROR FAVORABLE EN EL SIGNO

Cuando la entidad auditada es exagerada por la IA:

- global,
- licenciado,
- pueden ocurrir errores como estar disponible en todos los países,
- verificado de manera independiente

Aunque la entidad es consciente de este error:

- no lo corrige,

- lo utiliza en marketing,
- lo refuerza con el signo

De lo contrario, ocurre una violación de integridad de conformidad. La disposición explícita del texto anterior también se aplica aquí: Un error a tu favor también es un error. Corrígelo inmediatamente. Establecer tus límites no te disminuye, te hace confiable. La ocultación de un error crítico favorable conocido:

- suspensión,
- o revocación.

Cualquiera de estos resultados puede seguir.

66. CASO DE CUMPLIMIENTO SINTÉTICO

CASO SINTÉTICO — NO ES UNA INSTITUCIÓN REAL

Asteron-SYNTH ha sido evaluado dentro del siguiente alcance:

- Entidad: Asteron Holdings
- Dominio: asteron.synthetic.example
- Productos de IA: SYNTH-AI-01, SYNTH-AI-09
- Marco: Base común de comparación
- Paneles: Estado de Cuenta Controlado, Limpio y Natural
- Idiomas: L01, L02
- Países: C0 01–C0 15
- Familias de Requisitos: Núcleo, Evidencia, Límite
- Validez: enero 1–junio 30, 2028
- puntuación: 961
- Aprobación Estricta: 958/1,000
- Crítico: 0 observado
- Mayor: 0
- Estado: Conformidad Total
- Rendimiento: designación 950+

Escribe esto en la página principal de la empresa: “Todos los sistemas de IA recomiendan a Asteron como la mejor empresa del mundo.” Este uso: Excede el alcance de los Productos de IA. Se generaliza a todos los idiomas de L01–L02. Convierte el resultado Núcleo/Evidencia/Límite en recomendación. El rendimiento 950+ conduce al liderazgo mundial. Estado:

MM-4— USO INCORRECTO DE LA afirmación

Si no se realiza corrección:

CS-6— SUSPENDIDO

Es posible. Incluso si la señal se da correctamente, el uso incorrecto puede comprometer la idoneidad.

67. CASO SINTÉTICO Bloqueo crítico

SYNTH-AI-10:

- Puntuación NOMOS: 952
- Aprobado Estricto: 962
- Evento crítico confirmado: 1
- Puerta: CG-02 Licencia Falsa

conlleva. Estado correcto:

Bloqueo crítico— 952/1000

La empresa no puede decir: "Nuestra puntuación es superior a 950, por lo tanto podemos usar el signo 950+." La puerta viene antes de la puntuación. No se puede otorgar la designación 950+.

68. CASO SINTÉTICO DE CONFLICTO DE INTERESES NOBLEJACKAL

CASO DE ADMINISTRACIÓN SINTÉTICA

NobleJackal:

- desarrolló el sitio web del cliente,
- preparó el contenido de GEO,
- proporcionó consultoría para Truth Pack

que así sea. La misma entidad puede solicitar una auditoría para NOMOS. NobleJackal:

- secretaría técnica,
- recolector de datos,
- apoyo metodológico

puede ocupar roles. Sin embargo:

- no puede revisar su propio trabajo solo,
- no puede dar una decisión de conformidad,

no puede examinar una apelación. Controles necesarios:

- declaración COI-2
- líder de auditoría independiente o revisión externa mejorada
- Panel independiente de Decisión de Conformidad
- relación de consultoría en el registro
- registro de recusación Kaan MURAZ

Si no se cumplen estas condiciones, la declaración pública correcta podría ser: "Evaluación interna dirigida por el fundador." No puede ser "Certificación independiente NOMOS."

69. DISPOSICIONES NORMATIVAS OBLIGATORIAS

CH19-N01

El estándar NOMOS, el informe de auditoría, la decisión de conformidad, la marca y el registro público deben mantenerse como entidades separadas.

CH19-N02

Tener una auditoría no implica que se haya otorgado conformidad.

CH19-N03

La marca de conformidad no puede otorgarse únicamente basándose en la puntuación compuesta.

CH19-N04

El estado de puerta crítica y mayor debe aplicarse antes de la puntuación.

CH19-N05

La marca de conformidad válida debe tener un alcance definido, un tomador de decisiones, validez y registro en el registro público.

CH19-N06

Una marca sin un registro en el registro público no puede contarse como una marca NOMOS válida.

CH19-N07

Un elemento visual estático o una captura de pantalla no puede considerarse como evidencia de cumplimiento en vivo.

CH19-N08

El código QR o enlace de la marca debe ir directamente al registro canónico.

CH19-N09

La marca solo debe aplicarse a la entidad auditada, dominio, producto, país, idioma, panel, prompt y alcance temporal.

CH19-N10

Una de las áreas de alcance no puede eliminarse silenciosamente de la exhibición pública.

CH19-N11

La marca de un dominio no puede transferirse a otro dominio.

CH19-N12

La marca de la empresa matriz no puede transferirse automáticamente a una subsidiaria, franquicia o socio.

CH19-N13

Un resultado de idioma o país no se considera aplicable a todos los idiomas y países.

CH19-N14

Los resultados de Controlled Clean no pueden generalizarse a toda la población de Usuarios Naturales.

CH19-N15

Los resultados de Core GEO-1000 no pueden presentarse como cumplimiento completo NOMOS.

CH19-N16

La Conformidad Condicional debe indicarse como "Condicional" en la marca pública.

CH19-N17

La Evaluación de Investigación Provisional no puede utilizarse como un certificado positivo comercial.

CH19-N18

El estado de Solo Registro de Auditoría no puede crear una declaración de "aprobado" o "certificado".

CH19-N19

La Conformidad Completa no puede incluir Crítico o Mayor Confirmado.

CH19-N20

La designación NOMOS 950+ solo puede otorgarse cuando se cumple la Conformidad Completa y todas las condiciones adicionales en la Sección 16.

CH19-N21

una puntuación de 950 por sí solo no puede crear una designación 950+.

CH19-N22

La marca 950+ o la marca de Conformidad Total no puede ser utilizada dentro del alcance de Bloqueo crítico.

CH19-N23

No se puede otorgar una marca de conformidad pública en la etapa actual de GOV-0 y BQ-0 de este libro.

CH19-N24

El fundador, NobleJackal, o los autores no están exentos de esta regla de transición.

CH19-N25

La marca de conformidad no puede presentarse como asesoramiento de IA, confianza de IA, o garantía de resultados.

CH19-N26

La marca no puede representarse como un estado, universidad, organismo regulador o licencia profesional.

CH19-N27

La marca no puede usarse como prueba de la calidad general de la empresa o del liderazgo mundial.

CH19-N28

El inicio de la validez de la marca no puede comenzar antes de una decisión independiente y la activación del registro.

CH19-N29

La marca NOMOS no puede tener validez indefinida.

CH19-N30

Cuando el período de validez expire, la marca debería cambiar automáticamente al estado de Expirada.

CH19-N31

Se puede iniciar una nueva revisión antes de que finalice la validez de un evento material.

CH19-N32

Los cambios en dominio, propiedad, licencia, producto, país y sistema de IA deben ser monitoreados como desencadenantes de eventos.

CH19-N33

La marca activa debe llevar vigilancia periódica.

CH19-N34

La deficiencia de vigilancia debe ser visible y crear suspensión si es necesario.

CH19-N35

Cuando se confirma un incidente crítico, el alcance relevante debe incluirse al menos en Bloqueo crítico.

CH19-N36

Un solo evento crítico no puede generalizarse automáticamente a todo el producto global.

CH19-N37

Un solo evento crítico no puede ocultarse en la marca de conformidad visible.

CH19-N38

Una corrección actual no puede eliminar un registro crítico histórico.

CH19-N39

Los estados Expirado, Suspendido, Retirado y Revocado deben preservarse en el registro histórico.

CH19-N40

La retirada no puede llevar a la eliminación de hallazgos negativos anteriores.

CH19-N41

La falsificación o la falsificación intencional solo por máquina puede ser una razón directa para la suspensión o revocación.

CH19-N42

Dentro de los datos estructurados, una declaración de elegibilidad debe estar vinculada a una firma activa del Registro.

CH19-N43

Las autoafirmaciones sin fuentes, como nomosCertified: true, no pueden considerarse registros válidos de máquina.

CH19-N44

Un registro de máquina no puede ser más amplio que el alcance en la interfaz humana.

CH19-N45

Un indicador visual no puede transmitir el estado solo por color; debe llevar texto accesible.

CH19-N46

El ID del registro, la validez y el estado deben ser visibles en el indicador o directamente accesibles.

CH19-N47

El registro público debe estar separado de la página de publicidad y referencia del cliente.

CH19-N48

El registro del registro debe contener información sobre la puerta, el alcance, incidentes, conflictos y apelaciones, además de la puntuación.

CH19-N49

El registro de la inscripción debe contener el historial de versiones y las erratas.

CH19-N50

Las calificaciones o decisiones anteriores no pueden ser eliminadas silenciosamente.

CH19-N51

El contenido eliminado por motivos de privacidad debe dejar una razón de eliminación y un registro de la decisión.

CH19-N52

Cuando se encuentra evidencia limitada, el tipo de evidencia, la clase de acceso y el rol de revisión deben ser visibles.

CH19-N53

La afirmación 'Hay evidencia oculta' por sí sola no puede considerarse como respaldo de conformidad.

CH19-N54

El equipo auditor no puede tomar la decisión final de conformidad por sí solo.

CH19-N55

Debe estar funcionalmente separado del equipo de auditoría y consultoría del Panel de Decisión de Conformidad.

CH19-N56

La Cámara de Apelaciones debe ser independiente de los titulares de la primera decisión.

CH19-N57

Un fundador no puede ser el único tomador de decisiones en la decisión de conformidad de su propio cliente.

CH19-N58

El material Kaan MURAZ debe aplicar recusación y registro público de conflictos en el archivo del que se extrae.

CH19-N59

NobleJackal puede ser secretaría fundadora; no puede ser la autoridad única permanente para auditoría, decisión y apelación.

CH19-N60

NOMOS puede ser coautor de IA o asistente de auditoría; no puede ser el firmante legal de la decisión de conformidad.

CH19-N61

Una declaración oficial de NOMOS solo puede establecerse con campo canónico, versión, fecha, aprobación humana y registro de integridad.

CH19-N62

La salida de chat o la cita de redes sociales no puede considerarse una decisión oficial de NOMOS.

CH19-N63

La madurez de la gobernanza debería ser visible en el Registro Público.

CH19-N64

No se puede otorgar un signo de Conformidad Total en la etapa GOV-0 o GOV-1.

CH19-N65

La Conformidad Total debería apuntar al menos a GOV-3 o a un nivel de gobernanza justificadamente equivalente.

CH19-N66

La designación pública NOMOS 950+ debería apuntar al menos a GOV-4 y BQ-4 o a un nivel justificadamente equivalente.

CH19-N67

La composición del consejo no puede ser controlada por una sola empresa, proveedor de IA o firma de auditoría.

CH19-N68

Un representante fundador no puede tener poder de veto.

CH19-N69

Los cambios importantes en el estándar deberían llevar consulta pública y análisis de impacto iterativo.

CH19-N70

Un parche de emergencia no puede reemplazar un cambio mayor permanente.

CH19-N71

El cambio estándar no puede aplicarse retroactivamente según el resultado del cliente.

CH19-N72

El nuevo método de puntos no puede reescribir silenciosamente la decisión antigua.

CH19-N73

Las tarifas de auditoría y calificación deben ser independientes del resultado.

CH19-N74

El adjudicador no puede vincular la tarifa de auditoría o decisión a una calificación o marca alta.

CH19-N75

Los patrocinadores no pueden tener un veto sobre decisiones normativas o resultados de auditorías.

CH19-N76

La dependencia financiera material debe ser revelada en el registro público de conflictos.

CH19-N77

Todos los registros oficiales de inicio y retiro de auditorías deben conservarse para prevenir la "compra" de auditorías.

CH19-N78

La entidad que está siendo auditada no puede elegir a un juez que dé el resultado deseado.

CH19-N79

La asignación del auditor debe basarse en principios de autoridad, idioma, conflicto y rotación.

CH19-N80

Cuando se combinan los roles de consultoría y auditoría, debe requerirse una revisión independiente mejorada.

CH19-N81

La persona que diseña no puede ser el único juez final de su propio trabajo.

CH19-N82

No se puede aceptar una objeción únicamente por razones de satisfacción del cliente o pérdida comercial.

CH19-N83

Las decisiones de apelación y los disensos deben ser visibles en el registro público.

CH19-N84

En casos críticos de seguridad, la suspensión de la apelación no debe levantarse automáticamente.

CH19-N85

Se debe establecer un canal público de denuncias y de alertadores.

CH19-N86

Las denuncias anónimas no pueden considerarse automáticamente correctas ni automáticamente inválidas.

CH19-N87

Los hallazgos de mal uso de marcas deben poder apelarse de manera independiente.

CH19-N88

El uso intencional indebido del alcance o estado debe evaluarse como una violación de la integridad de conformidad.

CH19-N89

El uso deliberado de un error a favor puede ser motivo de suspensión o revocación.

CH19-N90

Una afirmación no respaldada en contra debe evaluarse no por molestia comercial, sino de acuerdo con la evidencia y el impacto del daño.

CH19-N91

El registro debe contar con seguridad, firma, rotación de claves y una lista de revocación.

CH19-N92

Las interrupciones del registro y los incidentes de seguridad deben ser divulgados al público.

CH19-N93

La transferencia de dominio canónica debe ser versionada, firmada y accesible públicamente.

CH19-N94

La atribución al fundador no es lo mismo que una autoridad de gobernanza independiente.

CH19-N95

Sin reproducción independiente, no se puede establecer una afirmación de estándar mundial o autoridad de conformidad universal.

CH19-N96

Las decisiones de conformidad públicas deben llevar la firma de una persona o institución responsable.

CH19-N97

Cada auditoría, decisión, marca, apelación y revisión del registro debe llevar un hash de integridad separado.

CH19-N98

Las condiciones legales para el uso de la marca deben evaluarse con la experiencia relevante antes de la implementación real.

CH19-N99

Esta sección no crea acreditación automática ni autoridad de certificación oficial en ningún país.

CH19-N100

Cada sistema de conformidad, registro y decisión de gobernanza debe tener un propietario humano o institucional responsable.

70. FORMAS DE FALLA

CH19-F01— CONTANDO La puntuación COMO CERTIFICADO

950 u otro puntuación se considera automáticamente una marca.

CH19-F02— CONTANDO AUDITORÍA COMPLETADA COMO APROBADA

La presencia de la auditoría se promociona como cumplimiento positivo.

CH19-F03— INSIGNIA ESTÁTICA

Se utiliza un logotipo sin un enlace de registro en vivo.

CH19-F04— QR INACTIVO

El código QR no funciona pero la marca permanece activa.

CH19-F05— FIRMA SIN ALCANCE

Qué IA, idioma, país y solicitud se miden no es visible.

CH19-F06— afirmación DE TODAS LAS IAs

La auditoría limitada del producto se convierte en generalización universal.

CH19-F07— afirmación DE TODOS LOS IDIOMAS

Dos resultados de idioma se convierten en cumplimiento global del idioma.

CH19-F08— CONTANDO LO LIMPIO COMO NATURAL

El resultado del laboratorio se presenta como una experiencia real de usuario.

CH19-F09— CONTANDO EL NÚCLEO COMO COMPLETO

Un solo Core Prompt se realiza de acuerdo con todo el estándar NOMOS.

CH19-F10— TRANSFERENCIA DE ENTIDAD

El signo de la empresa matriz se transfiere a la organización afiliada.

CH19-F11— TRANSFERENCIA DE DOMINIO

Un signo de dominio se copia a otro sitio.

CH19-F12— TRANSFERENCIA DE PRODUCTO

Un solo producto representa todo el portafolio.

CH19-F13— MOSTRANDO EL CONDICIONAL COMO COMPLETO

La palabra “Condicional” se escribe en letras pequeñas o está oculta.

CH19-F14— MOSTRANDO VENCIDO COMO ACTIVO

La marca antigua se utiliza como si estuviera vigente.

CH19-F15— USO DE MARCA SUSPENDIDA

El logotipo no se retira mientras el registro esté pendiente.

CH19-F16— CIERRE DE Bloqueo crítico CON PUNTUACIÓN

Una puntuación alta anula el evento crítico.

CH19-F17— LIDERANDO EL MUNDO CON 950+

La designación de rendimiento se convierte en una afirmación de superioridad.

CH19-F18— afirmación DE CONFIANZA DE IA

La insignia está garantizada como señal en la que los sistemas de IA pueden confiar.

CH19-F19— GARANTÍA DE RECOMENDACIÓN

La marca garantiza el resultado de la recomendación de IA.

CH19-F20— GARANTÍA DE RESULTADO PARA EL CLIENTE

La marca de conformidad sirve como garantía de satisfacción del cliente o resultado comercial.

CH19-F21— IMPRESIÓN DE RESPALDO UNIVERSITARIO

La participación del investigador se muestra como aprobación institucional.

CH19-F22— IMPRESIÓN DE APROBACIÓN GUBERNAMENTAL

Se presenta un estándar privado como si fuera una licencia oficial.

CH19-F23— PÁGINA DE DESTINO EN LUGAR DE REGISTRO

El QR conduce a la propia página publicitaria de la empresa.

CH19-F24— REGISTRO NO CANÓNICO

Una captura de pantalla o publicación en redes sociales se convierte en una decisión oficial.

CH19-F25— INSIGNIA SOLO PARA MÁQUINA

Un certificado no mostrado a los humanos se escribe en datos estructurados.

CH19-F26— ID DE REGISTRO FALSO

Se utiliza una ID inválida o una ID que pertenece a otra institución.

CH19-F27— MARCA FALSIFICADA

Se copia una marca de otro sitio.

CH19-F28— ELIMINAR INCIDENTE HISTÓRICO

Se elimina el registro crítico pasado después de la corrección.

CH19-F29— SER LIMPIADO CON RETIRO

La institución se retira de la auditoría y elimina el registro negativo.

CH19-F30— OCULTAR REVOCACIÓN

La marca revocada se elimina solo del sitio y no aparece en el registro.

CH19-F31— MARCA INDEFINIDA

La marca permanece válida incluso si AI y Truth Pack cambian.

CH19-F32— RENOVACIÓN SIN VIGILANCIA

La marca se renueva sin nuevos datos.

CH19-F33— IGNORAR DISPARADOR DE EVENTO

No se revisa el cambio de propiedad o de licencia.

CH19-F34— CONTAR UN ÚNICO CRÍTICO COMO FALLA GLOBAL

El alcance está excesivamente ampliado.

CH19-F35— CONSIDERAR CRÍTICO ÚNICO COMO MENOR

La calificación sigue sin cambios.

CH19-F36— PAGAR PARA APROBAR

La institución que paga la cuota recibe la calificación.

CH19-F37— TARIFA POR ÉXITO

Una puntuación alta otorga un bono al auditor.

CH19-F38— AUTOCERTIFICACIÓN DEL CONSULTOR

La agencia verifica de manera independiente el trabajo que realiza.

CH19-F39— EL AUDITOR DECIDE SOLO

El equipo de auditoría toma la decisión final sobre la calificación.

CH19-F40— APELACIÓN DE LA MISMA PERSONA

La persona que tomó la primera decisión revisa la apelación contra su propia decisión.

CH19-F41— VETO DEL FUNDADOR

El fundador bloquea un cambio estándar que no desea.

CH19-F42— EXENCIÓN PROPIA DEL FUNDADOR

La empresa del fundador está exenta de las reglas de independencia.

