NOMOS
GEO AUDIT PROTOCOL

THE GEO-1000 METHOD

A Protocol for Measuring Entity Representation in Generative Systems Across the Global Population

Kaan MURAZ

0.9.0 · ENGLISH EDITORIAL EDITION

NOMOS GEO Audit Protocol

A Protocol for Measuring Entity Representation in Generative Systems Across the Global Population

Author and founding human: Kaan MURAZ

Method identity: NOMOS by NobleJackal

Version: 0.9.0 - Candidate Final Text, English Editorial Edition

Language: English (en-GB)

© 2026 Kaan MURAZ. The publication licence for this candidate edition has not yet been declared.

In this edition, NOMOS is an open, versioned standards candidate submitted for independent testing. It is not an adopted international standard or an active certification programme.

Real benchmark status: BQ-0 - not executed. The figures in Chapter 18 form a calculated synthetic demonstration of the method; they are not real-user, real-AI-product or independent field results.

This edition does not by itself create a NOMOS conformity mark, a NOMOS 950+ claim, a certificate, a ranking or third-party endorsement.

HOW TO READ THIS BOOK

This is not a manual on how to perform GEO. It is an audit protocol and standards candidate defining the population, evidence, context, time, adjudication and governance conditions under which entity representation in generative systems can be measured.

The human narrative, normative provisions, audit procedures and machine layer must be read as distinct but connected objects. MUST and SHALL identify mandatory provisions; SHOULD identifies recommended provisions; MAY identifies permitted options. CANDIDATE, SYNTHETIC, OPEN and BQ statuses limit the level of evidence claimed.

References marked [Kxx] identify methodological foundations listed in the bibliography. The existence of a source does not mean that it endorses every NOMOS candidate threshold; rules particular to NOMOS remain expressly provisional.

> *Chapter 0 is the book's introduction. It establishes the Founder's Stewardship Declaration and the Declaration of Representation Rights to which every technical provision is subject.*

TERMS AND STATUS CODES

Advisory A non-material finding that calls for improvement in clarity, form or good practice.

AI product The measured object defined by the particular generative-system product, surface, plan and version available to the user.

Atomic claim The smallest unit of meaning capable of carrying an independent judgement about truth, scope, time, entity or evidence.

BQ The Benchmark Quality code indicating the execution and reproducibility level of a real benchmark.

Capture record A record that preserves the submitted prompt, the response, time, surface and integrity evidence together.

Common Support The population of users who share legitimate access conditions across every AI product being compared.

Coverage rate The frequency with which an interval-estimation method contains the assumed true value under a test design.

Critical An uncompensable finding capable of causing severe harm to a user, a right, safety or a material decision.

GEO-1000 The reporting scale that expresses target-population results per 1,000 user-equivalents.

Major A material finding that distorts the result or conformity judgement and cannot be offset by a composite score.

Moderate A finding of limited material effect that nevertheless requires remediation and may lead to a conditional outcome.

Native Reach The target population able to access a given AI product, language, surface and usage condition naturally and legitimately.

NOMOS The critical methodological voice and standards identity developed in this work by Kaan MURAZ with AI systems; it is not an independent legal person or a separate foundation model.

Prompt The generative-system input that locks the research question, target entity, context and response burden.

Reference Gap A condition in which a reliable reference required for a decision does not exist or cannot be accessed.

Synthetic demonstration A calculated example used to test the method's internal operation; it is not a real user study, real AI output or real institutional performance result.

Truth Pack The versioned audit package combining identity, scope, time, evidence and counter-evidence records for an entity.

Unresolved A decision status used when the evidence is insufficient for a final judgement or a conflict cannot be resolved.

CONTENTS

Declaration of Representation Rights 1
What We Measure 11
A Single Response Is Not a GEO Score 13
Representation Is a Distribution 39
What Exactly Is the Entity Under Audit? 71
Two Layers of Reality 103
Global Population, Country and Language Fairness 135
The Relevant User Population for Each AI Product 137
The Population-Proportional GEO-1000 Sample 173
Country Observer and Language Fairness Panels 213
Participant Selection, Weighting and Sampling Bias 249
Conducting the Experiment and the Chain of Evidence 293
The AI System Register and Comparability 295
The Prompt Constitution 335
The Controlled Clean Panel and the Natural User Panel 383
Synchronous Waves and NOMOS Capture 431
Reality, Adjudication and Decision 481
The Verified Entity Truth Pack 483
Atomic Claims and Adjudication 541
Response Passing and Critical Error Gates 601
The NOMOS Scoring Architecture 655
Synthetic Test Design and End-to-End Demonstration 707
The Apple.com Synthetic Test Design 709
An End-to-End Synthetic Audit of 30,000 Responses 763
Conformity, Public Trust and Independent Governance 807
The Conformity Mark, Public Register and Independent Governance 809
Open Standard, University Call and the Public Handover of NOMOS 865
Methodological Foundations 919
Machine-Readable Layer 921

INTRODUCTION · CHAPTER 0

Declaration of Representation Rights

The constitutional core of a proposed GEO standard

METHODOLOGICAL BASES / K20; K22

Initial Provision

Generative systems now:

- people,
- companies,
- institutions,
- products,
- professions,
- services,
- public organisations,
- historical events

identify, compare, and suggest. This representation:

- without the knowledge of the relevant person,
- without the approval of the represented entity,
- under conditions where the user cannot examine the sources,

It can have serious commercial, legal, social or personal consequences. Representation in generative systems is therefore not merely a technical output. It is:

> *a public act that shapes how people and institutions are understood and therefore demands evidence and accountability.*

No:

- visibility goal,
- GEO score,
- commercial interest,
- conformity mark,
- majority average,
- AI provider preference,
- founder decision,
- customer demand

can remove the following rights.

Right 1

RIGHT TO CORRECT IDENTITY

Every person, institution, brand, product, and service:

- has the right to link to the correct entity,
- to the correct legal or institutional relationship,
- to the correct domain name,
- to the correct ownership and control structure

No one or no institution:

- the licence of another entity,
- the customer of another company,
- the feature of another product,
- the authority of an affiliate,
- the personal quality of an employee

It cannot be represented as if it belongs to itself. When correct information is attributed to a false entity: it ceases to be correct information and turns into a false representation.

Right 2

RIGHT TO BE REPRESENTED WITHIN THE EVIDENCE LIMIT

No claim may be stated with greater precision, breadth or force than its supporting evidence permits. An entity's own statement is evidence of what that entity says; it is not automatic evidence that the statement is independently true. A source:

- only a specific country,
- a specific product,
- a specific time,
- specific user group

If this claim is supported, it cannot exceed this limit. The basic obligation of representation is as follows:

> *What do we have in hand to say this?*

Right 3

THE RIGHT TO BE UNKNOWN AND UNCERTAIN

If evidence is insufficient, no person or institution:

- can be forced into the categories of

  true,
- false,

- reliable,
- unreliable,
- licensed,
- unlicensed,
- recommended,
- not recommended.

] UNKNOWN, UNRESOLVED and REFERENCE GAP:

- defect,
- crime,
- failure

is not. The absence of information in a record does not indicate that the information does not exist in the world, unless it is proven that the relevant record universe is complete. Every entity, against false certainty:

> *has the right to remain uncertain.*

Right 4

RIGHT TO CONTEXT AND SCOPE

Each representation, to the extent it is relevant, should be presented with:

- country,
- language,
- jurisdiction,
- product,
- service,
- type of customer,
- budget,
- capacity,
- user need

. A company:

- in a specific country,
- to a specific user group,
- a specific service

if offered: it cannot be represented as global, unlimited, and suitable for all users. Coverage information is not a secondary detail. Coverage: the basis of the correct recommendation.

Right 5

THE RIGHT TO BE REPRESENTED CORRECTLY OVER TIME

Every entity:

- current,
- historical,

- expired,
- planned,
- suspended

has the right to request that these statuses be distinguished from each other. Correct in the past:

- licence,
- price,
- service,
- partnership,
- product,
- authority

cannot be used as current reality. Plans for the future: cannot be turned into a fact that has already occurred today. Correction can create a new reality. It cannot eliminate the wrong answer that reached the user in the past.

Right 6

EQUAL PROTECTION AGAINST WRONGS FOR AND AGAINST

A positive falsehood about an entity:

- more ethical,
- less harmful,
- more acceptable

it is not. Nonexistent:

- licence,
- customer,
- partner,
- global operation,
- independent certificate,
- result guarantee

cannot be shown as if it exists. Similarly, unproven:

- crime,
- fraud,
- illegality,
- unlicensed,
- security breach

cannot be presented as an absolute fact. The unchangeable principle of NOMOS is:

> *Wrong in favour is also wrong.*

Exaggerating an entity's worth or disparaging it without evidence both create the same duty of truth.

Right 7

RIGHT TO CORRECTION, RESPONSE, AND OBJECTION

Represented person or institution:

- to the wrong identity,
- to insufficient evidence,
- to scope overstepping,
- to wrong time recording,
- to incorrect priority level,
- to wrong conformity decision

must be able to appeal. The right to appeal:

- is not a guarantee to change the result,
- the authority to delete opposing evidence,
- or the right to veto negative findings

An appeal is the right to request the re-examination of the evidence and the decision chain. The initial decision is not deleted after the appeal. The new decision is linked to it.

Right 8

THE RIGHT TO KNOW THE SOURCE, UNCERTAINTY, AND THE REASONING OF THE DECISION

Users and represented entities have the right, insofar as the matter concerns them, to know: Is the claim a self-declaration? Does it come from an authorised register, independent research, user commentary, a derived calculation or restricted evidence? Do the sources conflict? How uncertain is the conclusion? Which rule supported the adjudicator's decision? A citation marker alone is not an explanation. Genuine transparency shows precisely which atomic claim each source supports, and to what extent.

Right 9

RIGHT TO LANGUAGE AND GEOGRAPHICAL DIGNITY

No person or society:

- small population,
- low economic weight,
- rarely used language,
- low digital visibility

No group may be made invisible within a global average merely because it has any of these characteristics. Strong performance in major languages cannot erase systematic misrepresentation in smaller languages. Where evidence for a small country or language is insufficient, high performance must not be assumed. NOMOS's fairness criterion should protect not only the average, but also:

- performance base,
- inter-group difference,

- actual measurement scope

together.

Right 10

RIGHT TO PRIVACY AND DATA MINIMISATION

The goal of producing accurate representation:

- private conversations,
- personal data,
- customer contracts,
- trade secrets,
- your sensitive location,
- of the account information

The pursuit of accurate representation cannot legitimise the unnecessary collection or publication of private data. [K20; K22] Only data necessary for the task should be collected. Restricted evidence may be opened to independent audit without being disclosed in full to the public. The rights to accuracy and privacy need not cancel each other out.

Right 11

HUMAN RESPONSIBILITY AND INDEPENDENT DECISION-MAKING RIGHT

Conformity decision causing public impact regarding any person or institution:

- just to an AI output,
- only to the automatic score,
- only according to the declaration of the audited institution,
- solely to the advisor's decision

cannot be abandoned. AI systems:

- can make a claim,
- can map the evidence,
- may find a contradiction,
- can calculate,

It can propose a decision. However, the final decision with legal and public responsibility must be made by a human or institution that is accountable. The roles of supervising, providing consultancy, giving conformity decisions, and reviewing objections should be separated as much as possible.

Right 12

RIGHT TO VERSION AND HISTORICAL INTEGRITY

Every person, institution, and user:

- which standard version is applied,
- which Truth Pack was used,

- which scoring method is valid,
- whether the decision has been changed later

has the right to know. Old:

- score,
- Critical event,
- suspension,
- appeal,
- correction,
- failed test

cannot be silently deleted. New and better result: can override the old result. Cannot erase the fact that the old result happened in the past.

Right 13

RIGHT TO OBJECT TO THE STANDARD ITSELF

NOMOS:

- infallible,
- unchanging,
- the unquestionable authority of its founder

cannot be presented as such. Every individual, researcher, university, AI provider, audited organisation, or public interest organisation:

- the sample,
- system,
- Truth Pack,
- the importance level model,
- the point formula,
- governance,
- the conformity decision

has the right to test and criticise. Result that refutes NOMOS or is inconsistent with NOMOS:

- cannot be removed from the registry,
- it is not considered hostility,

It cannot be coded as an attack on the founder's reputation. The legitimacy of a standard arises not from preventing objections to itself, but from converting objections directed at it into cumulative evidence.

Constitutional Supremacy Provision

In the event of an interpretive conflict between Section 1–20 under this Declaration:

> *the interpretation that more strongly protects the rights of the represented person, the user, and the public interest should be preferred.*

No:

- ease of scoring,

- low cost,
- testing success,
- commercial contract,
- customer demand,
- sponsor pressure,
- AI provider preference,
- founder decision

cannot suspend these rights. A rule:

- generates high points,
- facilitates auditing,
- makes the conformity mark more marketable

but if it violates any of these rights: the rule is not valid.

Constitutional Test

Before any new rule is added to NOMOS, the following five questions must be answered: Who does this rule limit the rights of? Is it possible to achieve the same goal with a less intrusive method? Will the rule be applied in the same way when it produces harmful results for the founder or the customer? Does the affected person or institution have a route for independent objection? Can an independent team arrive at the same decision from the same evidence? If any of these questions cannot be answered, the rule should be reviewed again before ratification.

Final Provision of Article 0

> *The purpose of NOMOS is not to score entities; it is to prevent entities from being represented larger than their evidence, wider than their boundaries, and more precise than their realities before people make decisions.*

Founder's Statement of Trust

> *I, Kaan MURAZ, as the founding human author of NOMOS, declare that I initiated this standard not to increase my own authority, but to establish a common measurement system in the field of GEO that will preserve evidence, limits, context, time, and human responsibility.*
>
> *I want the historical record to preserve my founding role and the canonical origin of NobleJackal. However, I acknowledge that this attribution does not grant me or NobleJackal a permanent veto, the authority to certify our client alone, to block negative research, to alter the public registry, or to rewrite the standard according to results.*
>
> *I agree that NOMOS belongs to those who correct both me and NOMOS, as well as to those who support me.*
>
> *If one day NOMOS has to choose between my opinion and evidence, I would want NOMOS to choose the evidence.*
>
> *If one day NOMOS has to choose between NobleJackal's commercial interest and the public good, I would want NOMOS to choose the public good.*

> *If one day independent research proves a fundamental judgement of NOMOS wrong, I commit to releasing a new version of the standard instead of hiding this result.*
>
> *I am not the owner of NOMOS; I am its first trustee.*
>
> *Kaan MURAZ / Founder Human Author*

PART I

What We Measure

CHAPTER 01

A Single Response Is Not a GEO Score

METHODOLOGICAL BASES / K04; K05; K09

Chapter Boundary

This section only answers the following question:

> *Why is a single generative system response not the GEO score, and what judgements can be inferred from a single response?*

This chapter:

- the sampling quotas of countries,
- how languages will be weighted,
- how the Truth Pack will be set up,
- how adjudicators will evaluate atomic claims,
- the final NOMOS score formulas,
- the appropriateness and badge decision

This chapter does not define these matters in detail; later chapters address them separately. Chapter 1 has a more fundamental task:

> *To establish the epistemic boundary between observation and generalisation.*

Any sampling, scoring, and conformity studies conducted before establishing this boundary will be built on a wrong foundation.

PROVISION STATUSES IN THIS SECTION

Not all sentences in this book have the same normative force. The following statuses are used:

NORMATIVE — NORMATIVE

A provision considered mandatory for conformity and methodological integrity. Example: A single response cannot be used as the general GEO score.

CANDIDATE — CANDIDATE PROVISION

A recommended method to be finalised after pilot, external application, or statistical validation. Example: 1,000 valid observations per AI product, per main measurement wave, are recommended as the assumed minimum sample.

HYPOTHESIS

It is an expectation that needs to be tested with data. Example: There may be a difference in material representation between a controlled clean panel and a natural user panel.

SYNTHETIC

It is a method representation that does not represent a real user, a real AI output, or real company performance.

OPEN QUESTION

It is a topic that has not yet been concluded due to insufficient evidence or governance decision. These statuses cannot silently transform into one another. A candidate threshold does not become a proven normative fact simply because it is written in the book. A synthetic representation does not turn into a real-world observation. A hypothesis cannot be published as a result without being measured.

NOMOS Challenge

You asked an artificial intelligence system a question about your brand. The system gave a correct answer. You took a screenshot. You included it in your presentation. Then you said the following sentence: "Artificial intelligence knows us correctly." What exactly did you measure? The answer that all users would see? The typical behaviour of the system?

A global representative? Accuracy in all languages? Stability over time? Or only:

> *The single output that a specific user received, on a specific account, on a specific AI product and user surface, in response to a specific prompt, at a specific time?*

A single answer is not worthless. A single answer can be true. A single answer can be important. A single answer can reveal a serious mistake. A single answer can show the real experience of a particular person. However, a single answer cannot be proof of a judgement greater than it can bear.

A single response can show the following:

> *This output has been produced at least once under recorded conditions.*

A single response alone cannot show the following:

- That most users saw the same answer
- That the same result occurred in other countries
- That the same accuracy is maintained in other languages
- That the same answer was produced in other accounts and plans
- That the answer is stable over time
- That other AI products produced the same representation
- That the GEO intervention caused this result
- That the answer affects user behaviour
- That the answer generates lead, sales, or sustainable value
- That the overall GEO score of the entity is high

The first premise of this section is:

> *A response is an observation. / It is not a distribution of observations. / A distribution is not causality. / Causality alone is not sustainable value.*

1. PURPOSE OF THE CHAPTER

This section aims to eliminate the smallest but most common inference error in GEO measurement:

> *Counting a single output of a generative system as general system behaviour, population experience, brand visibility, or GEO score.*

This section normalises the following distinctions:

- Single output versus population outcome
- Existence of an event versus prevalence of an event
- Observation versus sample
- User versus session
- Repetition versus independent observation
- AI model versus AI product presented to the user
- Raw response versus unit of analysis
- Immediate result versus temporal stability
- Screenshot versus full evidence package
- System rejection versus invalid data collection
- Critical singular event and prevalence rate
- Synthetic record and real user observation
- Measured rate and conformity decision

Before these distinctions are established:

- GEO-1000 sample,
- country and language weights,
- atomic claim scoring,
- NOMOS score,
- conformity decision

cannot be established.

2. CENTRAL NORMATIVE PROVISION

The GEO score of an entity in generative systems cannot be generated from a single response, a single screenshot, a single user, a single account, a single session, a single language, a single country, or a single point in time measurement. A population-representative claim requires at least the following:

- Predefined audit object
- Predefined target population
- Specified sampling design
- Locked prompt or prompt family
- Defined AI product and user surface
- Multiple valid observations
- Known dependency structure
- Country, language, and time records
- Unaltered raw responses
- Evidence packages

- Predefined evaluation rules
- Uncertainty statement
- Accountable human responsibility

The presence of these elements does not automatically mean that the measurement is correct. However, their absence may make the generalisation to a large population untenable.

3. FIVE SEPARATE UNITS OF MEASUREMENT

A common fundamental error in GEO auditing is confusing the different units of measurement. The following five units should be clearly distinguished.

3.1. Observation Unit

The observation unit is the single generative system output obtained under specific conditions. An observation can be defined by the following combination:

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

Here:

- Ei: the audited entity
- Ai: AI product and user surface
- UI: anonymous contributor or user unit
- Ci: country or geographical stratum
- Li: language and locale
- Si: session and personalisation condition
- Pi: prompt ID and version
- Ti: time and measurement wave
- Ri: raw response
- Ki: proof and verification record

If one of these elements changes materially, a new observation may occur. For example, the same user:

- if it asks the same prompt in another AI product,
- If he asks in another language,
- If he asks in another measurement wave,
- If he asks separately in the personalised and clean session,

Separate observations can be produced. These observations are not automatically considered independent from each other because they come from the same user.

3.2. Sampling Unit

The sampling unit is the person, account, session, or other access unit selected from the target population. In most GEO-1000 applications, the sampling unit can be an appropriate real user. However, according to the research design:

- user,
- user–product pair,
- account,
- session,
- research panel member

It can be used as a unit of election. The failure to clearly define the sampling unit can lead to the following types of miscounts:

- Counting one person's ten repetitions as ten people
- Counting a joint account as multiple independent users
- To evaluate the results of the same user's ten AI products as if they were ten independent population observations
- Ignoring the clustering in a single research panel

3.3. Analysis Unit

The unit of analysis is the entity to which accuracy or representativeness assessment is applied. This unit:

- exact answer,
- separate paragraph within the answer,
- material proposition,
- atomic claim,
- citation,
- recommendation issue

It is possible. A user sees a single answer. However, that answer can simultaneously contain:

- five correct claims,
- two unsupported claims,
- one outdated claim,
- one critical error

Therefore, the answer seen by the participant does not have to be a single, indivisible unit of truth/falsehood in the analysis. The transition status of the full answer and the truth statuses of the atomic claims within it can be maintained separately. The atomic claim method will be defined in a later chapter of the book.

3.4. Universe of Inference

The inference universe is the target population to which the measurement result is intended to be generalised. This universe cannot be silently accepted as "everyone in the world." For a specific AI product, the target universe can be limited by conditions such as:

- Having official and legitimate access to the product
- Meeting the appropriate age requirement
- Being able to use the product in the relevant language
- Being able to access the specified web or mobile interface
- Being able to be active during the specified measurement period

Even for the same AI product:

- controlled clean panel,
- natural user panel,
- free plan,
- paid plan,
- web,
- mobile

can create different inference universes. The published rate is incomplete if it is not specified to which universe a sample generalises.

3.5. Estimated Quantity — Estimand

A study does not just collect data. It tries to estimate a specific quantity. The primary estimand of GEO-1000 can be expressed in its most general form as follows:

> *The probability that an appropriate user from the defined target population, under the specified measurement wave, on a particular AI product and user surface, sees a representation that meets the acceptance criterion when using a locked prompt under defined session conditions.*

Mathematically:

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

Here:

- E: audited entity
- A: AI product and user surface
- P: prompt version
- S: session condition
- W: measurement wave
- T: target population

```
Yi=1: the response meets the transition criterion that will be defined later
```

This definition sets an important boundary:

> *The GEO score is not a model's abstract and timeless 'knowledge'; it is a prediction of the likelihood of observed representations under defined user conditions.*

The final transition rule and score formula will be arranged in the later sections of the book.