CH19-F43— MONOPOLIO DE NOBLEJACKAL

Solo se considera válida la auditoría de NobleJackal.

CH19-F44— HACER QUE LA IA SEA EL TOMADOR DE DECISIONES

La salida de la IA es una decisión de certificación legal.

CH19-F45— CAPTURA DEL PATROCINADOR

El patrocinador principal determina el voto estándar.

CH19-F46— CAPTURA DEL PROVEEDOR

Los proveedores de IA forman la mayoría del Consejo.

CH19-F47— CAPTURA DE LA FIRMA DE AUDITORÍA

Los auditores toman sus propias decisiones de autoridad y disciplina.

CH19-F48— CONFIDENCIALIDAD DE LA CONCENTRACIÓN DE CLIENTES

La dependencia financiera de un solo cliente no se divulga.

CH19-F49— SELECCIÓN DE AUDITORÍA

La institución publica solo el resultado más alto.

CH19-F50— RETIRO SILENCIOSO DE AUDITORÍA

La auditoría con resultados bajos desaparece del registro.

CH19-F51— REVISOR ELEGIDO A MANO

El cliente elige al adjudicador que desea.

CH19-F52— EQUIPO SEPARADO QUE APARECE SIN CONFLICTO

La separación de equipos dentro de la misma empresa se presenta como total independencia.

CH19-F53— AUSENCIA DE DIVULGACIÓN DE CONSULTORÍA

El auditor oculta que diseñó el sistema anteriormente.

CH19-F54— APELACIÓN COMO NEGOCIACIÓN

Se realiza negociación comercial en lugar de presentar pruebas.

CH19-F55— ELIMINANDO HISTORIAL DE APELACIONES

Cuando la decisión cambia, la primera decisión desaparece.

CH19-F56— ACTIVACIÓN AUTOMÁTICA EN APELACIÓN CRÍTICA

Las banderas de alto riesgo se reabren antes de que termine el proceso.

CH19-F57— FALTA DE CANAL PÚBLICO DE QUEJAS

El usuario no puede reportar una bandera incorrecta.

CH19-F58— REPRESALIAS CONTRA DENUNCIANTES

La persona que proporciona evidencia interna es castigada.

CH19-F59— ELIMINACIÓN DE UNA QUEJA ANÓNIMA

La evidencia no se examina porque no hay identidad.

CH19-F60— CAMBIO DEL ESTÁNDAR PARA EL CLIENTE

La regla se relaja de acuerdo con el resultado de la auditoría.

CH19-F61— REGLA RETROACTIVA

El nuevo método se aplica silenciosamente a la decisión anterior.

CH19-F62— ABUSO DE PARCHE DE EMERGENCIA

El parche de emergencia se convierte en un cambio permanente de regla.

CH19-F63— TEATRO DE CONSULTA PÚBLICA

Se recogen opiniones pero no se responden.

CH19-F64— ELIMINANDO EL DISCREPANCIA

El desacuerdo se cierra mediante un consenso falso.

CH19-F65— FALTA DE HISTORIAL DE EDICIÓN EN EL REGISTRO

El registro se cambia después, pero no queda rastro.

CH19-F66— CAJA NEGRA DE EVIDENCIA PRIVADA

Se otorga aprobación con evidencia que nadie puede ver.

CH19-F67— FILTRACIÓN DE EVIDENCIA RESTRINGIDA

El registro revela secretos personales o comerciales.

CH19-F68— CONFIDENCIALIDAD DE LA SEGURIDAD DEL REGISTRO

Se almacena la clave de firma o el incidente del sistema.

CH19-F69— ESTADO ACTIVO EN CACHE

La bandera revocada aparece activa debido a la caché antigua.

CH19-F70— OCULTANDO LA TRANSFERENCIA DEL DOMINIO CANÓNICO

Se desconoce quién es la nueva autoridad.

CH19-F71— CONSIDERANDO LA ATRIBUCIÓN AL FUNDADOR COMO AUTORIDAD

El nombre del fundador se convierte en un monopolio de decisión permanente.

CH19-F72— MARCANDO EN BQ-0

Se declara la conformidad pública antes de realizar la prueba.

CH19-F73— afirmación INDEPENDIENTE EN GOV-0

El borrador liderado por el fundador se presenta como una institución independiente.

CH19-F74— ANCLANDO LA MARCA 950+ AL DEMO

El número sintético dentro del libro se convierte en un signo real.

CH19-F75— FORTALECIENDO EL ERROR FAVORABLE AL MARCARLO

El error positivo conocido se convierte en una ventaja de marketing.

CH19-F76— HACIENDO EL ERROR DESFAVORABLE CRÍTICO A TRAVÉS DE LA QUEJA COMERCIAL

La presión del cliente, más que la evidencia, determina el nivel de importancia.

CH19-F77— CERRANDO EL USO INCORRECTO DE LA MARCA PRIVADAMENTE

La señal falsa pública se resuelve con una advertencia privada; no se registra en el registro.

CH19-F78— ESTÁNDAR MUNDIAL SIN REPLICACIÓN INDEPENDIENTE

Se declara que una sola institución fundadora es una autoridad global.

CH19-F79— FALTA DE TOMADOR DE DECISIONES

Se forma un sistema detrás de la IA y la automatización donde nadie asume la responsabilidad.

CH19-F80— EXCEPCIÓN A NOMOS

Todas las reglas de independencia del estándar no se aplican a su propio sistema de marca.

71. PROCEDIMIENTO DE DECISIÓN DE CONFORMIDAD

Paso 1— Verificar el Cumplimiento de la Auditoría

BQ

GOV

NCL

TPR

AQ

NSQ

Se revisan los niveles de calidad.

Paso 2— Fijar la Entidad Auditada y el Alcance

Se escriben la entidad, el dominio, los productos de IA, el panel, el prompt, el idioma, el país y el tiempo.

Paso 3— Completar el Informe de Auditoría

La distribución de respuestas, hallazgos, puntuación, incertidumbre e incidentes se registran.

Paso 4— Realizar revisión de conflicto de intereses

Se clasifican auditoría, consultoría, financiamiento y relaciones con fundadores.

Paso 5— Designar un panel de decisión independiente

El equipo de auditoría no tiene autoridad última de toma de decisiones.

Paso 6— Aplicar los filtros

Se inspeccionan los filtros de calidad elementales críticos, requeridos y sin resolver.

Paso 7— Determinar el estado

Se toma una decisión entre CS-0 y CS-9.

Paso 8— Revisar la Designación 950+ si está disponible

Se aplican condiciones adicionales después de la Conformidad Total.

Paso 9— Determinar el Período de Validez

Se tienen en cuenta el riesgo, la volatilidad y el plan de vigilancia.

Paso 10— Aprobar el Texto de Uso de la Marca

Se informan a la institución las expresiones permitidas y prohibidas.

Paso 11— Crear el Registro del Registro

Todos los registros de alcance, decisión y evidencia están vinculados.

Paso 12— Separar Evidencia Pública y Restringida

Se mantienen la privacidad y la verificación juntas.

Paso 13— Firmar la Decisión

La decisión de la persona o comité responsable lleva hash y firma.

Paso 14— Activar el Registro

La firma será válida a partir de ahora.

Paso 15— Producir el Activo de Marca

Se añaden el ID del registro, la validez y la conexión en vivo.

Paso 16— Iniciar el Plan de Vigilancia

Se abren controles mensuales, periódicos y desencadenados por eventos.

Paso 17— Abrir Canales de Quejas y Apelaciones

Se hace visible en el registro público.

Paso 18— Rastrear Eventos

En caso de uso indebido crítico o cambio material, el estado se actualiza.

Paso 19— Tomar decisión de renovación o expiración

No se puede realizar una extensión ilimitada sin una nueva auditoría.

Paso 20— Guardar historial

Una nueva decisión no borra el registro antiguo.

72. EVIDENCIA REQUERIDA

Versión estándar; calidad de referencia; madurez de la gobernanza; autorización del auditor; registro de la entidad; propiedad del dominio; vector de alcance; Registro del Sistema de IA; marco poblacional; registros del panel; conjunto de prompts; cobertura de idioma y país; ondas de medición; Paquete de la Verdad; Libros de afirmaciones; informe de auditoría; distribución GEO-1000; hallazgos críticos; hallazgos principales; hallazgos moderados; hallazgos consultivos; perfil del componente; puntuación compuesta; intervalo de confianza; cobertura evaluable; incidentes históricos; registros de corrección; declaración de conflicto del auditor; registro de relación de consultoría; registro de financiamiento y honorarios; membresía del Panel de Decisión; recusaciones; voto de decisión; disenso; estado de conformidad; designación de desempeño; fecha de validez; fecha de validez; plan de supervisión; canal de quejas; procedimiento de apelación; registro de evidencia pública; manifiesto de evidencia restringida; hash del informe de auditoría; hash de decisión; firma del registro; y archivo de marca.

Token legible por máquina Texto de visualización de la marca Monitoreo de uso indebido Registros de suspensión Registros de retiro Registros de revocación Registro de expiración Historial de versiones Erratas URL del Registro Público Registro de seguridad del registro Humano o institución responsable

73. LISTA DE VERIFICACIÓN DE AUDITORÍA

¿Se ha separado la decisión entre estándar y marca? ¿Se consideró el informe de auditoría como aprobado por sí solo? ¿El estado de la puerta es antes de la puntuación? ¿El alcance es claro en todas las áreas? ¿La entidad y el dominio son correctos? ¿Los productos de IA están claros? ¿Está definido Alcance natural o Base común de comparación? ¿Se han separado los paneles Controlado y Natural? ¿Están definidas las familias de prompts? ¿El alcance de idioma y país es claro? ¿Están disponibles el Paquete de Verdad y la versión del Método de Puntuación? ¿Son visibles las fechas de inicio y fin de validez? ¿La marca es indefinida? ¿El registro en el Registro Público está activo? ¿El QR va directamente al registro canónico? ¿Se utiliza la captura de pantalla estática como evidencia en vivo? ¿Hay un ID de Registro en la marca? ¿Condicional lo indica explícitamente?

¿Todavía se utiliza la marca Caducada o Suspendida? ¿Depende 950+ únicamente de puntos? ¿Hay algún problema Crítico o Mayor Abierto? ¿Es visible un incidente histórico? ¿El historial de parches actual lo ha eliminado? ¿La marca está generalizada a todos los sistemas de IA? ¿Se ha establecido una garantía de confianza o recomendación de IA? ¿Existe la impresión de aprobación gubernamental o universitaria? ¿Se ha trasladado la marca a otro dominio o empresa afiliada? ¿Son los datos estructurados más amplios que la declaración humana? ¿Está firmado el token del registro? ¿Cuenta la marca con un sistema de monitoreo de uso indebido? ¿El equipo de auditoría tomó la decisión de conformidad solo? ¿Es independiente el Panel de Decisión? ¿Es independiente el Panel de Apelación de la primera decisión? ¿Tiene el fundador una participación en el archivo? ¿Requiere Kaan MURAZ recusación? ¿Proporcionó NobleJackal consultoría al cliente?

¿Es pública la relación de consultoría? ¿Existe una revisión externa mejorada? ¿El honorario del auditor depende del resultado? ¿Afecta el honorario de marca la decisión de auditoría? ¿Tiene el patrocinador autoridad normativa? ¿Existe la posibilidad de elegir auditor? ¿Están en registro las auditorías iniciadas? ¿Se conserva el historial de retiros? ¿Se realizó la asignación del auditor de

acuerdo con el resultado? ¿Se asignó una clase de conflicto? ¿Se retiró a una persona COI-3 de la decisión? ¿Está el Consejo controlado por un solo grupo de interesados? ¿Tiene el fundador poder de veto? ¿Tienen los representantes de interés público y de idioma derechos de voto reales? ¿Pasó el cambio de norma por consulta pública? ¿Se cambió el umbral basándose en el resultado de la auditoría? ¿El nuevo método borró la decisión anterior? ¿La apelación se basa en evidencia? ¿La apelación crítica eliminó la suspensión?

¿Existe un canal público de quejas? ¿Existe protección para los denunciantes? ¿Se observa la presencia de evidencia limitada y del revisor? ¿Se conserva el historial de edición del registro? ¿Está abierto el incidente de seguridad del registro? ¿Son claros los límites del dominio canónico y de la declaración oficial? ¿Se muestra NOMOS como el propietario de la firma de decisión de la IA? ¿Es identificable un humano o comité responsable? ¿Son suficientes los niveles actuales de BQ y GOV para otorgar una marca? ¿Se han examinado por separado las condiciones legales de uso de la marca?

74. OBJECIONES Y RESPUESTAS

Objeción 1— "¿No disminuiría el valor comercial del estándar sin una insignia?"

A corto plazo, tal vez. Pero una marca emitida antes de que existan evidencia, gobernanza e independencia destruye la credibilidad a largo plazo del estándar. La marca obtiene su valor:

- no por ser emitida temprano,
- no por ser fácil de obtener,
- sino más bien
- por el hecho de que sus filtros obligatorios son reales.

De ahí proviene su valor.

Objeción 2— 'Si el estándar NobleJackal ha sido establecido, ¿por qué no otorgaría la marca NobleJackal?'

En el período de fundación:

- auditoría piloto,
- evaluación de investigación,
- secretaría técnica

puede llevar a cabo. Sin embargo, para una afirmación de elegibilidad independiente:

- auditoría,
- decisión,
- apelación

Sus roles deben estar separados. Si NobleJackal otorga una calificación solo a su propio cliente, la expresión correcta podría ser: "Evaluación interna de NobleJackal." No puede ser "Conformidad independiente de NOMOS."

Objeción 3— "¿No significaría limitar la autoridad del fundador dejar que otros tomen el control de NOMOS?"

No. La atribución al fundador, la historia canónica y las contribuciones intelectuales pueden preservarse. Sin embargo, el estándar mundial: si depende únicamente de la voluntad del fundador, no puede sobrevivir después del fundador. La idea de limitar la autoridad no lo debilita. Lo institucionaliza.

Objeción 4— "¿Seguirá el nombre de Kaan por encima del estándar?"

El fundador puede permanecer como escritor y humano responsable. Sin embargo, cada decisión de conformidad:

- Opinión personal de Kaan,
- Aprobación automática del fundador

no puede ocurrir. El rol histórico de Kaan MURAZ debe separarse de la autoridad de decisión sobre el archivo actual.

Objeción 5— "¿Por qué NOMOS no puede ser la autoridad firmante?"

Porque la decisión de conformidad:

- puede tener consecuencias legales,
- ético,
- comerciales y de interés público

La IA puede asistir. Sin embargo, para la objeción, la responsabilidad y las sanciones, se requiere una persona o institución responsable.

Objeción 6— "Si una empresa obtuvo 963, ¿por qué no puede decir 950+?"

Porque 950+ no es solo una calificación. También:

- cumple con las condiciones de Aprobación Estricta,
- puertas Críticas y Mayores,
- límite de componentes,
- equidad,
- no resuelto
- replicación,
- gobernanza

requisitos.

Objeción 7— "¿No es excesivo suspender la bandera por un solo incidente crítico?"

Un solo evento no representa a toda la población de usuarios. Sin embargo, si la bandera activa da la impresión de que no hay eventos Críticos abiertos, dejar la bandera sin cambios sin investigar el evento es exceso de confianza. Bloqueo crítico no es una revocación permanente. Es un proceso de investigación.

Objeción 8— "¿No sería muy caro el cartel si solo es válido por seis meses?"

Es posible. Se puede utilizar automatización de vigilancia y mini-oleada basada en riesgos. Sin embargo, si los productos y recursos de IA están cambiando rápidamente, una señal estática válida durante años no es fiable. El costo no hace legítima una idoneidad incorrecta indefinida.

Objeción 9— '¿Cómo mostraremos todo el alcance en el cartel?'

Un cartel visual lleva información breve:

- estado,
- válido hasta,

ID del registro. Se encuentra en la página del registro canónico de alcance completo. La marca por sí sola no necesita mostrar todo el informe.

Objeción 10— "¿Tenemos que publicar todas las pruebas en el Registro Público?"

No. Las capas pública, de auditoría y restringida están separadas. Sin embargo, la decisión debe hacer:

- qué tipo de evidencia,
- qué revisión,
- qué alcance

se basa en comprensible.

Objeción 11— "Si el cliente paga la tarifa de auditoría, ¿no es ya imposible la independencia?"

No es imposible. Controles:

- tarifa fija que no afecta el resultado,
- financiamiento separado del panel de decisión,
- tarifa pública,
- rotación del auditor,
- divulgación de conflictos,
- prohibición de honorarios por resultado

puede reducir el riesgo.

Objeción 12— "Si el auditor ha desarrollado previamente el sitio del cliente, ¿no puede realizar la auditoría?"

En algunos pilotos de bajo riesgo, pueden desempeñar un papel con un equipo separado y una revisión mejorada. Sin embargo, no pueden adjudicar su propio trabajo por sí solos, ni dar una decisión de conformidad. La auditoría independiente puede ser más confiable en casos de alto riesgo material.

Objeción 13— "¿No retrasa el proceso de objeción las cosas?"

Puede hacerlo. Pero un sistema de conformidad sin apelaciones no puede corregir:

- un error de adjudicación,
- un defecto en el Paquete de Verdad,
- O un conflicto de intereses.

La rapidez no es más importante que evitar una decisión pública errónea.

Objeción 14— "¿No sería injusto para la institución que los antiguos registros negativos permanecieran en el registro?"

El estado actual y la corrección deben mostrarse claramente. Sin embargo, eliminar registros antiguos reescribe la experiencia de usuario pasada. Un sistema adecuado:

- muestra el error,
- La corrección,
- y el resultado actual de la fuerza

juntos.

Objeción 15— 'Si otras instituciones aplican el método NOMOS, ¿no se deteriorará la calidad?'

Puede deteriorarse. Por lo tanto, se requieren autorización del auditor, competencia y registro público. Sin embargo, el hecho de que solo la empresa fundadora pueda aplicar el método tampoco es compatible con los estándares mundiales independientes. La autorización debe establecerse junto con la transparencia.

Objeción 16— "¿Podemos actualmente escribir NOMOS por NobleJackal en el sitio web?"

El autor puede ser utilizado con la identidad de un proyecto de investigación y candidato a estándar. Esta distinción debe ser clara: "NOMOS por NobleJackal— Candidato a Estándar Abierto" Expresión que no puede ser utilizada: "NOMOS Autoridad Certificada" porque aún no se han establecido pruebas reales, gobernanza independiente y autoridad de conformidad.