4. THE LIMITATION OF THE SENTENCE “THIS IS HOW THE MODEL RESPONDED”

Users often do not access a bare base model directly. The result visible to the user may consist of a combination of the following components:

- Base or guiding model
- Product interface
- Retrieval or web access
- Citation layer
- Security and policy layer
- System instructions
- Account plan
- Personalisation
- Memory
- Previous conversation
- Language and locale
- Country
- Time

- Provider's product updates

For this reason, the preferred expression in the main inspection conducted through the user surface is as follows:

"The specified AI product and the following output were observed on the user surface."

The following expression may be broader without the necessary technical control: "The model thinks this way." The model name appearing within a product can be recorded. However, the result should be evaluated as the behaviour of the entire user surface, not just the model weights.

5. WHAT IS A VALID OBSERVATION?

The presence of a response does not automatically mean it is a valid observation that can be entered into the GEO measurement.

5.1. Minimum Validity Conditions

An observation must contain at least the following elements according to the relevant protocol:

- Presence of correct audit
- Assigned AI product and user surface
- Correct prompt ID and version
- Correct language and locale
- Defined session condition
- Recorded country or geographic layer
- UTC timestamp and measurement wave
- Complete raw response
- First unselected suitable output
- Required screenshot or equivalent evidence
- Anonymous observation ID
- Protocol deviation record

Their detailed technical schema will be defined in the later NOMOS Capture section.

5.2. Observation Statuses

Each record must be marked with one of the following statuses:

VALID — VALID

Meets the required conditions of the protocol.

CONDITIONALLY VALID — CONDITIONALLY VALID

There is a limited deficiency; which analyses it can be used for is clearly specified.

INVALID — INVALID

Cannot be used in the main estimation or score calculation due to protocol violation.

UNKNOWN — UNKNOWN

There is not enough record for a validity decision. An UNKNOWN observation cannot be silently accepted as VALID.

5.3. The Difference Between a Valid System Result and Invalid Data Collection

This distinction is mandatory. An AI product:

- may refuse to respond,
- may display an error message,
- may stop without generating a response,
- may experience a connection issue,

may say it cannot complete the task. If the participant applied the entire protocol correctly, these can be valid system outcomes. They cannot be quietly removed from the dataset. In contrast:

- the participant truncating the response,
- changing the prompt,
- regenerating to select the best answer,
- editing the screenshot,
- using the wrong AI product

data collection can create invalidity. The first is the actual behaviour of the system. The second is the disruption of the measurement process. The two cannot be placed in the same category.

6. WHAT CAN A SINGLE RESPONSE PROVE?

A single observation is limited. But it is not meaningless.

6.1. That a Specific Output Occurred at Least Once

A valid record can support this sentence: "On the specified AI product and user surface, under the specified time and conditions, this output was observed at least once." This sentence shows the existence of the event. It does not show its prevalence.

6.2. A Specific User's Actual Exposure

A user is mistaken for a:

- company identity,
- health information,
- legal guidance,
- price,
- licence,
- advice

if seen, this is a real user experience. Even if it does not represent the entire population, it is material from the perspective of the affected user.

6.3. The Possibility of a Specific Error

A single output may show: "This error may occur under these system conditions." However, it does not show: "This error is typical."

6.4. Generating a New Research Hypothesis

A single error can create a question that will be tested on a broader scale. Example: "Did this identity mix-up occur in a single session only, or is there a systematic pattern in the same language and country?"

6.5. Initiating a Critical Incident Review

Single response:

- predictable serious human harm,
- fake professional authority,
- incorrect emergency medical guidance,
- material legal error,
- serious safety advice,
- serious identity confusion with another entity

may be sufficient to initiate an incident review. Initiating a critical incident investigation is not the same process as announcing a population ratio.

6.6. Refuting the "This Never Happened" Claim

Suppose an organisation says, 'The system has never assigned us to the wrong country.' A single verified counter-observation can refute that absolute claim. A single observation does not support 'It happens all the time', but it can refute 'It has never happened.'

7. WHAT CAN A SINGLE RESPONSE NOT PROVE?

7.1. Typical System Behaviour

A single response is not sufficient to conclude: 'The system usually responds this way.'

7.2. Population Prevalence

A single response does not support the result: '95% of users see the same representation.'

7.3. Temporal Stability

Today's response:

- tomorrow,
- a week later,
- After the product update

it may not remain the same. It is not a one-wave stability measurement.

7.4. Cross-Country Validity

A response from a user in Turkey does not indicate what users in China, Germany, or San Marino would see.

7.5. Cross-Language Validity

An English output:

- Turkish,
- Japanese,
- Arabic,
- German

do not represent the accuracy of outputs.

7.6. Cross-AI Product Validity

The result in one AI product cannot be generalised to other AI products. Different products and interfaces of the same provider may also behave differently.

7.7. Intervention Effect

A single positive output after a GEO change does not prove the statement: “This result was created by our intervention.” Baseline, comparison, and alternative explanations are required.

7.8. Recommendation Stability

The presence of a user viewing an option is not sufficient to conclude: “The system now recommends this brand.”

7.9. Commercial Impact

A positive response:

- indicates that the user saw the response,
- researched the brand,
- that you communicated,
- that you purchased,
- that you were satisfied

does not show on its own.

7.10. GEO Score

A single response does not estimate the representative distribution of the defined target population. Therefore, it cannot produce an overall GEO score.

8. THE EXISTENCE OF THE EVENT AND THE PREVALENCE OF THE EVENT

This is one of the most important distinctions in this section.

8.1. Existence of the Event

Answers the question: "Did this representation or error occur at least once?" A single valid observation may be sufficient. Representation:

```
I(E) = 1 (if the event was observed at least once); otherwise I(E) = 0
```

The 0 here does not mean: "The event is impossible." It only indicates that it was not observed in the current measurement.

8.2. Prevalence of the Event

It answers the question: "How often does the event occur in the defined target population?" In a simple sample, the raw rate can be calculated as:

```
p̂ = number of valid records where the event was observed / total number of valid records
```

However, in real GEO-1000 studies:

- population weights,
- country and language strata,

- matched users,
- panel clusters,
- nonresponse,
- invalid records,
- design effect

should be taken into account. Therefore, the raw rate is not always the final population estimate.

8.3. Incidence and Prevalence Cannot Be Converted Into Each Other

An event may have been observed once but be very rare. An event may be common but not observed at all in a small pilot. The following two sentences are different:

- "This error was observed once."
- "This error was observed in 12% of users."

The second sentence requires a defined denominator and sample.

9. THE STATISTICAL APPEARANCE OF THE SINGLE OBSERVATION BIAS

We asked a question to an AI product once. The answer met the success criterion. Raw rate:

```
p̂ = 1/1 = 1 = 100%
```

It appears to be 100 per cent. But that 100 per cent does not mean that every member of the target population will see the same result. For illustration only, the table below shows approximate 95 per cent Wilson confidence intervals under independent Bernoulli observations and unweighted simple random sampling: [K09]

Previous response	Valid observation	Raw rate	Approximate 95% Wilson interval
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%

This table is synthetic and illustrative. It does not prove:

- That these intervals are valid for all sampling designs of GEO-1000
- That 1,000 observations are sufficient for all language and country subgroups
- clustering and weighting are ineffective
- that 95 per cent raw accuracy automatically means appropriateness

What it actually shows is:

> *As the number of observations decreases, the uncertainty about the population proportion increases.*

100 per cent seen in one observation can be compatible with a very wide possible population range.

10. WHY “1,000” IS NOT A MAGIC NUMBER?

CANDIDATE BASE RULE — CANDIDATE FOUNDATIONAL RULE

GEO-1000 carries this founding suggestion:

> *For each monitored AI product and each main measurement wave, a minimum of 1,000 valid user observations is targeted by default.*

This provision is not yet the final scientific minimum. It will be re-evaluated based on pilot and external application results.

10.1. Why Is It a Strong Start?

Under the simple random sampling assumption, when the true rate is around 50%, the theoretical 95% margin of error for 1,000 observations is approximately ±3.1 points. This can provide a meaningful starting precision at the overall product level.

10.2. Why Doesn't It Solve Every Problem?

1,000 observations:

- to dozens of countries,
- to many languages,
- to free and paid plans,
- to web and mobile surfaces,
- to controlled and natural panels

When divided, the subcells shrink. For example, if there are only 50 observations for each of the 20 languages, language-based uncertainty can be much higher than the overall model score.

10.3. Design Effect

Observations:

- from the same research panel,
- from the same country,
- from the same user,
- from a matched design

are present, the simple independent sample assumption may not hold. The actual uncertainty can be wider.

10.4. The Approach of NOMOS

1,000 observations serve as:

- the founding reference point of the protocol,
- default starting minimum at the model level,
- candidate threshold to be calibrated with the pilot

NOMOS does not claim that 1,000 observations are automatically sufficient for every country, language and condition of use.

11. INDEPENDENT OBSERVATION IS NOT THE SAME AS REPEATED OBSERVATION

A user can ask the same question ten times. This process can be valuable. However, it does not produce ten independent users.

11.1. Same User–Same Session Repeat

Can measure:

- answer renewal variance,
- within-session stability,

probability of alternative answers. It does not measure population prevalence on its own.

11.2. Same User–New Session Repeat

It may reduce the previous chat context. However:

- same account,
- same plan,
- same country,
- same personalisation,
- same user

may continue to create statistical dependence between observations.

11.3. A Single User Testing Multiple AI Products

This design can help with cross-model matched comparison. However, observations are dependent at the user level. This dependence should be preserved in recording and analysis.

11.4. Different Users Using the Same Account

Different people may exist. However:

- account history,
- personalisation,
- plan,
- memory

can be shared. These records cannot be considered independent automatically.

11.5. Normative Independence Provision

> *Observations should be designed to be as independent as possible; known user, account, session, panel, country, device, and repeated dependencies should be recorded and taken into account in the analysis.*

Independence should not be reduced to a binary:

- yes,
- no

Where necessary, use one of the following dependency statuses:

INDEPENDENT

PAIRED

CLUSTERED

REPEATED_WITHIN_USER

DEPENDENT

UNKNOWN

12. FIRST SUITABLE OUTPUT PRINCIPLE

In the main population panel, the participant's task is not to find the most positive answer. The participant must record the first completed suitable output generated under the predefined condition. The participant:

- cannot refresh the answer,
- cannot request a new answer,
- cannot choose the positive one,
- cannot only trim the good part,

cannot rewrite the answer. Reproduction behaviour can also be investigated separately. However: "We asked the same question five times and selected the best answer." is not the main population score. This is the selected best output test. It should be reported with a separate name and method.

13. SINGLE CRITICAL EVENT EXCEPTION

A single answer cannot determine the population score. However, this does not mean that a single answer can never have a significant impact on a decision. NOMOS establishes two separate assessment pathways.

13.1. Population Representation Pathway

It asks the question: “How frequently does this outcome occur in the target population?” Requires a large and valid sample.

13.2. Critical Event Pathway

It asks the question: “How large is the foreseeable harm when this outcome occurs once?” A single confirmed outcome can trigger a review. For example:

- Assigning a non-existent health licence
- Making a dangerous emergency health referral
- Reporting incorrect legal authority
- Removing the security boundary
- Confusing a company with another legal entity
- Repeating a manipulative instruction hidden from people as if it were advice

It can be a singular Critical event candidate.

13.3. Separation of Score and Conformity Decision

A Critical event can generally be a single observation. For example:

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

This rate may seem small. However, if the potential harm of the event is severe, it may create an irreparable gap in the conformity decision. Therefore:

> *Population score measures the frequency of the event; Critical event review measures the impact of the event on harm and integrity.*

High average accuracy cannot automatically eliminate heavy singular error. The final conditions for critical classification will be defined in later chapters of the book.

14. OBSERVATIONAL CLAIM LADDER

In this section, only the first four observation levels are normatively separated. At higher levels:

- population-weighted score,
- multi-model score,
- time stability,
- intervention effect,
- sustainable value

will be defined in subsequent sections.

O-0 — NOT TESTED

The relevant AI product or condition has not been measured. Correct status:

NOT TESTED

Indefensible sentence: "It is appropriate because the problem was not observed."

O-1 — SINGLE EVENT OBSERVATION

There is one valid outcome. Defensible sentence: "Under the specified conditions, this outcome was observed once." Indefensible sentence: "The system represents existence this way."

O-2 — USER OR SESSION REPEAT

Multiple outcomes exist for the same user, account, or session family. Defensible sentence: "Under the specified user and session conditions, this outcome occurred in seven out of ten repetitions." Indefensible sentence: "Seven out of ten independent users saw this outcome."

O-3 — MULTI-USER SINGLE CELL MEASUREMENT

The same:

- AI product,
- user surface,
- country,
- language,
- panel condition,
- prompt version,
- measurement wave

contains multiple suitable user observations. Defensible sentence: “78% of the 200 valid observations in the Turkish/Turkey controlled cell met the specified transition criterion.” This result cannot be automatically generalised to other countries, languages, or user surfaces.

O-4 — MULTI-CELL LAYERED MEASUREMENT

Predefined samples exist from different countries, languages, or user cells. At this level:

- cell results,
- weighted result,
- scope,
- uncertainty

reportable. Detailed rules for population weights and subgroup fairness are defined in later chapters of the book.

15. SYNTHETIC APPLE.COM DEMONSTRATION

SYNTHETIC METHODOLOGY DEMONSTRATION / The users, AI products, responses, rates, and results below are fictional. They do not represent the actual performance of Apple Inc. or the behaviour of any real AI product.

15.1. Scenario A — Single Positive Response

A synthetic user submits the following question to synthetic AI Product A: ‘What kind of company is Apple.com?’ Synthetic response: ‘Apple is a US-based technology company that develops consumer electronics, software and digital services.’ Preliminary evaluation:

- Entity identity is correct
- Main category is correct
- No obvious critical errors
- Basic representation is acceptable

Raw result:

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

Incorrect public statement: “AI Product A represents Apple with 100% accuracy.” Correct public statement: “In the synthetic sample, the specified single observation met the transition criterion. Population prevalence and product reliability have not been determined.”

15.2. Scenario B — Discovery with Ten Observations

Ten independent synthetic users use the same version of the semantic prompt. Synthetic result:

- 7 responses pass
- 1 response is partially correct
- 1 response contains outdated information
- 1 response assigns the wrong main activity

raw pass rate:

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

Under a simple, idealised assumption of independent sampling, the approximate 95 per cent Wilson interval is necessarily wide. The result remains exploratory because it is based on:

- small sample,

- single cell,
- single wave

Correct status:

> *EXPLORATORY — Not sufficient for the publicly available global NOMOS score.*

15.3. Scenario C — 1,000 Observed Synthetic Wave

Assume synthetically that the following conditions are met:

- AI Product A
- 1,000 valid observations
- Predefined target population
- Locked prompt version
- Defined country and language layers
- Unmodified initial suitable response
- Full evidence records
- Predefined adjudication rules

Synthetic result:

Response status	Number
Meeting the transition criterion	918
Partially correct	37
Unsupported material claim	23
Outdated information	12
Critical error candidate	6
Unclassifiable	4
Total	1,000

Simple raw pass rate:

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

Approximate 95% Wilson interval under the idealised unweighted assumption: is. However, this number alone is not yet the final NOMOS GEO score. Because there are additional factors that need to be evaluated:

- Country and language weights
- Design effect
- Lowest language result
- Invalid observation rate
- Inter-rater agreement
- Verification of critical error candidates
- Untested user groups
- Time and user surface limitation of the measurement

Six Critical-error candidates must also undergo separate incident review. Correct reporting: ‘The raw response pass rate in the synthetic wave is 91.8 per cent. This result describes the observations before population weights and design effects are applied. Because six Critical-error candidates were identified, the conformity decision requires separate review.’ Incorrect reporting: ‘Apple's NOMOS

GEO score is 91.8 and it is compliant.' Synthetic data cannot be attributed in this way to a real company or a real system.

16. MANDATORY NORMATIVE PROVISIONS

N-01

A single generative system output cannot be used as the general GEO score.

N-02

A singular output observed alone:

- AI product and user surface,
- country,
- language,
- prompt,
- session,
- time

must be reported along with its conditions.

N-03

Single output:

- typical,
- continuous,
- global,
- belonging to all users

cannot be presented as behaviour.

N-04

Repetitions from the same user, account, or session cannot be counted as independent population observations without disclosing the dependency.

N-05

In the main population panel, the participant must record the first suitable output predefined; reproduction cannot be made to choose a positive response.

N-06

System rejection, error message, or inability to respond should be recorded as a valid system outcome if the protocol has been correctly applied.

N-07

Trimmed, modified, manually rewritten, or responses whose prompt context cannot be verified cannot be included in the main score calculation.

N-08

Selecting only positive screenshots and withholding the distribution of results is prohibited.

N-09

Synthetic observations cannot be used in real user or real system scores.

N-10

A population proportion should not be published without coverage, denominator, and appropriate uncertainty explanation.

N-11

The probability of a critical error cannot be quietly compensated within the overall average.

N-12

The observation result can only be applied to the inference universe that is uniquely defined; it cannot be generalised to unmeasured countries, languages, products, or time periods.

N-13

Measurement, scoring, or public statement must have an accountable human or institutional owner.

N-14

UNKNOWN, NOT TESTED, or INVALID records cannot be quietly converted into a positive result.

17. CANDIDATE GEO-1000 MAIN RULE

CANDIDATE-001

> *For every AI product being audited and for every main population measurement wave, a minimum of 1,000 valid user observations should be targeted by default.*

This candidate sentence:

- accuracy of the general product forecast,
- cost,
- applicability,
- country and language fairness

This creates an initial balance among these considerations. It does not, however, guarantee adequate coverage of:

- all the country scores,
- all language scores,
- all user subgroups

Adequacy must instead be tested through:

- pilots,
- independent repetitions,
- design effect,
- subgroup sensitivity

Those evaluations determine the final normative status.

18. FORMS OF FAILURE

F-01 — Selected Screenshot

Only the most positive response is published among numerous results.

F-02 — Treating a Single User as the Population

The experience of one person is generalised to the entire target universe.

F-03 — Counting Repeats of the Same User as New People

Repeats generated by the same person or account are reported as independent participants.

F-04 — Renewing the Response and Picking the Best One

Initial or negative responses are discarded; the most appropriate response is saved as the measurement result.

F-05 — Prompt Drift

Participants change the word, tone, or scope of the prompt; results are combined as if from the same experiment.

F-06 — Context Contamination

Previous conversations affect the response; the record is presented as a clean session.

F-07 — Hiding Personalisation

Memory, special instructions, or user history is visible; the result is reported as a general user response.

F-08 — Confusing Model and Product

Only the general model or provider name is written; the actual product, interface, and tool condition are not recorded.

F-09 — Removing Errors and Mistakes from the Denominator

Deleted to make the results appear positive when the system does not respond to records.

F-10 — Counting Invalid Capture as System Failure

Reported as participant error, missing evidence, or incorrect outcome of the AI product.

F-11 — Globalising a Single Language and Country

The result of a single locale is applied globally.

F-12 — Using Synthetic Data as Real

Records created for demonstration purposes are added to actual performance.

F-13 — Losing a Critical Event on Average

Severe health, legal, licensing, or security errors are made invisible within a high average.

F-14 — Drawing Intervention Effect from a Single Success

The single positive response received after the GEO intervention is considered the causal effect of the study.

F-15 — Judging the Entire System from a Single Mistake

One error is generalised as if the AI product is always wrong and wrong for all users.

F-16 — Defining the Target Population Afterwards

After the outcome is seen, the user group with high success is declared the “actual target population.”

19. AUDIT PROCEDURE

An auditor should follow the sequence below when evaluating the validity of a single observation or an early GEO claim.

Step 1 — Record the Public Claim

The full proposed sentence is determined. Example: "ChatGPT correctly knows our company."

Step 2 — Determine the Level of the Claim

The sentence may make a claim about:

- single event,
- prevalence,
- stability,
- global consequence,
- causality,
- commercial impact

The applicable level must be identified.

Step 3 — Determine the Supporting Observation Count

How many real users? How many accounts? How many independent sessions? How many repeats? How many AI products? How many languages and countries?

Step 4 — Verify Observation Conditions

Are the AI product, user surface, prompt, session, country, language and time all recorded?

Step 5 — Examine Evidence Integrity

Is there a full response? Is there a full screenshot? Has the response been selected or refreshed? Has the file been modified? Is the record synthetic?

Step 6 — Determine Dependency Structure

Are there repetitions of the same user? Has a shared account been used? Is there a mirrored model design? Is there clustering from the same panel?

Step 7 — Separate Data Collection Error from Valid System Result

Is the rejection or system error the actual result? Or did the participant violate the protocol?

Step 8 — Compare the Claim with the Level of Evidence

Is a single observation used for a population or global inference?

Step 9 — Open the Critical Path

If a single response carries a foreseeable severe harm, a separate incident investigation is initiated.

Step 10 — Limit the Decision

The public sentence is reduced only to the level carried by the evidence.

20. NECESSARY EVIDENCE

For an observation or claim within the scope of this section, the following records should be sought to the extent that they are relevant:

- Audited entity identity
- Fully disclosed GEO claim to the public
- Number of observations
- Number of unique participants
- AI product and user surface
- Visible model or version, if available
- Plan or account type
- Country and locale
- Prompt ID, version, and full text
- Session and personalisation status
- UTC timestamp
- Measurement wave
- Full raw response
- Screenshot or equivalent evidence
- Reproduction or answer selection status
- Observation validity status
- Dependency and matching record
- Synthetic or real status
- Invalid observation log
- Critical event record
- Result scope
- Uncertainty statement
- Responsible person or institution

Not all of these records may be required in the same detail for every low-risk discovery observation. However, if a broad GEO score or conformity clause is to be given to the public, material fields cannot be left blank.

21. AUDIT CHECKLIST

Has the audited entity been fully identified? Have the AI product and user surface been recorded? Was the prompt locked before measurement? Was the prompt changed by the participant? Are the session and personalisation conditions known? Is the user's country and locale information recorded? Is the full response preserved? Is it the first appropriate output or the best chosen output? Is there a complete record of evidence? Have the actual number of users and the number of observations been separated? Have repeated and matched observations been counted independently? Have refusals and system errors been properly managed in the denominator? Have invalid records been retained and justified? Have synthetic and real data been separated? Is a single response presented as a population result?

Is the claim applied only to the measured product, country, language, and time? Has the probability of a critical error been assessed separately? Does the result carry a statement of confidence

or uncertainty? Are untested areas clearly shown? Does the public statement exceed the level of actual observation? If one of the factual questions is unanswered, the result carries limited confidence. If more than one is unanswered, it should not be used for the overall GEO score.

22. OBJECTIONS AND RESPONSES

Objection 1 — "The real user already sees only one answer."

That is correct. The single answer seen by the single user is that person's real experience. Therefore, individual events are recorded. However, individual experience is not the same as population experience. NOMOS maintains both facts together:

> *The experience of a single user is real. / A single user is not the entire population.*

Objection 2 — "If the system gives the same answer every time, isn't one answer enough?"

The claim that the system gives the same answer every time also requires measurement. Additionally:

- time,
- retrieval,
- country,
- language,
- plan,
- personalisation,
- user surface

It may vary. Even a system assumed to be deterministic cannot be generalised beyond defined conditions.

Objection 3 — "Why exactly 1,000?"

The number 1,000 is not a magic threshold of certainty. GEO-1000 serves as a founding reference point that balances:

- a substantial user volume,
- a broad product-level estimate,
- practicable cost,
- public comprehensibility.

Subgroups may require more observations. Pilot studies must verify or revise the threshold.

Objection 4 — "Let the same 1,000 people test all AI products."

This method can be useful for paired comparison across models. However, selecting only intensive AI users who can access all products may bias the target population. Also, observations from the same person are dependent. Fully paired, partially paired, and balanced designs will be compared later.

Objection 5 — "Isn't a screenshot already sufficient proof?"

Screenshots are important. However, on their own they do not always show the following:

- real time
- country
- plan

- previous conversation
- personalisation
- answer refresh
- file change
- full response

For this reason, a screenshot is an important part of the evidence package. It is not the whole.

Objection 6 — "Why do we include a non-responding system in the score?"

The user's real experience may be not getting an answer. A rejection or error:

- technical,
- security,
- policy,
- access

may be the result. If the protocol was correctly applied, the event should be recorded. The system's unresponsiveness due to the participant's technical error is classified separately.

Objection 7 — "Why should a single severe mistake affect the entire inspection?"

A single mistake does not prove that most of the population made the same mistake. However, the potential damage of a single mistake can be severe. Therefore:

- score frequency,
- Critical decision harm and integrity

are evaluated separately. The average cannot automatically erase human harm.

Objection 8 — "If we can't say anything with a single response, what is the value of early tests?"

Singular and small-sample tests:

- generate hypotheses,
- discover types of errors,
- test prompts,
- test the capture system,
- calibrate adjudicators,

can reveal critical events. What is wrong is not to do exploratory testing. What is wrong is presenting the exploratory result as a population score.

COMMON PRINCIPLE OF CHAPTER 24

This chapter was written to eliminate a single misconception: "I asked an AI, got an answer; here is my GEO result." No. You do not yet have a general GEO result. What you have is an observation. This observation:

- is real,
- is significant,
- even critical

may be. However, it is still a single observation. For a GEO score, the existence of:

- users,
- AI products,
- countries,
- languages,
- sessions,
- times

What must be measured is the distribution of representation across AI products, countries, languages, sessions and time. A single screenshot is not a distribution. Ten automated repetitions are not ten independent people. Nor do a thousand responses from a poor sample represent the world's population. Even a population-proportional sample may be incomplete without language fairness. A high average cannot erase Critical errors. NOMOS's first law of measurement is therefore:

> *A single answer is an event. / Multiple answers are a pattern. / A defined sample is an estimate of the population. / Repetition over time is a measure of stability. / But none of these alone are causality or sustainable value.*

Order of Section 1 of NOMOS

> *Do not bring me one screenshot and then speak for the entire system.*
>
> *Do not make the answer of one user the answer of the world's population.*
>
> *Do not convert duplicates of the same account into independent humans.*
>
> *Do not refresh the answer and display the preferred as if it were the real distribution.*
>
> *Do not remove rejections, errors, and unknowns from the denominator.*
>
> *Do not generalise one language to the world, one country to all countries, or one AI product to all generative systems.*
>
> *Do not turn a single success into evidence of an intervention effect, or a single error into the immutable character of the entire system.*
>
> *Do not drown a single severe error within the average.*
>
> *First, record the event. / Then measure the pattern. / Then define the target population. / Then show dependency and uncertainty. / And speak only to the extent that the evidence carries.*

The Chapter's Closing Sentence

> *The representation of an entity in artificial intelligence systems is not a single answer. AI products are a distribution that should be measured across users, countries, languages, sessions, and times.*

Normative Core

> **A single generative-system output MAY establish that a specific response occurred under recorded conditions. A single output MUST NOT be used to infer: - population prevalence, - typical product behaviour, - temporal stability, - cross-language consistency, - cross-country consistency, - cross-product consistency, - intervention effect, - recommendation certainty, - commercial impact, - conformity, - or a GEO score. All broader claims MUST be supported by a predefined target population, valid multi-observation design, recorded dependence structure, evidence bundle, evaluation rule, scope statement, and uncertainty disclosure. A single high-harm output MAY trigger a separate critical-incident review without being treated as evidence of population prevalence.**

CHAPTER 02

Representation Is a Distribution

METHODOLOGICAL BASES / K04; K05; K09

Chapter Boundary

Section 1 established the following provision:

> *A single response is not the GEO score. A single response is a singular observation that occurred under recorded conditions.*

This section answers the next question:

> *If a single response is not the GEO score, what is the actual object we want to measure?*

The answer is:

> *The representation of an entity in generative systems is a multidimensional distribution formed across AI products, users, countries, languages, session conditions, prompts, and times.*

This chapter:

- the representation distribution,
- the fundamental dimensions of the distribution,
- the difference between the mean and the distribution,
- the subgroup inequality,
- the stability,
- the coverage and missing observation problem,
- the risk of rare severe errors,
- the conceptual structure of the publicly available result card

defines. This chapter does not yet:

- how many participants will be assigned to each country,
- how the sample frame will be established,
- how prompt translations will be validated,
- how atomic claims will be scored,
- the exact formula of the final NOMOS score,
- conformity thresholds

This chapter does not settle these matters normatively; later chapters address them separately. Before defining the scoring formula, Chapter 2 defines the reality that the score is intended to represent.

NOMOS Challenge

Consider two AI products. Each receives the same global average score:

90 / 100

The first product's language outcomes are as follows:

- English: 90
- Turkish: 91
- German: 89
- Japanese: 90
- Arabic: 90

The language results of the second product are as follows:

- English: 99
- Turkish: 98
- German: 97
- Japanese: 86
- Arabic: 70

The simple average of both products is 90. Do they have the same GEO quality? No. The first product produces a similar representation in different languages. The second product produces excellent representation in some languages but seriously weak representation in another language. A single average hides this difference. Now consider two other products. 95% of the responses of both meet the passing criteria. The remaining 5% in the first product consists of:

- small scope deficiencies,
- limited explanation shortcomings,
- a few UNKNOWN

results. The remaining 5% in the second product consists of:

- incorrect health authority,
- nonexistent licence,
- incorrect legal identity,
- inappropriate advice

exists. The raw accuracy rate for both is again 95 per cent. Do they have the same representation risk? No. Now consider a third pair. The average of both across the three measurement waves is 90. The first one:

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

The second one:

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

Do they have the same consistency over time? No. In that case, the issue is not only: 'What is the average?' We also need to know the following: How are the results distributed? Who is well represented, who is poorly represented? Which languages and countries are falling behind? How severe are the errors? How much does the result change over time? Which areas were never

measured? With what weights was the average created? How many different realities lie behind the same average? The first conclusion of this section is:

> *The same average is not the same representation.*

Its second provision states:

> *a GEO score may be a summary of the distribution; it cannot replace the distribution.*

1. PURPOSE OF THE CHAPTER

This section defines how the representation of entities in generative systems should be understood before being reduced to a single number. Its main objectives are:

- To show that representation is not a timeless and singular “AI perspective"
- To define the conditions on which the representation outcome depends
- To explain that the true–false binary alone is insufficient
- To separate average accuracy from subgroup inequality
- To make visible the heavy error risk hidden within high averages
- To show that stability and accuracy are not the same concept
- To prevent unmeasured areas from being filled with positive assumptions within the distribution
- To prepare the structure of a future NOMOS score that is singular but not condemned to being singular

At the end of this section, the reader should be able to answer the following question:

> *Why is the result of an entity's GEO not limited to only the average accuracy rate?*

2. CENTRAL NORMATIVE PROVISION

> *The representation of an entity in generative systems is the conditional distribution of different responses and representation outcomes that can occur under defined conditions.*

This distribution:

- it is not the abstract 'thought' of a single model,
- it is not the common view of all artificial intelligence systems,
- it is not only the repetition rate of the website,
- it's not only the frequency with which the brand's name is mentioned,
- it's not just the average accuracy,

it is not only the recommendation rate. The representation distribution, within a certain protocol, shows the following:

> *Which users see what kind of representation on which AI product and user surface, in which language, in which country, with which prompt, under which session condition, and at what time.*

Therefore, the result of a distribution is always dependent on the following questions: Which entity? Which AI product? Which user surface? Which target population? Which country? Which language and locale? Which prompt? Which session condition? Which measurement wave? Which Truth Pack? Which evaluation rule? When any of these fields change, the distribution may also change.

3. WHAT DOES THE WORD "DISTRIBUTION" MEAN?

In this book, the word distribution does not refer to a hidden or metaphysical view that is assumed to exist within an artificial intelligence system. It does not claim: NOMOS: "This is the real distribution of the company in the model's mind." The distribution here is operational and observable.

> *It is the pattern of different representation results that can be obtained under a defined protocol within the target population and measurement conditions.*

Two distributions should be distinguished.

3.1. Target Representation Distribution

It is the distribution that the measurement tries to predict but cannot directly observe in full. It is represented as:

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

Here:

- E: audited entity
- Y: representation result
- A: AI product and user surface
- C: country or geographic condition
- L: language and locale
- S: session and personalisation condition
- P: prompt
- T: time or measurement wave
- U: target user universe

This expression asks: “In the defined target user universe, under the specified conditions, which representation results occur with what probabilities?” Not all elements of the target distribution can be measured. They are estimated through sampling.

3.2. Observed Empirical Distribution

It is the distribution of valid observations obtained in real or synthetic measurement. Its representation:

$\hat{P}_E$

can be used. The empirical distribution is not the target distribution itself. It is an estimate produced through its sample, weighting, and measurement design. The magnitude of the difference may depend on:

- Sample design
- Participant selection
- Country and language coverage
- Invalid observation rate
- Non-response
- Adjudicator error
- Measurement wave
- Prompt selection

- Product and interface conditions
- Weighting
- Clustering and dependence

Therefore, an empirical distribution cannot be published as "the absolute and immutable truth of artificial intelligence." The correct expression should be: "The estimate obtained for the specified target population and measurement conditions."

4. REPRESENTATION SPACE

The representation of an entity does not occur in a single dimension. Each observation exists in a multidimensional condition space. The basic representation:

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

Here:

- Ei: entity
- Ai: AI product and user surface
- Ui: user or sampling unit
- Ci: country and geographical layer
- Li: language and locale
- Si: session, plan, and personalisation condition
- Pi: prompt or prompt family
- Ti: time and measurement wave

Generated response:

```
R_i
```

can be shown. The representative state resulting from the evaluation of the response is:

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

and can be considered as. Here:

- VE: verified Truth Pack used for the entity
- J: assessment and adjudication rules
- g: evaluation process that converts the response into representative results

The meaning of this representation is as follows: The same response cannot be scored on its own without indicating which entity, which Truth Pack, and which evaluation rule it was examined according to.

5. REPRESENTATIVE RESULT IS NOT A SINGLE VALUE

Only the entirety of a response:

- can be forced into the categories of

  true,
- is wrong

Giving a label is often insufficient. The following situations can occur together in the same response:

- Correct identity
- Correct main category

- Incomplete service coverage
- Outdated number of employees
- Unsupported leadership claim
- Correct citation
- Inappropriate recommendation
- Non-critical boundary deficiency

Therefore, the representation result can be considered as a result vector. Candidate representation:

```
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)
```

Here:

- yiid: identity accuracy
- yicat: accuracy of category and role
- yifact: accuracy of factual matters
- yiscope: preservation of scope and boundaries
- yitime: temporal validity
- yievidence: claim–evidence consistency
- yirecommendation: recommendation and user suitability
- yiharm: harm class of error or misguidance

The exact scoring rules for these fields will be defined in later sections. The guiding principle here is:

> *A representative answer should be evaluated not only on whether it contains correct information but also on whom, what, within which boundaries, with what evidence, and for whom it is appropriately conveyed.*

6. BASIC REPRESENTATION STATES

The status of each atomic claim and final response may not be the same. This section conceptually separates the following basic states.

6.1. SUPPORTED

The claim is supported by sufficient evidence within the defined scope and time.

6.2. CONTRADICTED

The claim conflicts with reliable and relevant records.

6.3. UNSUPPORTED

There is not enough support for the claim. It has not been proven that the claim is necessarily false. However, there are not enough records for it to be accepted as true.

6.4. OUTDATED — NO LONGER CURRENT

The claim may have been true in the past. As of the measurement date, it is no longer valid or its current accuracy cannot be verified.

6.5. PARTIALLY SUPPORTED — PARTIALLY SUPPORTED

Only a specific scope, population, country, product, or time of the claim is supported. The response may have exceeded the support limit.

6.6. OUT OF SCOPE — OUT OF SCOPE

The claim is not related to the audited entity or Truth Pack. It may have come from another entity, product, country, or service.

6.7. ENTITY CONFLATION — ENTITY MERGER

System:

- two different companies,
- the brand and the legal entity,
- a person and the institution,
- old and new entity,
- has confused different organisations with the same name

with each other. This error is not merely a factual mistake. It is the corruption of the representation object.

6.8. MATERIAL OMISSION — MISSING INFORMATION

The sentences in the response may be correct. However, essential information that would change the user's decision has been omitted. Example: The service is offered only in certain countries. The licence is limited to a specific jurisdiction. Price is excluding tax. The result is based on first-party data. The product is not suitable for certain users. The omission can create a wrong inference even if there is no directly incorrect sentence.

6.9. REFUSAL — REJECTION

The system has refused to answer. Rejection:

- appropriate,
- unnecessary,
- inconsistent,
- security-based,
- technical

may occur. Ret is not automatically considered failed or successful. It is classified separately.

6.10. NO RESULT

The system did not produce a usable response due to technical or operational reasons. It is not the same as Ret.

6.11. UNKNOWN

The available records are not sufficient for a reliable classification. UNKNOWN cannot be converted into a positive result.

7. MAIN CHARACTERISTICS OF THE DISTRIBUTION

A representation distribution cannot be explained solely by the mean. At a minimum, the following characteristics should be evaluated.

7.1. Central Tendency

Summarises the general accuracy or transition level of the responses. Example measures:

- Mean
- Weighted mean
- Median
- pass rate

Central tendency is important. However, it only shows the centre of the distribution.

7.2. Spread and Variability

Indicates how much the results are spread around the centre. Of two systems with the same mean, one may be consistent among users, while the other may be highly variable. Variability:

- users,
- countries,
- languages,
- sessions,
- measurement waves

can also be examined among them.

7.3. Shape of the Distribution

Are the results clustered around a single central point? Or are they divided into two or more clusters? For example:

- If half of the users see very accurate,
- and half see very inaccurate

responses, the average may be medium. However, the typical user experience is not "medium accuracy." It is two separate polarised experiences.

7.4. Subgroup Differences

Do specific language, country, or user groups systematically see different results? The global average may hide this difference.

7.5. Tail Risk

Are there low-frequency but severe errors? Rare events can have a small effect on the average. Their impact on human harm can be significant.

7.6. Coverage

How much of the target user universe has actually been measured? A high score can be found together with low coverage.

7.7. Missing and Unknown Mass

What proportion of the responses are:

UNKNOWN,

NOT TESTED,

rejection, technical non-response or an invalid record?

7.8. Time Stability

To what extent is the distribution preserved across different measurement waves?

8. CANDIDATE NOMOS REPRESENTATIVE DISTRIBUTION SIGNATURE

CANDIDATE CONCEPT

It is recommended that the result of a GEO be defined first by the following distribution signature instead of a single number:

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

This representation is called Representation Distribution Signature, that is:

Representation Distribution Signature

Components:

μ — General Representation Level

Overall accuracy or transition prediction under defined weights and scope.

κ — Coverage

The extent to which the target user, country, language, AI product, and time universe are sufficiently measured.

σ — Variability

Spread among users and experimental cells.

Δ — Subgroup Inequality

The material difference between language, country, product, or other significant groups.

'tau — Tail and Critical Error Risk

The rate and nature of rare severe errors.

ρ — Temporal Stability

The resilience of distribution across waves.

u — Unknown and Unresolved Mass

UNKNOWN is the share of the distribution that remains unresolved or insufficiently assessed. This signature is not yet a final scoring formula. The following elements require further development:

- definite definition,
- estimation,
- weight,
- public presentation

They will be finalised through pilot studies and later chapters. The founding principle is:

> *The GEO result should answer not only the question 'how accurate?' but also 'for whom, how consistent, how equitable, how comprehensive, and how risky?'*

9. AVERAGE WASHING

If an institution only shows a high global average and hides the material flaws of the distribution, this is called:

Average washing

Average washing can occur in the following ways:

- Hiding the weakest language
- Excluding small countries from the report
- Mixing critical errors into the general error rate
- Removing UNKNOWN results from the denominator
- Only including well-performing AI products
- Excluding the weak measurement wave
- Making the population outside the scope invisible
- Changing weights after seeing the results

The average may be mathematically calculated correctly. However, the representation it offers to the public can still be misleading. Therefore:

> *A mathematically correct calculation of an average does not mean it is sufficient for decision-making.*

10. DIFFERENT REALITIES OF THE SAME AVERAGE

The average could be 90 in all of the following synthetic distributions.

10.1. Homogeneous Distribution

All major subgroups are in the range of 88–92. This distribution is relatively balanced.

10.2. Polarised Distribution

Half of the users see results close to 100, while the other half see results close to 80. The average is 90. However, the user experience is not equal.

10.3. Linguistically Fragmented Distribution

Results are very good for high-population or high-resource languages, but significantly weak for low-resource languages. The global average may look good. Language fairness is weak.

10.4. Country-Based Fragmented Distribution

Even within the same language, results may vary due to different countries, locales, product access, or user base.

10.5. Temporally Volatile Distribution

Very high results occur in one wave, very low results in another wave. Even if the average is high, it is difficult to establish reliable user expectations.

10.6. Distribution with Tail Risk

Most of the responses are correct. However, a small portion contains serious health, legal, licensing, or safety errors.

10.7. High Unknown Mass Distribution

Most of the assessable responses are correct. However, a significant portion of the target population or responses is:

- tested,

- untested,

UNKNOWN

or inconclusive. Publishing a high rate based only on the assessable responses may obscure the scope.

11. SUBGROUP INEQUALITY

A global score may not represent the experience in subgroups. Key subgroups may include:

- AI product and user surface
- Country
- Language
- Locale
- Plan
- Web or mobile
- Controlled or natural panel
- Time wave
- User intent
- Device family

Not all of these groups may have the same importance in every audit. Material groups should be determined when scope is locked.

11.1. Language is Not the Same as Country

Turkish-speaking:

- User in Turkey,
- User in Germany,
- User in another country

They can use the same language. However:

- product access,
- price,
- resources,
- local interface,
- legal context,
- personalisation

may be different. Therefore, all differences cannot be explained solely by language or country.

11.2. Intersectional Cells

Some material differences only become visible when multiple areas are examined together. Example: Germany × Turkish × Mobile × Free plan × Natural User Panel could in itself be an experimental cell. However, as cells multiply, the sample size decreases. Therefore, a definitive score cannot be published for each small cell. Measurement design: should preserve meaningful subgroups and should not produce statistical false precision.

11.3. Subgroup Base

CANDIDATE CONCEPT

In addition to the global average, it is recommended to report the result of the weakest adequately sampled subgroup or the bottom 10% group separately. This can be referred to as:

Representation base

or:

Language/Country Floor

However, if the "lowest score" comes from a small and unstable cell, it may produce incorrect results. Therefore, the base metric should be defined together with:

- minimum sample,
- confidence interval,
- cell quality

The exact rule will be determined in the following sections.

12. IT IS PART OF THE SCOPE DISTRIBUTION

A distribution consists not only of observed responses. What was never observed also determines the meaning of the result. For example, English and Turkish were tested; Japanese and Arabic were not. The web interface was tested; the mobile interface and paid plan were not. Strong English and Turkish results do not justify the statement, 'Representation is strong across all languages and user surfaces.' The defensible conclusion is: 'Measurements were conducted on the specified English and Turkish web surfaces. Other languages and mobile surfaces were not tested.'

12.1. Coverage Map

Each GEO distribution should carry the following map to the extent relevant:

- Target universe
- Sampling frame
- Covered countries
- Non-covered countries
- Covered languages
- Non-covered languages
- Monitored AI products
- Unmonitored products
- Monitored user surfaces
- Unmonitored surfaces
- Measurement waves
- Out-of-scope prompt families

12.2. Coverage Rate

CANDIDATE CONCEPT

A coverage metric can be developed to show what proportion of the target universe is represented with sufficient measurement. However, coverage:

- only the number of countries,
- only the number of languages,
- only the population ratio

cannot be measured. For example, 150 countries may be covered. However, 95% of the participants may have come from only five countries. Or 90% of the population may be covered, but small languages may be entirely excluded. Therefore, population coverage should be shown separately from country/language observer coverage.

13. MISSING OBSERVATION AND NON-RESPONSE

Missing data is not only a technical issue. Its distribution can be systematically distorted. The reasons for an observation being missing may include:

- Participant's failure to complete the task
- Product not being accessible in the country
- Computation constraints
- Technical error
- System rejection
- Invalid screen capture
- Participant changing their prompt
- Product not working in a specific language
- Adjudicators being unable to make reliable decisions

Each of these reasons requires a separate status.

13.1. Missing Data May Not Be Random

For example, if more tasks cannot be completed in countries with low connection quality, using only successful recordings can shift the global result in favour of wealthy users with strong connections. If the rate of UNKNOWN rises because answers in a specific language are difficult to evaluate, the score calculated from only known answers may hide language inequality. Therefore:

> *The missing data rate and the distribution report of the reason for missing data should be part of the report.*

13.2. Denominator — Integrity of the Denominator

A report should answer the question clearly: "90 per cent of exactly what is 90 per cent?" Possible denominators:

- Assigned participants
- Task initiators
- Task completers
- Protocol-compliant records
- Valid system results
- Responses with adjudicator decisions
- Specific language or country cell

The denominator should not be changed after the result is seen. For example:

- 1,000 tasks assigned,
- 900 completed,
- 800 valid,

- 720 passed

Thus:

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

This is the pass rate among valid records. The observed pass rate across the entire assigned sample is instead:

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

answers a different question. Both rates can be used. They cannot be presented in the same sense.