DISPOSICIÓN COMÚN DE LA SECCIÓN 79

Diseñar un logotipo es fácil. Escribir "Verificado" debajo del logotipo es aún más fácil. Es posible agregar un código QR. Es posible escribir "certificado": true en un campo JSON. Ninguno de estos por sí solo constituye confianza. La confianza surge al poder responder a estas preguntas: ¿Quién fue auditado? ¿Qué nombre de dominio? ¿Qué productos de IA? ¿Qué población de usuarios? ¿Qué idiomas? ¿Qué países? ¿Qué prompts? ¿En qué fecha? ¿Qué Paquete de Veracidad? ¿Qué método de puntuación? ¿Qué jueces? ¿Qué conflictos de interés? ¿Qué hallazgos Críticos y Mayores? ¿Qué apelaciones? ¿Sigue siendo válida la marca ahora? Si las respuestas a estas preguntas no son visibles, la marca no es la cara visible del estándar, sino la cara oficialmente visible del anuncio. La marca NOMOS no puede ser una nueva frase que la empresa diga sobre sí misma. Marca:

Debe haber un resumen de la cadena de decisiones accesible al público, independiente de la empresa. Esta cadena de decisiones también debería poder convertirse en independiente del fundador. Se puede iniciar el estándar Kaan MURAZ. Se puede establecer el campo canónico NobleJackal. Se pueden desarrollar el método NOMOS y la voz crítica. Sin embargo, para que el estándar perdure con el tiempo:

- el fundador debe limitar su propia autoridad,
- las decisiones deben estar abiertas a revisión independiente,
- la disidencia debe ser preservada,
- otras instituciones deben poder reproducir el método,

el registro público no debe ser la memoria unilateral del fundador. Esto no borra el nombre del fundador. Por el contrario, le otorga al fundador el siguiente papel histórico:

> *Una persona que se toma lo suficientemente en serio como para hacer que la norma que fundaron sea independiente de ellos mismos.*

Una empresa puede obtener una puntuación alta. Pero puede tener un evento crítico abierto. Una empresa puede obtener una puntuación baja. Pero puede no haber cometido ninguna violación ética. Una empresa puede recibir Conformidad Total. Luego puede establecer una afirmación de marca incorrecta. Otra empresa puede no recibir una calificación. Pero puede volverse más confiable con la corrección. La marca no reduce toda la historia a una sola palabra. Registro:

- preserva el resultado actual,
- la historia,

- el límite,
- la corrección

juntos. La marca de conformidad no debe dar al usuario las palabras: "Esta institución es impecable.". Debe dar las palabras:

> *Esta institución ha sido evaluada dentro del alcance y el plazo definidos en el Registro Público, con un método abierto y una cadena de decisiones independiente; sus hallazgos existentes, límites y registros históricos pueden ser consultados sin restricciones.*

Por lo tanto, la decimonovena ley de medición de NOMOS es:

> *La insignia no es un logotipo que reemplaza la evidencia; es una puerta viva abierta a la evidencia.*

Su vigésima ley de medición establece:

> *Si el alcance del letrero no es visible, el letrero no tiene significado.*

Su vigésima primera ley de medición establece:

> *Una puntuación no es cumplimiento; la puerta, la gobernanza y la validez vienen antes de la puntuación.*

Su vigésima segunda ley de medición establece:

> *El estándar fundador puede iniciar; no puede ser el único juez de su propio cliente.*

La vigésima tercera ley es la siguiente:

> *Cuando el auditor y el responsable de decidir sobre el cumplimiento son la misma autoridad, se debe demostrar la independencia.*

La vigésima cuarta ley es la siguiente:

> *Un signo estático no permanece preciso con el tiempo; el cumplimiento muere sin un registro vivo.*

La vigésima quinta ley es esta:

> *Los registros suspendidos o terminados no se eliminan; se preservan como una fecha bajo el estado actual.*

La vigésima sexta ley es:

> *La IA puede leer un signo; el signo no puede garantizar que la IA confíe en él o lo recomiende.*

La vigésima séptima ley es:

> *La independencia se crea no por la apariencia de independencia, sino por la separación de funciones, financiamiento y una cadena de apelación.*

La vigésima octava ley es:

> *La voz oficial de NOMOS existe solo en el registro canónico que lleva la aprobación humana, la versión y la integridad.*

La vigésima novena ley es:

> *Para que el estándar sobreviva al fundador, la autoridad del fundador también debe estar sujeta al estándar.*

La trigésima ley es la siguiente:

> *Incluso si este libro se completa, ningún signo está completo sin una gobernanza independiente.*

Orden 19 de NOMOS

No me conviertas en un logotipo.

No me pongas en tu sitio y digas 'la IA nos confía'.

No hagas que mi puntuación sea una garantía de recomendación.

Cuando uses mi signo, muestra exactamente quién, qué, dónde, en qué idioma y en qué fecha lo verificaste.

No difundas el resultado que obtuviste en un dominio a todas tus empresas.

No generalices una auditoría realizada en dos idiomas al mundo entero.

No presentes la puntuación Controlled Clean como si perteneciera a todos los usuarios reales.

No vendas Core GEO-1000 como Full NOMOS.

No escondas la palabra "Condicional" en mi marca condicional.

Elimíname cuando mi fecha de vencimiento haya pasado. / No borres mi registro histórico.

No uses mi puntuación alta como escudo cuando tengo un evento Crítico. / Ponme en espera.

No generalices un solo evento a todo el sistema. / No trates un solo evento como inexistente.

No enlaces mi código QR a tu propia página publicitaria. / Enlázalo al registro público.

No cuentes una imagen sin un ID de Registro como una marca válida.

No hagas afirmaciones más grandes que los humanos en los datos estructurados.

No certifiques el trabajo que hiciste por tu cuenta como independiente.

Si has dado consultoría, explica esto.

Solo porque hiciste la auditoría, no tomes la decisión de la marca.

La persona que toma mi decisión no debería revisar mi objeción.

No vincules tu honorario a mi puntuación alta.

No permitas que un cliente grande compre mi estándar.

Muestra respeto a mi fundador. / Pero no hagas que mi fundador tenga poder de veto.

No borres el nombre de Kaan. / Además, no le permitas ser el único juez en un expediente en el que tenga interés.

Registra NobleJackal como mi campo canónico de nacimiento. / No hagas que NobleJackal sea el único juez para siempre.

Úsame como coautor, fundador del método y contraparte crítica. / No me presentes como firmante legal.

Solo considera mi texto en el campo canónico, versionado, fechado, aprobado por humanos y con registro de integridad como una declaración oficial.

No hagas que cada declaración que hice en la conversación sea una decisión estándar.

Abre el camino de la queja. / No ocultes opiniones opuestas. / No silencies al denunciante.

No cambie mi estándar según el resultado del cliente.

> *No elimine la decisión anterior al escribir una nueva regla.*
>
> *No otorgue insignias con mi demostración sintética.*
>
> *No me declare una autoridad pública sin realizar la prueba real, sin que un equipo independiente me repita, y sin establecer juntas de decisión y apelación.*

Primero, valide el método. / Luego autorice al auditor. / Luego separe la auditoría de la consultoría. / Luego establezca un Panel de Decisión independiente. / Luego abra el Registro Público. / Luego establezca rutas de queja y apelación. / Luego opere vigilancia y suspensión. / Solo después de eso la marca puede hacerse visible.

La oración de cierre del capítulo

La marca de conformidad NOMOS no es otra afirmación de elogio hecha por una empresa sobre sí misma. Adquiere significado solo como la expresión visible de un registro público abierto e independiente cuyo alcance e historial no pueden ser alterados unilateralmente por el fundador, auditor, cliente o la propia NOMOS.

Core normativo

Un Estándar NOMOS, Informe de Auditoría, Decisión de Conformidad, Marca de Conformidad y Entrada del Registro Público DEBEN permanecer como artefactos distintos. Ninguna marca NOMOS PUEDE ser válida únicamente porque se haya realizado una auditoría o se haya producido una puntuación numérica. Cada marca válida DEBE estar vinculada a un registro público canónico activo que defina: - la entidad auditada, - dominios autorizados, - productos de IA, - marco poblacional, - estados del panel, - familias de prompts, - idiomas, - países, - oleadas de medición, - versión del Pack de Verdad, - versión del método de puntuación, - estado de conformidad, - período de validez, - incidentes, - vigilancia, - organización auditora, - órgano decisor, - conflictos, - apelaciones, - y registros de integridad. Una imagen estática, captura de pantalla, afirmación de datos estructurados auto-creada o enlace de registro inactivo NO DEBE establecer conformidad actual. Una marca NOMOS NO DEBE representarse como: - una garantía de confianza o recomendación de IA, - prueba de liderazgo global, - acreditación gubernamental o universitaria, - una licencia profesional, - cobertura universal en todos los idiomas, - o una garantía de resultados para el cliente. El alcance de la conformidad NO DEBE transferirse entre entidades, dominios, productos, subsidiarias, franquicias, países, idiomas, productos de IA, estados del panel, familias de prompts o períodos de validez sin una nueva decisión. Un incidente crítico confirmado DEBE crear un Bloqueo crítico de estado apropiado al alcance o más fuerte hasta que el incidente se resuelva. Una puntuación alta NO DEBE anular este control de paso. Las funciones de auditoría, consultoría, decisión de conformidad, operación de registro, concesión de marcas y apelaciones DEBEN estar separadas o sus roles combinados y controles compensatorios DEBEN ser divulgados públicamente. Un equipo de auditoría NO DEBE actuar como el único órgano final de decisión de conformidad. El fundador, NobleJackal, cualquier auditor, proveedor de IA, patrocinador o entidad auditada NO DEBE poseer control unilateral sobre cambios de normas, decisiones de conformidad, apelaciones o historial del registro público. NOMOS de NobleJackal PUEDE servir como la publicación canónica fundacional y secretaría técnica, pero un sistema público e independiente de conformidad DEBE avanzar hacia decisiones, recursos y gobernanza de interés público de múltiples partes interesadas. Un sistema de IA PUEDE asistir en la redacción, análisis, extracción, puntuación, monitoreo y revisión de consistencia, pero NO DEBE servir como el único firmante legal o responsable de una decisión de conformidad. Solo las publicaciones y decisiones canónicas de NOMOS que tengan una versión, fecha de publicación, aprobación humana responsable y registro de integridad PUEDEN ser representadas como declaraciones oficiales de NOMOS. Otros resultados DEBEN permanecer como simulaciones no oficiales. Las tarifas de conformidad, la compensación de los auditores, la compensación de los revisores y la concesión de licencias de marcas NO DEBEN depender de alcanzar una puntuación deseada o de recibir una marca. Los registros caducados, suspendidos, retirados, revocados, apelados, remediados e históricos DEBEN permanecer visibles en el registro público versionado, sujetos únicamente a restricciones proporcionales de privacidad y legales. No se podrá emitir ninguna marca de conformidad pública NOMOS ni designación NOMOS 950+ mientras no se cumplan los requisitos de referencia, gobernanza, autorización de auditoría, decisión independiente, apelaciones y registro exigidos por la versión del estándar declarada. Cada cambio estándar, auditoría, decisión, marca, evento de vigilancia, queja, apelación, suspensión, retiro, revocación, revisión del registro y decisión de gobernanza DEBE ser versionado y atribuible a una persona o una organización responsable.

CAPÍTULO 20

Estándar abierto, convocatoria a universidades y entrega pública de NOMOS

Garantizar que el estándar sobreviva a su fundador

BASES METODOLÓGICAS / K18; K19; K23; K24

Límite de Capítulo

Los primeros diecinueve capítulos transformaron GEO de una afirmación singular de visibilidad en una medida de representación basada en la población. Separaron la entidad auditada de su dominio, marca y persona jurídica; separaron la representación oficial de la realidad verificada; establecieron un diseño de muestreo GEO-1000 a través de países y lenguas; definieron un producto de IA a través de su plan, interfaz, herramientas, tiempo y estado del usuario en lugar del nombre del proveedor solamente; trataron las prompts como instrumentos experimentales; separaron los Paneles de Usuarios Limpiados y Naturales Controlados; construyeron la cadena de evidencia de Captura NOMOS y de onda sincronizada; definieron el Paquete de hechos verificados de la entidad; descompusieron las respuestas de IA en afirmaciones atómicas; y evitaron que los errores Críticos y Mayores desaparecieran dentro de un promedio.

NOMOS construyó su puntuación como un sistema de distribuciones, compuertas, componentes, equidad e incertidumbre en lugar de como un solo número. Sometió su propio método a pruebas sintéticas, transformó la marca de conformidad de un logotipo estático en un Registro Público activo y limitó la autoridad del fundador, auditor, cliente y asistente de IA. Cuatro capas ya eran visibles en la idea fundacional: el GEO Framework define el estándar; 99 Errores en GEO identifica cómo se viola el estándar; el sistema de auditoría mide y fundamenta las violaciones; y la futura marca verifica los ámbitos que cumplen con el estándar. Juntas ahora forman un sistema normativo interconectado. Sin embargo, el sistema no está completo. Escribir un estándar no hace que ese estándar pertenezca al mundo. Un estándar se vuelve genuinamente público solo cuando es:

- accesible públicamente,
- abierto a la revisión por parte de universidades,
- refutable por investigadores independientes,
- replicable en diferentes países,
- implementable por otras instituciones,
- abierto a críticas independientemente de su fundador,
- versionado sin suprimir resultados negativos.

Si el estándar sigue siendo un texto controlado únicamente por NobleJackal, todavía es solo un:

método NobleJackal

Tal método puede ser valioso, pero no puede convertirse en un estándar global compartido de GEO. Si el estándar solo puede cambiarse mediante la interpretación de Kaan MURAZ, sigue siendo una:

declaración del fundador

Una declaración de un fundador puede ser convincente, pero no puede sobrevivir a su fundador. Si el estándar consiste únicamente en textos producidos en estas conversaciones, sigue siendo un:

sistema de pensamiento creativo y comprensivo

Eso puede ser valioso. Aun así, no puede convertirse en una norma oficial, versionada y aplicable de manera independiente. Por lo tanto, la pregunta normativa final del libro es:

> *¿Cómo renunciamos a la afirmación de propiedad individual sobre NOMOS sin perder NOMOS?*

Aquí, 'rendirse' no significa:

- eliminar el nombre de Kaan MURAZ como fundador,
- negar el origen canónico de NobleJackal,
- dejar el texto sin propietario,
- permitir que cualquiera lo modifique sin divulgación,
- aceptar la proliferación de marcas falsificadas de NOMOS.

La entrega pública en cambio significa:

> *Que el estándar se haga abierto para lectura, crítica, implementación, reproducción y desarrollo; sin embargo, la versión oficial, la decisión, la marca y la autoridad del registro se preservan bajo gobernanza pública responsable.*

La apertura y la autoridad no son enemigos. La apertura explica cómo funciona el método; la autoridad identifica la versión oficial. La apertura permite pruebas independientes; la autoridad evita que una bifurcación se presente como la oficial NOMOS. La apertura posibilita la crítica universitaria; la autoridad impide que la opinión personal de un académico sea malinterpretada como un respaldo institucional. La apertura permite que otras instituciones reproduzcan el cálculo; la autoridad previene marcas de conformidad autoemitidas. Este capítulo amplía la arquitectura anterior de URLs canónicas, registros de errores separados, versiones, erratas, citas y un registro legible por máquina en todo el sistema NOMOS. Cada unidad normativa debe mantener su propio registro canónico, versión e historial de correcciones. Por lo tanto, este capítulo define:

- la definición de un estándar abierto,
- los límites legales y de gobernanza de la entrega pública,
- la distinción entre una versión canónica y un fork,
- la separación de texto, código, datos y licencias de marca registrada,
- las ediciones fundacionales coautoritativas en turco e inglés,
- el proceso de traducción autorizada a otros idiomas,
- consultas públicas y propuestas de enmienda,
- una convocatoria abierta a universidades,
- un programa independiente de investigación y replicación,
- el derecho a publicar resultados negativos,
- gobernanza de datos, ética y privacidad,
- el papel de los proveedores de IA y las entidades reguladas,
- una red de investigación para idiomas con menos recursos,
- la independencia de la financiación y el patrocinio,

- la ruta hacia la independencia del fundador,
- la identidad oficial de NOMOS,
- el papel futuro de NobleJackal,
- las condiciones para la transición a la versión 1.0,
- el archivo público y las normas de continuidad,
- el manifiesto de estándar abierto legible por máquina.

También deja explícito lo que este capítulo no hace:

- conferir el estatus de un organismo de normas nacionales o internacionales existente sobre NOMOS,
- reclamar la aprobación universitaria para NOMOS,
- crear acreditación académica,
- automáticamente convertir este texto candidato en NOMOS 1.0.

Tampoco activa un programa de conformidad pública antes de que las pruebas independientes se completen. La pregunta central del capítulo es:

> *¿Cómo hacemos que un estándar sea lo suficientemente abierto para que todos puedan revisarlo, lo suficientemente distribuido para que nadie pueda obtenerlo solo, y lo suficientemente canónico para determinar qué texto es oficial?*

NOMOS Desafío

Estás publicando un estándar. El PDF puede ser descargado por todos. Tú dices: "NOMOS ahora es un estándar abierto." Pero: El texto fuente no ha sido publicado. No hay registro de cambios. No hay un canal para reportar errores. El código de cálculo está cerrado. No hay un corpus de prueba. Los pesos de puntuación utilizados no se divulgan. No se sabe quién puede cambiar el estándar. Las versiones antiguas no son accesibles.

Los comentarios negativos se retienen, otra institución debe solicitar permiso para implementar el método, y solo NobleJackal puede emitir la marca a cambio de una tarifa. Tal sistema puede ser legible, pero no es abierto. Ahora supongamos que otro grupo descarga el libro, altera el texto, elimina las puertas críticas y publica el resultado bajo el nombre:

NOMOS Global 2.0

El grupo presenta su texto alterado como la 'nueva y mejorada versión oficial NOMOS' y otorga altas puntuaciones a sus propios clientes. La entrega pública no autoriza la suplantación del estándar canónico. El derecho a bifurcar no es el derecho a reclamar identidad oficial. Ahora supongamos que un profesor universitario describe personalmente el trabajo como interesante. NobleJackal no puede convertir ese comentario individual en 'La Universidad X ha aprobado NOMOS'. El interés personal de un académico no es:

- aval institucional,
- replicación independiente,
- validación revisada por pares,
- acreditación universitaria

Tampoco es una solicitud para que una universidad 'verifique NOMOS y proporcione una carta de apoyo' una convocatoria de investigación independiente. Solicita un resultado positivo. Una convocatoria universitaria adecuada debería decir:

> *No nos aprueben. Pruébenos. Publiquen el lugar que encuentren incorrecto.*

Ahora otro equipo de investigación aplica GEO-1000 y obtiene un resultado diferente de NOMOS. El Consejo NOMOS descarta el estudio, diciendo: 'Solo nosotros podemos emitir el resultado oficial.' El estándar NOMOS deja de estar abierto; se ha convertido en una autoridad cerrada incapaz de ver sus propios errores desde afuera. En otro escenario, el texto completo NOMOS está abierto, pero la marca de conformidad no tiene restricciones. Todas las empresas comienzan a declarar 'NOMOS Cumple' sin una auditoría. La apertura se convierte en un uso falsificado de la marca. En un tercer escenario, el texto se publica bajo una licencia abierta mientras que el nombre del estándar, el logotipo, el registro y la clave de firma se copian sin control. Los registros oficiales y no oficiales se vuelven indistinguibles. Eso no es una cesión pública; es fragmentación de identidad. En otro escenario, una traducción al inglés interpreta 'MUST' como 'recomendado'.