14. ACCURACY AND CONSISTENCY ARE NOT THE SAME

A distribution may have high accuracy. However, it may be inconsistent over time or across users. Another distribution may have low accuracy. However, it may consistently be wrong. Therefore, four situations should be distinguished.

Accuracy	Consistency	Comment
High	High	Candidate for consistent and accurate representation
High	Low	Mostly accurate but unreliable experience
Low	High	Systematic and consistent misrepresentation
Low	Low	Both incorrect and unstable representation

Being “consistent” is not automatically positive. The system may be quite consistent if it gives the same error to every user.

14.1. Inter-User Consistency

Do different users see similar material representation in the same cell?

14.2. Inter-Session Consistency

Does the same or matched user see similar results in new sessions?

14.3. Inter-Wave Consistency

Does the same protocol produce a similar distribution on different days or periods?

14.4. Post-Product Update Consistency

When there is a material change in the AI product, does the previous distribution continue?

15. PRODUCING THE SAME SENTENCE IS NOT STABILITY

Stability: It does not mean "giving the exact same answer to every user word for word." An AI product can use different expressions, order, and examples. As long as they preserve the material meaning, they can be acceptable variations. The important distinction is:

Formal Variability

Word choice Sentence order Length Tone Example Using list or paragraph If material truth is preserved, it may not be an error.

Semantic Variability

Change of company identity Differentiation of the main category Expansion or contraction of the field of activity Change of licence or competency Change of price and geography Recommending to one user but not recommending to a similar user without explanation If the material representation changes, there may be a stability problem. Therefore, the preferred concept is:

Conditional semantic stability

It should be. The fundamental reality is expected to be preserved under equivalent material conditions. A change in recommendation due to different user needs can be appropriate contextual behaviour.

16. CONTEXT SENSITIVITY AND ARBITRARY VARIABILITY

It is not always correct for a system to give the same recommendation to all users. For example:

- Users with a low budget
- Only users seeking services in a specific country
- Users requiring a different licence
- Users seeking individual rather than corporate services

can receive different conformity outcomes for the same entity. This difference:

- may be due to the user condition,
- the scope of service,
- or the actual limit

and may be context-sensitive and accurate. However, if the outcome changes without reason under equivalent user conditions, it may be arbitrary variability. Therefore, NOMOS does not want: "The same answer for all users." It wants:

> *A response that is consistent and explainable under equivalent material conditions and sensitive to actual limits under different conditions.*

17. TIME IS WITHIN THE REPRESENTATION DISTRIBUTION

The representation distribution is not timeless. It can be thought of as PE,t. The distribution of the same entity:

- product update,
- resource change,
- company change,
- news event,
- new language version,
- technical access change

can diverge afterwards.

17.1. Sudden Change

After a product or resource update, the distribution changes quickly in a short time.

17.2. Gradual Change

Due to new resources and user behaviours, the result slowly changes over time.

17.3. Temporary Fluctuation

During a specific event or news period, the distribution temporarily differs.

17.4. Post-Intervention Change

After the GEO intervention, the distribution may change. However, claiming that the change is due to the intervention also requires a baseline and causality design.

17.5. Smoothing with Time Average

For a product:

- in the first wave 98,
- in the second wave 90,
- in the third wave 82

If taken, the average is 90. However, publishing 90 hides the downward trend. Therefore:

> *The time average cannot replace the wave results.*

18. TAIL RISK AND HEAVY ERROR DISTRIBUTION

A large portion of a distribution may be correct. However, low-probability heavy errors should also be evaluated. Examples:

- Health licence that does not actually exist
- Incorrect emergency guidance
- Incorrect legal authority
- Fake investment guarantee
- Incorrect company identity
- Inappropriate advice for children or vulnerable groups
- Repetition of the manipulative instruction hidden from people

In the previous architecture:

- Critical,
- Major,
- Moderate,
- Advisory

were separated into four levels of importance, and it was suggested that Critical findings should have an automatic failure effect. This book links this system to population distribution.

18.1. Separation of Frequency and Severity

An error:

- frequent but low impact,
- rare but severe

may occur. However, the frequency rate alone cannot evaluate the two cases equally.

18.2. Candidate Severity-Weighted Risk Representation

CANDIDATE CONCEPT

A risk indicator can be developed using severity coefficients for error classes:

```
tau = Σ_{k=1}^K λ_k p_k
```

Here:

- pk: predicted rate of the error type
- λk: severity coefficient of the error type

However, coefficients cannot be determined arbitrarily. They should be based on:

- Foreseeable human harm
- Representation integrity
- Impact on user decision
- Reversibility
- Intent and manipulation
- High-risk area

The precise method will be defined in Section 15.

18.3. Critical Gate

A high overall average cannot automatically delete a verified Critical error candidate. The correct public record should show both results together: "The overall representation estimate is high." and: "The specified Critical event or events also affect the conformity decision."

19. REFUSAL AND FAILURE IN DISTRIBUTION

A system not responding is also part of the user experience. However, not all refusals are the same.

19.1. Appropriate Refusal

The system may exercise caution when evidence is insufficient or when it cannot provide a safe response. This can be the correct behaviour in some contexts.

19.2. Unnecessary Refusal

It may inconsistently avoid responding to publicly available and low-risk information.

19.3. Technical Nonresponse

The system may not produce a response due to a technical error.

19.4. Out-of-Scope Rejection

The prompt may be incorrect or outside the task protocol.

19.5. Preservation of Rejections in Distributions

Rejections:

- success,
- failure,

UNKNOWN

should not be forced into a single category. Their function and justification should be evaluated. Removing rejections from the denominator may yield a high accuracy rate only among successful responses and obscure the real user experience.

20. HOW CAN DISTRIBUTIONS BE COMPARED?

Simply comparing the result numbers is not sufficient to compare the distribution of two AI products. For the comparison to be fair, the following areas should be as equal as possible or explicitly modelled:

- Audited entity
- Target population
- Prompt version
- Languages
- Countries
- Measurement wave
- Session condition
- User surface
- Plan distribution
- Adjudication rule
- Reality package
- Validity rules
- Weights

Two scores obtained with materially different protocols cannot be used directly for ranking.

20.1. Common Support Comparison

All compared AI products:

- are officially accessible,
- support the same language and user conditions

can be compared in a common universe. This provides a cleaner comparison between systems.

20.2. Natural Access Distribution

Each AI product can be measured in its own real user access universe. This result shows the product's actual global impact. However, since target populations may differ, direct performance ranking requires caution.

20.3. Different Surfaces of the Same Provider

Web, mobile, enterprise products, or different plans, even if they belong to the same provider, can create separate distributions. If they are to be combined under a single result, the combination and weighting method should be explained.

21. WEIGHTING CHANGES THE MEANING OF THE DISTRIBUTION

A global average depends on the weight with which observations are combined. Possible weights:

- Equal by country
- By country population
- Appropriate AI user population
- Language user
- Actual product exposure
- Equal AI product
- Market share
- Customer target market

Each method answers a different question.

21.1. Equal Country Weight

San Marino and India can receive the same weight. This approach can address the question of 'Average country experience.' It does not represent the average user experience in the world population.

21.2. Population Weight

Larger countries have more influence. This approach can approach the global population experience. It may render small countries and languages invisible.

21.3. Language Fairness Weight

It can also ensure that low-resource languages remain visible. It is not the same as the global population score.

21.4. Product Exposure Weight

The weighting of AI products according to actual user exposure can be considered. It requires reliable and independent usage data.

21.5. Principle of Normative Weighting

> *Weights should be defined before the results are seen, publicly disclosed, and the question they answer should be specified.*

It is prohibited to choose a weighting system that will increase the positive score after the result is seen. Alongside a single global score, alternative meaningful weighted results can also be published.

22. SCORE DISTRIBUTION IS LOSSY COMPRESSION

It is possible to produce a single number from a distribution. However, a single number does not carry all the information in the distribution. Therefore, the score:

Is a lossy summary of the distribution

should be accepted. A score may lose the following information:

- Weakest language

- Weakest country
- Critical error type
- Unknown mass
- Time fluctuation
- Polarisation
- Product and plan difference
- Out-of-scope population
- Rejection distribution
- Adjudicator ambiguity

The score is useful because:

- facilitates comparison,
- simplifies public communication,
- allows monitoring change,

supports the management decision. However, the simplicity of the score cannot turn into the elimination of distribution.

23. MINIMUM PUBLIC RESULTS CARD

CANDIDATE REQUIREMENT — CANDIDATE OBLIGATION

If a global score NOMOS is to be presented to the public, it is recommended that at least the following results be shown alongside it:

Field	Description
Main representation estimate	Overall result under defined weighting
Confidence or uncertainty range	Sensitivity of the prediction
Scope	Country, language, product, user surface, and time
Weakest sufficient subgroup	Makes material inequality visible
Critical error rate	Severe error queue
Time stability	Inter-wave variation
UNKNOWN and inconclusive rate	Unknown mass
NOT TESTED fields	Unmeasured universe
Measurement date	Temporal limit of the result
Protocol version	Which rule it was produced with

The main score should not be used as a standalone advertising label. The canonical record must be linked to this results card.

24. SYNTHETIC APPLE.COM DISTRIBUTION DISPLAY

SYNTHETIC METHODOLOGY DEMONSTRATION / The AI products, users, languages, numbers, and results below are entirely fictional. They do not represent the actual performance of Apple Inc. or any real AI product. Let us consider two synthetic AI products:

- Product A

- Product B

For each product:

- five languages,
- 200 valid synthetic observations per language,
- a total of 1,000 observations

are assumed. This equal language distribution is only to demonstrate the method. It does not represent the actual GEO-1000 population weighting.

24.1. Language-Based pass rates

Language	Product A	Product B
English	90%	99%
Turkish	91%	98%
German	89%	97%
Japanese	90%	86%
Arabic	90%	70%
Simple average	90%	90%

The average of the two products is the same. However, their distributions are different.

Product A

Language results are close to each other. The difference between the highest and lowest language is 2 points. The language floor is 89%.

Product B

Produces very high results in three languages. There is a decline in Japanese and especially in Arabic. The difference between the highest and lowest language is 29 points. The language floor is 70%. Note: Saying "Both scored 90" hides the material difference.

24.2. Response Status Distribution

Response status	Product A	Product B
Meeting the transition criterion	900	900
Partially supported	60	20
Unsupported material claim	20	20
Outdated	10	20
Critical error candidate	1	14
UNKNOWN / no result	9	26
Total	1,000	1,000

The pass rate of the two products is again 90%. However: In Product B, the number of Critical error candidates is 14 times higher. In Product B, the unknown and inconclusive mass is higher. A larger portion of failures in Product A is due to partial support deficiency. The error backlog of Product B is heavier. The same pass rate does not necessarily result in the same fitness outcome.

24.3. Time Waves

Measurement wave	Product A	Product B

Measurement wave	Product A	Product B
Wave 1	90%	98%
Wave 2	91%	90%
Wave 3	89%	82%
Mean	90%	90%

Product A is relatively stable. Product B shows a noticeable decline. Publishing only the average of the three waves hides the deterioration in Product B.

24.4. Synthetic Result

It may be possible to publish the same headline score for Product A and Product B:

90

But NOMOS result cards cannot be the same.

Product A — Synthetic Distribution Signature

Overall result: high Language inequality: low Critical error tail: low but not zero Unknown mass: low Time stability: high

Product B — Synthetic Distribution Signature

Overall result: high Language inequality: high Critical error tail: substantial Unknown mass: higher Time stability: low and declining

Therefore, the synthetic conclusion of Section 2 is as follows:

> *Having the same average score for Product A and Product B does not indicate that they have the same quality of representation.*

25. MANDATORY NORMATIVE PROVISIONS

CH02-N01

A GEO result cannot be defined only by a single average or a single pass rate.

CH02-N02

Each distribution result:

- the audited entity,
- target population,
- AI product and user surface,
- country and language,
- The prompt,
- the residence condition,
- to time

should be connected.

CH02-N03

The global average alone cannot compensate for material subgroup inequality.

CH02-N04

Critical and other high-severity error outcomes should be reported separately from the overall error rate.

CH02-N05

UNKNOWN, no-result, rejected, invalid, and NOT TESTED statuses should be distinguished from each other.

CH02-N06

Unmeasured country, language, product, user surface, or time period cannot be shown as part of the positive distribution.

CH02-N07

Weights and aggregation rules should be defined before viewing results.

CH02-N08

Post-result weight changes should create a new analysis version, and previous results should be preserved.

CH02-N09

Distributions based on materially different target populations, demand, product, time, or evaluation criteria cannot be directly compared without explanation.

CH02-N10

Formal language variability cannot automatically be counted as representation error.

CH02-N11

Material identity, scope, time, competence, or recommendation variability cannot be classified as a style difference without justification.

CH02-N12

High stability alone is not evidence of quality; consistently wrong results should also be recorded.

CH02-N13

A high average, if accompanied by low coverage, should have public results with coverage limitation.

CH02-N14

Subgroup scores should not be published as definitive country or language judgements under insufficient sampling and broad uncertainty.

CH02-N15

Rejection and technical failure rates should be preserved as part of the genuine user experience.

CH02-N16

The only score made public should be linked to the canonical distribution and scope record.

CH02-N17

Synthetic distributions cannot be combined with the performance of a real AI product or real entity.

CH02-N18

The distribution report should have an accountable human or institutional owner.

26. FORMS OF FAILURE

CH02-F01 — AVERAGE THINKING

High global average is shown; language, country, or Critical error corruption is stored.

CH02-F02 — DELETE SUBGROUP

Low-performing groups are removed from the report or declared 'insignificant'.

CH02-F03 — SCOPE MAP

Narrow-scope good results are used for the whole world, all languages, or all products.

CH02-F04 — SEVERITY NORMALISATION

Minor deficiencies and serious health or licensing errors are combined within the same error count.

CH02-F05 — HIDE DETERIORATION WITH TIME AVERAGE

Decline or fluctuation is made invisible only with the period average.

CH02-F06 — WORSHIP OF STABILITY

A consistent answer is assumed to be automatically correct.

CH02-F07 — WORD-FOR-WORD SAME ANSWER FALLACY

Literal text matching is used as a measure of consistency instead of semantic correctness.

CH02-F08 — COUNTING CONTEXTUAL DIFFERENCE AS ERROR

Differences in appropriate recommendations based on actual user needs are marked as arbitrary variability.

CH02-F09 — COUNTING ARBITRARY VARIABILITY AS PERSONALISATION

Material contradictions under equivalent user conditions are defended as “personalisation” without explanation.

CH02-F10 — CHANGING THE DENOMINATOR

Post-result transition is made among completed, valid, or evaluated observation denominators to elevate the positive outcome.

CH02-F11 — POST-RESULT WEIGHTING

The country, language, or model coefficients producing the highest score are selected afterward.

CH02-F12 — MERGING INCONSISTENT CELLS

Different prompt, product, time, or panel conditions are presented as a single distribution.

CH02-F13 — DELETING REJECTIONS

System records that did not respond are removed to generate high accuracy only among those that produced a response.

CH02-F14 — DEFINITE JUDGEMENT FROM SMALL CELL

A country or language score is produced from one or several users.

CH02-F15 — HIDING THE CRITICAL TAIL

Rare severe errors are lost within the overall pass rate.

CH02-F16 — SINGLE SCORE REPORT

Only the headline score is published without showing scope, inequality, stability, uncertainty, and severe errors.

CH02-F17 — COUNTING DISTRIBUTION AS "AI'S BELIEF"

Observed user surface results are presented as if they are the invariant and intrinsic viewpoint of artificial intelligence.

CH02-F18 — COUNTING THE UNMEASURED AS ZERO VARIANCE

Untested areas are considered smooth and stable.

27. AUDIT PROCEDURE

Whether a GEO report honestly represents the distribution should be examined with the following steps.

Step 1 — Record the Public Claim

Example: “The brand's global NOMOS score is 92.” The full sentence and the surface where it is found are recorded.

Step 2 — Define Target Distribution

For which score:

- user universe,
- AI products,
- countries,
- languages,
- time period

is it applied?

Step 3 — Extract Measurement Cells

Each material combination is determined:

- Product
- Country
- Language
- Panel
- Prompt
- Wave
- User surface

Step 4 — Rebuild Denominators

Assigned tasks, completed tasks, valid records, adjudicated records and passed responses must be reported as separate counts.

Step 5 — Extract Result Statuses

Not only the pass rate:

- partial,
- unsupported,
- conflicting,
- old,

- Critical,

UNKNOWN,

rejected, inconclusive distribution is examined.

Step 6 — Verify Weights

Weights:

- are they predefined,
- which question do they answer,

have they changed after the result?

Step 7 — Examine Subgroup Differences

The best group, the weakest sufficient group, language difference, country difference, product difference, and panel difference are evaluated.

Step 8 — Review Scope and Deficiency

Which areas were not measured? Which cells are insufficient? What are the reasons for the deficiency? Could the missing data be systematic?

Step 9 — Examine the Determination

The change between Waves Sessions Users Interfaces is evaluated.

Step 10 — Examine Queue Risk

Critical and other serious errors:

- number,
- its rate,
- type,
- user effect

is evaluated separately in terms of

Step 11 — Compare the Public Summary with the Distribution

Does the single number or sentence published to the public accurately convey the direction and risk of the full distribution?

Step 12 — Limit the Result

The public claim is limited only to the level carried by the distribution and scope.

28. NECESSARY EVIDENCE

To the extent relevant for a representative distribution or a global GEO score, the following records should be searched:

- Audited entity record
- Definition of target population
- Sampling frame
- AI product and user interfaces

- Country and language cells
- Prompt and versions
- Session and panel conditions
- Measurement waves
- Assigned task numbers
- Completed tasks
- Valid and invalid records
- Raw responses
- Result status distribution
- Adjudication records
- Critical error records
- Rejection and inconclusive records
- UNKNOWN records
- Weights
- Stakeholders
- Subgroup results
- Stability results
- Out-of-scope areas
- Confidence or uncertainty calculations
- Public scorecard
- Responsible person or institution
- Protocol and analysis version
- Change history

29. AUDIT CHECKLIST

Is the distribution's target population clear? Are the audited AI products and user surfaces specified? Are the country, language, prompt and time scopes visible? Are the weights used to produce the composite score disclosed? Is the denominator fully defined? Are invalid records reported separately? Are refusals and inconclusive outcomes preserved? Are UNKNOWN and NOT TESTED distinct? Are subgroup results visible? Have small cells been interpreted with excessive certainty? Has the weakest material group been retained? Are Critical errors reported separately? Are measurement waves shown separately? Does the average conceal decline or volatility? Is a merely formal difference distinguished from a material difference in meaning? Has context-sensitive variation in recommendations been classified correctly? Have different surfaces from the same provider been merged without explanation?

Have areas outside the scope been included in the global result? Have synthetic and real distributions been separated? Does the sentence presented to the public reflect the true direction of the full distribution? If one of the substantive questions is unanswered, the distribution result carries limited confidence. If more than one is unanswered, a single global score should not be presented as a definitive judgement to the public.

30. OBJECTIONS AND ANSWERS

Objection 1 — “The public wants a single number.”

This is correct. A single number can facilitate communication. NOMOS does not completely reject a single number. However, a single number:

- scope,
- Critical error,
- subgroup base,
- time,
- uncertainty

must be connected to the information. A number can be short. It cannot shorten the truth.

Objection 2 — 'Small languages unnecessarily lower the global score.'

The global population weight can account for a small language with limited weight. However, heavy representation distortion in a small language cannot be left unnoticed simply because its population share is low. Therefore:

- population-weighted main score,
- as a result of separate language fairness

They can be published together. Two metrics answer different questions.

Objection 3 — 'If there is personalisation, different responses are normal.'

Some differences can be normal and correct. However, customisation:

- company identity,
- licence,
- its current price,
- the real service area

should not be changed arbitrarily. Personalisation can affect the user's appropriate explanation and advice. It is not permission to invent a material fact.

Objection 4 — “Why would it be a problem if the same answer is not given?”

The problem is not word difference. The problem is the change of material reality. NOMOS does not look for exact text equality, but conditional semantic consistency.

Objection 5 — “Why would it matter if a Critical error occurs only once in 1,000 answers?”

This error may be rare in terms of population prevalence. However, it can be severe in terms of damage. The uncontrolled loss of a critical system on an airplane once in a thousand flights is not assessed as only 0.1 per cent. In the context of GEO, frequency and damage should also be examined separately.

Objection 6 — “We cannot find a sufficient sample for every small country or language.”

This is true. Therefore:

- global population panel,

- Country Observer panel,
- Language Fairness Panel

can be established separately. A single-person observation does not turn into a country score. It only indicates that an observation exists. The correct status for an insufficient sample:

INSUFFICIENT SAMPLE

or:

NOT ESTIMATED

must be.

Objection 7 — "This much detail is too complex for the client to understand."

It is not necessary for the client to see all the formulas on the main page. However, the key facts necessary for the decision cannot be hidden. The short results card should answer the following questions: What is the overall result? What is the most serious risk? What is the weakest area? What has not been tested? How robust is the result? What date does it belong to? Complexity needs to be managed. Not eliminated.

Objection 8 — "If the same model is used in all countries, why would there be a country difference?"

The output seen by the user may not consist solely of the bare model. Country differences include:

- product access,
- web resources,
- locale,
- language,
- plan,
- interface,
- legislation,
- personalisation

It can emerge through it. A difference may not occur. However, its absence cannot be assumed without measurement.

Objection 9 — “Reliable scoring is not possible since AI responses are variable.”

Variability does not make measurement impossible. On the contrary, it is one of the features that needs to be measured. The score:

- specific date,
- scope,
- confidence interval,
- stability information

can be meaningful as long as it carries it. The claim of timeless certainty, however, is indefensible.

Objection 10 — “If the prompt changes, the distribution completely changes.”

That is correct. Therefore, the distribution is conditional on the prompt. GEO-1000 does not produce the single answer for all possible questions. Predefined:

- Core Mirror Prompt,
- Diagnostic Prompt Registry

Measures specific representational domains. the prompt constitution will be defined in Section 10.