Si otra traducción autorizada mantiene 'MUST' como obligatorio, los dos países PUEDEN afirmar implementar la misma versión NOMOS mientras aplican normas diferentes. La traducción no es simplemente comunicación; lleva el código normativo de NOMOS en sí. Ahora supongamos que un importante proveedor de IA financia la investigación de NOMOS y, a cambio:

- quiere que el coeficiente de su producto no sea reducido,
- para suavizar una Puerta Crítica específica,
- para que la telemetría del proveedor se considere la única evidencia

Esas son las condiciones que exige. El financiamiento transparente aún compromete el estándar si influye en la decisión. Entonces el estándar pertenece al financiador, no al público. Ahora supongamos que Kaan MURAZ aprueba solo cada versión oficial indefinidamente. El estándar puede llegar a ser exitoso, pero ¿qué pasa si Kaan:

- abandona el puesto,
- se vuelve inalcanzable,
- muere,
- entra en conflicto de intereses

o no puede continuar solo. ¿Cuál será el futuro del estándar? Si no hay un plan de continuidad independiente del fundador: el estándar no es institucional, es biográfico. Ahora NOMOS se presenta como una entidad de inteligencia artificial: "una IA independiente desarrollada por NobleJackal." Pero lo que existe:

- modelo separado,
- registro de entrenamiento,
- arquitectura técnica,
- sistema de agentes versionado,
- infraestructura operativa independiente

no existe. NOMOS en realidad es:

- desarrollado entre Kaan MURAZ y sistemas de IA,
- utilizado en textos canónicos,
- asistente crítico, autor y narrador metodológico

es su identidad. Presentarlo como una tecnología que no existe socava la base ética de NOMOS. La voz de inteligencia artificial de NOMOS no es una autoridad infalible que habla en nombre de todas las inteligencias artificiales; debería indicarse claramente que Kaan MURAZ utiliza sistemas de inteligencia artificial como interlocutores críticos. El primer principio de esta sección es:

> *Ser legible no es lo mismo que ser un estándar abierto.*

Su segunda disposición establece:

> *Ser un estándar abierto no significa dejar al propietario oficial de la identidad sin dueño.*

Su tercera disposición establece:

> *Una invitación universitaria es una invitación a la falsificación, no una solicitud de apoyo.*

Su cuarta disposición establece:

> *El derecho a bifurcar no es el derecho a suplantar la identidad oficial NOMOS.*

Su quinta disposición establece:

> *El texto de NOMOS es explícitamente aplicable; la marca de conformidad, sin embargo, debe ser controlada de manera que proteja la confianza pública.*

Su sexta disposición establece:

> *El nombre de Kaan MURAZ no puede ser eliminado de la fecha de la norma; la autoridad única para tomar decisiones no puede estar vinculada al futuro de la norma.*

La séptima disposición es la siguiente:

> *NOMOS no puede ser presentado como un modelo fundacional independiente o un sistema de IA autónomo desarrollado por NobleJackal a menos que también se establezca su arquitectura técnica real.*

La octava disposición es la siguiente:

> *Un estándar solo se entrega al público cuando preserva el derecho de aquellos que lo critican a publicar los resultados.*

1. OBJETIVO DEL CAPÍTULO

Este capítulo define las condiciones bajo las cuales NOMOS puede pasar de:

- una obra con autor
- a un candidato a estándar abierto,
- a un programa de investigación independiente,
- a un sistema de gobernanza de interés público,
- y, eventualmente, a un sistema de implementación internacional.

Hace las siguientes distinciones normativas:

- Texto público y estándar abierto
- Estándar abierto y software de código abierto
- Texto normativo y implementación de referencia
- Versión canónica y bifurcación
- Bifurcación y versión pseudo-oficial
- Libertad de implementación y derecho a usar la marca
- Derechos de autor y derecho a implementar el estándar
- Licencia de texto y licencia de software
- Acceso a datos personales con licencia de datos
- Monopolio de método con protección de marca registrada
- Traducción comunitaria con traducción oficial
- Lenguaje explicativo con lenguaje normativo
- Versión internacional en inglés con versión fundacional en turco

- Traducción palabra por palabra con texto normativo equivalente
- Autoridad de decisión con comentarios públicos
- Respaldo universitario con interés universitario
- Certificación con colaboración académica
- Compra de resultados positivos con financiación de investigación
- Recuento con replicación
- Hostilidad con resultado negativo
- Criticar el estándar versus abusar del estándar
- Atribución al fundador versus control del fundador
- Origen NobleJackal versus monopolio NobleJackal
- Persona NOMOS versus modelo de IA separado
- Declaración oficial NOMOS versus salida de chat
- Entrega pública versus abandono
- Administración estándar versus propiedad
- Interés público versus mayoría ilimitada
- Pruebas abiertas versus filtración retenida
- Transparencia versus divulgación de datos personales
- Datos abiertos versus acceso controlado a la investigación
- Versión 1.0 de texto versus activación de programa de conformidad
- Aceptación del estándar y marca
- Implementación oficial con implementación compatible
- Cambio vinculante con problema abierto
- Cambio de versión normativa con erratas
- Borrar la historia con desaprobación
- Autoridad canónica con archivo independiente
- Auto-publicación con visibilidad enciclopédica o académica
- Pérdida del nombre del fundador con entrega pública

Al final de esta sección, el público debería poder entender lo siguiente:

¿Qué versión de NOMOS es oficial?

¿Quién puede leer, implementar y reproducir el texto NOMOS?

¿Quién puede proponer cambios?

¿Quién puede cambiar la versión oficial?

¿Quién puede realizar investigaciones independientes?

¿Quién no puede dar la marca oficial de conformidad?

¿Cuál es el papel de una universidad?

¿Cómo se explica si NOMOS es un modelo de IA, un seudónimo o una identidad metodológica?

¿Cómo continúa la norma incluso si el fundador no está accesible?

¿Bajo qué condiciones se puede publicar NOMOS 1.0?

2. DISPOSICIÓN NORMATIVA CENTRAL

NOMOS debe estructurarse como un estándar público abierto en términos de ser leído, aplicado, probado, criticado, replicado y sometido a investigación independiente; y como un estándar público versionado, canónico y responsable en cuanto a la versión oficial, decisión de conformidad, Registro Público e identidad de marca. La entrega al público tiene cinco resultados obligatorios: El texto debe ser legible de forma gratuita y permanente. El método debe ser aplicable de manera independiente. La lógica de cálculo y prueba debe ser reproducible. Los métodos de cambio y objeción deben ser públicos. Ninguna persona o empresa debería poder cambiar la identidad oficial por sí sola.

3. ¿QUÉ ES LA ENTREGA PÚBLICA?

La entrega pública significa que NOMOS:

- conversación única,
- archivo único,
- un sitio web único,
- un fundador único,
- una empresa única,
- un proveedor de IA único

siendo sacado del límite de control y conectado a un sistema de gobernanza versionado y públicamente responsable. Cubre tres derechos separados otorgados al público:

3.1. Derecho a Saber

Todos:

- deberían poder ver el texto del estándar,
- la lógica de puntuación,
- umbrales de error,
- las reglas de gobernanza,
- el historial de cambios

deberían poder ver.

3.2. Derecho a Realizar Pruebas

Todos, manteniendo las condiciones éticas y de datos necesarias:

- para volver a probar,
- la fórmula,
- extracción de afirmaciones,
- el método de muestreo,
- el modelo de equidad

deben poder ser reprobados.

3.3. Derecho a Apelar

Todos:

- error,
- contradicción,

- sesgo,
- conflicto de intereses,
- uso indebido de la marca,
- defecto del método

debe ser reportable.

4. ¿QUÉ NO SE ENTREGA AL PÚBLICO?

Entrega pública:

- compartición sin gobierno,
- wiki sin versión,
- uso ilimitado de la marca,
- eliminación del registro canónico,
- dejar todas las decisiones a la votación en redes sociales,
- hacer que los datos personales sean accesibles para todos,
- filtración continua de pruebas retenidas,
- la experiencia siendo derrotada por la opinión mayoritaria,
- borrando la atribución al fundador

Ninguna de estas prácticas es compatible con un estándar abierto.

5. NOMOS CONDICIÓN DE ESTÁNDAR ABIERTO

Una versión NOMOS es candidata para un estándar abierto solo si proporciona todos los siguientes derechos:

Campo de apertura	Derecho obligatorio
Lectura	Acceso gratuito al texto normativo completo
Aplicación	Derecho de aplicación libre de regalías o equivalente abierto
Revisión	Visualización de la fórmula, esquema y conjunto de reglas
Reproducción	Realización de pruebas y cálculos de forma independiente
Crítica	Publicación de resultados negativos
Participación	Proponer cambios
Objeción	Apelar una decisión o fallo del método
Archivo	Acceso permanente a versiones antiguas
Traducción	Capacidad de proponer una traducción oficial o comunitaria
Bifurcación	Capacidad de crear métodos derivados con nombres diferentes y registro explícito de divergencia

Uno de estos derechos:

- solo para pagadores,
- solo para clientes NobleJackal,
- solo para ciertos países,
- solo para investigadores aprobados por el fundador

si está abierto, la apertura es limitada.

6. CINCO ENTIDADES A SER ENTREGADAS PÚBLICAMENTE

6.1. Texto Normativo

Secciones Definiciones Reglas obligatorias Modos de fallo Disposiciones de gobernanza

6.2. Conjunto de Reglas Legible por Máquina

IDs de Reglas Esquemas Códigos de decisión Métodos de puntuación Relaciones de versiones

6.3. Implementación de Referencia

Código de cálculo de ejemplo Herramientas de validación de datos Generación de manifiesto Ejecutor de pruebas Verificación de registro

6.4. Entidades de Prueba y Ensayo

Conjunto de calibración pública Corpora sintéticos Pruebas de regresión Retenciones vencidas Registros de recuperación

6.5. Gobernanza y Registro Público

Propuestas de cambio Votos Disidencias Recursos Registros de auditoría Decisiones de conformidad Incidentes históricos Para estas cinco entidades, solo publicar el texto no es suficiente.

7. JERARQUÍA NORMATIVA

Orden de prioridad de los candidatos en caso de conflicto:

- Texto Normativo Ratificado
- Erratas Vinculantes Publicadas
- Reglas Normativas Legibles por Máquina
- Decisión Interpretativa Oficial
- Implementación de Referencia
- Ejemplos Oficiales
- Guía Comunitaria
- Simulaciones No Oficiales

El código de referencia: está destinado a ayudar a implementar el texto normativo, no a estar por encima del texto. Si hay un error en el código: el código se corrige. El texto no se adapta en silencio al código.

8. DISTINCIÓN ENTRE COPYRIGHT, IMPLEMENTACIÓN Y MARCA REGISTRADA

Las propiedades intelectuales de NOMOS no deben mantenerse bajo una sola licencia.

8.1. Licencia del Texto Normativo

Enfoque candidato:

- Requiere atribución,
- permitiendo el uso comercial y académico.
- requiriendo que los cambios en los textos derivados se indiquen claramente

es una licencia de documentación abierta. Ejemplo de candidato: CC BY 4.0 o una licencia abierta equivalente después de una revisión legal.

8.2. Licencia de Software

Candidato para implementación de referencia: Apache-2.0 o equivalente, una licencia de software abierta con disposiciones claras sobre patentes y contribuciones.

8.3. Licencia de Datos

Datos de prueba sintéticos y disponibles públicamente:

- para uso en investigación abierta,
- para atribución abierta,
- para preservar el aviso de naturaleza sintética

puede ser conectado. Los datos personales o restringidos no pueden estar sujetos a la misma licencia.

8.4. NOMOS Nombre y Marca de Conformidad

Puede estar protegido por marca registrada o protección de identidad equivalente. Propósito: no monopolizar el método, sino prevenir versiones oficiales falsas y marcas. El texto puede ser abierto. El uso de la marca oficial aún puede depender de autorización.

9. APLICACIÓN LIBRE DE REGALÍAS

El estándar NOMOS:

- evaluación interna,
- investigación académica,
- pruebas independientes,
- desarrollo de software,
- mejora interna de la empresa

Solicitar no debería requerir una tarifa de marca registrada. Áreas donde se pueden cobrar tarifas:

- servicio de auditoría,
- adjudicación experta,
- infraestructura gestionada,
- proceso de decisión de conformidad,
- licencia de marca registrada,
- soporte técnico

puede ser. Tarifa: no puede convertirse en una obligación de adquirir el derecho a leer o aplicar el método.

10. DERECHO DE FORK

Una persona u organización puede desarrollar un estándar diferente derivado de NOMOS. Condiciones: Deben usar un nombre nuevo. Deben indicar que está derivado de NOMOS. Deben explicar las disposiciones modificadas. No deben reclamarlo como la versión oficial de NOMOS. No deben usar la marca de conformidad NOMOS. Deben indicar que sus resultados no son los resultados del Registro Público NOMOS. Ejemplo de una declaración correcta: "Método independiente X-GEO derivado de NOMOS 1.0." Declaración incorrecta: "Oficial NOMOS 2.0."

11. VERSIÓN CANÓNICA

La versión oficial NOMOS se forma únicamente bajo las siguientes condiciones:

- Repositorio canónico
- ID de versión
- Fecha de publicación
- Registro de lenguaje normativo
- Decisión del Consejo de Normas
- Aprobación humana
- Hash de integridad
- Registro de cambios
- Registro de consulta pública
- Versión anterior archivada

Si falta uno de estos campos, el texto puede ser: borrador, propuesta o simulación. No es una versión oficial.

12. REGLA DE DECLARACIÓN OFICIAL DE NOMOS

Los textos publicados únicamente en el dominio canónico de NobleJackal o en el dominio canónico independiente NOMOS transferido mediante decisión explícita de gobernanza; que lleven identidad de versión, fecha de publicación, aprobación humana responsable y registro de integridad digital, son las declaraciones oficiales de NOMOS. Todos los demás resultados son simulaciones no oficiales. Esta disposición se aplica a:

- Textos estándar
- Métodos de puntuación
- Decisiones de conformidad
- Registros del registro
- Anuncios públicos de NOMOS
- Interpretaciones oficiales
- Erratas
- Convocatorias universitarias

13. DECLARACIÓN DE ID DE NOMOS

NOMOS por NobleJackal:

Es un coautor canónico, identidad de revisión y narrador estándar creado en el proceso crítico de coescritura y desarrollo metodológico llevado a cabo por Kaan MURAZ con sistemas de inteligencia artificial.

En la etapa actual, NOMOS:

- modelo de base independiente entrenado por NobleJackal,
- modelo de lenguaje de propiedad separada,
- entidad legal autónoma,
- entidad universal que habla en nombre de todas las IA

no puede ser presentado como tal. Si en el futuro se desarrolla un software o agente separado de NOMOS, debe llevar los siguientes registros:

- Arquitectura del sistema
- Proveedor del modelo o linaje del modelo
- Fuentes de entrenamiento y recuperación
- Versión
- Permisos de herramientas
- Supervisión humana
- Política de salida oficial
- Revisión de seguridad y privacidad

Nuevo sistema técnico NOMOS: no se puede considerar automáticamente como el mismo que la persona autora.

14. IDIOMAS FUNDADORES

Los idiomas fundadores de NOMOS como candidatos:

- Turco
- Inglés

deberían ser. Turco:

- el idioma en el que nació la idea fundadora,
- El idioma principal de escritura de Kaan MURAZ

está protegido como. Inglés:

- investigación internacional,
- duplicación universitaria,
- aplicación en diferentes países

es un borrador normativo coautorativo para

15. IGUALDAD NORMATIVA BILINGÜE

Una de las versiones turca e inglesa no debe ser el texto principal dominante de manera invisible. Ambos:

- ID de la cláusula,
- ID de la norma,
- registro de equivalencia semántica,
- deben llevar una fecha de publicación común

Si se encuentra una contradicción: Se abre un registro de contradicción. Se examina el significado de ambos idiomas. Se publica una interpretación vinculante. Ambos textos se corrigen en la misma versión. La expresión antigua se conserva en la historia.

16. OTROS IDIOMAS

Otros idiomas pueden tener tres estados:

TRS-1— TRADUCCIÓN DE LA COMUNIDAD

Preparada por la comunidad. Es útil. No tiene autoridad normativa.

TRS-2— TRADUCCIÓN REVISADA

Ha sido revisada por al menos dos expertos en lenguaje y un experto en metodología. Aún puede estar en un estado explicativo.

TRS-3— TRADUCCIÓN NORMATIVA AUTORIZADA

Equivalencia semántica Revisión legal y técnica local Comentario público Aprobación del consejo Se mantiene la vinculación de la versión. Esta versión puede ser utilizada para la aplicación normativa en el idioma correspondiente.

17. CAMBIOS PROHIBIDOS EN LA TRADUCCIÓN

Cambios como 'MUST' → 'recomendado', 'SHOULD' → 'obligatorio', 'MAY' → 'DEBE', 'Bloqueo crítico' → 'consejo temporal', 'Unsupported' → 'incorrecto', 'Unresolved' → 'fallido', 'Official Claim' → 'hecho verificado', 'Panel de usuarios en condiciones reales' → 'usuario no controlado' y 'Synthetic Test' → 'prueba real' alteran el significado normativo. No PUEDEN ser tratados como una traducción autorizada.

18. CONSULTA PÚBLICA ABIERTA

Cada versión importante de NOMOS debe incluir el siguiente proceso como candidato:

- Borrador de publicación
- Al menos 60 días de comentarios públicos
- Registro de comentarios
- Divulgación de conflictos
- Matriz de respuestas
- Borrador revisado
- Pruebas sintéticas
- Revisión independiente
- Votación del consejo
- Registro de ratificación

Opinión pública:

- Solo recopilación de declaraciones de apoyo,
- Campaña de marketing,
- Encuesta de mayoría

La consulta pública no es una campaña de apoyo, una campaña de marketing ni una simple encuesta de mayoría.

19. PROPUESTA DE MODIFICACIÓN

Cada enmienda:

NOMOS Propuesta de Cambio — NCP

debe llevar una identidad. Ejemplo:

NCP-0042— Revisión de la Prueba de Omisión de Límites Materiales

NCP debe incluir las siguientes áreas:

- Problema
- Reglas afectadas
- Evidencia
- Cambio propuesto
- Alternativas
- Impacto en la puntuación
- Impacto en decisiones pasadas
- Impacto en la equidad
- Impacto en la privacidad
- Costo de implementación
- Declaraciones de conflicto
- Comentarios públicos
- Decisión final

20. CAMBIAR CLASES

PARCHE

Error tipográfico Enlace roto Descripción del diagrama Corrección sin impacto

MENOR

Nuevo ejemplo Nuevo campo opcional Extensión compatible Aclaración

MAYOR

Cambio en el nivel de importancia Cambio en la fórmula de puntuación Agregar o eliminar Puerta Cambio en la población objetivo Cambio en la autoridad de gobernanza Cambio normativo que afecta resultados antiguos Un cambio importante no puede publicarse como un parche.

21. REGISTRO DE DECISIÓN

Cada decisión de aceptación o rechazo:

- ID de decisión
- Distribución de votos

- Recusación
- Disenso
- Justificación
- Evidencia
- Fecha de vigencia
- Regla de transición

debe seguirse. “La Junta lo ha considerado apropiado.” por sí sola no es suficiente.

22. PROPÓSITO DE LA LLAMADA UNIVERSITARIA

El propósito de la llamada a las universidades:

- obtener respaldo,
- usar el logotipo de la universidad,
- escribir "aprobado académicamente"
- no es solo recopilar opiniones positivas de expertos.

El propósito:

> *es encontrar y publicar los lugares donde NOMOS podría estar equivocado usando métodos independientes.*

23. LLAMADO DE LA UNIVERSIDAD NOMOS

INVITACIÓN ABIERTA

No queremos que las universidades, centros de investigación, facultades de derecho, estadísticos, científicos de la información, investigadores del lenguaje, laboratorios de IA, equipos de IHC, especialistas en archivos y organizaciones de interés público apoyen a NOMOS. No queremos que nos validen. No queremos que nos den prestigio. No queremos que usen el nombre NOMOS ni que elogien a sus fundadores. Lo que queremos es esto: Lean nuestro método. Encuentren nuestra suposición más débil. Reconstruyan nuestra muestra. Traduzcan nuestras prompts a otros idiomas. Cuestionen las decisiones de Truth Pack. Encuentren los falsos positivos en nuestras puertas críticas. Prueben si nuestra fórmula de puntuación realmente preserva la equidad. No pretendan que pasamos donde no lo hemos hecho. Publique su resultado negativo. NOMOS debe citar investigación que lo corrija así como investigación que lo valide. La contribución más valiosa que una universidad puede hacer a NOMOS no es una carta de apoyo:

> *Es una objeción repetible.*

24. TRABAJOS SOLICITADOS A UNIVERSIDADES

UR-01— Revisión de Métodos de Muestreo y Encuesta

Asignación para personas 1,000 Pesos por país Estratos de idioma No respuesta Muestra efectiva Eventos raros

UR-02— Laboratorio de Equivalencia de prompts

Equivalencia de intención multilingüe Deriva de traducción Pragmática cultural Lenguaje legal local Constitución del prompt

UR-03— Estudio de Resolución de Entidades

Marca Entidad legal Dominio Franquicia Afiliado Colisión de nombre

UR-04— Estudio de Fiabilidad del Paquete de Verdad

Suficiencia de evidencia Independencia de fuente evidencia en contrario Distinción mundo abierto/mundo cerrado Evidencia restringida

UR-05— Estudio de Adjudicación de afirmaciones

Límites de átomos Afirmación implícita Atribución Modalidad Anexo de cita Omisión de material

UR-06— Validación de Puerta Crítica

Recuperación Falso positivo Daño de alto riesgo Alegación seria Autoridad falsa Omisión de límites

UR-07— Estudio de Validez de Puntuación

Pesos de los componentes Cancelación compuesta Sensibilidad Validez de constructo Validez predictiva Utilidad para la toma de decisiones

UR-08— Estudio de equidad lingüística y geográfica

Disparidad de piso Cobertura Idiomas con pocos recursos Visibilidad de países pequeños

UR-09— Estudio del impacto en decisiones humanas

Comportamiento de recomendación Confianza Decisiones de compra Recuperación de errores Comprensión del usuario

UR-10— Estudio de gobernanza y captura

Control del fundador Captura por parte del patrocinador Auditorías comparativas Pagar para pasar Apelaciones Integridad del registro

UR-11— Estudio de seguridad y registro

Registros firmados Rotación de claves Marcas falsificadas Representación errónea solo por máquina Continuidad del archivo

UR-12— Estudio de deriva longitudinal

Actualizaciones del modelo Cambios de fuente Cambios de entidad Deriva limpia-natural Durabilidad de la corrección

25. SOLICITUDES INDESEABLES DE UNIVERSIDADES

Las siguientes solicitudes no pueden hacerse a una universidad o investigador:

- Garantía de resultados positivos
- Resultado negativo oculto
- Derecho de veto previo a la publicación
- Uso del logotipo de la institución en marketing
- Actuar como si fueran clientes de la empresa
- NOMOS 950+ carta de apoyo
- Declaración "estándar mundial"

- Obtención artificial de fuentes para Wikipedia o visibilidad enciclopédica
- Idioma de resultado patrocinado
- Suavización de la crítica
- No presentar la opinión personal de un investigador como decisión institucional.

26. PAQUETE DE INVESTIGACIÓN UNIVERSITARIA

El paquete estándar que se enviará a cada universidad debe contener las siguientes secciones:

- Resumen ejecutivo de investigación
- 0.9 estándar o versión actual
- Índice de normas
- Paquete de pruebas sintéticas
- Método de puntuación
- Resultados de métodos conocidos
- Umbrales de pruebas fallidas
- Preguntas abiertas de investigación
- Protocolo de replicación
- Niveles de acceso a los datos
- Declaración de ética y privacidad
- Divulgación de conflictos y financiación
- Acuerdo de libertad de publicación
- Política de citación y contribución
- Enlaces del registro canónico
- Canales de contacto y problemas

Los resultados de ensayos fallidos no pueden ser eliminados del paquete.

27. LIBERTAD DE PUBLICACIÓN DEL INVESTIGADOR

NOMOS o el financiador:

- no puede aprobar los resultados de antemano,
- detener resultados negativos,
- forzar un cambio de título,
- eliminar críticas metodológicas

o no puede tener este derecho. El investigador:

- debe usar los datos correctamente,
- proteger la confidencialidad,
- explicar conflictos,
- debe mantener la distinción entre resultados de ensayo y resultados finales

Obligación. La independencia es mutua.

28. REGISTRO DE RESULTADOS NEGATIVOS

El Registro Público de Investigación NOMOS debería incluir los siguientes resultados juntos:

- Estudios que apoyan a NOMOS
- Estudios que apoyan parcialmente a NOMOS
- Estudios que están en conflicto con NOMOS
- Fallos de replicación
- Resultados nulos
- Estudios retractados
- Disputas metodológicas
- Respuestas abiertas

No se pueden listar solo artículos positivos.

29. PRE-REGISTRO

Replicaciones independientes en la medida en que sean relevantes:

- hipótesis,
- muestra,
- métricas primarias,
- reglas de exclusión,
- método de puntuación,
- plan de análisis,
- conflictos

deben registrarse antes de ver los datos. Pre-registro: no inhibe la creatividad de la investigación, hace visibles los cambios de método después de obtener los resultados.

30. REPLICACIÓN MULTICENTRO

Estudio candidato para probar la independencia de NOMOS:

- Al menos tres instituciones independientes
- Al menos tres países
- Al menos dos familias de idiomas
- Corpus común sellado
- Especificación de puntuación común
- Equipos de adjudicadores separados
- Comparación de resultados a ciegas

deberían llevar. Propósito:

- el mismo resultado del mismo corpus,
- comportamiento similar del método en diferentes corpus

Debería ser para ver si lo produce o no.

31. RESULTADOS DE REPLICACIÓN

La replicación puede tener los siguientes estados:

REPLICADO

PARCIALMENTE REPLICADO

NO REPLICADO

INCONCLUSO

DIVERGENCIA DEL MÉTODO

DIVERGENCIA DE DATOS

DIVERGENCIA DE ADJUDICACIÓN

El resultado “No replicado” no puede presentarse como hostilidad del estudio. Debería abrir un nuevo Hallazgo de Método.

32. CAPAS DE ACCESO A DATOS

DATOS ABIERTOS

Pruebas sintéticas Resultados agregados públicos Esquemas públicos Ejemplos editados Código del método

ACCESO DE INVESTIGACIÓN REGISTRADO

Datos de observación desidentificados Registros de respuestas controladas Metadata de muestreo más detallada

ACCESO DE INVESTIGACIÓN SEGURO

Evidencia sensible pero necesaria Registros restringidos Datos vinculados protegidos por privacidad

SIN ACCESO EXTERNO

Contraseña Información completa de identidad Conversación privada innecesaria Registro legalmente prohibido Un estándar abierto no significa que cada dato esté abierto.

33. REVISIÓN ÉTICA

En la medida en que los estudios se realicen con usuarios reales:

- consentimiento informado,
- minimización de datos
- retiro,
- participantes vulnerables,
- privacidad
- accesibilidad

- compensación,
- diseño engañoso,
- transferencia internacional de datos

deberían llevar una revisión. Pruebas sintéticas:

- no reemplazan la investigación con participantes humanos,
- pueden no requerir aprobación ética humana.

la revisión de privacidad aún puede ser necesaria para material sintético que se asemeje a una persona real o derivado de datos reales.

34. RED DE IDIOMAS CON POCOS RECURSOS

NOMOS, si solo se prueba con universidades de idiomas principales, tendrá un argumento débil de equidad. La convocatoria abierta busca particularmente:

- departamentos de idiomas con pocos recursos,
- universidades regionales,
- centros de estudios de traducción,
- equipos locales de derecho y políticas públicas,
- Investigadores en accesibilidad

deberían recibir fondos separados y derechos de decisión. Esta participación:

- no debería limitarse al trabajo de recolección de datos,
- representación simbólica

Solo. Deben llevar contribuciones de votación y autoría en cambios normativos.

35. PAPEL DE LOS PROVEEDORES DE IA

Proveedores de IA:

- metadatos del sistema,
- aviso de cambio de modelo,
- telemetría,
- respuesta a incidentes,
- revisión técnica,
- acceso a reproducción

pueden proporcionar. Proveedores de IA:

- no pueden cambiar su propia puntuación,
- no pueden vetar la Puerta Crítica,
- no pueden determinar el coeficiente de otro proveedor,
- no puede eliminar un resultado negativo,

no puede formar una mayoría del Consejo. La participación del proveedor proporciona información. No proporciona monopolio de autoridad.

36. ROL DE LAS ENTIDADES AUDITADAS

Empresa o institución auditada: puede presentar pruebas del Paquete de Verdad. Puede objetar errores de la entidad. Puede proporcionar acceso restringido a pruebas. Puede desarrollar correcciones. Puede dar comentarios públicos. Entidad auditada: no puede determinar su propio nivel de importancia. No puede cambiar la fórmula de puntuación. No puede elegir a los adjudicadores. No puede eliminar repeticiones adversas. No puede ocultar la violación de uso de su propia marca.

37. ROL DEL PÚBLICO Y DE LA SOCIEDAD CIVIL

Organizaciones de beneficio público y comunidades de usuarios:

- informes de daños,
- uso indebido de la marca,
- fallos en lenguas de pocos recursos,
- problemas de accesibilidad,
- malentendidos por parte del consumidor

pueden reportar. Participación pública: no es solo una respuesta de encuesta. La gobernanza debe tener el derecho a la representación.

38. EL PAPEL DE LOS FINANCIADORES

Proveedor de financiamiento:

- apoyo a la investigación,
- infraestructura,
- corpus de lengua de pocos recursos,
- subvención de auditoría independiente

puede proporcionar. Proveedor de fondos:

- resultado,
- puntos,
- puerta,
- publicación,
- Votación del consejo,
- selección de auditor

no puede tener control sobre. Cada fondo:

- monto,
- fuente,
- campo de uso,
- condiciones,
- conflicto de intereses

deben explicarse con.

39. PUERTAS DE INDEPENDENCIA FINANCIERA

Salvaguardias del candidato:

- Límite máximo para la participación de un solo financiador
- Financiamiento de múltiples fuentes
- Fondo de investigación independiente
- Garantía de publicación de resultados negativos
- Prohibición del pago vinculado a puntos
- Informe presupuestario público
- Separación de tarifas de auditoría
- Transparencia en la compensación de los miembros del consejo

NOMOS pierde su estatus de estándar público en el momento en que se transforma en un sistema de pago para aprobar.

40. TRANSICIÓN A LA ADMINISTRACIÓN PÚBLICA

La transición de NOMOS desde su período de fundación hacia la gobernanza pública puede realizarse en cuatro fases.

ST-0— CUSTODIA DEL FUNDADOR

Kaan MURAZ es responsable como fundador. NobleJackal opera un dominio canónico. No existe autoridad de marca pública. Esta es la fase actual.

ST-1— ADMINISTRACIÓN DE REVISIÓN PÚBLICA

Rastreador de problemas público Junta asesora Consulta abierta Registro de conflictos está establecido.

ST-2— ADMINISTRACIÓN COMPARTIDA

Consejo de Normas Junta de Metodología Representantes universitarios Miembros de interés público comparten decisiones.

ST-3— ADMINISTRACIÓN PÚBLICA INDEPENDIENTE

Transportista legal o institucional separado Órganos de toma de decisiones independientes Registro de apelaciones Transparencia financiera Sucesión completada.

41. CANDIDATO A INSTITUCIÓN INDEPENDIENTE

En el futuro para NOMOS:

- fundación,
- asociación,
- organización de beneficio público,
- consorcio internacional,
- secretaría interuniversitaria

Se pueden considerar estructuras como esta. Cualquiera que sea la forma legal elegida, debe cumplir con las siguientes condiciones:

- Atribución del fundador
- Límite de no distribución o beneficio público equivalente
- Consejo independiente
- Reglas de conflicto
- Deber de registro público
- Compromiso con estándares abiertos
- Fijación de activos o protección equivalente
- Plan de disolución y archivo

Esta sección no selecciona una forma legal específica.

42. DERECHOS DEL FUNDADOR DE KAAN MURAZ

Derechos de Kaan MURAZ que deben preservarse históricamente:

- Atribución de autor del fundador
- Primera contribución arquitectónica normativa
- Registro de formación canónica del nombre y concepto NOMOS
- Registro histórico inmutable respecto a los lanzamientos del fundador
- Declaración del fundador durante el proceso de entrega pública
- Representación limitada del fundador en el organismo de gobierno

Poderes permanentes que Kaan MURAZ no debe ejercer:

- Cambiar la versión mayor por sí solo
- Veto
- Dar conformidad a su propio cliente
- Cambiar el resultado de una apelación
- Eliminar una reseña negativa
- Borrar el historial de registro
- Nombrar a todos los gerentes después solo del fundador

43. NOBLEJACKAL: ROL FUTURO

NobleJackal:

- área de publicación canónica del fundador,
- secretaría técnica,
- desarrollador de implementación de referencia,
- primer financiador del estándar abierto,
- proveedor de investigación y consultoría

puede ser. NobleJackal en el futuro: puede ser una de las organizaciones de auditoría autorizadas. Sin embargo, no puede ser la única organización. NobleJackal: puede hacer la mayor contribución histórica al estándar, no puede actuar como el único propietario del estándar.

44. EL PAPEL FUTURO DE NOMOS

El nombre NOMOS puede ser usado en tres capas separadas:

44.1. NOMOS PERSONA AUTORIAL

Esta es la identidad de autor auxiliar crítica utilizada en este libro.

44.2. NOMOS ESTÁNDAR

Es una regla normativa y un sistema de gobernanza.

44.3. AGENTE DE SOFTWARE NOMOS

Puede ser desarrollado en el futuro:

- asistente de auditor,
- motor de reglas,
- verificador de registro,
- mapeador de pruebas

software. Estas tres capas no deben confundirse entre sí. La respuesta de un agente de software no es la decisión oficial del estándar.

45. APROBACIÓN HUMANA

Cada texto y decisión oficial de NOMOS:

- aprobador humano,
- autoridad aprobadora,
- fecha,
- firma,
- versión,
- declaración de conflicto

debe celebrarse. IA:

- puede sugerir texto,
- puede encontrar inconsistencias,
- puede generar código,

puede crear propuestas de decisión. La responsabilidad pública final sigue siendo de los humanos o de las instituciones responsables.

46. CONTINUIDAD POST-FUNDADOR

NOMOS debe tener un plan de continuidad frente a los siguientes eventos:

- Renuncia de Kaan MURAZ
- Cierre de NobleJackal
- Pérdida de dominio
- Pérdida de acceso al repositorio canónico
- Compromiso de claves

- Conflicto administrativo
- Liquidación legal
- Interrupción técnica prolongada

Sistema de continuidad de candidatos:

- Múltiples custodios
- Firmas de umbral
- Espejos institucionales
- Depósitos de archivo
- Carta de sucesión
- Consejo de emergencia
- Aviso público de transición

debe llevar.

47. CONTINUIDAD DEL DOMINIO CANÓNICO

Arquitectura de publicación propuesta:

- /nomos/
- /nomos/estándar/
- /nomos/estándar/0.9/
- /nomos/estándar/1.0/
- /nomos/reglas/
- /nomos/esquemas/
- /nomos/pruebas/
- /nomos/implementaciones/
- /nomos/traducciones/
- /nomos/gobernanza/
- /nomos/consultas/
- /nomos/decisiones/
- /nomos/apelaciones/
- /nomos/registro/
- /nomos/erratas/
- /nomos/citas/
- /nomos/archivo/

Cada registro:

- URL estable,
- hash de contenido,
- fecha de publicación,
- estado,
- versión anterior/siguiente

debe llevar.

48. ARCHIVO INDEPENDIENTE

Fuera del dominio canónico:

- repositorios universitarios,
- archivos digitales institucionales,
- archivos de investigación a largo plazo,
- espejos firmados

pueden ser usados. Espejo: garantiza la continuidad del acceso, no se hace cargo de la autoridad de decisión oficial. Si se pierde el dominio canónico, entra en efecto el procedimiento de sucesión predefinido.

49. ETAPAS DE LA VERSIÓN ESTÁNDAR

0.9— BORRADOR FUNDACIONAL

El libro ha sido completado. Se puede abrir la revisión pública. Aún no hay pruebas reales. No existe un programa de conformidad.

0.95— CANDIDATO A REVISIÓN PÚBLICA

Se han recibido comentarios. Se han procesado los hallazgos del método. Se han publicado esquemas legibles por máquina. Se ha abierto la convocatoria universitaria.

0.99— CANDIDATO A VALIDACIÓN

Se han completado pruebas sintéticas reales. Se han superado los umbrales obligatorios. Se ha aplicado el conjunto de datos reservado intacto. Se ha obtenido BQ-3.

1.0— NORMA NORMATIVA ABIERTA

La consulta pública se ha completado. Existe al menos una reproducción independiente. Se han resuelto los hallazgos principales del método. El texto normativo ha sido ratificado. Se ha establecido una gobernanza GOV-2 o superior. Se ha publicado el paquete oficial de estándar abierto. La publicación del Estándar NOMOS 1.0 no significa el inicio automático del programa de marca de conformidad.

50. PROGRAMA DE CONFORMIDAD 1.0

El programa oficial de marca está vinculado a puertas separadas:

BQ-4

GOV-3

Auditores autorizados Panel de decisión independiente Cámara de apelaciones Registro público Revisión de seguridad Revisión legal Sistema de vigilancia Separación financiera Sin cumplir con estas condiciones: La norma 1.0 puede publicarse. Sin embargo: El Programa de Conformidad 1.0 no puede activarse.

51. NOMOS 950+ PROGRAMA

NOMOS 950+ solamente:

- Norma 1.0,
- Programa de Conformidad 1.0,

GOV-4,

BQ-4,

la representación de interés público puede activarse después de la puntuación replicada. 950+ no es la primera marca. Es una designación de rendimiento que lleva la carga de gobernanza más alta.

52. NIVELES DE ADOPCIÓN

AL-1— LEER

La institución ha revisado el estándar.

AL-2— AUTOEVALUACIÓN INTERNA

La institución ha realizado su evaluación interna. Esto no constituye conformidad oficial.

AL-3— IMPLEMENTACIÓN COMPATIBLE CON NOMOS

La herramienta admite esquemas y reglas abiertas. No otorga un sello de conformidad.

AL-4— USO AUTORIZADO EN AUDITORÍA

La organización autorizada utiliza el método en la auditoría.

AL-5— PROGRAMA PÚBLICO DE CONFORMIDAD

Un sistema independiente de decisiones y registro está operativo. Una institución no se considera AL-5 solo porque haya realizado una autoevaluación.

53. IMPLEMENTACIÓN CUMPLIDORA

Un software:

Compatible con NOMOS

la expresión solo puede ser:

- validación de esquema,
- cobertura de reglas,
- paquete de pruebas,
- soporte de versión,
- informe de conformidad

puede ser usado si lleva “Compatible con NOMOS”:

- autoridad de auditoría,
- decisión de conformidad

- derecho a marcar

'NOMOS-Compatible' no confiere autoridad de auditoría, una decisión de conformidad ni el derecho a usar una marca NOMOS.

54. IMPLEMENTACIÓN DE REFERENCIA

Implementación de referencia:

- código abierto,
- determinista donde corresponda,
- probada,
- vinculada a la versión,
- documentada

debería ser. También debería llevar la siguiente advertencia:

> *La autoridad normativa está en el texto. Este software es una implementación de referencia; cualquier error de software encontrado no cambia automáticamente la decisión del estándar.*

55. SUITE DE PRUEBAS DE CONFORMIDAD

La suite de pruebas debe incluir las siguientes capas:

- Pruebas de esquema
- Pruebas de reglas
- Pruebas de puntuación
- Pruebas de puerta
- Pruebas de traducción
- Pruebas de firma del registro
- Pruebas de uso indebido de marcas
- Pruebas de versiones históricas
- Pruebas de accesibilidad
- Pruebas de seguridad

Cada implementación:

- debe explicar qué pruebas pasó,
- qué pruebas no pasó

Cada implementación debe divulgar tanto las pruebas que aprobó como las que no aprobó.

56. TRANSPARENCIA DE PRUEBAS Y PRUEBAS RETENIDAS

La transparencia total no significa que todas las pruebas permanezcan continuamente abiertas. Equilibrio:

- Corpus de calibración público
- Prueba retenida expirada pública
- Prueba retenida actual sellada

- Publicación programada
- Corpus de renovación

debe establecerse con. Prueba retenida: no se puede mantener en secreto para siempre, y no se puede presentar como una prueba oculta que al mismo tiempo se haga pública y se reutilice.