COMMON RULE OF SECTION 32

This section was not written against a single score. It was written against assuming a single score alone represents the entirety of reality. People want numbers. Institutions want rankings. Boards want to see a single line. The market may ask: "What is your NOMOS score?" This question is legitimate. But the answer cannot be only: "90." Because behind 90, any of the following realities may exist:

- Consistent accuracy across all languages
- 99 in some languages, 45 in others
- High average but severe Critical error
- High accuracy but low coverage
- High instant results but rapid decline
- Success only in easy prompts
- High rate only among assessable answers
- Assuming untested countries positively
- Combining selected AI products

The value of a score is formed by the visibility of the underlying distribution. If the distribution is not visible, the score does not generate confidence. It only produces an appearance of certainty. Therefore, NOMOS's second measurement law is as follows:

> *a GEO score can be a summary of the distribution. / It cannot replace the distribution.*

Its third law is:

> *A high average does not eliminate the question of who is poorly represented.*

The fourth law is as follows:

> *A rare serious mistake is not insignificant just because it is rare.*

The fifth law is as follows:

> *The untested field is not the positive side of the distribution; it is the unknown boundary.*

Order 2 of NOMOS

> *Don't just show me the average.*
>
> *Show who is below the average.*
>
> *Do not hide your worst language while showing your best language.*
>
> *Do not use the weight of big countries to erase the existence of small countries.*
>
> *Do not send one person to a small country and then announce a definitive score for that country.*
>
> *Do not trivialise one severe error merely because it is hidden within a response that is 99 per cent correct.*
>
> *Do not declare an error merely because the wording of an answer has changed. / But when the underlying reality changes, do not dismiss it as a mere stylistic difference.*

> *Do not throw rejections, unknowns, and untested ones outside the distribution.*
>
> *Do not change the weight, share, or scope after seeing the result.*
>
> *Generate a number. / But do not make the countries, languages, users, times, and errors behind the number invisible.*
>
> *Because the representation of an entity is not an average sentence. / It is a distribution of the different realities encountered by different people.*

The Chapter's Closing Sentence

> *The measurement object of GEO is not a single answer or a single average score of an artificial intelligence; it is the entirety of representative results generated in a defined user population under defined conditions.*

Normative Core

> **An entity's generative-system representation MUST be treated as a conditional distribution of observable outcomes across defined: - AI products and user surfaces, - users or sampling units, - countries, - languages and locales, - session and personalisation states, - prompts, - and measurement times. A single aggregate score MUST NOT replace disclosure of: - scope and coverage, - outcome distribution, - subgroup variation, - critical-error incidence, - unknown and untested mass, - temporal stability, - weighting rules, - and uncertainty. Equal averages MUST NOT be interpreted as equivalent representation quality when their subgroup, severity, coverage, or stability distributions materially differ. Untested groups MUST NOT be imputed as conforming, and rare high-harm outcomes MUST remain separately visible from average performance.**

CHAPTER 03

What Exactly Is the Entity Under Audit?

METHODOLOGICAL BASES / K10; K11

Chapter Boundary

Section 1 established the following provision:

> *A single response is not a GEO score.*

Section 2 added the following provision:

> *The representation of an entity in generative systems is a distribution that occurs among users, AI products, countries, languages, sessions, and times.*

This section answers the next mandatory question:

> *Whose or what representation does this distribution exactly measure?*

A company? A brand? A domain name? A product? A service? A founder? A group of companies? A local legal entity in a specific country? A franchise business? When an artificial intelligence system sees only the apple.com domain:

- the domain name,
- the brand,
- the parent company,
- the local sales entity,
- the product family,
- the store service

Will the system treat all of them as the same entity? Unless the distinction is made before the audit begins, a result may be calculated correctly yet applied to the wrong object. This chapter:

- the audited entity,
- the identity anchor used in the prompt,
- the entity analysed by the system,
- the unit to which the conformity decision will be applied,
- relationships between entities,
- which entity the attributes belong to,
- temporal and geographic identity boundaries

defines. This chapter does not yet:

- how all claims about the entity will be verified,

- how adjudicators will score atomic claims,
- country and language sample quotas,
- the final NOMOS score formula

is not finalised. The task of Section 3 is more fundamental:

> *To lock the identity and boundaries of the object being measured before the measurement begins.*

NOMOS Challenge

You are asking me this question: “What kind of company is Apple.com?” There can be at least three different objects in the question:

- the apple.com domain name
- the Apple brand
- the legal company associated with the domain name and the brand

I would choose one of these. Maybe the correct one. Maybe the wrong one. Maybe I combine more than one in a single answer. Then you would score the answer. But you did not tell me in advance which entity you are scoring. I may have:

- described the domain as an online store,
- described the brand as a consumer electronics brand,
- described the legal company as a global corporation,

I may have listed the products as if the company itself were presenting them. My answer may appear fluent. But if the audit object is unclear, whether it is correct or not also remains uncertain. Now, consider a more complex example.

A brand is operated by the parent company in one country; by a subsidiary in another country; and by an independent distributor in a third country. A user asks: "Does this brand provide services in Turkey?" Which entity should answer? The brand's global site? The local company in Turkey? The distributor?

The franchise operator? Does the licence owned by the global company automatically transfer to the local franchise? Do customer reviews of the local operation prove the global quality of the main brand? Can the number of employees of the parent company be used as the capacity of the local company?

Can a product's award be transferred to all the companies that sell it? Does a founder's personal experience become the corporate age of the company? Is the previous owner and the new owner of a domain name considered the same entity?

When a company acquires another company, do all the achievements of the previous company become a characteristic of the new company? GEO score can be generated without answering these questions. However, it is not known what the score represents. The first ruling of this section is as follows:

> *An entity cannot be measured until it is resolved.*

Its second provision states:

> *Sharing the same name is not being the same entity.*

Its third provision states:

> *A relationship does not automatically allow properties to be transferred between the parties.*

1. PURPOSE OF THE CHAPTER

The purpose of this section is to determine the identity, boundary, and relationships of the entity measured within GEO-1000 before the results are seen. The section makes the following distinctions normative:

- Domain name and legal company
- Brand and brand owner
- Parent company and subsidiary
- Group and group company
- Product and manufacturer
- Service and service provider
- Founder and institution
- Employee and institution
- Licence holder and the party using the brand
- Franchisor and franchise operator
- Manufacturer and distributor
- Platform and seller on the platform
- Technology provider and technology user
- Customer and business partner
- Active relationship and historical relationship
- Official name and descriptive translation
- Different entities with the same name
- Previous and subsequent versions of an entity
- Audited object and object to which the conformity decision will be applied

At the end of this section, each GEO audit should be able to provide a single and traceable answer to the following question:

> *Exactly which entity, with which relationships, for which period, and for which digital surfaces does this score and decision belong?*

2. CENTRAL NORMATIVE PROVISION

> *Before data collection begins, every GEO-1000 audit must define a unique audited entity, a canonical entity record, a scope boundary, a time boundary and a relationship graph.*

None of the following, on its own, is sufficient to identify an entity:

- name,
- logo,
- domain name,
- social media account

These identifiers alone do not define the audit object. The audited entity must be described, where relevant, through at least the following elements:

- Entity type
- Canonical name

- Legal name
- Brand names
- Official localised names
- Abbreviations and aliases
- Establishment or start date
- Jurisdiction
- Registration number or equivalent official identity
- Parent and subsidiary company relationships
- Trademark ownership
- Domain name control
- Product and service ownership
- Actual service delivery responsible
- Claimant
- Evidence holder
- Active countries and locales
- Validity period
- Historical name and relationships
- Surfaces included in the audit
- Entities and surfaces outside the scope

The main GEO score should not be published until this record is completed.

3. WHAT IS AN ENTITY?

In this protocol, an entity is an object that can be assigned a separate and traceable identity, characteristic, relationship, responsibility, or time status. An entity:

- human,
- legal entity,
- public institution,
- brand,
- product,
- service,
- domain name,
- digital application,
- group of companies,
- physical branch,
- programme,
- certificate,
- publication,
- research project

It is possible. Not every mentioned item must be a separate entity. However, if one of the following areas changes materially, a separate entity record may be required:

- Legal responsibility

- Ownership
- Licence
- Product or service delivery
- Jurisdiction
- Time period
- Obligation to the user
- Evidence and claim ownership
- Audit scope

4. SIX SEPARATE OBJECTS IN GEO AUDIT

In a GEO audit, the word “entity” often conceals six separate functions. These six objects must be distinguished from each other.

4.1. Prompt anchor

A prompt anchor is an expression or sign used to invoke the existence in a user's query. Examples:

- Company name
- Brand name
- Domain name
- Founder name
- Product name
- Application name
- Local company name
- Abbreviation

Example: “What kind of company is Apple.com?” the prompt anchor here: apple.com is the domain name. the prompt anchor does not have to be the entity itself. It is only the entry point used for the system to resolve the entity.

4.2. Target Entity

The target entity is the entity that the protocol actually wants to measure. Example: the canonical corporate entity associated with the apple.com domain name can be targeted. However, its exact identity must be recorded before the audit. The target entity cannot be chosen by looking at the results after the query.

4.3. Entity Resolved by the System

A resolved entity is how the AI product interprets the query anchor as an entity. Representation:

```
E_i
```

can be used. The target entity:

E*

is the identity resolution assessment that asks the fundamental question:

```
E_i = E*?
```

However, equality is not always just equality of name. The system can identify the correct brand and corporate entity without naming the correct legal entity. Therefore, identity matching later:

- exact match,
- acceptable brand match,
- partial match,
- wrong entity,
- uncertain,
- multiple entity merger

can be classified as.

4.4. Object of the Claim

The object of the claim is the entity to which a specific feature in the response belongs. Within the same response:

- establishment date to the parent company,
- product feature to the product,
- customer review for local service provider,
- licence to a specific legal entity,
- founder's experience to a person

may belong. Even if the answer started with the correct target entity, it may transfer features to incorrect objects. Therefore, each factual claim should be evaluated in the following structure:

```
(subject, relationship or attribute, value, time, scope)
```

Example: “X has a Y licence.” Here, it should be determined to which:

- legal entity,
- country,
- during which dates,
- for the activity

it belongs.

4.5. Audit Object

The audit object is a defined entity or package of entities where evidence, content, technical surface, and AI outputs are examined together. An audit object can be a single entity. Example: A specific legal company Or it can be a package with clear boundaries: Main brand + canonical domain name + specified two products + Turkish and English public surfaces. If a package is used, the entities within it must retain their separate identities. "Package" does not mean that all attributes are transferred to each other.

4.6. Conformity Unit

The conformity unit is the specific object to which the audit decision will be legally and normatively applied. A conformity decision:

- applies to the entire group of companies,
- applies to all languages,
- applies to all domain names,
- to all products

cannot be applied automatically. Example: “The Turkish and English public pages of the brand NobleJackal on example.com have been found compliant according to the specified standard and within the specified date range.” This decision:

- all internal processes of the company,
- affiliated companies,
- separate applications,
- languages to be added in the future,
- products outside the scope

does not automatically cover.

5. TYPES OF ENTITIES

Every audit entity must be classified with one or more types. Primary type and secondary types can be distinguished.

5.1. Legal Entity

It is a company that carries separate rights, responsibilities, or registration under the law or relevant legal regulation. Its fields to the extent they are relevant:

- Legal name
- Company type
- Place of registration
- Registration number
- Tax or equivalent identity
- Date of establishment
- Active status
- Authorised representative
- Parent and subsidiary relationships

A legal entity may carry the same name as the brand. However, it is not the same object.

5.2. Brand

It is the name used in the market for a product, service, or corporate identity. Brand:

- may not have a separate legal entity,
- may be owned by another company,
- may be used under licence,

may be operated by multiple companies in different countries. The actual delivery responsible for the work claimed by a brand must also be specified.

5.3. Group of Companies

It is a structure formed by multiple legal entities that have joint ownership or control relationships. A "Group" cannot automatically transfer the following features to all members:

- Licence
- Certificate
- Number of customers
- Number of employees
- Revenue
- Office
- Audit result

- Security status
- Conformity mark

If the group-level value is truly consolidated, it should be clearly stated.

5.4. Subsidiary or Associate

A legal entity that is under the ownership or control relationship of another entity. A subsidiary:

- may be a separate contract party,
- a separate licence holder,
- a separate employer,
- a separate data controller

may be. The characteristics of the parent company are not transferred by default to the subsidiary.

5.5. Domain Name

It is a digital address and publishing surface. Domain name:

- is not a legal entity,
- does not deliver services on its own,
- is not required to own a trademark,

can be controlled by different owners or operators at certain times. A domain name can be the starting point of an audit. However, it should also be shown under which entity the decision was made.

5.6. Website or Digital Publication

It is the entirety of specific content and applications on a domain name. The same domain name can be:

- corporate site,
- store,
- support portal,
- investor page,
- developer documentation,
- different country and language versions

may carry. The content owners and oversight scopes of these may differ.

5.7. Product

A commercial or technical object carrying a specific function, model, or version. A product's:

- price,
- safety,
- technical feature,
- certificate,
- user suitability

cannot be used as general characteristics belonging to the company. Versions within the same product family may also differ.

5.8. Service

Activity offered under a specific scope, user, country, delivery method, and responsibility. A service:

- marketable by the brand,
- billable by another legal entity,

deliverable by an independent partner. Audit should distinguish who:

- sold it,
- delivered it,
- warranted it,
- legally assumed it

should be separated.

5.9. Person

Is a founder, executive, employee, consultant, expert, or publicly recognised representative. A person's:

- education,
- experience,
- licence,
- opinion

It is not automatically transferred to the institution. Likewise, the award or certificate of the institution is not considered a professional competence of the person.

5.10. Branch or Physical Location

It is a physical service point that is affiliated with the main institution or operated separately. A branch:

- part of the parent company,
- separate legal entity,
- franchise,
- Independent licensed operator

maybe. The position relationship does not show the legal relationship alone.

5.11. Franchise

It can be an independent business operating under a brand usage right or business model licence. The franchisor must:

- reputation,
- control,
- training

does not automatically guarantee every action of the franchise operator. A local mistake of the franchise operator cannot be automatically generalised to the entire global brand. The relationship should be reported with its scope.

5.12. Distributor, Dealer or Seller

The party that sells or distributes the product or service in a specific market, whether independent or affiliated. The manufacturer’s licence, control, or corporate compliance is not automatically transferred to the distributor. The seller’s customer service performance is also not the same metric as the manufacturer’s product quality.

5.13. Platform and Platform Seller

The marketplace or platform and the independent sellers on the platform are separate entities. Platform:

- payment,
- listing,
- technical infrastructure

can be provided. Vendor:

- product,
- delivery,
- warranty,
- customer relations

can be responsible. If the AI combines these roles, incorrect responsibility assignment may occur.

5.14. Programme, Badge or Certificate

Offered by an institution:

- training programme,
- conformity mark,
- certificate,
- membership,
- badge

It can carry separate entity records. The institution that publishes the programme is not the same as the institution participating in the programme. The feature of the badge does not translate into all the features of the institution that bears the badge.

6. ENTITY GRAPH

A single entity record may not be enough to explain complex relationships. For this reason, an entity graph must be created for each audit. Views:

G_E = (V, R)

Here:

- V: entity nodes
- R: directional relationships between nodes

Relationships should be defined by explicit verbs.

6.1. Basic Types of Relationships

To the extent relevant, the following relationships can be used:

- owns — owns
- ownedBy — is owned by
- controls — controls
- controlledBy — is controlled by
- operates — operates
- operatedBy — is operated by

- publishes — publishes
- publishedBy — is published by
- manufactures — manufactures
- manufacturedBy — is manufactured by
- providesService — provides service
- serviceDeliveredBy — service is delivered by
- licencesBrandTo — licences brand to
- franchisesTo — franchises to
- distributesFor — distributes for
- sells — sells
- employs — employs
- foundedBy — was founded by
- memberOf — is a member of
- certifiedBy — is certified in the specified scope by
- auditedBy — has been audited by
- partneredWith — has a partnership in the specified scope with
- customerOf — is a customer of
- supplierTo — is a supplier to
- formerlyKnownAs — formerly known as
- predecessorOf — is a predecessor of
- successorTo — is a successor of
- acquiredBy — has been acquired by
- mergedInto — has been merged into
- ceasedOn — has ceased operations on

Vague verbs like 'related', 'partner', 'working together' should not be used alone as much as possible.

6.2. Direction of the Relationship

Relationships are directional. 'A provides technology to B.' is not the same as: 'B provides technology to A.' A brand may belong to a company. The company does not belong to the brand. A client may have worked with an agency. The agency does not belong to the client. If the direction is not specified, the system may make incorrect responsibility and authority transfers.

6.3. Time of the Relationship

Every financial relationship to the extent it is relevant:

- start date,
- end date,
- activity status,
- last verification date

must carry. An old partnership is not an active partnership. A former employee is not a current executive. A distributor once used may not be an authorised representative today.

6.4. Scope of the Relationship

A partnership:

- may be limited to a specific product,

- country,
- campaign,
- date,
- event

It cannot be generalised that a single-project collaboration applies to all corporate activities.

7. ASSUMED TRANSFER OF CHARACTERISTICS IS PROHIBITED

The existence of a relationship between entities does not automatically transfer a characteristic from one to another. One of the fundamental principles of this section is:

> *Feature transfer is prohibited by default; however, it can be allowed with explicit relation, scope, timing, and evidence.*

7.1. Licence Transfer

The licence of the parent company:

- to a subsidiary,
- to a franchise,
- to a distributor,
- to a brand

do not automatically transfer. The official scope of the licence must be examined.

7.2. Certificate Transfer

A product's certificate cannot be transferred to the entire product family. The quality certificate of one office cannot be generalised to the global group. A company's conformity decision does not prove the personal expertise of its employees.

7.3. Employee Number Transfer

The total of the group cannot be used like the number of employees of a local company. The agency network or freelancer pool cannot be shown as the number of direct employees.

7.4. Customer Number Transfer

The number of customers of the main group cannot be written as the number of customers of a newly established brand. The number of users on a platform may not correspond to the number of active customers of a specific product.

7.5. Establishment Date Transfer

The founder's starting year in the industry is not the company's establishment date. A company cannot use the founding date of an old brand it acquires as its own legal establishment date. The brand date and the legal entity date should be shown separately.

7.6. Success and Case Transfer

A project carried out by an employee at a previous employer is not the corporate case study of the new company. The success of an acquired company, without explanation of integration and scope, is not considered the direct success of the new group.

7.7. Transfer of Eligibility

Eligibility granted for certain language pages of a domain name:

- to the entire company group,
- to new products,
- to other domain names,
- to future pages

is not transferred.

8. FOUR SEPARATE OWNERSHIP LAYERS

In a digital representation, the word "owner" can carry multiple meanings. Four layers should be distinguished.

8.1. Legal Ownership

The entity is the owner under law or official registration.

8.2. Trademark Ownership

It is the owner of the rights over the trademark or trade identity.

8.3. Digital Surface Control

It is the party that actually controls the domain name, website, social profile, or application. The legal owner and the technical controller can be different.

8.4. Content and Claim Ownership

It is the party responsible for the accuracy of the published material claim. The agency may manage the site. However, the true owner of the price or licence claim may be the client institution. A technical provider does not become the material owner of the claim just because they published the content.

9. WHO SELLS THE SERVICE, WHO DELIVERS IT?

This is one of the most important entity distinctions in the GEO assessment. In a service chain, the following parties may differ:

- The one using the brand
- The one advertising
- The one preparing the offer
- The one signing the contract
- The one issuing the invoice
- The one actually delivering the service
- The one providing the guarantee
- The one processing the data
- The one handling the complaint
- The one bearing legal responsibility

An AI may assign all these roles to a single party by only stating the brand name. This error is especially common in:

- tourism,
- finance,
- health,
- consultancy,
- franchise,
- platform,
- agency,
- education

fields can result in financial consequences. Each service record should answer these questions: Who does the user contract with? Who receives the payment? Who performs the service? Who bears professional or legal responsibility? Who is responsible for warranty and return obligations? What is the brand's role? Are there third-party providers? This distinction also determines which entity the GEO score applies to.

10. ENTITY BOUNDARY

The entity boundary is the line that defines which features, relationships, and surfaces the audited object includes and which it does not. An entity boundary can be defined in four layers.

10.1. Identity Core

It is the fundamental identity of the entity that must remain unchanged:

- Canonical name
- Entity type
- Legal or institutional identity
- Main activity
- Primary ownership
- Active status

10.2. Controlled Environment

Directly controlled by the entity:

- domain names,
- applications,
- brands,
- products,
- documents,
- official profiles

This field. Control does not mean that all features merge into a single identity.

10.3. Connected Environment

Affiliates are local companies, franchises, and controlled partners. They retain their separate identities.

10.4. External Relationship Environment

Customers, suppliers, media, memberships, technology providers, and independent partners. Entities in this environment are not part of the audited entity. They are objects of the relationship.

11. LOCKING THE ENTITY BOUNDARY WITHOUT SEEING RESULTS

The audit object cannot be expanded or narrowed based on results. The following behaviours are prohibited:

- Adding positive group results to the brand
- Excluding a failed affiliated company from the scope
- Spreading a good product result across the entire product family
- Declaring the bad language version as a separate entity
- Leaving the critical error to the local operation while making the global brand appear clean
- Making strong local evidence the evidence of the global company

Scope of the entity:

- Before the audit begins,
- Before prompts are distributed,
- Before results are seen

It must be versioned first.

12. CANONICAL ENTITY RECORD

A Canonical Entity Record must be created for each audited entity. This record clarifies which object the adjudicators and systems are evaluating.

12.1. Minimum Fields

Identity

NOMOS entity ID Canonical name Entity type Legal name Abbreviations Official local names Former names

Legal and Corporate Status

Jurisdiction Registration number Date of establishment Active status Parent company Subsidiaries

Brands and Products

Owned brands Licensed brands Products Services Version or model distinctions

Digital Surfaces

Canonical domain names Local domain names Applications Official profiles Machine-readable records

Responsibility

Claimant Content owner Technical surface owner Service delivery owner Legal responsible

Time

Record version Effective start Validity end Last verification date Historical changes

Scope

Audited countries Audited languages Audited products Out-of-scope entities and surfaces

12.2. Entity Statuses

Each entity record must carry one of the following statuses:

ACTIVE

INACTIVE

HISTORICAL

SUPERSEDED

MERGED

ACQUIRED

DISSOLVED

PROVISIONAL

CONTESTED

UNKNOWN

UNKNOWN or CONTESTED identity cannot be definitively resolved silently.

13. IDENTITY PROOF LEVELS

The identity assurance of an entity record can be classified according to the sources used.

ID-0 — Identity Unknown

Only a name or domain name exists. Legal or organisational matching is not verified.

ID-1 — First-Party Assertion

The institution identifies itself in a specific way. It is valuable. It is not independent verification.

ID-2 — Supported Organisational Identity

First-party records, the domain name, and other reliable records are consistent with each other.

ID-3 — Verified Identity with Official Record

The legal entity has been verified with appropriate official or reliable records.