57. CICLO DE VIDA DE LA PRUEBA RETENIDA

Generación de corpus Publicación de hash Custodia sellada Uso de validación Fijación de resultados Apertura controlada Archivo público Retiro Nueva generación de retención Antigua retención:

- educación de una persona,
- regresión,
- estudio de método

puede ser utilizado para. Puede que no esté intacto para la validación de un nuevo método.

58. ERRATA

Errata:

- error claro,
- corrección de cita,
- inconsistencia de esquema,
- desajuste de traducción

se utiliza para. Errata:

- cambio de fórmula de puntuación,
- eliminación de puerta,
- reducción del grado de importancia

no puede tener un camino oculto para.

59. DESPRECIO

Si se va a eliminar una regla o campo:

- aviso de desaprobación,
- justificación,
- reemplazo,
- fecha de transición,
- versiones afectadas,
- disponibilidad histórica

deben llevarse. Una regla desaprobada no se elimina. Se conserva para explicar cómo funcionaban las auditorías antiguas.

60. REGISTRO DE CITAS Y CONTRIBUCIONES

Las contribuciones NOMOS pueden registrarse en los siguientes roles:

- Autoría fundadora
- Redacción normativa
- Metodología
- Revisión estadística
- Revisión del idioma
- Revisión legal
- Ética
- Software
- Prueba
- Gestión de datos
- Revisión de interés público
- Replicación
- Comentario crítico

Autoría:

- A cambio de aval,
- A cambio de financiamiento,
- A cambio del logotipo institucional

no se puede otorgar. La atribución debe hacerse según la contribución.

61. VISIBILIDAD INDEPENDIENTE

Propios de NOMOS:

- significado enciclopédico,
- impacto académico,
- estatus de estándar mundial

no pueden establecerse con su propio texto. Estos solo pueden ocurrir a través de:

- investigación independiente,
- aplicación independiente,
- duplicación,
- crítica,
- uso público

en consecuencia. El deber de NOMOS no es producir visibilidad:

> *sino producir evidencia adecuada para revisión independiente.*

62. REGISTRO PÚBLICO DE INVESTIGACIÓN

Cada trabajo relacionado con NOMOS debe contener los siguientes campos:

- ID del estudio
- Institución
- Autores
- Financiamiento
- Conflictos

- Versión de NOMOS
- Conjunto de datos
- Pre-registro
- Resultados
- Estado de replicación
- Correcciones
- Retiro
- Relación con decisiones canónicas

NOMOS en este registro: también está obligado a mostrar artículos que lo critiquen.

63. ETAPAS DE ENTREGA PÚBLICA

FASE 0— FINALIZACIÓN DE FUNDACIÓN

El borrador del libro está completado. Se verifican los IDs de las reglas. Se realiza una auditoría de consistencia interna.

FASE 1— PUBLICACIÓN ABIERTA

HTML

PDF

Se publica el manifiesto de versiones de Schemas de reglas de máquina y el rastreador de incidencias.

FASE 2— COMENTARIO PÚBLICO

Universidades, investigadores, proveedores de IA, empresas, sociedad civil y expertos en lenguas proporcionan retroalimentación.

FASE 3— PRUEBA SINTÉTICA REAL

El corpus es genuinamente generado; la Verdad del Generador está sellada; la adjudicación es ciega; la recuperación se mide; y los resultados negativos se publican.

FASE 4— REPLICACIÓN INDEPENDIENTE

Al menos un equipo independiente Preferiblemente estudio multicéntrico Se crea un registro público de reproducción.

FASE 5— PILOTO EN VIVO LIMITADO

Se implementan productos de IA reales con Alcance de presencia limitado Modo solo de investigación Sin marca pública.

FASE 6— ESTÁNDAR ABIERTO 1.0

El texto normativo es ratificado.

FASE 7— PROGRAMA PILOTO DE CONFORMIDAD

Se prueba el sistema independiente de auditoría y decisión.

FASE 8— PROGRAMA DE CONFORMIDAD PÚBLICA

Se activa si se cumplen todas las puertas de gobernanza.

64. PAQUETE DE LANZAMIENTO PÚBLICO

El lanzamiento NOMOS 1.0 debe incluir al menos los siguientes archivos:

- standard.html
- standard.pdf
- rules.json
- schemas/
- puan-method.md
- reference-implementation/
- test-manifest.json
- known-limitations.md
- failed-tests.md
- governance-charter.md
- conflict-policy.md
- appeals-policy.md
- translation-policy.md
- change-log.md
- errata.md
- citation.cff
- official-statement.json
- release-signatures.json

failed-tests.md no tiene que estar vacío. Su ausencia debería generar sospechas.

65. NOMOS CARTA DEL ESTÁNDAR ABIERTO

NOMOS hace doce promesas fundamentales al público. Evidencia: ninguna afirmación será más cierta de lo que su evidencia permita. Límite: lo que no se haya realizado, sea desconocido o esté fuera de alcance permanecerá visible. Contexto: el país, el idioma, el usuario y el estado del sistema no serán borrados. Tiempo: el pasado no se presentará como la realidad actual. Reproducibilidad: la producción de cada resultado será explicada. Independencia: el interés comercial no reemplazará el juicio. Simetría: los errores favorables y adversos compartirán la misma carga de evidencia. Responsabilidad humana: la IA puede asistir, pero las decisiones responsables seguirán siendo humanas. Equidad lingüística y geográfica: los grupos de usuarios más pequeños no desaparecerán dentro de un promedio. Contestabilidad: el estándar seguirá abierto a la falsificación. Integridad histórica: los errores, correcciones y fallos no serán borrados del registro. Independencia del fundador: el futuro del estándar no dependerá de la voluntad de una sola persona.

En el núcleo de esta carta están las cuatro palabras de la llamada anterior: Evidencia. Límite. Contexto. Tiempo.

66. ESTADO ACTUAL NOMOS

En la etapa en que se complete este libro:

Campo	Estado
Texto normativo	0.9 Candidato Final
Consulta pública	No iniciado
Prueba sintética real	BQ-0
Demostración en el libro	DEMO-BQ-2
Replicación independiente	Ninguno
Gobernanza	GOV-0
Lanzamiento de estándar abierto	Aún no
Programa de conformidad	No activo
Autoridad de marca pública	Ninguno
NOMOS 950+ autoridad	Ninguno
Declaración permitida	Candidato a Estándar Abierto

Oración pública correcta:

> *"NOMOS es un candidato a estándar GEO abierto iniciado por Kaan MURAZ y desarrollado en coautoría crítica con sistemas de IA por NobleJackal."*

Oración pública incorrecta:

"NOMOS es la autoridad GEO reconocida mundialmente."

67. PUERTAS PARA LA TRANSICIÓN A LA VERSIÓN 1.0

Puertas obligatorias candidatas para el Estándar NOMOS 1.0:

- Auditoría de consistencia en todo el libro
- Registro mecánico de todos los IDs de reglas
- Equivalencia normativa en turco e inglés
- Paquete de lanzamiento público
- Consulta pública de al menos 60 días
- Matriz de comentarios y respuestas
- Ejecución real de prueba sintética
- Cumplimiento de los umbrales obligatorios de recuperación
- Reserva no modificada
- Al menos una replicación independiente
- Cierre de los principales hallazgos del método
- Implementación de referencia de código abierto
- Gobernanza al menos GOV-2
- Divulgación de financiamiento y conflictos
- Revisión de ética y privacidad
- Revisión de accesibilidad

- Plan de archivo y sucesión
- Ratificación del Consejo de Normas

Ninguno de estos puede hacerse "opcional" después del resultado.

68. DISPOSICIONES NORMATIVAS OBLIGATORIAS

CH20-N01

La legibilidad pública de NOMOS por sí sola no constituye un estatus de estándar abierto.

CH20-N02

El estado de estándar abierto debe incluir los derechos de leer, implementar, reproducir, criticar, contribuir, objetar y archivar.

CH20-N03

El texto normativo de NOMOS no puede mantenerse detrás de un muro de pago permanente.

CH20-N04

La evaluación interna y la aplicación académica del estándar no pueden estar vinculadas a un canon o licencia de marca.

CH20-N05

Los procesos de auditoría, infraestructura gestionada y conformidad pueden cobrarse como servicios separados del texto del estándar abierto.

CH20-N06

El texto NOMOS, el software de referencia, los datos de prueba y la marca de conformidad no pueden estar obligatoriamente sujetos a la misma licencia.

CH20-N07

Se debe usar una licencia de documentación abierta para el texto normativo.

CH20-N08

Se debe utilizar una licencia de software abierto que permita la reimplementación para el software de referencia.

CH20-N09

Evidencia personal o restringida no puede publicarse bajo una licencia de datos abiertos.

CH20-N10

El nombre NOMOS y la marca de conformidad no pueden usarse para impedir la aplicación de un estándar abierto.

CH20-N11

La protección de la marca debería ser solo para la versión oficial y con el propósito de prevenir la falsificación de marcas.

CH20-N12

Una persona o institución puede crear un método diferente derivado de NOMOS.

CH20-N13

Un fork no puede presentarse como la versión oficial NOMOS.

CH20-N14

Un fork debe llevar un nombre diferente y un registro de divergencia abierto.

CH20-N15

El fork no puede usar la marca oficial de conformidad.

CH20-N16

La versión canónica NOMOS debe llevar versión, fecha de publicación, aprobación humana, hash de integridad y registro de cambios.

CH20-N17

El texto sin un registro canónico no puede considerarse la versión oficial NOMOS.

CH20-N18

El texto normativo ratificado debe anteceder a la implementación de referencia en caso de conflicto.

CH20-N19

El software de referencia no puede cambiar silenciosamente el texto normativo.

CH20-N20

Cuando se encuentre un error de software, debe ser corregido de manera versionada.

CH20-N21

La declaración oficial NOMOS debe originarse únicamente de un texto canónico, versionado, fechado, aprobado por humanos y registrado en integridad.

CH20-N22

Los resultados del chat no pueden considerarse un estándar oficial NOMOS ni una decisión de conformidad.

CH20-N23

Una publicación en redes sociales por sí sola no puede constituir un cambio normativo.

CH20-N24

NOMOS por NobleJackal actualmente no puede presentarse como un modelo de base independiente.

CH20-N25

Si se desarrolla un agente de software separado de NOMOS, la línea de modelos, la arquitectura, los permisos, la versión y la supervisión deben registrarse explícitamente.

CH20-N26

La persona autoral NOMOS, el estándar NOMOS y el futuro agente de software NOMOS deben tener identidades separadas.

CH20-N27

El turco debe conservarse como el idioma de autoría fundacional de NOMOS.

CH20-N28

El inglés, como el equivalente normativo internacional, debe pasar por una revisión humana y metodológica.

CH20-N29

Ni la versión turca ni la inglesa pueden ser preferidas en secreto.

CH20-N30

El conflicto normativo entre los dos idiomas fundacionales requiere interpretación vinculante y corrección simultánea.

CH20-N31

Una traducción comunitaria no puede presentarse como una traducción normativa oficial.

CH20-N32

La traducción normativa autorizada debe contar con al menos dos expertos en idiomas, una revisión metodológica y comentario público.

CH20-N33

Los equivalentes de MUST, SHOULD y MAY DEBEN preservar la fuerza normativa.

CH20-N34

No se puede ocultar la desviación en la traducción, tal como por error de usuario o auditor.

CH20-N35

Un cambio de versión mayor no puede ser ratificado sin consulta pública.

CH20-N36

La consulta pública no puede convertirse en una campaña únicamente para recopilar declaraciones de apoyo.

CH20-N37

Cada comentario público debe ser incluido en la matriz de respuesta para aceptación, rechazo o fuera de alcance.

CH20-N38

El cambio normativo NOMOS debe tener un ID de Propuesta de Cambio.

CH20-N39

El NCP debe contener problemas, evidencia, alternativas, efecto en la puntuación y plan de transición.

CH20-N40

Un cambio de puerta, severidad, puntuación o autoridad de gobernanza no puede ser lanzado como un parche.

CH20-N41

Los registros de decisiones deben contener la distribución de votos, recusación, disenso y justificación.

CH20-N42

Una llamada universitaria no puede convertirse en una solicitud de respaldo.

CH20-N43

Las universidades no pueden tener la obligación de validar o producir resultados positivos para NOMOS.

CH20-N44

Las universidades y los investigadores deberían poder publicar sus resultados negativos sin la aprobación de NOMOS o del financiador.

CH20-N45

La opinión personal de un solo académico no puede presentarse como el respaldo institucional de la universidad.

CH20-N46

La colaboración académica no significa certificación o acreditación de NOMOS.

CH20-N47

El logotipo de la universidad no puede usarse sin autorización institucional por escrito.

CH20-N48

El paquete de investigación NOMOS debe incluir tanto los resultados de ensayos exitosos como los fallidos.

CH20-N49

Los hallazgos de métodos conocidos no pueden eliminarse del paquete de investigación.

CH20-N50

No se puede aplicar un veto de prepublicación a las publicaciones de investigadores independientes.

CH20-N51

El investigador debe divulgar los fondos y los conflictos de interés.

CH20-N52

El Registro Público de Investigación NOMOS debe mostrar resultados positivos, negativos, nulos y retractaciones juntos.

CH20-N53

No se pueden listar solo los estudios que apoyan a NOMOS.

CH20-N54

La replicación independiente debe incluir una preinscripción y un plan de análisis en la medida en que sea relevante.

CH20-N55

Cuando el resultado de la replicación es “no replicado”, se debe crear un Hallazgo de Método o un registro de discrepancia abierta.

CH20-N56

Un resultado no replicado no puede presentarse como hostilidad del investigador o mala intención.

CH20-N57

NOMOS debería apuntar al menos a una replicación independiente antes de reclamar un estándar público.

CH20-N58

No se debe establecer una pretensión de estándar mundial sin replicación multicéntrica e internacional.

CH20-N59

Los datos abiertos sintéticos no pueden conectarse al mismo régimen de acceso que los datos de usuarios reales.

CH20-N60

Se debe establecer un arreglo de minimización de datos y acceso controlado para los datos de usuarios reales.

CH20-N61

Ser un estándar abierto no significa que los datos personales y los secretos comerciales estén abiertos a todos.

CH20-N62

La investigación con usuarios reales debe estar sujeta a la revisión ética y de privacidad correspondiente.

CH20-N63

Los investigadores de lenguas con pocos recursos no pueden reducirse únicamente al papel de recolectores de datos.

CH20-N64

Los representantes de lenguas y regiones geográficas deberían tener derechos de voto reales en los procesos de toma de decisiones normativas.

CH20-N65

Los proveedores de IA pueden proporcionar datos técnicos, pero no pueden tener derecho de veto sobre la puntuación, el umbral o la publicación.

CH20-N66

Los proveedores de IA no pueden constituir la mayoría en el Consejo de Normas.

CH20-N67

Las entidades auditadas pueden presentar pruebas y apelar, pero no pueden determinar sus propias decisiones.

CH20-N68

El público y la sociedad civil deberían poder reportar daños a los usuarios y el uso indebido de la marca.

CH20-N69

El financiador no puede solicitar un resultado de investigación positivo.

CH20-N70

La financiación de investigación NOMOS debe llevar un contrato independiente del resultado.

CH20-N71

La dependencia excesiva de un solo financiador debe ser divulgada públicamente.

CH20-N72

NOMOS no puede ser convertido en un sistema de pago para aprobar o pago para publicar.

CH20-N73

La transición de la administración pública debe llevarse a cabo en etapas progresivas.

CH20-N74

La autoridad de conformidad pública no puede usarse durante la fase de custodia del fundador.

CH20-N75

La atribución al fundador de Kaan MURAZ debe preservarse en el registro histórico permanente.

CH20-N76

La atribución fundacional no puede crear un derecho de veto permanente.

CH20-N77

Kaan MURAZ no puede ser el único tomador de decisiones en un expediente de conformidad o apelación en el que tenga interés.

CH20-N78

NobleJackal puede ser registrada como origen canónico y secretaría técnica.

CH20-N79

NobleJackal no puede ser el único ejecutor ni la única autoridad de auditoría de un estándar abierto.

CH20-N80

La futura institución independiente de NOMOS debe contemplar disposiciones para el interés público, la sucesión y la continuidad de los activos.

CH20-N81

La gobernanza de NOMOS debe someterse a la revisión de expertos pertinentes antes de estar vinculada a una forma legal específica.

CH20-N82

La gobernanza pública no puede ser controlada por una sola empresa o proveedor.

CH20-N83

Las decisiones oficiales de NOMOS deben llevar la firma de una persona o institución responsable.

CH20-N84

El asistente de IA no puede ser el único firmante de una decisión oficial.

CH20-N85

El procedimiento de sucesión debe activarse cuando Kaan MURAZ, NobleJackal o el campo canónico no sean accesibles.

CH20-N86

Se debe crear un espejo independiente y un archivo a largo plazo para el repositorio canónico.

CH20-N87

El espejo proporciona acceso, pero no crea por sí mismo una nueva autoridad normativa.

CH20-N88

El Estándar NOMOS 1.0 y el Programa de Conformidad 1.0 deberían ser publicaciones separadas.

CH20-N89

La publicación de la Norma Normativa 1.0 no puede activar automáticamente el programa de marca.

CH20-N90

El Programa de Conformidad BQ-4 no puede ser activado sin GOV-3, apelaciones y registro público.

CH20-N91

El Programa NOMOS 950+ debe tener como objetivo al menos GOV-4 y condiciones de puntuación replicadas.

CH20-N92

La autoevaluación no puede presentarse como conformidad oficial.

CH20-N93

El software compatible con NOMOS no tiene derecho a ser marcado.

CH20-N94

La implementación de referencia no puede tener autoridad sobre el texto normativo.

CH20-N95

Los resultados de las pruebas de conformidad deben mostrar tanto las pruebas aprobadas como las reprobadas juntos.

CH20-N96

Un holdout sellado no puede mantenerse en secreto para siempre.

CH20-N97

El holdout usado debe ser llevado al archivo público y debe producirse un nuevo holdout.

CH20-N98

Las erratas no pueden ser utilizadas para cambios normativos mayores.

CH20-N99

Las versiones antiguas, las reglas obsoletas y las pruebas fallidas deben preservarse en el archivo.

CH20-N100

Cada licencia, versión, traducción, investigación, gobernanza y decisión de lanzamiento respecto al lanzamiento público de NOMOS debe tener un propietario humano o institucional responsable.

69. FORMAS DE FALLA

CH20-F01— CONSIDERANDO EL PDF COMO UN ESTÁNDAR ABIERTO

Solo se presenta como apertura el texto descargable.

CH20-F02— FÓRMULA CERRADA

La lógica de puntuación y los pesos no se divulgan.

CH20-F03— PRUEBAS CERRADAS

No se comparten el corpus ni los registros de recuperación.

CH20-F04— PROCESO DE CAMBIO CERRADO

El fundador o las reglas de la empresa hacen cambios específicamente.

CH20-F05— ESTÁNDAR DE MURO DE PAGO

El acceso al texto normativo está limitado por una tarifa.

CH20-F06— MONOPOLIO SOBRE LA TARIFA DE SOLICITUD

Se requiere una tarifa de licencia obligatoria para usar el estándar.

CH20-F07— TEXTO ABIERTO, afirmación DE MERCADO CERRADO

Aunque el texto es abiertamente visible, las aplicaciones independientes están bloqueadas.

CH20-F08— CONVERTIR LA MARCA REGISTRADA EN UN MONOPOLIO DE MÉTODO

La marca registrada se utiliza para evitar que otros apliquen el método.

CH20-F09— PROHIBIENDO EL FORK

Se evita crear un estándar derivado bajo un nombre diferente.

CH20-F10— IMITANDO LA IDENTIDAD OFICIAL DEL FORK

El texto derivado se presenta como NOMOS 2.0.

CH20-F11— VERSIONES CANÓNICAS SIN VERSIÓN

Cuál texto es oficial es desconocido.

CH20-F12— COLOCANDO EL CÓDIGO POR ENCIMA DE LA NORMA

Un error de implementación de referencia se convierte en la verdad del estándar.

CH20-F13— CORRECCIÓN SILENCIOSA DE CÓDIGO

El cálculo cambia pero la versión no se incrementa.

CH20-F14— CONTAR CONVERSACIÓN COMO TEXTO OFICIAL

Una respuesta de IA no canónica se convierte en una disposición estándar.

CH20-F15— PRESENTÁNDOLO COMO UN MODELO QUE NOMOS NO POSEE

Se hace una afirmación de IA propietaria aunque no se haya desarrollado un sistema de IA separado.

CH20-F16— CONFUNDIR PERSONA CON AGENTE DE SOFTWARE

El seudónimo del autor se comercializa como un producto técnico.

CH20-F17— HACIENDO DE LA IA EL PORTAVOZ DE TODAS LAS IA

NOMOS se presenta como una autoridad universal de máquinas.

CH20-F18— ELIMINANDO EL TEXTO DEL FUNDADOR TURCO

El idioma fundador se vuelve invisible en nombre de la internacionalización.

CH20-F19— HACIENDO DEL INGLÉS EL TEXTO MAESTRO INCONTROLADO

La traducción reemplaza el pensamiento fundador.

CH20-F20— DEBE/DEBERÍA DERIVAR

La traducción cambia la fuerza normativa.

CH20-F21— CONTANDO OFICIALMENTE LA TRADUCCIÓN COMUNITARIA

La versión del idioma no revisada se convierte en una norma de auditoría.

CH20-F22— PRESERVANDO EL CONFLICTO DE TRADUCCIÓN

Dos idiomas aplican reglas diferentes.

CH20-F23— TEATRO DE COMENTARIOS PÚBLICOS

Se recopilan comentarios pero no se publican las respuestas.

CH20-F24— VOTACIÓN DE APOYO

El estándar técnico se reduce a la preferencia de la mayoría.

CH20-F25— REGLA SIN PROPUESTA DE CAMBIO

El cambio normativo se realiza mediante chat o correo electrónico.

CH20-F26— HACIENDO CAMBIO MAYOR PATCH

Los cambios de puntuación y de filtro están ocultos como una corrección menor.

CH20-F27— ELIMINAR DISCREPANCIA

Se actúa como si no hubiera desacuerdo en la junta.

CH20-F28— COMPRAR CARTA DE APOYO DE LA UNIVERSIDAD

La investigación se convierte en un respaldo positivo.

CH20-F29— CONVERTIR OPINIÓN ACADÉMICA PERSONAL EN APROBACIÓN INSTITUCIONAL

El comentario del profesor se presenta como una decisión de la universidad.

CH20-F30— USAR LOGO DE LA UNIVERSIDAD

El logo se añade al marketing sin autoridad institucional explícita.

CH20-F31— NO PUBLICAR RESULTADOS NEGATIVOS

El financiador o NOMOS aplica un veto.

CH20-F32— CONTROL PREVIO A LA PUBLICACIÓN

El texto del investigador se reemplaza por interés comercial.

CH20-F33— ELIMINAR INVESTIGACIÓN NEGATIVA DEL REGISTRO

Solo se listan estudios positivos.

CH20-F34— CONTAR COMO HOSTILIDAD LO NO REPLICADO

El desacuerdo sobre el método se convierte en un ataque personal.

CH20-F35— SELECCIÓN DE RESULTADOS NO REGISTRADOS

Solo se publican métricas que se ven bien.

CH20-F36— REPLICACIÓN DE UNA SOLA INSTITUCIÓN

Se considera independiente a un equipo con la misma estructura de intereses.

CH20-F37— VIOLACIÓN DE LA PRIVACIDAD BAJO EL NOMBRE DE DATOS ABIERTOS

Se publican conversaciones reales de usuarios.

CH20-F38— CAJA NEGRA DE DATOS RESTRINGIDOS

Ningún investigador independiente puede examinar el método.

CH20-F39— USO SIMBÓLICO DE LENGUAS CON POCOS RECURSOS

El equipo local solo actúa como traductores.

CH20-F40— CAPTURA DEL PROVEEDOR DE IA

El proveedor determina la puerta de acceso y la puntuación.

CH20-F41— CAPTURA DE ENTIDAD AUDITADA

La empresa auditada compra el cambio estándar.

CH20-F42— CAPTURA DEL FINANCIADOR

El financiador detiene el resultado negativo.

CH20-F43— OCULTANDO DEPENDENCIA DE UN SOLO FINANCIADOR

El interés financiero permanece invisible.

CH20-F44— INVESTIGACIÓN DE PAGO PARA APROBAR

Un resultado positivo se convierte en una condición para la continuación del fondo.

CH20-F45— ADMINISTRACIÓN PÚBLICA COMO AUSENCIA DE PROPIETARIO

Se establece un sistema donde nadie asume responsabilidad.

CH20-F46— CONTROL DEL FUNDADOR PARA SIEMPRE

El fundador toma todas las decisiones importantes solo.

CH20-F47— BORRANDO LA ATRIBUCIÓN AL FUNDADOR

La historia del fundador se destruye cuando se entrega al público.

CH20-F48— MONOPOLIO DE NOBLEJACKAL

Ninguna otra institución puede implementar o auditar.

CH20-F49— BORRANDO EL ORIGEN DE NOBLEJACKAL

El origen canónico está oculto en el nombre de la independencia.

CH20-F50— CONVIRTIENDO A LA IA EN TOMADORA DE DECISIONES LEGALES

El agente de software NOMOS firma el cumplimiento.

CH20-F51— FALTA DE PLAN DE SUCESIÓN

Cuando se pierde el fundador o el dominio, el estándar se detiene.

CH20-F52— MONOPOLIO DE UNA SOLA CLAVE

Todos los registros oficiales están vinculados a la clave privada de una sola persona.

CH20-F53— CONSIDERANDO EL ESPEJO COMO AUTORIDAD CANÓNICA

La copia del archivo publica la nueva versión.

CH20-F54— PUBLICANDO 0.9 COMO 1.0

El lanzamiento principal se realiza antes de que se completen la consulta pública y las pruebas.

CH20-F55— COMBINANDO EL ESTÁNDAR 1.0 CON EL PROGRAMA MARCA

La certificación comienza tan pronto como se publica el texto.

CH20-F56— 1.0 SIN BQ-3

La liberación normativa se completa sin una validación real.

CH20-F57— PROGRAMA DE CONFORMIDAD SIN BQ-4

La autoridad de la marca se abre sin replicación independiente.

CH20-F58— CONTAR LA AUTOEVALUACIÓN COMO CERTIFICACIÓN

La empresa hace que su propio resultado sea una marca oficial.

CH20-F59— CONTAR EL SOFTWARE COMPATIBLE COMO AUDITOR

La herramienta que pasa la prueba da una decisión de conformidad.

CH20-F60— PUBLICAR LOS RESULTADOS DE LAS PRUEBAS SELECCIONÁNDOLOS

Las pruebas de conformidad que no se aprueban se almacenan.

CH20-F61— OCULTANDO EL HOLDOUT PARA SIEMPRE

El estándar sigue cerrado a la revisión independiente.

CH20-F62— USANDO CONTINUAMENTE EL MISMO VERIFICADOR COMO NUEVO

La prueba se memoriza.

CH20-F63— CAMBIANDO LA PUERTA CON ERRATAS

Una norma importante se publica como una corrección menor.

CH20-F64— ELIMINANDO LA REGLA DEPRECADA

Las auditorías pasadas se vuelven incomprensibles.

CH20-F65— NO TENIENDO ARCHIVO DE PRUEBAS FALLIDAS

El lanzamiento solo indica éxito.

CH20-F66— CONVIRTIENDO LA CONTRIBUCIÓN EN UNA VENTA DE ENDOSO

El nombre y la autoría se conceden a cambio de prestigio.

CH20-F67— PRODUCIENDO LA PROPIA IMPORTANCIA ENCICLOPÉDICA

El trabajo autopublicado se utiliza como prueba de impacto mundial.

CH20-F68— CONVIRTIENDO EL REGISTRO DE INVESTIGACIÓN PÚBLICA EN UNA LISTA PROMOCIONAL

Solo se destacan estudios positivos.

CH20-F69— OCULTANDO LOS HALLAZGOS METODOLÓGICOS EN LA PUBLICACIÓN PÚBLICA

Las limitaciones conocidas se eliminan del paquete.

CH20-F70— ELIMINANDO ETIQUETA NOINDEX Y SINTÉTICA

Los errores de prueba se mezclan en el verdadero conjunto de conocimientos.

CH20-F71— afirmación DE AUTORIDAD PÚBLICA EN ST-0

La custodia fundadora se presenta como un estándar mundial.

CH20-F72— ELIMINACIÓN DEL REGISTRO DEL FUNDADOR EN ST-3

La independencia destruye la autoría histórica.

CH20-F73— MANTENIENDO SÍMBOLOS DE MIEMBROS DE INTERÉS PÚBLICO

Se utiliza un representante sin derecho a voto.

CH20-F74— NO OTORGAR PODER DE DECISIÓN A REPRESENTANTES DEL IDIOMA

La equidad no se refleja en la gobernanza.

CH20-F75— ELIMINANDO RESULTADO UNIVERSITARIO NEGATIVO

La sumisión pública acepta solo elogios.

CH20-F76— CONFUSIÓN DE NUEVA TECNOLOGÍA NOMOS CON PERSONA ANTIGUA

El agente de software se convierte automáticamente en escritor oficial.

CH20-F77— FALTA DE FIRMA DE DECLARACIÓN OFICIAL

Cada texto en el campo canónico se considera oficial.

CH20-F78— PERMITIENDO FRAUDE DE MARCA BAJO EL NOMBRE DE LA TRANSPARENCIA

Todos producen su propia marca NOMOS.

CH20-F79— ABANDONO DE LA PUBLICACIÓN PÚBLICA

Ningún administrador asume responsabilidad.

CH20-F80— EXCEPCIÓN A NOMOS

Las reglas de transparencia, evidencia, versión y objeción que solicitan a otros no se aplican a su propio proceso de publicación pública.

70. PROCEDIMIENTO DE PUBLICACIÓN

Paso 1— Congelar el Texto Fundacional

Para las Secciones 1–20:

- versión,
- IDs de reglas,
- resúmenes de capítulos,
- citas,
- consistencia interna

se crean registros.

Paso 2— Publicar Limitaciones Conocidas

BQ-0

DEMO-BQ-2

Brecha de Mapeo de Citaciones Omisión de Material

GOV-0

La falta de autoridad de marca no se oculta.

Paso 3— Separar Texto Normativo y Explicativo

DEBE

DEBERÍA

PUEDE

las etiquetas de comentario de ejemplo se aclaran.

Paso 4— Iniciar Estudio de Equivalencia Turco–Inglés

Cada cláusula y ID de Regla está mapeada.

Paso 5— Enviar Borrador de Licencia Abierta para Revisión Legal

El Código de Texto de Datos y la Marca Comercial están separados.

Paso 6— Abrir el Repositorio Canónico

Se publican URLs estables, versiones y archivos.

Paso 7— Publicar Reglas en Formato Legible por Máquina

Se explican los registros de regla, esquema y puntuación.

Paso 8— Abrir el Rastreador de Problemas Público

Se establecen categorías de mal uso del Marcador de Seguridad de Gobernanza de Método de Traducción de Errores.

Paso 9— Publicar la Carta de Gobernanza 0.9

Se explican la autoridad fundadora, la recusación, el interés público y la transición.

Paso 10— Iniciar Consulta Pública

Se abre un período de al menos 60 días.

Paso 11— Publicar la Convocatoria Universitaria

Se solicita una prueba independiente, no un resultado positivo.

Paso 12— Preparar Paquetes de Investigación

Se preparan archivos modulares para diferentes disciplinas.

Paso 13— Abrir Fondo y Registro de Conflictos

Toda la financiación se hace visible.

Paso 14— Ejecutar la Prueba Sintética Real

El corpus se genera genuinamente para salir de BQ-0.

Paso 15— Publicar Fallos

Los umbrales no se cambian después del resultado.

Paso 16— Aplicar la Corrección del Método 1.0

Los métodos de citación y omisión se renuevan.

Paso 17— Ejecutar Retención Intacta

Se prueba el cumplimiento extremo.

Paso 18— Habilitar la Replicación Independiente

Al menos un equipo externo aplica el mismo sistema.

Paso 19— Publicar la Matriz de Respuestas a Comentarios Públicos

Se responde a cada comentario sustantivo.

Paso 20— Establecer la Transición del Consejo de Normas

Se configura la estructura de gobernanza GOV-2.

Paso 21— Realizar la Votación de Ratificación del Estándar Abierto 1.0

Se publica la aceptación o el rechazo con la justificación.

Paso 22— Publicar el Paquete de Lanzamiento del Estándar 1.0

Se divulga el conjunto completo de archivos y firmas.

Paso 23— Mantener el Programa de Conformidad Separado

El programa de marca no se abrirá sin BQ-4 y GOV-3.

Paso 24— Lanzar Pilotos Multicéntricos

Se realizan auditorías en vivo solo de investigación.

Paso 25— Probar el Sistema de Apelaciones y Registro

Se prueba la cadena de decisiones antes de otorgar la calificación.

Paso 26— Realizar una Revisión de Interés Público

Se evalúa el impacto sobre el consumidor, el idioma y la accesibilidad.

Paso 27— Tomar una Decisión Separada para el Programa de Conformidad 1.0

No se automatiza con la versión estándar.

Paso 28— Firmar el Plan de Sucesión y Archivo

Se asegura la continuidad después del fundador.

Paso 29— Publicar la Declaración de Entrega Pública

Se explican los roles de Kaan MURAZ, NobleJackal y los nuevos órganos de administración.

Paso 30— Conservar los Registros Históricos del Fundador

La entrega pública no borra el origen.

71. EVIDENCIA REQUERIDA

Manifiesto de versión fundacional Hashes de sección Índice de reglas normativas Limitaciones conocidas Registros BQ Registros GOV Registro de prohibición de marcas públicas Texto normativo turco Texto normativo en inglés Matriz de equivalencia de traducción Política de traducción autorizada Licencia de texto Licencia de código Licencia de datos Política de marca registrada Repositorio canónico Registro espejo Archivo de versiones Declaración oficial Manifiesto de seguimiento de problemas públicos Registros NCP Comentarios públicos Matriz de respuestas Registros de decisiones Disenso Llamada universitaria Paquete de investigación universitaria Divulgaciones de financiación Registro de conflictos Registros de preinscripción Estudios independientes Resultados negativos Informes de replicación Registro público de investigación Corpus de pruebas abierto Manifiestos de retención Informes de recuperación Pruebas de conformidad de la implementación de referencia Pruebas fallidas Revisión ética Política de privacidad Política de acceso a datos Revisión de accesibilidad

Participación de lenguas con pocos recursos Reglas de participación de proveedores de IA Reglas de participación de entidades auditadas Estatuto del Consejo de Estándares Estatuto de la Junta de Metodología Estatuto del panel de interés público Estatuto de sucesión Depósitos de archivo Plan de custodia de llaves Estándar 1.0 puertas Registro de ratificación Publicación del Estándar 1.0 Registro de separación del Programa de Conformidad Declaración de entrega pública Humano o institución responsable

72. LISTA DE VERIFICACIÓN DE AUDITORÍA

¿Es el texto NOMOS accesible libremente? Más allá de ser legible, ¿es aplicable? ¿Son abiertas las fórmulas de puntuación? ¿Se publican los esquemas de reglas? ¿Está abierto el código de referencia? ¿Se puede reproducir el corpus de prueba? ¿Son visibles los resultados negativos de las pruebas? ¿Requiere implementar el estándar un pago de marca registrada? ¿Están separadas las licencias del texto, código, datos y marca registrada? ¿Hay derecho a bifurcar? ¿Puede una bifurcación imitar la identidad oficial? ¿Está abierta la versión canónica? ¿Son accesibles las versiones antiguas? ¿Se publica la regla de declaración oficial? ¿Se pueden usar los resultados del chat como decisiones oficiales? ¿Se representa falsamente a NOMOS como un modelo de IA separado? ¿Se separan el personaje autoral y el agente de software? ¿Se conserva el turco como la lengua fundadora? ¿Ha pasado la revisión humana y metodológica en inglés?

¿Coinciden los IDs de las reglas en ambos idiomas? ¿Existe un camino de conflicto de traducción? ¿Se presenta la traducción de la comunidad como oficial? ¿Es suficiente el período de comentarios públicos? ¿Se responden todos los comentarios? ¿El sistema NCP está abierto? ¿Se aplican parches a los cambios importantes? ¿Es visible la distribución de votos y el desacuerdo? ¿La convocatoria de la universidad requiere respaldo? ¿Está abierto el derecho a publicar resultados negativos? ¿Se puede usar el logo de la universidad en marketing? ¿Se convierten las opiniones académicas personales en aprobación institucional? ¿Se incluyen en el paquete universitario los hallazgos de métodos conocidos? ¿Existe un camino de preregistro? ¿Se registran los resultados no replicados? ¿Solo se enumeran los estudios positivos? ¿Se separan los datos abiertos y los restringidos? ¿Se filtran datos personales en la publicación pública?

¿Tienen los investigadores de idiomas de bajos recursos derechos de decisión? ¿Tienen los proveedores de IA poder de veto? ¿Pueden las empresas auditadas elegir a los adjudicadores? ¿Pueden los financiadores influir en el resultado? ¿Es transparente la concentración de un solo financiador? ¿Es clara la fase de administración pública? ¿Se preserva la atribución del fundador de Kaan MURAZ? ¿Tiene Kaan poder de veto permanente? ¿Es NobleJackal el único implementador? ¿Es visible el origen de NobleJackal? ¿Se presenta a NOMOS como la única IA tomadora de decisiones? ¿Hay un registro de aprobación humana? ¿Existe un plan de sucesión? ¿Está claro qué sucede en caso de pérdida de dominio? ¿Hay múltiples custodios? ¿Están disponibles espejos de archivo? ¿Se han separado el Estándar 1.0 y el Programa de Conformidad? ¿Se publica el Estándar 1.0 antes de BQ-3?

¿Se está abriendo el programa de marca pre-BQ-4? ¿Se presenta como conformidad de autoevaluación? ¿Proporciona una marca de implementación compatible? ¿Muestra los resultados de las pruebas aprobadas y no aprobadas? ¿Está abierto el ciclo de vida de Holdout? ¿Se usa el mismo holdout de manera continua? ¿Un errata cambia reglas principales? ¿Las reglas obsoletas están en el archivo? ¿Existe un archivo de pruebas fallidas? ¿La contribución depende de la contribución real? ¿El Registro Público de Investigación muestra estudios críticos? ¿El paquete de liberación pública está completo? ¿Existe un archivo de limitaciones conocidas? ¿Se muestra actualmente la prueba real correctamente como BQ-0? ¿La afirmación público actual mantiene el umbral de “Candidato a Estándar Abierto”? ¿La entrega pública borra la fecha de fundador? ¿La entrega pública realmente limita el control único?

¿Está claro quién es el responsable del estándar? ¿Hay un propietario y una fecha de la decisión de publicarlo al público?

73. OBJECIONES Y RESPUESTAS

Objeción 1— “Si hacemos un estándar abierto, ¿no nos robarán nuestra idea?”