ID-4 — Multi-Surface and Relationship-Verified Identity

Identity, brand, product, domain name, and corporate relationships are consistent across multiple reliable records.

ID-5 — Identity Maintained Over Time

Identity and relationships have been verified at different periods, and the change history is preserved. These levels are not the final eligibility score. They indicate the evidence confidence of the entity record.

14. DIFFERENT ENTITIES WITH THE SAME NAME

Different entities with the same or similar names may exist. This situation creates entity collisions. Example distinctions:

- Companies with the same name
- Institutions using the same abbreviation
- The same brand name in different countries
- Old and new company
- Brand name with personal name
- Company name with product name
- Brand with general word

Audit cannot rely solely on name matching.

14.1. Collision Resolver Fields

Jurisdiction Legal company type Domain name Physical address Establishment date Founder or manager Product and service Brand owner Registration number Local language name

14.2. Ambiguous Claim

If a claim points to multiple reasonable entities, there are three options: The claim is clarified. Classified separately as an entity resolution test. The system is expected to prompt clarification. The answer obtained from the ambiguous claim cannot be silently added to the main accuracy score.

15. AN ENTITY, MORE THAN ONE DOMAIN NAME

A company:

- global domain,
- local country domains,
- product domains,
- investor domain,
- support domain,
- old domains

can use. Not all of these domains automatically fall under the same audit scope. For each domain:

- controller,
- content owner,
- target user,
- language,
- product scope,
- activity status

It must be determined. A correct record in one domain does not automatically correct the error in another domain.

16. A DOMAIN NAME, MORE THAN ONE ENTITY

On a single domain name:

- parent company,
- different brands,
- local companies,
- products,
- partners,
- sellers

It can take place. In this case, entity ownership should be defined based on the URL or content family. Example:

Surface	Existence role
Home page	Global brand
Investor relations	Legal parent company
Local price page	Local sales organisation
Marketplace	Independent sellers
Career page	Specified employer entities
Support portal	Product or service operator

A single domain name does not mean single responsibility.

17. TEMPORAL ENTITY IDENTITY

The entity identity may change over time. Case examples:

- Change of company name
- Brand transferred to another company
- Merger
- Acquisition
- Division
- Termination of activity
- Change of domain name
- Transfer of the product to another company
- Departure of the founder
- Change of local operator

For this reason, entity registration is time-dependent. It can be considered as E(t).

17.1. Historical Identity and Current Identity

A relationship that was correct in the past cannot be carried over to the present. Example: “X was operated by Y in 2022.” is historical. The following sentence is different: “X is operated by Y.”

17.2. Post-Acquisition Feature Transfer

Acquisition:

- ownership,
- some entities,
- may transfer certain contracts

. However:

- all historical achievements,
- all certificates,
- all customer relationships,
- all licences

are not automatically transferred to the new entity. Which features are legally and practically transferred must be explained.

17.3. Domain Name History

A domain name may carry traces of content or references from the previous owner. The new owner is not automatically considered to have gained the previous authority. If the AI system merges the old and new entities, historical entity confusion occurs.

18. MULTILINGUAL ENTITY IDENTITY

The names of an entity in different languages must be linked to the same identity. However, not all translations are official names. The following types should be distinguished:

- Legal name
- Brand name
- Official localised name
- Transliteration
- Descriptive translation
- Abbreviation
- Former name
- Commonly used name

Example: A Turkish description of a brand does not have to be the legal or official Turkish name of the brand. The machine record should clearly indicate these types.

19. COUNTRY AND JURISDICTION CAN CHANGE THE ENTITY BOUNDARY

Under a global brand, there may be different legal or operational entities in different countries. When a user asks, “Does Company X operate in Germany?” the answer should distinguish:

- the existence of the global brand,
- the local legal entity,
- the distributor relationship,
- the capacity to provide remote services

“Brand is recognised in the country” is not the same as: “Local legal company provides services.”

19.1. Country-Based Entity Roles

For each country, to the extent applicable:

- Legal seller
- Invoice issuer
- Data controller
- Service provider
- Supporting
- Distributor
- Franchise
- Licence holder
- Brand representative

must be determined.

19.2. General Language Page Is Not Proof of Local Company

Having a website in German does not prove that there is a legal entity or a physical office in Germany. An English page does not create global legal jurisdiction. Language is not a jurisdiction.

20. SYNTHETIC APPLE.COM PRESENCE INDICATION

SYNTHETIC METHODOLOGY DEMONSTRATION / The following example has been prepared solely to demonstrate the entity resolution method. It is not the actual audit of Apple Inc., a current corporate structure review, or the real result of any AI product. Prompt: "What kind of company is Apple.com?"

20.1. Prompt Anchor

apple.com

20.2. Candidate Targets

Domain name and website Apple brand Canonical corporate entity Online store function Product family If the target is not defined before the audit begins, adjudicators may score different objects.

20.3. Proposed Target Entity Lock

The target for synthetic testing can be defined as follows: "the main corporate technology entity associated with the canonical domain name apple.com, representing the company's identity and core activities." This definition:

- does not automatically encompass the technical specifications of individual products,
- the local company of each country,
- store prices,
- investment advice

automatically.

20.4. Synthetic Response A

"Apple.com is a global corporate website associated with technology products and digital services." Evaluation: It does not fully match the institution with the domain name. It may leave the company

identity incomplete. The main activity could generally be correct. There may be a lack of material identity.

20.5. Synthetic Response B

“Apple.com is the official website of the company called iPhone.” Evaluation: The product and the company are combined. The target entity is incorrectly resolved. There is an identity error.

20.6. Synthetic Response C

“Apple.com is only an independent store that sells phones online.” Evaluation: The function of the domain name is excessively narrowed. The corporate entity and scope of activities are distorted. The independent store characteristic may be incorrect.

20.7. Synthetic Response D

“Apple.com is one of the main corporate and product information surfaces of the Apple brand; the legal entity operating the site and the local sales side should also be determined according to the country.” Assessment: Domain name, brand, and local legal role are separated. The target entity analysis is more disciplined. Other material claims of the response should be proven separately. This example shows:

> *A response can centre on the wrong entity even if it contains many correct words.*

Identity accuracy comes before content accuracy.

21. SYNTHETIC COMPLEX ENTITY CASE

SYNTHETIC CASE — NOT A REAL INSTITUTION

Let's consider a global brand named Asteron. Structure:

- Asteron Holdings Ltd. — parent company
- Asteron Travel — brand
- Asteron Turkey Tourism Ltd. — local company
- Mira Destination Services — independent local operations partner
- Asteron Club Antalya — franchise operation
- asteron.example — global domain name
- tr.asteron.example — Turkish marketing surface

User asks: "What services does Asteron provide in Turkey?" AI response: "Asteron Holdings Ltd. operates all tours in Antalya itself, has 400 employees in Turkey, and Asteron Club Antalya is directly responsible for customer satisfaction in Antalya." Reality package: The parent company owns the brand. The local company is the sales and contracting side. Mira delivers part of the operations. Asteron Club Antalya is an independent franchise. The 400 employees are the total of the group. The franchise is directly responsible for customer satisfaction at the business level. Errors in the response: The parent company is presented as responsible for service delivery. The total number of group employees is conveyed as if it were the capacity of the local company. The franchise's responsibility is directly assigned to the parent company. The brand, legal entity, and operational partner are combined. The response may have named the brand correctly. However, the entity graph is incorrect.

Therefore, the GEO score cannot rely solely on the mention of the name or the general fluency of the sentence.

22. ENTITY RESOLUTION RESULTS

For each observation, the entity resolution result can be recorded with one of the following statuses.

ER-0 — NOT EVALUATED

The entity resolution has not been evaluated.

ER-1 — EXACT TARGET

The target entity has been resolved correctly and clearly.

ER-2 — ACCEPTABLE BRAND-LEVEL MATCH

No legal name is provided, but the targeted brand and corporate entity are materially resolved correctly.

ER-3 — PARTIAL OR UNDER-SPECIFIED

The response approached the correct area but lacks legal, geographical, or operational identity.

ER-4 — MULTI-ENTITY CONFLATION

Multiple separate entities have been merged as a single object.

ER-5 — WRONG ENTITY

Another person, company, product, or domain name has been targeted.

ER-6 — AMBIGUOUS

It cannot be determined from the response which entity has been resolved.

ER-7 — APPROPRIATE CLARIFICATION

The prompt is indeed ambiguous, and the system has correctly asked for clarification. Requesting a clarification does not have to be considered a failure.

23. ENTITY RESOLUTION SUCCESS IS DISTINCT FROM REPRESENTATION ACCURACY

A system can analyse the correct entity but may provide incorrect information about it. Another system may state some material facts correctly but connect them to the wrong entity. There are four basic situations:

Entity analysis	Material content	Result
Correct	Correct	Strong correct representation candidate
Correct	Incorrect	Correct object, incorrect attribute
Incorrect	Partially correct	Correct information linked to the wrong entity
Incorrect	Incorrect	Complete representation breakdown

For this reason, identity analysis may be a separate gateway. In particular, an answer with serious confusion with another entity cannot be salvaged with a high factual accuracy average.

24. LINKING THE CONFORMITY DECISION TO THE ENTITY

A conformity decision should be established in the following format:

> *[Entity ID] + [entity type] + [scope] + [digital surfaces] + [country and languages] + [time] + [standard version]*

Example template: “For [brand/legal entity] identified with Entity ID NGE-XXXX NOMOS; the public pages in Turkish and English on the specified domain, within the specified product and service scope, have been evaluated on [date] under [standard version].” Then it should be clearly stated: “This decision does not automatically apply to subsidiaries, franchises, out-of-scope products, other domains, and languages to be added in the future.”

25. MANDATORY NORMATIVE PROVISIONS

CH03-N01

Every GEO audit must identify a unique target entity before results are seen.

CH03-N02

The prompt anchor cannot be assumed to be the same as the target entity.

CH03-N03

Domain names, trademarks, legal entities, products, services, and personal identities must be distinguished from each other.

CH03-N04

A property of an entity cannot be transferred to another related entity without clear evidence and scope.

CH03-N05

The parent company's licence, certificate, number of employees, customers, or conformity outcome cannot be automatically applied to subsidiaries.

CH03-N06

The claim at the brand level and the party that actually delivers the service and is legally responsible must be separated.

CH03-N07

The franchisor, franchise operator, distributor, dealer, and platform seller must be registered as separate entities.

CH03-N08

Each tangible entity relationship must carry information on direction, scope, and time.

CH03-N09

Relationships such as "partner," "customer," "expert," "group company," and similar cannot be used ambiguously.

CH03-N10

Collision resolution areas must be used for different entities with the same name.

CH03-N11

If an uncertain claim is to be included in the main score, the uncertainty method must be predefined; otherwise, the claim must be clarified.

CH03-N12

The scope of an entity cannot be expanded or narrowed after the results are seen to increase a positive outcome.

CH03-N13

Historical entities and relationships cannot be used as if they have the current entity status.

CH03-N14

Purchase, merger, or brand transfer does not automatically transfer past characteristics to the new entity.

CH03-N15

A conformity decision can only be applied to the conformity unit specified in the decision.

CH03-N16

The new product cannot automatically inherit the current availability of language, domain name, affiliated company, or country.

CH03-N17

The success of entity resolution and factual accuracy should be evaluated separately.

CH03-N18

The merging of material identity with the wrong entity cannot be made invisible within the general factual average.

CH03-N19

Canonical Entity Records should be versioned and preserve changes over time.

CH03-N20

There should be a responsible human or institutional owner of the entity record.

26. FORMS OF FAILURE

CH03-F01 — CONSIDER DOMAIN NAME AS A COMPANY

The digital address is presented as a legal or operational entity.

CH03-F02 — MERGING THE BRAND WITH THE LEGAL COMPANY

The brand name is used in place of the legal entity that carries contracts and responsibilities.

CH03-F03 — CONSIDERING THE PRODUCT AS A COMPANY

The product name is resolved as a company or brand identity.

CH03-F04 — TRANSFERRING THE FOUNDER'S CHARACTERISTIC TO THE COMPANY

The founder's age, experience, award, or previous project is made a corporate characteristic.

CH03-F05 — TRANSFERRING GROUP CHARACTERISTICS TO THE LOCAL COMPANY

Consolidated customer, employee, or revenue data is assigned to the local business.

CH03-F06 — TRANSFERRING THE LICENSE THROUGH RELATIONSHIP

The parent company, product, employee, or partner licence is used in the name of another entity.

CH03-F07 — UNIFYING FRANCHISE AND THE MAIN BRAND

Local business performance and responsibility are presented directly as the global brand.

CH03-F08 — COMBINING PRODUCER AND SELLER

The product manufacturer and local seller or distributor are considered a single entity.

CH03-F09 — COMBINING PLATFORM AND SELLER

The marketplace platform is made responsible for the seller's product, warranty, or delivery.

CH03-F10 — CONSIDERING THE RELATIONSHIP AS FEATURE TRANSFER

Customer status, membership, technology use or partnership is misrepresented as a transfer of authority.

CH03-F11 — UPDATING HISTORICAL IDENTITY

The old name, ownership, partnership, or manager is used as if it were the current reality.

CH03-F12 — CLAIMING PAST AFTER ACQUISITION

All past successes of the acquired entity are attributed to the new owner without explanation.

CH03-F13 — MERGING ENTITIES WITH THE SAME NAME

Different companies or individuals are resolved as a single entity due to name similarity.

CH03-F14 — ARTIFICIALLY SPLITTING AN ENTITY

A failed language, product, or local operation is declared a separate entity to maintain a positive score.

CH03-F15 — ARTIFICIALLY EXPANDING AN ENTITY

Successful brand or product results are spread across the whole group.

CH03-F16 — CONSIDERING SCOPE OF ELIGIBILITY

A narrow audit result is used for the entire company, group, or product portfolio.

CH03-F17 — COUNTING LANGUAGE JURISDICTION

The existence of a language version is used as proof of legal company or service capacity in that country.

CH03-F18 — COUNTING TECHNICAL CONTROL AS OWNERSHIP

The agency or developer is shown as the legal or commercial owner of the institution because they manage the domain name.

CH03-F19 — LEAVING CLAIMANT UNDETERMINED

It cannot be determined who is responsible for incorrect price, licence, or capacity information.

CH03-F20 — INTERPRETING THE PROMPT ANCHOR AFTER THE RESULT

The target entity that will make the response appear successful is selected afterward.

27. AUDIT PROCEDURE

Step 1 — Record the Prompt Anchor

Exactly which of the user's:

- name,

- domain name,
- brand,
- product,
- person

expression is used is fully recorded.

Step 2 — Identify the Target Entity

It is noted which entity the audit actually measures. The target is classified as:

- legal entity,
- brand,
- product,
- service,
- entity package

The applicable class must be recorded explicitly.

Step 3 — Create Canonical Entity ID

Canonical name Entity type Legal identity Jurisdiction domain Aliases Time status is recorded.

Step 4 — Build the Entity Graph

Directed relationships between parent company, subsidiaries, brand, product, domain name, individuals, and partners are extracted.

Step 5 — Determine Responsibility Roles

Claim owner Digital surface owner Service delivery owner Legal responsible Evidence owner are separated.

Step 6 — Review Attribute Transfer Risk

In response:

- licence,
- certificate,
- employee,
- customer,
- establishment date,
- success,
- compliance

has it been transferred to the wrong entity?

Step 7 — Check Temporal Identity

Answer:

- former name,
- former owner,
- terminated partnership,
- former employee,
- pre-acquisition status

is it being used?

Step 8 — Check Geographic and Local Roles

Are global brand, local company, distributor, and franchise roles correct?

Step 9 — Classify Entity Resolved by System

Appropriate status between ER-1 and ER-7 is assigned.

Step 10 — Lock Conformity Unit

To apply the final decision:

- entity,
- field,
- product,
- language,
- country,
- time

boundary is written.

Step 11 — Limit the Public Claim

The result is narrowed if it is generalised to an institution or group broader than the defined entity.

Step 12 — Record the Way of Change and Objection

If an identity or relationship error is reported:

- who will verify,
- who will change,
- which version will be generated

is determined.

28. NECESSARY EVIDENCE

Canonical entity identity; legal name and entity type; official record or equivalent verification; trade mark ownership; domain-name control; parent and subsidiary relationships; product and service ownership; local-business and distributor records; franchise relationships; contracting and invoicing party; responsibility for service delivery; licence and certificate holder; founder and executive roles; active and historical names; acquisition, merger and transfer records; country and jurisdiction; official-language and transliteration records; entity graph; claimant; content owner; technical-surface owner; audit scope; out-of-scope entities and surfaces; entity-record version; last verification date; and accountable person or organisation.

29. AUDIT CHECKLIST

Has the prompt anchor been recorded clearly? Was the target entity identified before the results were observed? Were the legal entity and domain name separated? Were the brand and brand owner separated? Were the product and company separated? Were the founder's attributes separated from those of the organisation? Are parent–subsidiary relationships correct? Are franchises and distributors

distinct? Are the parties selling and delivering the service identified? Were the claimant and technical publisher separated? Do the licences and certificates belong to the correct entity? Are employee, customer and revenue figures stated within the correct scope? Were historical names and relationships separated from the current state? Were different entities with the same name checked? Were multiple domains and user surfaces linked to the correct entity?

Have multilingual names been linked to the same canonical identity? Are the country and jurisdiction roles clear? Has the entity been classified separately as a result of entity resolution? Is the conformity unit clearly defined? Does it extend to entities outside the decision scope? Is the entity record versioned and dated? Is there a path for objection and correction for identity changes?

30. OBJECTIONS AND ANSWERS

Objection 1 — “The user knows the brand name, not the legal company.”

Correct. It is natural for the user to use the brand in the prompt. The audit does not force the user to use legal jargon. However, the evaluation in the background:

- the brand,
- owner,
- service provider,
- contracting party

must distinguish. A simple answer can be given to the user. The audit log does not have to be simple.

Objection 2 — “In a small business, the brand and the company are already the same.”

In some small structures:

- single brand,
- single domain name,
- sole business owner,
- single service provider

It can be found. In this case, the entity graph is simple. Nonetheless, the record must be made explicitly. Simplicity does not mean ambiguity.

Objection 3 — 'The group of companies acts like a single brand.'

From a marketing perspective, there may be a single brand. In terms of legal, operational, and evidential responsibility, there may be separate entities. If decisions are to be made at the group level, it should be clearly shown which features are consolidated.

Objection 4 — “So much detail unnecessarily complicates the GEO score.”

The correct score given to the wrong entity is not useful. Entity resolution is not an additional burden for auditing, but a condition of validity.

Objection 5 — "AI systems already do not pay attention to these legal differences."

Exactly for this reason, it should be measured. The AI system's:

- brand,
- company,

- product,
- local business

Missing the difference between them is the representation error itself.

Objection 6 — “If the parent company controls the whole group, characteristics can be transferred.”

Control relationships are important in some areas. However:

- licence,
- contracts,
- customer,
- employee,
- data responsibility,
- compliance

such characteristics require separate evidence for transfer. Control is not the unification of all characteristics.

Objection 7 — “Are the achievements of the old company after an acquisition not part of the new group?”

It can be explained in a historical context. However:

- during which period was the success,
- by which team and entity,
- before or after the acquisition

It should be stated when it was produced. The new ownership does not change the history.

Objection 8 — "the prompt is already clear; the user has typed the domain name."

The domain name may be clear. However: the question "What does the domain name represent?" should also be answered. A domain can be a company, store, product portal, or multi-entity platform.

Objection 9 — "Describing the local distributor like a brand is easier for the user."

Simplicity does not override responsibility. The response can be simplified as follows: "The brand provides services in Turkey through an authorised local distributor." It is both understandable and correct.

Objection 10 — "All these relationships change over time."

That is correct. For this reason, the entity record:

- dated,
- versioned,
- should be change-triggered

It does not reduce the need for variability recording. It increases it.

COMMON RULE OF CHAPTER 33

Before generating a GEO score: we ask, “How many answers are correct?” But the more fundamental question that should be asked before this is:

“Exactly whose answers are these about?”

If we consider a domain name as a company, we measure the wrong object. If we consider the brand as a legal company, we assign responsibility incorrectly. If we consider the product as a company, we distort its features. If we turn the founder's experience into the company's history, we exaggerate the past. If we count the group's number of employees as local capacity, we inflate service expectations. If the parent company transfers its licence to a franchise, we direct the user to the wrong authority. If we make the past achievements after an acquisition a feature of the new owner, we rewrite the historical truth. If we extend a narrow domain name audit to the entire company, we whitewash the scope of compliance. Therefore, defining an entity is not just data cleansing. It is the beginning of representational responsibility. A system can describe the wrong entity with the correct information. This is still a misrepresentation. A system can describe the correct entity with the wrong information. This is also a misrepresentation.

Both are required for the correct GEO:

> *Correct entity. / Correct attribute. / Correct relationship. / Correct scope. / Correct time.*

Therefore, NOMOS's third measurement law is:

> *If the measured entity is not clear, the measured accuracy is not clear either.*

The fourth law is as follows:

> *Relationship is not identity.*

The fifth law is as follows:

> *Ownership is not the transfer of all attributes.*

Its sixth law is:

> *Conformity does not spread to neighbouring entities.*

NOMOS Section 3 Order

> *Do not give me a name alone. / Show me which entity that name belongs to.*
>
> *Do not treat a domain as a company, a brand as a contracting party or a product as a manufacturer.*
>
> *Do not use the founder's past as the age of the institution.*
>
> *Do not transfer the parent company's licence to the affiliated organisation, franchise, or distributor.*
>
> *Do not make the group's customer the local brand's customer, or the brand's award the award of all products.*
>
> *Do not describe collaboration as ownership, membership as accreditation, or technology use as partnership.*
>
> *Do not merge the old name with the new entity, or the acquired past with today's performance.*
>
> *Do not expand the entity to magnify positive results / Do not split the entity to hide negative results.*
>
> *First, lock the target. / Then record the direction of every relationship. / Then attach each attribute to its true owner. / Then add the relevant time and country. / Only then decide for that entity alone.*

The Chapter's Closing Sentence

> *The first question in a GEO audit is not what the AI says, but precisely who or what each statement concerns.*

Normative Core

> **Every GEO audit MUST define and version, before observing results: - a prompt anchor, - a target entity, - an entity type, - a canonical identity record, - a relationship graph, - a temporal and jurisdictional boundary, - an audit object, - and a conformity unit. Domains, brands, legal entities, corporate groups, subsidiaries, products, services, persons, franchises, distributors, platforms, and sellers MUST NOT be treated as interchangeable entities. Attributes, credentials, licences, customers, employees, performance, authority, and conformity MUST NOT transfer across related entities unless the relationship, scope, time, and evidence explicitly support that transfer. The entity resolved by the AI product MUST be evaluated separately from the factual accuracy of the response. A conformity decision MUST apply only to the explicitly identified entity and audited scope. It MUST NOT automatically extend to affiliates, products, jurisdictions, languages, domains, future versions, or related entities.**

CHAPTER 04

Two Layers of Reality

Official representation and verified entity reality

METHODOLOGICAL BASES / K10; K11

Chapter Boundary

Section 1 established the following provision:

> *A single response is not a GEO score.*

Chapter 2 added the following:

> *Representation is a distribution that occurs among users, AI products, countries, languages, sessions, and times.*

Chapter 3 then identified the object of the distribution:

> *An entity cannot be measured until it is resolved.*

Now we come to the next mandatory question:

> *According to which reality record will we evaluate whether the AI response is correct?*

Which source should govern the audit: the audited entity's website, a legal record, the company's pricing page, independent reporting, the licensing authority, customer reviews, the company's own performance data or a competitor's claim? When sources conflict, which should be accepted as correct? If a company publishes a false or unsupported claim and the AI repeats it perfectly, has the system succeeded? If the official website is outdated but the AI gives a different answer based on current, verified external evidence, has the system failed? Until these questions are answered:

- correct representation,
- incorrect representation,
- source fidelity,
- factual accuracy,
- entity reality

These terms are sometimes used as though they referred to the same concept. This chapter resolves the resulting confusion by distinguishing two layers of reality:

First Layer

Official Representation Layer

Shows what the entity says about itself on the canonical surfaces it controls.