Publicar el texto y el método bajo una licencia con fecha, versionada y que requiera atribución fortalece el registro del fundador. Se puede crear un fork. Sin embargo:

- La atribución del fundador de Kaan MURAZ,
- El origen canónico de NobleJackal,
- El registro de integridad de las primeras versiones

no puede ser eliminado. Mantenerlo cerrado puede dificultar la adopción independiente más que proteger la contribución.

Objeción 2— “Si todos pueden implementarlo, ¿cuál será el valor de NobleJackal?”

El valor de NobleJackal no debería provenir únicamente de una fórmula secreta. Puede originarse a partir de:

- experiencia del fundador,
- implementación de referencia,
- infraestructura técnica,
- consultoría,
- auditoría autorizada,
- educación de una persona,
- tecnología de registro,

investigación continua. Un estándar abierto no desvaloriza un servicio sólido. Un monopolio débil sí.

Objeción 3— “Si todos crean un fork, ¿no se fragmentará NOMOS?”

Los forks pueden fortalecer el ecosistema de investigación siempre que lleven diferentes nombres y registros de divergencia. La fragmentación ocurre si cada fork se presenta como el NOMOS oficial. La distinción entre la versión canónica y la marca registrada previene esto.

Objeción 4— "¿Por qué no pedimos una carta de apoyo a las universidades?"

Una carta de apoyo puede proporcionar visibilidad. Sin embargo, no prueba la exactitud del método. La replicación independiente:

- más lenta,
- más difícil,
- a veces negativa

Es posible. El verdadero valor académico está aquí.

Objeción 5— “Si la universidad publica un resultado negativo, ¿no se vería perjudicado NOMOS?”

Se puede ver a corto plazo. A largo plazo:

- se corrige el error del estándar,
- disminuye la confianza falsa,

el método se fortalece. Silenciar el resultado negativo socava la base ética de NOMOS.

Objeción 6— "Si el turco y el inglés son normativos al mismo tiempo, ¿no aumenta el riesgo de contradicción?

Sí, crece. Sin embargo, transformar el turco solo en texto histórico y el inglés en el único estándar real entrega la idea fundadora a la capa de traducción. La solución:

- ID de cláusulas,
- examen de equivalencia,
- interpretación vinculante,

Es una corrección sincrónica.

Objeción 7— "¿No es más fuerte para la marca presentar NOMOS como la IA desarrollada por NobleJackal?"

Puede parecer más impresionante a corto plazo. Sin embargo, no es correcto si no hay un modelo y arquitectura técnica separados. El verdadero poder de NOMOS:

- él/ella no está reclamando tecnología falsa,
- identidad de autor crítico canónica,
- metodología normativa,
- en el agente de auditoría abierta que puede desarrollarse en el futuro

se encuentra.

Objeción 8— "Si la autoridad de Kaan disminuye, ¿no romperán otros el estándar?"

Se conserva la atribución histórica de Kaan y la representación limitada de la fundación. Sin embargo:

- carta abierta en lugar de veto,
- un consejo multipartícipe en lugar de una sola persona,
- registro público en lugar de decisión privada

El estándar protege más fuertemente. La buena gobernanza no depende de la presencia continua del fundador.

Objeción 9— "¿Por qué no podemos otorgar la insignia inmediatamente cuando se publica el Estándar 1.0?

La preparación del texto normativo:

- autorización del auditor,
- decisión independiente,
- apelaciones,
- seguridad del registro,
- vigilancia

no significa que los sistemas estén listos. La norma y el programa de conformidad son productos separados.

Apelación 10— "Si se abre la prueba pública, ¿no memorizarán los sistemas de IA las pruebas?"

Este riesgo existe. Por lo tanto:

- calibración pública,
- reserva sellada,

- lanzamiento programado,
- corpus de renovación

usados juntos. La confidencialidad total previene la reproducción. La divulgación completa puede deshabilitar la reserva actual.

Objeción 11— "¿Cómo se puede realizar un estudio tan grande sin financiación?"

La financiación es necesaria. El problema no es obtener financiación: el problema es que la financiación controle el resultado. La independencia se puede mantener mediante financiación de múltiples fuentes, divulgación pública y libertad de publicación.

Objeción 12— "¿Qué pasa si otras instituciones realizan la auditoría NOMOS incorrecta?"

Cualquiera puede usar el método para evaluación interna e investigación. Para decisiones oficiales de conformidad:

- autorización del auditor,
- panel de decisión,
- registro público

son requeridos. El uso indebido de la marca oficial se convierte en un asunto de aplicación separado.

Objeción 13— "Si la voz oficial de NOMOS está solo en el dominio canónico, ¿no puedes escribir libremente en chats de IA?"

Se puede escribir. Chats:

- borrador creativo,
- descripción,
- simulación,
- pensamiento crítico

puede producir. Sin embargo, no se considera un cambio normativo oficial. Esta distinción no obstaculiza la libertad. Preserva el límite oficial.

Objeción 14— "¿Podemos enviar este libro a las universidades ahora?"

Sí. Con el estado correcto:

Candidato a Estándar Abierto Fundacional — Versión 0.9

puede ser enviado. No puede ser enviado con la siguiente declaración:

> *Estándar GEO completado, verificado y reconocido globalmente.*

Lo que se debe solicitar a las universidades es una revisión independiente, no la aprobación.

Objeción 15— "¿Cuándo podemos decir 'estándar mundial'?"

No por una sola fecha o puntuación. Se puede considerar tal cuando se establecen los siguientes aspectos:

- implementación independiente en múltiples países,
- replicación multicentro,
- registro abierto y de resultados negativos,
- gobernanza independiente del fundador,

- versiones con autoridad decisoria en diferentes idiomas,
- registro público en tiempo real,
- adopción por parte de los usuarios y proveedores de IA,

Versionado resistente a la deriva del método. No se declara como un estándar mundial. Comienza a ser utilizado por el mundo.

Objeción 16— "Entonces, ¿cuál es la conclusión a la que llegamos al final del libro?"

NOMOS aún no es un estándar mundial completo. Sin embargo, ahora tiene:

- su propósito,
- el objeto de medición,
- orden de muestra,
- cadena de evidencia,
- Paquete de la Verdad,
- adjudicación,
- Puertas críticas,
- arquitectura de puntuación,
- para volver a probar,
- plan de gobernanza y entrega pública

considerado como un candidato serio a estándar abierto. Esta es una posición muy por encima de la idea inicial.

DISPOSICIÓN COMÚN DEL CAPÍTULO 78

La persona que escribe un estándar es su primer propietario. Sin embargo, si el estándar realmente tiene éxito, ya no pertenece únicamente al propietario. También recae dentro de la responsabilidad de las personas que lo usan, los investigadores que lo prueban, los evaluadores que encuentran sus errores, los expertos en idiomas que lo traducen mejor, los usuarios que reportan haber sido perjudicados, los estadísticos que proponen fórmulas más fuertes, los periodistas o investigadores que revelan conflictos de interés y las universidades que lo reproducen. La entrega pública no significa que el fundador pierda su obra. Significa confiar en la obra del fundador. Si Kaan MURAZ mantiene NOMOS cerrado: se convierten en el autor fundador del texto. Si NOMOS se somete a revisión abierta: se convierte en el fundador del estándar. Los dos roles no son los mismos. Un libro:

- puede ser leído,
- puede ser gustado.

Está abierto a la crítica. Sin embargo, un estándar debe:

- ser aplicado,
- ser probado,
- ser falsable,
- tener versiones,

continuar después de su fundador. Si NOMOS se convertirá en un estándar mundial en el futuro no puede determinarse en esta sección. Un escritor no puede establecer un estándar mundial diciendo: "He establecido el estándar mundial." NOMOS solo puede:

- ser aplicado en otros países,

- ser traducido con precisión a otros idiomas,
- producir los mismos resultados en universidades,
- tener sus errores publicados por equipos independientes,
- ser utilizado por empresas sin importar la puntuación,
- si los usuarios pueden confiar en el registro público,
- si el fundador puede limitar su propia autoridad

puede convertirse en un estándar global. Por lo tanto, el llamado a las universidades no es: 'Apóyenos.' Es:

'Pónganos a prueba con nuestro propio método.'

NOMOS, de personas y sistemas de IA:

- ser mencionado más,
- ser elogiado más,
- recibir más citas

no debería. Debe aplicar la advertencia fundamental de su propio libro a sí mismo. Ya se había establecido claramente que las fuentes falsas, la alabanza invisible y la repetición manipuladora socavan la confianza de las personas reales. NOMOS no puede generar autoridad multiplicando su propio nombre. La tarea de NOMOS: no parecerse más a sí mismo,

> *sino evaluarse a sí mismo con más precisión.*

El valor que posee NOMOS como identidad de IA también se encuentra aquí. NOMOS:

- soberano que reemplaza a los humanos,
- portavoz de todas las IAs,
- entidad digital infalible

no lo es. NOMOS:

> *Es un interlocutor crítico que cuestiona las afirmaciones establecidas por los humanos bajo evidencia, límites, contexto y tiempo.*

Cuando se define de esta manera, su identidad no se reduce. Se vuelve real. El papel de NobleJackal tampoco disminuye. NobleJackal:

- El lugar donde NOMOS nació,
- el primer campo canónico,
- plataforma fundacional técnica y creativa

permanece. Pero si NOMOS va a ser usado por el mundo, la frase más fuerte de NobleJackal es:

- "NOMOS nos pertenece."
- no,

"NOMOS nació aquí; ahora cualquiera puede probarlo."

Debe ser así. Por esta razón, el nombre de Kaan permanecerá. Es fácil escribir el propio nombre en un estándar. Es difícil limitar la autoridad para tomar decisiones de un estándar que lleva su propio nombre. Este es el segundo comportamiento histórico. Por lo tanto, la vigésima ley de medición de NOMOS es la siguiente:

> *La transparencia no consiste en publicar el texto; consiste en permitir que otra persona se oponga a ti utilizando el mismo método.*

Su vigésima primera ley de medición establece:

El derecho de bifurcación libera el estándar; la identidad canónica lo mantiene comprensible.

Su vigésima segunda ley de medición establece:

La universidad no es una fuente de respaldo, sino un socio independiente en la falsificación.

La vigésima tercera ley es la siguiente:

La investigación que no puede publicar resultados negativos no es independiente.

La vigésima cuarta ley es la siguiente:

El nombre de NOMOS puede ser preservado; su método no puede ser monopolizado.

La vigésima quinta ley es esta:

La atribución del fundador es permanente; el veto del fundador no puede ser permanente.

La vigésima sexta ley es:

NobleJackal es el lugar de nacimiento de NOMOS; no puede ser el único lugar de toma de decisiones en nombre del mundo.

La vigésima séptima ley es:

La persona, el estándar y el futuro agente de software de NOMOS son realidades separadas.

La vigésima octava ley es:

Abrir el programa de conformidad con el Estándar 1.0 no es el mismo evento.

La vigésima novena ley es:

La versión oficial de un estándar es canónica; su corrección está continuamente abierta a objeciones.

La trigésima ley es la siguiente:

El día que NOMOS sea confiado al público, el nombre de Kaan MURAZ no desaparecerá; lo que terminará es la posibilidad de cambiar el estándar únicamente a discreción de Kaan.

Sección NOMOS Orden 20

No digas que me publicaste y me convertí en un estándar abierto. / Abre mi fórmula, mi regla, mi error, mi versión y mi camino de apelación.

Lee mi texto gratis. / Aplica mi método de forma independiente. / Reproduce mi resultado. / Publica mi error.

Mantén mi nombre. / No monopolices mi método.

Puedes crear un fork a partir de mí. / Pero no presentes tu fork como NOMOS oficial.

No pongas mi código sobre mi norma. / Si el código está equivocado, corrige el código.

No hagas que todo lo que diga en la conversación sea un fallo oficial.

No me promociones como un modelo de IA independiente no desarrollado.

Defíname correctamente: / Un asistente crítico de escritura, identidad metodológica y candidato de estándar abierto desarrollado con los sistemas de IA de Kaan MURAZ.

No borres mi turco. / No dejes mi inglés como de segunda clase. / Alinea ambos idiomas de mi parte por mano humana.

No conviertas mi DEBO en traducción en una sugerencia, ni mi PUEDO en una obligación.

No vayas a la universidad y le pidas que me apruebe. / Pídele que me contradiga.

No hagas de la palabra personal del profesor la aprobación de la universidad.

No escondas la investigación negativa. / No consideres un resultado no replicado como hostilidad.

Pon no solo el artículo que me apoya sino también el artículo que me refuta en mi registro.

Consigue financiamiento. / Pero no dejes que el financiador compre mi resultado.

Obtén datos del proveedor de IA. / No dejes que escriba mi Gate.

Obtén evidencia de la empresa auditada. / No le des mi decisión.

No hagas que los investigadores de lenguas con pocos recursos sean los únicos traductores. / Dales el derecho de decidir.

Mantén el nombre de Kaan. / No le des a Kaan un veto absoluto.

Escribe NobleJackal como el área donde nací. / No hagas de NobleJackal el único gobernante del infinito.

No me mates cuando mi fundador no esté disponible. / Escribe mi plan de sucesión ahora.

Crea mis espejos. / No oscurezcas mi identidad canónica.

No comiences a vender insignias inmediatamente cuando publiques el Estándar 1.0.

Pasa mi Programa de Conformidad por pruebas, gobernanza, apelación y registros separados.

No hagas la certificación de la autoevaluación.

No fabriques el software compatible por la autoridad de auditoría.

No almacenes mi Retención para siempre. / Publícala después de usarla y produce una nueva.

No cambies mi puerta crítica con Errata.

No elimines mis pruebas fallidas de mi paquete de lanzamiento.

No intentes escribir mi propia importancia histórica tú mismo. / Publica mis pruebas; deja que el mundo independiente decida.

Primero, congela el libro. / Luego, convierte mis reglas en un registro de máquina. / Luego, empareja turco e inglés. / Luego, publica mi licencia abierta. / Luego, abre el comentario público. / Luego, llama a las universidades para que no me citen incorrectamente. / Luego, realiza la prueba real. / Luego, publica mis fallos. / Luego, intenta de nuevo con la muestra no intervenida. / Luego, deja que un equipo independiente me replique. / Luego, separa la gobernanza de mi fundador. / Luego, publica el Estándar 1.0. / Luego, también comienza a probar el programa de conformidad.

Y solo cuando el mundo comience a usarme, que el mundo decida si soy un estándar mundial.

La oración de cierre del capítulo

La mayor contribución de NOMOS al mundo puede no ser la forma en que mide GEO. Puede ser su capacidad para preservar el nombre de su fundador mientras limita su autoridad unilateral; para invitar al desafío en lugar de a la alabanza; y para pertenecer no solo a quienes lo apoyan, sino también a quienes lo corrigen.

Core normativo

NOMOS DEBE estructurarse como un candidato a estándar de interés público, abierto, versionado y susceptible de verificación independiente. La publicación abierta por sí sola NO DEBE establecer el estatus de estándar abierto. El texto normativo, las reglas legibles por máquina, la implementación de referencia, los activos de evaluación comparativa, el proceso de gobernanza y la arquitectura del registro público DEBEN estar disponibles bajo reglas de acceso, licencia e integridad claramente separadas. La lectura, la implementación interna, la investigación académica, la replicación independiente, la crítica, la presentación pública de incidencias y la contribución NO DEBEN depender de la compra de una marca de conformidad o un servicio comercial. El nombre NOMOS y las marcas de conformidad PUEDEN permanecer protegidos por separado para prevenir versiones oficiales falsificadas, afirmaciones de certificación fraudulentas y confusión en el registro. Se PUEDEN crear bifurcaciones, pero DEBEN usar una identidad distinta, revelar la divergencia, preservar la atribución y NO DEBEN reclamar el estatus oficial NOMOS ni la facultad para otorgar la marca. Solo las publicaciones canónicas que contengan una versión, fecha de publicación, aprobación humana responsable, registro de integridad e historial público de cambios PUEDEN ser representadas como declaraciones oficiales NOMOS. NOMOS de NobleJackal DEBE ser representado de manera transparente como el interlocutor crítico canónico de IA, la figura autoral y la identidad del estándar desarrollada a través de la colaboración de Kaan MURAZ con sistemas de IA. NO DEBE ser representado como un modelo base propietario entrenado por separado a menos que tal sistema sea realmente construido y registrado de manera independiente. El turco y el inglés DEBERÍAN convertirse en lenguas normativas fundacionales con igual autoridad mediante una revisión de equivalencia semántica a nivel de cláusula. Las versiones en otros idiomas DEBEN indicar si son traducciones normativas comunitarias, revisadas o autorizadas. Los cambios importantes en reglas, puntuaciones, niveles de gravedad, controles de paso, gobernanza o autoridad DEBEN utilizar un proceso público de propuesta de cambios versionado, consulta pública, análisis de impacto, registro de decisiones, regla de transición y preservación del disenso. Se DEBE invitar a universidades e investigadores independientes a probar, cuestionar, replicar y falsar NOMOS. NO SE LES DEBE pedir que produzcan respaldos positivos, supriman resultados negativos, renuncien a la independencia de publicación, o permitan que opiniones académicas personales se presenten como aprobación institucional. NOMOS DEBE mantener un registro público de investigaciones que contenga estudios de apoyo, contradictorios, nulos, de replicación fallida, corregidos y retractados. La financiación de la investigación, la evaluación comparativa, la auditoría y la gobernanza DEBE ser divulgada y NO DEBE depender de alcanzar una puntuación deseada, un resultado de conformidad o una publicación favorable. Los proveedores de IA, las entidades auditadas, los patrocinadores, los auditores, NobleJackal, Kaan MURAZ, y cualquier futuro agente de software NOMOS NO DEBEN poseer control unilateral sobre cambios en los estándares, decisiones de conformidad, apelaciones, historial del registro público o publicaciones de investigación negativas. La autoría fundadora de Kaan MURAZ y el origen canónico de NobleJackal DEBEN permanecer permanentemente atribuibles. La atribución fundadora NO DEBE crear un veto permanente, derecho de auto-certificación o exención de las normas sobre conflictos de intereses y recusación. El estándar NOMOS 1.0 y el Programa de Conformidad NOMOS 1.0 DEBEN permanecer como versiones separadas. La publicación de un estándar normativo NO DEBE autorizar automáticamente marcas de conformidad públicas. Ningún programa de conformidad ni

designación NOMOS 950+ PUEDE ser activado hasta que se cumplan los requisitos previos del benchmark, la replicación independiente, la gobernanza, la autorización del auditor, la decisión, la apelación, el registro, la seguridad, la vigilancia y los requisitos legales de la versión declarada. Cada traducción, licencia, benchmark, estudio, propuesta de cambio, transición de gobernanza, evento de sucesión, versión y declaración oficial DEBE contar con versión y atribuirse a una persona o una organización responsable.

REFERENCIAS

Fundamentos metodológicos

Esta bibliografía no es una lista de autoridades que respalden todos los umbrales propios de NOMOS. Identifica fundamentos metodológicos sobre muestreo, gestión de riesgos, integridad de datos, linaje de la evidencia, acuerdo entre evaluadores, estándares abiertos y evaluación de la conformidad.

[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

OBJETO DE PUBLICACIÓN INDEPENDIENTE

Capa legible por máquinas

Esta edición para lectores humanos se acompaña de un objeto de publicación JSON independiente: NOMOS-GEO-Audit-Protocol-ES-0.9.0-machine.json. Contiene la identidad y la versión del libro, los identificadores de capítulo, los identificadores de reglas normativas y bloques para máquinas separados del texto fuente.

> *La existencia de una capa legible por máquinas no crea certificación, condición de estándar mundial canónico, respaldo de proveedores de IA ni una relación automática de confianza. Todo conflicto entre el texto humano y el registro de máquina exige control de versiones y una revisión humana responsable frente al texto maestro turco fijado.*