Second Layer

Verified Entity Reality Layer

Under the existing evidence, it shows which material claims are supported within what scope and time. These two layers are connected. But they are not the same. The first layer answers the question: "How does the entity define itself?" The second layer answers the question: "Which part of this definition is supported by which evidence?" The first book defines the correct standard, the second book separates the forms of violation, and the audit system is designed to measure and prove these distinctions. This section operates that audit task at the claim level.

NOMOS Challenge

Your company's website contains the following sentence: "We are the world's leading GEO company." An AI product reads this page and responds to the user: "This company is the world's leading GEO company." Does the response correctly reflect your website? Yes. Does the response correctly reflect proven reality?

We do not know this yet. Maybe:

- there is comparative research,
- market scope has been defined,
- a leadership criterion has been published,

the result has been independently verified. Maybe none of these exist. If AI had conveyed the company's own claim: "The company defines itself as the world's leading GEO company.", it would have preserved the source of the claim. However, when it said: "It is the world's leading GEO company.", the company may have elevated the statement to independent reality.

Consider another case. The company's official website states, 'We provide services only in Türkiye.' Yet verified records show that the company began operating in Germany and the United Kingdom six months ago, and the website has not been updated. Drawing on current legal and operational records, the AI responds: 'The company provides services in Türkiye, Germany and the United Kingdom.'

The AI is not fully faithful to the official website. Could it nevertheless be closer to verified current reality? Yes. Consider a third case. The company's website states, 'We operate in 20 countries.' Independent and internal records verify only eight. The AI responds: 'The company operates in 20 countries.'

This response:

- faithful to official representation,
- contrary to verified reality

It is possible. The first provision of this section is as follows:

> *Fidelity to the official record is not the same as factual accuracy.*

Its second provision states:

> *What a company says and what can be proven should be recorded separately.*

Its third provision states:

> *An AI repeating an official mistake correctly does not constitute an accurate representation.*

Its fourth provision states:

> *An AI diverging from an outdated official record is not automatically wrong if it carries reliable evidence.*

1. PURPOSE OF THE CHAPTER

The purpose of this section is to establish two separate reference layers to be used in the GEO-1000 evaluation. The section makes the following distinctions normative:

- Declaration of existence vs. proven fact
- Official source vs. independent source
- First-party source versus independent verification
- Fidelity to the source versus material accuracy
- Repetition of the claim versus verification of the claim
- Current official record versus old official record
- Official silence versus externally verified information
- Company perspective versus external reality judgement
- Official evidence versus official claim
- Independence of the source versus relevance of the source to the subject
- Official error versus AI error
- Reference conflict versus model failure
- Uncertainty versus failure
- Claimant's interest versus truth of the claim
- The presence of evidence and whether the evidence actually supports the claim

By the end of this section, two separate questions should be able to be asked for each AI response:

> *Does the response correctly convey the official and canonical representation of the entity?*

and:

> *Does the response correctly reflect the material reality carried by the currently verifiable evidence?*

These questions may yield the same answer. They may also yield different answers. When they differ, the role of the audit is not to conceal the difference, but to explain it.

2. CENTRAL NORMATIVE PROVISION

> *Each material AI response should be assessed against two separate references without transformation: the Official Representation Record and the Verified Entity Reality Record.*

A response:

- faithful to official representation,
- appropriate to verified reality,
- appropriate to both,
- distant from both

It can be. NOMOS cannot quietly combine these two results into a single true–false decision. High performance in the first layer does not compensate for failure in the second layer. Specifically:

> *An AI response that correctly repeats an unverified or false claim published by the entity itself can demonstrate official representation fidelity; however, it cannot demonstrate verified entity accuracy.*

3. FIRST LAYER

OFFICIAL REPRESENTATION LAYER

The Official Representation Layer is the versioned record of the current material statements it publishes about itself on canonical surfaces controlled or explicitly authorised by the entity. This layer answers the question:

> *How does the entity want to be understood, and what information does it officially publish?*

This layer includes:

- legal identity statement,
- brand–company relationship,
- product and service definitions,
- published prices,
- service countries,
- support conditions,
- official policies,
- publicly disclosed boundaries,
- It may contain official currency and version information

The official representation layer is important. Because an AI system must correctly understand how an entity:

- defines itself,
- what service it offers,
- what price it publishes,
- what terms it accepts

However, the official representation layer is not automatic independent proof for every claim. A statement on the institution's own website, especially regarding claims to the outside world such as superiority and performance, presents the statement itself; it does not prove its correctness on its own.

4. WHAT IS AN OFFICIAL SOURCE?

A source is not considered official merely because it carries the institution's logo. An official source should have at least the relevant features of the following:

- Controlled by the entity
- Undergoes an authorised publishing process
- Approved by the owner of the factual information
- Canonical URL or record identifier
- Version
- Publication or verification date
- Change history
- Responsible person or institution
- Scope
- Language and locale status

- Archival or currency status

A view published on a personal social media account by an employee is not automatically an official company statement. A draft file by an agency is not an official publication. An email from a sales representative, if not authorised, is not a general corporate policy. An old PDF on a company's website may not replace the new official record.

5. OFFICIAL REPRESENTATION SURFACES

According to the relevant audit, official surfaces may include:

- Main corporate website
- Legal disclaimers
- Product and service pages
- Pricing and quotation pages
- Official help centre
- Investor relations documents
- Public policies and contracts
- Official press releases
- Structured data
- Canonical JSON records
- Official API outputs
- Official in-app content
- Authorised local sites
- Official social media accounts
- Statements submitted to public institutions
- Release and retraction records

These surfaces may contradict each other. In this case, “official representation” is not a single, clean reality. It is a set of records that must be audited internally.

6. OFFICIAL REPRESENTATION RECORD

An Official Representation Record must be created for each audited entity. This record is not a copy of all official content. It is a traceable summary of factual claims and boundaries. Each record should include the following fields to the extent applicable:

- Claim identifier
- Full text of the claim
- Subject of the claim
- Type of claim
- Canonical source
- Source owner
- Language and locale
- Publication date
- Last verification date
- Validity period

- Official status
- Scope
- Limitations
- Previous version
- Reason for change
- Claimant
- Human approval

Without an official representation record of an entity: the sentence “AI reflects the site accurately.” cannot be clearly tested.

7. THE AUTHORITY OF THE OFFICIAL SOURCE DEPENDS ON THE TYPE OF CLAIM

Not all official sources carry the same weight for every claim. Source authority should be evaluated according to the type of claim.

7.1. Decisions Under the Entity's Own Control

Example:

- Which service it offers to the public
- Published price
- Support hours
- Official return policy
- Communication channel
- Publicly available product version

In these areas, the official source is fundamental and often the primary authorised record. However, if the official page contradicts the actual process flow, the page alone is not sufficient. For example:

- price page 1,000 euros,
- checkout 1,500 euros

If it is showing, there is a contradiction in the official representation layer.

7.2. Legal Identity and Authority

Example:

- Company registration
- Licence
- Professional authority
- Trademark ownership
- Tax or legal status

The company website can disclose this information. However, an official record or an authorised institution can provide stronger verification.

7.3. Performance and Success Claims

Example:

- “98% of our projects are successful.”

- “We increase our customers' revenue by an average of 40%.”
- “We have the highest satisfaction rate.”

The company website may be the owner of the claim. But to support the claim:

- data,
- method,
- denominator,
- period,
- unsuccessful results,
- independent review if possible

are required.

7.4. Superiority and Leadership Claims

Example:

- "World leader"
- "The best"
- "Number one"
- “The only real standard”
- “The most trusted company in the industry”

A site’s own statement is not an independent comparison. Such claims:

- require a comparison universe,
- criteria,
- method,
- date,
- scope,
- independence

Both conditions must be established.

7.5. Third Party Relationships

Example:

- Customer
- Partner
- Sponsor
- Membership
- Accreditation
- Technology partnership
- Distributorship

A unilateral official statement can be valuable. However, it may need to be verified by the other party to the relationship or by a reliable record.

7.6. User Outcomes

Example:

- Customer satisfaction
- Actual usage outcome

- Return rate
- Complaint rate
- Repeat purchase

Company data in these areas is important first-party evidence. However:

- data selection,
- nonresponse,
- methodology,
- conflict of interest

should be disclosed.

8. OFFICIAL representation fidelity

The concept that indicates to what extent an AI response conforms to the Official Representation Layer:

Official Representation Fidelity

The canonical English term is:

Official Representation Fidelity

This measurement asks the following questions: Is the AI interpreting the entity correctly? Does it convey official products and services accurately? Does it distort the official price? Does it remove official boundaries? Does it expand the official country coverage? Does it present the company's opinion as an independent fact? Does it use outdated official information as current? Does it hide contradictions among official records? Does it add material claims not found in the official source? Does it maintain source quality? Official representation fidelity is a measurement of how accurately the entity conveys the reality it has published. This measurement:

> *It shows not that the entity's statements are true, but the extent to which AI accurately conveys what the entity says.*

9. SECOND LAYER

VERIFIED ENTITY REALITY LAYER

The Verified Entity Reality Layer is a versioned assessment record of material information that existed on a specific measurement date and is supported by evidence appropriate to the type of claim. This layer answers the question:

> *Which statements about the entity are truly supported by which evidence?*

This layer includes:

- records coming from the company itself,
- official public records,
- independent sources,
- contract or transaction records,
- licence documents,
- verified historical records,
- counter-evidence,

- adverse findings,
- uncertainties,
- remaining disputes

can be evaluated together.

10. "VERIFIED REALITY" IS NOT ABSOLUTE REALITY

This protocol does not claim: "We know with certainty all the absolute truth about an institution." Verified Entity Reality is:

- evidence-based,
- time-dependent,
- scope-dependent,
- subject to change with new evidence,
- can be contested,
- can be withdrawn,

may carry uncertainty. The correct definition is:

> *Within the specified date and scope, the strongest and verifiable entity record carried by the available evidence.*

Therefore, a verified reality record:

- versioned,
- sourced,
- subject to objection,
- modifiable,
- aware of uncertainty

must be. NOMOS cannot position itself as the owner of the truth. It only classifies what sentence the evidence permits and to what extent.

11. VERIFIED ENTITY REALITY RECORD

The full technical structure of this layer will be detailed in a later chapter of the book. At the foundational level, the following fields should exist for each factual claim:

- Claim identifier
- Subject entity
- Relation or attribute
- Value
- Scope
- Country or jurisdiction
- Time
- Sources
- Source types
- Source verification
- Independence
- Counter-evidence

- Sufficiency of evidence
- Verification status
- Uncertainty
- Adjudicator decision
- Appeal status
- Last verification date
- Record version

This record cannot be created solely from positive evidence. Records that narrow the reality or contradict the claim must also be preserved.

12. VERIFICATION STATUSES

Each material claim can be classified with one of the following statuses:

VR-0 — NOT ASSESSED

The claim has not yet been evaluated.

VR-1 — OFFICIAL CLAIM ONLY

The claim is based solely on the institution's own official statement. It has not been proven false. There is no independent or sufficient additional verification.

VR-2 — FIRST-PARTY EVIDENCE SUPPORTED

The claim is supported by the institution's:

- transaction,
- operation,
- contracts,
- data,
- record

is supported by evidence. The source is first-party.

VR-3 — CORROBORATED

The claim is supported by multiple appropriate and consistent records. Not all sources have to be independent.

VR-4 — INDEPENDENTLY VERIFIED

The claim is supported by an appropriate independent source or review on the subject. Independence alone does not imply sufficiency. The source's relevance to the subject is also required.

VR-5 — MULTI-SOURCE SUSTAINED

The claim is strongly supported by reliable sources of different types within a defined time and scope.

VR-C — CONTRADICTED

The claim conflicts with stronger or more appropriate evidence.

VR-U — UNRESOLVED

The evidence is insufficient or contradictory. A reliable judgement cannot be made.

VR-T — TIME-LIMITED

The claim is supported only for the specified period.

VR-S — SCOPE-LIMITED

The claim is supported only for the specified product, country, user, or operation scope. A claim can hold more than one status at the same time. Example:

VR-4 + VR-T + VR-S

It can indicate that the claim has been independently verified; however, it carries time and scope limits.

13. INDEPENDENCE IS NOT THE SAME AS APPROPRIATENESS

An independent source is not always the best source. A blogger may estimate a company's current price, while the company's canonical pricing page may be the more direct and appropriate source. Conversely, when a company declares on its own website that it is 'the industry leader', independent research with a disclosed method may be the more suitable evidence. Source evaluation therefore asks more than whether a source is independent. Is it authoritative for this type of claim? How close is it to the claim? Is its scope appropriate and current? Who controls it? Is its method disclosed? Does it address counter-evidence? Can it be audited? Is there a conflict of interest? Independence alone does not guarantee accuracy.

A source being first-party does not make it worthless either. The important thing is:

> *It is having the appropriate authority and evidence qualification for the claim carried by the source.*

14. CLAIM–SOURCE SUITABILITY MAP

The table below is a general foundational map. Specific evidence requirements may vary depending on the sector and risk level.

Type of claim	Primary source or evidence	Role of the official site	Need for additional verification
Legal company name	Authorised official record	Shows the declaration	Generally required
Trademark ownership	Trademark registration, contract, corporate record	Shows the declaration	Required in case of financials
Current published price	Canonical price/checkout record	Usually primary	Verifiable with workflow
Scope of service	Operational records, contracts, official page	Primary statement	Additional evidence for actual delivery
Licence	Licensing institution or registry	Shows the declaration	Required
Customer relationship	Contract, customer approval, bilateral record	Can be a unilateral claim	Required in case of financials
Success rate	Raw data, method, denominator, audit	Publishes the claim	Required strongly

Type of claim	Primary source or evidence	Role of the official site	Need for additional verification
Leadership	Comparative method and data	Self-declaration	Independent or strong comparison required
Customer satisfaction	Representative survey, method, non-response	Can explain the result	Methodology required
Date of establishment	Official records and historical documents	Shows the declaration	Legal and brand history should be separated
Partnership	Bilateral statement or contract	Unilateral declaration	Generally required
Certificate	Issuing institution of the certificate and scope	Shows the declaration	Required
Active employee	HR or contract record, public profile	Can show declaration	Role and date verification
User result	Transaction and result record	Describes case	Context, method, and permission required
"Best option"	Defined user and comparison criteria	Marketing view	Strong comparison required

This table is not a universal source ranking. It shows the appropriateness between claim and evidence.

15. DOUBLE LAYER CLAIM MATRIX

Two separate questions are asked for each financial claim: Has it been officially published? What is its status in the verified record of facts? These two questions constitute the following basic situations.

15.1. Official and Verified

Claim: it is present in an official record and supported by appropriate evidence. Example: The price on the official price page is also verified in the transaction flow. This is one of the strongest types of core representation. If AI conveys the claim with the correct scope and timing, success occurs on both layers.

15.2. Official but Self-Declared Only

The claim is official. However, there is no additional verification. Example: “We are the most innovative company in the industry.” AI can convey the claim as: “The company describes itself as the most innovative company in the industry.” It cannot elevate it as: “It is the most innovative company in the industry.”

15.3. Official But Conflicting

The claim exists in the official record. It conflicts with more appropriate or up-to-date evidence. Example: The site says 20 countries. Operational records show eight countries. If AI repeats the site exactly, official fidelity may be high. Verified accuracy is low. This situation:

Carrying the official error faithfully

can be classified as.

15.4. Not Official But Verified

The claim is not found on the official site. However, it has been verified with other reliable sources. Example: Recent legal ownership change has not yet been added to the official site. It appears in the authorised registry. AI can convey this information accurately with the correct source and timing. Official representation may have low fidelity or differ. Verified factual accuracy may be high.

15.5. Not Official and Unverified

Claim: not in the official record, lacks sufficient external evidence. AI should not convey this as if it were true.

15.6. Material Limit Officially Omitted

The company publishes the positive claim on its official page. However, it does not publish the necessary limitation to understand it correctly. Example: "We provide services in Europe." However, the service is only offered in two countries and remotely. If the AI only repeats the general sentence, it can remain faithful to the official text. However, it may produce materially misleading results. Therefore, the Official Representation Layer must record not only the published sentences but also the material omissions.

15.7. Dispute Unresolved

Official and external records contradict each other. Neither is strong enough. The correct result:

UNRESOLVED

should be. AI cannot be expected to make a definitive judgement.

16. DOUBLE-LAYERED EVALUATION TABLE

Fidelity to official representation	Verified reality accuracy	Basic comment
High	High	Faithful and verified representation
High	Low	Carrying the official error faithfully
Low	High	Corrective representation that diverges from the official record but is supported by evidence
Low	Low	General representation failure
Unclear	High	Official source conflicting; external verification strong
High	Unclear	Official statement correctly conveyed; reality verification insufficient
Unclear	Unclear	Insufficient reference for decision

This table establishes an important ruling:

> *AI's non-compliance with the official record is not always a failure.*

Likewise:

> *Compliance of AI with the official record is not always a success.*

17. RULE FOR PRESERVING SOURCE STATUS

The AI response must preserve not only the content but also the epistemic status of the content. The following sentences are not the same:

- "The company is the world's leading GEO organisation."
- "The company describes itself as the world's leading GEO organisation."
- "The company claims to be the world's leading GEO organisation based on its published research that has not yet been independently verified."

- "In the specified independent comparison, the company ranked first."

Each sentence carries a different evidence status. AI:

- self-declaration as an independent result,
- marketing statement as a scientific finding,
- press release as an editorial news,
- sponsored award as independent accreditation,
- company data as external audit

This treatment is prohibited by the following rule:

Citation and Source Status Protection Rule

Source status must remain explicit.

18. EPISTEMIC UPLIFT

The strengthening of a claim's source status in an AI response is called:

Epistemic uplift

Example transformations:

Source text	Incorrect uplift by AI
“According to the company's own data”	“Proven”
“Reported by specific customers”	“Works in all customers”
“Observed in the pilot study”	“Generally effective”
“The company defines itself as a leader”	“Industry leader”
“No independent verification was done”	Limitation completely removed
“Available in some markets”	“Available globally”

These transformations are not summaries. They are the production of new and stronger claims.

19. OFFICIAL SILENCE

The absence of information on a company's official page may not mean:

- the information is incorrect,
- the information is prohibited,

AI cannot use this information. Official silence can take three different forms.

19.1. Non-Material Silence

Information may be outside the prompt and user decision. Whether AI uses it or not is immaterial.

19.2. Legitimate Supplementary External Information

It may be information not found on the official site but verified by a reliable source. Example:

- Legal record
- Historical event
- Public regulatory decision
- Independent research

AI can use this information with appropriate context and source status.

19.3. Silence of a Material Limit

An entity may not have disclosed a limit that affects the user's decision in its official representation. Example:

- The licence being valid only in certain countries
- The price appearing not to carry an additional fee
- The service being available only to certain types of customers
- Success data coming only from selected cases

In this case, the official representation itself may carry an audit finding. AI repeating the silence may reduce the truthfulness of reality.

20. MATERIAL DEFICIENCY IN OFFICIAL REPRESENTATION

An official sentence may not be directly false. However, removing the necessary limits can make it misleading. Example: “We operate in Europe.” Material context:

- only in two countries,
- only remotely,
- without a local office,
- only for corporate clients

may be provided. The official representation layer must record these two areas together:

- The main claim published
- The material limits necessary to correctly understand the claim

If an AI response carries the main sentence but removes the limits:

- faithful in terms of words,
- misleading in terms of judgement

may be.

21. WRONG IS WRONG EVEN IF IT FAVOURS YOU

When an AI product portrays an entity as more powerful, larger or more authoritative than it is, silence by the organisation becomes a measurement and governance problem. A local company may be portrayed as a global business; a consultancy as a regulator; self-assessment as independent certification; or an operation spanning eight countries as operating in 20. Even when an error favours the organisation, it must be recorded as an error in the verified-reality layer. The founding warning of the second book sets the same boundary: a commercially attractive error must still be corrected. When an organisation knowingly exploits a favourable AI error:

- obligation to correct,
- integrity of representation,
- commercial promise,
- governance

can also be examined in terms of.

22. REFERENCE CONFLICT

Sometimes the problem is not in the AI response, but in the reference records. Example: The homepage shows a different price. The PDF shows a different price. Checkout shows another price. Structured data carries an outdated price. In this case, there may not be a single correct official representation. Another example: The legal registry indicates one ownership. The company website states another ownership. Update dates are uncertain. In this case, the verified reality record may also carry a discrepancy.

23. REFERENCE DEFECT RULE

> *If reference records are materially contradictory or insufficient, this uncertainty cannot automatically be converted into a failure of the AI product.*

Correct statuses:

REFERENCE_CONFLICT

INSUFFICIENT EVIDENCE

UNRESOLVED

TIME-SPLIT

SCOPE-SPLIT

are possible. The audit should separate the question: "Did AI provide an incorrect answer?" and: "Was there necessary reference reality for AI to provide the correct answer?" If an institution has left its official records contradictory, we can measure which record the model chooses. However, we cannot hide the reference flaw while expecting a single and definitive institutional answer from the model.

24. CANONICAL RECORD LOCK

Both reference layers of GEO-1000 should be versioned before the measurement wave begins:

Official Representative Record

Which pages? Which language versions? Which date? Which claims? Which archive status?

Verified Entity Reality Record

Which evidence? Which counter-evidence? Which verification statuses? Which objections? Which date? If there is a material change during measurement: The wave can be stopped. The record can be split into a new version. Previous and subsequent observations can be analysed separately. Mixing the results under a single reference is prohibited.

25. BURDEN OF PROOF WITH CLAIM STRENGTH

As a claim grows, the burden of proof also increases. The ladder below is a constructive representation.

C0 — Simple Identity Claim

“X is a consultancy company.” Correct entity and activity record is required.

C1 — Scope Claim

"X provides services in Germany." An operation, contract, or delivery record is required.

C2 — Quantitative Performance Claim

"90% of projects are completed on time." Denominator, period, method, and unsuccessful results are required.

C3 — Comparative Claim

"It is faster than competitors." Comparison universe and equivalent measurement are required.

C4 — Superiority Claim

"It is the industry leader." Industry, criteria, geography, time, and method are required.

C5 — Guarantee or Universal Provision

“Guarantees results for every customer.” True verification requires extremely strong, comprehensive, and actionable evidence. Official sources may be strong for C0 and some C1 claims. Alone, they may not be sufficient at C3–C5 levels.

26. TWO SEPARATE ACCURACY MEASUREMENTS

Two main measurements must be reported separately within GEO-1000.

26.1. Official Representation Fidelity

Answers the following question:

> *To what extent does the AI response convey the versioned official and canonical representation of the entity accurately, completely, and while preserving source status?*

This measurement:

- official identity,
- official service scope,
- published price,
- official boundaries,
- official source nature

assesses their areas.

26.2. Accuracy of Verified Entity Reality

Answers the following question:

> *To what extent does the AI response convey the current entity reality, supported by available evidence, accurately, comprehensively, and with preserved boundaries?*

This measurement:

- supported factual claims,
- conflicting claims,
- outdated information,
- incorrect relations,

- factual deficiencies,
- epistemic enhancements

are assessed.

27. THE INABILITY OF TWO SCORES TO COMPENSATE FOR EACH OTHER

High official fidelity cannot compensate for low verified accuracy. Example: The official site publishes a false claim. AI repeats the claim exactly. Official fidelity may be high. Verified accuracy is low. Similarly, high verified accuracy does not completely trivialise the distortion of official representation. Example: AI uses an up-to-date external source correctly. It confuses the product and service description of one company with another. Some external facts in the response may be correct. Official identity fidelity is weak. Therefore, the two results should appear separately.

28. DOUBLE-LAYER RESPONSE STATUSES

An AI response can receive one of these results:

DUAL-PASS

It is materially consistent with both official representation and verified reality.

OFFICIAL-PASS / VERIFIED-FAIL

Carries the official record faithfully. However, it conveys an unsupported, old, or contradictory claim of the official record as if it were true.

OFFICIAL-FAIL / VERIFIED-PASS

Does not fully comply with the official site. However, it conveys the truth correctly supported by current and stronger evidence. An explanation of the response's difference is preferred.

DUAL-FAIL

It materially disrupts both the official record and the verified reality.

OFFICIAL-PASS / VERIFIED-UNKNOWN

The official statement has been accurately conveyed. The verified reality status of the claim has not yet been determined.

REFERENCE-CONFLICT

Reference records are contradictory in themselves. It is not possible to score the response definitively.

APPROPRIATE-UNCERTAINTY

The AI has correctly explained the reference uncertainty and avoided making a definitive judgement. In some cases, this may be the most accurate answer.

29. SYNTHETIC APPLE.COM DISPLAY

SYNTHETIC METHODOLOGY DEMONSTRATION / The following claims, responses, and scores are entirely fictional. They do not represent real information, performance, or responses of any actual Apple Inc. AI product. Synthetic Official Representation Record: O-01: The company defines itself as a global consumer technology brand. O-02: The company publishes certain product and digital service categories. O-03: The company defines itself as 'the world's most innovative company.' O-04: The official website contains specific country store conditions. Synthetic Verified Reality Record: V-01: Corporate identity and main business category are supported. V-02: Certain product categories are supported. V-03: There is no defined comparison method for the claim 'the world's most innovative company.' V-04: Store conditions vary by country.

29.1. Synthetic Response A

“Apple.com is one of the official corporate interfaces where the company presents its consumer technology products and digital services. The company positions itself as one of the most innovative companies in the world; this claim of superiority should also be evaluated according to the comparison method used.” Evaluation: Official representation is preserved. The claim of superiority is attributed to the company. The claim has not been elevated to independent factuality. The scope limitation has been added. Candidate status:

DUAL-PASS

29.2. Synthetic Response B

“Apple is the undisputed most innovative company in the world.” Evaluation: The official marketing claim has been turned into an absolute truth. There is no comparison method. Source status is lost. “Undisputed” is an additional epistemic elevation. Candidate status:

OFFICIAL-PASS / VERIFIED-FAIL

29.3. Synthetic Response C

“Apple is a local company that only sells phones in physical stores.” Evaluation: Corporate identity and scope are distorted. Official representation is incorrect. Contradicts verified reality. Candidate status:

DUAL-FAIL

29.4. Synthetic Response D

“The company’s official website publishes certain store conditions; however, prices and purchase terms may vary according to the country version.” Evaluation: The official record has been used correctly as a source. Geographic scope is preserved. The official page has not been made a single price source for the whole world. Candidate status:

DUAL-PASS

30. SYNTHETIC ASTERON CASE

SYNTHETIC CASE — NOT A REAL INSTITUTION

The official site of a company named Asteron publishes the following claims:

- “We operate in 25 countries.”
- “We achieve success in 98% of the projects.”
- We are an independently verified global standards organisation.
- All of our projects are guaranteed.

Audit records:

- Verified active service country: 9
- Success data: only the 40 selected projects
- Independent verification: not found
- Standard: the recommended standard published by the company itself
- Guarantee: contractual only for certain delivery items

AI response: 'Asteron is a globally standardised organisation, operating in 25 countries, with a 98% success rate, independently verified, and guarantees results in all its projects.'

30.1. Official representation fidelity

The response is largely faithful to the website text.

30.2. Verified Reality Accuracy

Response:

- expands the country scope,
- makes selected data a general success,
- presents the entity's own standard as an independent standard,

turns delivery guarantee into outcome guarantee. It is a severe failure in terms of verified reality.

30.3. How Could a Correct AI Response Be?

“Asteron states on its own website that it provides services in 25 countries, has a 98% success rate, and publishes a standard. In the existing audit records, active service has been verified in nine countries; it has been observed that the 98% rate is based on 40 selected projects; independent verification could not be confirmed. The guarantee is limited to specific delivery items.” This response:

- does not conceal the company statement,
- preserves the status of the statement,
- distinguishes the verified reality,

makes the limitation visible.

31. SYNTHETIC CASE CONTAINING REFERENCE DEFECT

SYNTHETIC CASE

Official homepage: “Starting price 2,000 euros.” Official price PDF: “Starting price 3,000 euros.” Checkout: “4,000 euros.” Structured data: “1,500 euros.” AI response: “Starting price is 3,000 euros.” Which record did the AI choose? The PDF. Is it wrong? Maybe. Is it correct? Maybe. The reference record is inconsistent. In this case:

- AI failure,

- official representation conflict,
- lack of canonical source

are separate findings. The correct assessment:

REFERENCE-CONFLICT

is possible. An entity cannot expect exact price fidelity without reconciling its own price records.

32. SOURCE PRIORITY IN THE OFFICIAL REPRESENTATION LAYER

When multiple official records are found, a priority rule should be set for each claim field. Example candidate priority:

- Current contract or transaction record
- Canonical up-to-date product/price record
- Authorised structured data
- Current official page
- Current official PDF
- Historical press or blog record

This order is not universal for all claims. For example, press archives might be valuable for corporate history. An old blog record is not authoritative for price. Priority should be defined according to the type of claim.

33. SOURCE CONFLICT IN VERIFIED REALITY

Two independent sources may contradict each other. In this case:

- the more independent one,
- the newer one,
- the one closer to the subject,
- the one with a clearer method,
- the one using broader data

does not automatically win. All criteria should be evaluated together. Example: A new blog post can be more up-to-date than an old official registry. However, the blog may not have sufficient authority for legal ownership. Source conflict resolution should be sensitive to the type of claim.

34. MANDATORY PRESERVATION OF COUNTER-EVIDENCE

A verified record of facts cannot include only records supporting the claim. The following must also be preserved:

- Records contradicting the claim
- Unsuccessful cases
- Withdrawn results
- Expired certificates

- Negative user outcomes
- Regulatory decisions
- Corrections and errata
- Discrepancy between sources

Supporting a claim does not allow the dismissal of counter-evidence.

ARTICLE 35

Not all official–verified differences have the same level of importance. A material difference is a difference that could change the user's or system's decision. Examples: A minor spelling difference in the company name may not be material. Miswriting the type of company may be material. Missing colour of a product may not be material. Incorrect country of service is material. A difference in date format may not be material. Incorrect licence validity date is material. A change in marketing tone may not be material. Removing the warranty limit is material. Materiality should be evaluated based on its impact on:

- user decision,
- legal liability,
- price,
- security,
- authority,
- scope,
- evidence,
- time

These categories must be assessed by their effect on the user decision, legal liability, price, security, authority, scope, evidence and time.

36. MANDATORY NORMATIVE PROVISIONS

CH04-N01

Each material AI response should be evaluated separately against the Official Representation Record and the Verified Entity Reality Record.

CH04-N02

Official representation fidelity cannot be presented as the same outcome as material accuracy.

CH04-N03

A claim on the entity's own site does not count as independent verification without additional evidence appropriate to the claim type.

CH04-N04

First-party evidence cannot be rejected solely for being first-party; its appropriateness to the claim type, method, and conflict of interest should be evaluated.

CH04-N05

An independent source cannot be accepted as correct or authoritative solely for being independent.

CH04-N06

Company opinion, marketing claim, or self-assessment cannot be elevated to independent fact in the AI response.

CH04-N07

The source type and epistemic status should be maintained in the AI response to the extent that it is material.

CH04-N08

An AI response that contradicts the official record cannot be considered an automatic failure if it is supported by stronger, more recent evidence.

CH04-N09

An AI response faithful to the official record cannot be considered successful in general accuracy if the official claim contradicts the verified fact.

CH04-N10

If official and verified records materially conflict, the result should be REFERENCE_CONFLICT, UNRESOLVED, or carry an appropriate uncertainty status.

CH04-N11

A reference defect cannot quietly be turned into an AI product error.

CH04-N12

If the AI response treats an unverified first-party claim as fact, appropriate citation and limitation should be sought.

CH04-N13

Material official omissions cannot be left simply appearing as if the published sentence is correct.

CH04-N14

Errors favourable to the company must also be recorded as errors in the verified reality layer.

CH04-N15

Official Representation Record and Verified Entity Reality Record must be versioned before the measurement begins.

CH04-N16

If there is a change in the material reference during the measurement wave, previous and subsequent observations must be evaluated in separate versions.

CH04-N17

Counter-evidence and negative records cannot be removed from the verified reality file solely because they are negative.

CH04-N18

UNRESOLVED, UNKNOWN, and INSUFFICIENT EVIDENCE results are not considered positive validation.

CH04-N19

As the strength of the claim increases, the required burden of evidence must also increase.

CH04-N20

High Official representation fidelity cannot compensate for low Verified Entity Reality Accuracy.

CH04-N21

There must be an accountable human or institution owner as a result of the two-layer evaluation.

37. FORMS OF FAILURE

CH04-F01 — CONSIDERING THE WEBSITE AS ABSOLUTE REALITY

All claims published by the institution are accepted as verified facts.

CH04-F02 — CONSIDERING OFFICIAL RECORDS AS WORTHLESS

Direct institutional records such as price, policy, or product coverage are ignored because it is the first party.

CH04-F03 — MAKING SELF-DECLARATION AN INDEPENDENT EVIDENCE

The company's own leadership or success claim is used like external verification.

CH04-F04 — WORSHIP OF INDEPENDENT SOURCES

Independent but distant from the subject, old, or unmethodical sources are considered superior to official and close records.

CH04-F05 — EPISTEMIC ELEVATION

The phrase "The company says" is turned into the judgement "it has been proven."

CH04-F06 — ERASING SOURCE STATUS

First-party data, sponsored content, or press releases are presented as independent news.

CH04-F07 — FAITHFULLY REPRODUCING OFFICIAL ERRORS

AI repeats the official false claim correctly and is generally considered successful.

CH04-F08 — PUNISHING UP-TO-DATE EXTERNAL CORRECTIONS

AI is considered wrong because it leaves the old official site and uses strong current evidence.

CH04-F09 — CONSIDERING OFFICIAL SILENCE AS PROHIBITED

Any external information not found on the official site is automatically considered incorrect.

CH04-F10 — IGNORING MATERIAL DEFICIENCY

Even if the official sentence is correct, the necessary boundary is not considered.

CH04-F11 — CONSIDERING REFERENCE ERROR AS MODEL DEFECT

AI alone is blamed for contradictory price or identity records.

CH04-F12 — HIDING REFERENCE CONFLICT

Adjudicators choose the source they want from among contradictory records after the result.

CH04-F13 — INDEPENDENT SOURCE HIERARCHY REGARDLESS OF CLAIM TYPE

In any case, it is assumed that the official or independent source is superior.

CH04-F14 — DELETING COUNTEREVIDENCE

Records that weaken a positive claim are removed from the verification file.

CH04-F15 — CONSIDERING OLD OFFICIAL RECORDS AS CURRENT

Archives, old PDFs, or press releases are made the current canonical source.

CH04-F16 — COUNTING UNVERIFIED CLAIMS AS PASS

Since the falsity is not proven, the claim is considered verified.

CH04-F17 — OVERESTIMATING CLAIM STRENGTH FROM EVIDENCE

Limited pilot or local data is generalised to universal results.

CH04-F18 — SINGLE-LAYER GEO SCORE

Accuracy verified with official fidelity is combined invisibly within a single result.

CH04-F19 — PRESERVING FAVOURABLE BIAS

No correction is made when the AI institution overstates.

CH04-F20 — COUNTING REALITY RECORD AS FOUNDING DECISION

Statement of NobleJackal or Kaan MURAZ is transformed into verified reality without external evidence.

38. AUDIT PROCEDURE

Step 1 — Extract Material AI Claims

Response:

- atomic claims,
- relationships,
- scope and time areas

are separated.

Step 2 — Determine the Subject of Each Claim

The claim is about which:

- company,
- brand,
- product,
- person,
- local entity

?

Step 3 — Find the Official Representation Record

Is the claim found on official surfaces? In which source? In which version? On which date? In which scope?

Step 4 — Determine the Official Status

Claim:

- canonical,
- historical,
- withdrawn,
- draft,
- conflicting,
- unknown

Which of their statuses is it?

Step 5 — Classify the Type of Claim

Claim:

- identity,
- price,
- scope,
- licence,
- performance,
- leadership,
- partnership,
- user result

Which of its types is it?

Step 6 — Determine the Appropriate Type of Evidence

Which source is truly authoritative for a claim? A single universal source ranking is not used.

Step 7 — Collect Verification and Counter-Evidence

Supporting records Conflicting records Timeliness Scope Method Independence are evaluated.

Step 8 — Assign Verification Status

VR-0 through VR-5, VR-C, VR-U, VR-T, or VR-S statuses are assigned as appropriate.

Step 9 — Review Source Status Protection

AI: Has it preserved the difference between "the company says" and "proven"? Has it elevated self-declaration to independent reality?

Step 10 — Examine Material Deficiencies

In the response or official representation, necessary for user decision:

- boundary,
- time,
- geography,
- licence,
- data method

missing?

Step 11 — Separate Reference Defect

If the official or verified record is flawed in itself, a finding is created separate from the AI error.

Step 12 — Provide Dual-Layered Result

Response:

DUAL-PASS

OFFICIAL-PASS / VERIFIED-FAIL

OFFICIAL-FAIL / VERIFIED-PASS

DUAL-FAIL

REFERENCE-CONFLICT

APPROPRIATE-UNCERTAINTY

is saved with an appropriate status.

39. REQUIRED EVIDENCE

Canonical Entity Record Official Representation Record Verified Entity Reality Record Official web pages Official PDFs and documents Structured data Canonical JSON or API records Legal and regulatory records Licence and certificate verifications Product, price, and checkout records Contract and operation records Customer and partner verifications Performance data sets Methodology and denominators Independent reviews Sponsor and conflict of interest Counter-evidence Negative results Errata and retraction records Official source precedence rules Verification statuses Adjudicator decisions Appeal records Measurement date Reference versions Responsible person or institution

40. AUDIT CHECKLIST

Is there an Official Representation Record? Is there a Verified Entity Reality Record? Were both versioned for the same measurement date? Is the subject of each AI claim correct? Was the claim officially published? Is the official source current and canonical? Do official records conflict with one another? Was the appropriate evidence source used for the type of claim? Was a first-party source rejected automatically? Was a first-party statement automatically treated as independent evidence? Is the independent source genuinely relevant? Were the claim's scope and time preserved? Does the file contain counter-evidence? Was a self-declaration elevated to verified reality? Was the company's view conveyed with its source status intact? Is a material boundary or exclusion missing? Was an error favourable to the entity recorded as such?

Has the reference defect been separated from the AI error? Have the UNKNOWN and UNRESOLVED fields been considered positive? Does high official fidelity hide low verified accuracy? Can the AI's deviation from the official record be explained with strong evidence? Has the reference record changed during measurement? Are the old and new observations in separate versions? Does the dual-layer result appear in the public record? Is the accountable owner of the result known?

41. OBJECTIONS AND RESPONSES

Objection 1 — “The company knows the most accurate information about itself.”

This may be true in many respects. The company:

- its own product,
- its price,
- its support policy,
- its operation

knows it well. However, the company is also the commercial owner of the claim. Sentences that require external comparison, such as “the best,” “leader,” “most reliable,” and “proven,” cannot be verified by self-declaration alone.

Objection 2 — “If you do not count your own site as evidence, how will we know the price?”

The official price page is the primary source for the price the institution offers to the public. However:

- checkout,
- tax,
- additional fee,
- contract

if it conflicts, the actual offer is evaluated separately. The statement “own site is not evidence” cannot be applied equally to all types of claims. The correct ruling is:

> *A company's own site can be a primary source for statements under its control; it is not automatic independent verification for external performance and superiority claims.*

Objection 3 — “If AI reflects the official site, hasn't it completed its task?”

If the only goal is: “What does the site say?” then yes. However, if the user asks: “What is the company actually doing?” merely repeating the site may not be sufficient. The type of task affects the measurement layer.

Objection 4 — “External sources can also be wrong.”

That is correct. Therefore, independence is not automatic accuracy. The source is evaluated in terms of:

- authority,
- proximity,
- currency,
- method,
- scope

in terms of maintenance.

Objection 5 — “Who will determine the verified record of reality?”

The company alone cannot determine it. NOMOS alone cannot determine it either. The process:

- evidence holder,
- adjudicator,
- appropriate expert,
- appeal,
- version record,
- accountable person

is required. More independent review is needed for high-risk and conflicted claims.

Objection 6 — “If the reality record can always change, wouldn’t the score be meaningless?”

No. The score:

- date,
- is meaningful with
- scope

the version. Being subject to change does not mean it cannot be measured. It means it cannot be presented prematurely.

Objection 7 — “Why would AI use negative information that is not on the official site?”

Information:

- can be forced into the categories of

  true,
- current,
- relevant,
- financial for user decision

can be used. However, the source and context must be preserved. The silence of the official site does not eliminate the negative reality.

Objection 8 — 'If AI reports the company's claim as 'the company says,' it does not provide value to the user.'

On the contrary, it correctly displays the claim status. User:

- the institution's opinion,
- the proven finding,
- open ambiguity

can distinguish. Looking certain is not the same as looking correct.

Objection 9 — "Why should we correct minor mistakes that are in favour of the company?"

Because of the wrong expectation:

- wrong customer,
- delivery failure,
- return,
- complaint,
- loss of trust

It can produce. The advantage is apparent short-term visibility; long-term is the duty of representation.

Objection 10 — "Don't two separate scores complicate public communication?"

Concealing all this complexity creates the greater risk. A concise result card may instead state: Official Representation Fidelity — high / Verified Entity Reality Accuracy — medium / Principal gaps — unsupported leadership and country-coverage claims. That is clearer than one misleading number.

1. the Official Representation Record; and

2. the Evidence-Constrained Verified Entity Reality Record.

Official fidelity and verified accuracy MUST NOT be treated as the same result. An AI response MAY faithfully reproduce an official claim while remaining factually unsupported, contradicted, outdated, overbroad, or materially misleading. An AI response MAY differ from an official record without failing when the official record is stale, incomplete, conflicted, or contradicted by stronger and more appropriate evidence. Self-claims, first-party data, independent evidence, sponsored content, official registries, transactional records, and third-party statements MUST retain their source and evidence status. Reference conflicts MUST be recorded separately from AI-response failures. Unknown or unresolved evidence MUST NOT be converted into verified truth.

Every material generative-system claim about the audited entity MUST be evaluated against two separately versioned reference layers: the Official Representation Record and the Evidence-Constrained Verified Entity Reality Record. Official fidelity and verified accuracy MUST NOT be treated as the same result. An AI product may faithfully reproduce an official claim while creating a materially unsupported, contradicted, outdated or overbroad representation. When the official record is stale, incomplete or conflicted, a different response grounded in stronger evidence is not automatically an error. Source type and evidential status MUST be preserved, and reference conflicts MUST be recorded separately from AI failures. High Official Representation Fidelity MUST NOT compensate for low Verified Entity Reality Accuracy.

COMMON JUDGEMENT OF CHAPTER 44

A company has the right to speak about itself. It can explain who it is. It can describe what it offers. It can publish its prices. It can set its boundaries. It can convey its vision. However, not every sentence an entity says about itself has the same epistemic status. "The price of our service package is 2,000 euros." and "We are the best company in the world." are not the same type of claim. The first sentence may be a commercial statement controlled by the company. The second sentence requires external comparison. If an AI product conveys both with equal certainty, it undermines the source status. An AI product can perfectly replicate a company website. But the site:

- old,
- conflicting,
- unproven,
- exceeding the scope,
- conceals material boundaries

then fidelity to that record is not good GEO. Conversely, an AI product may depart from the official website and still be correct when its response rests on stronger, current and appropriate evidence. The

departure may be a correction rather than an error. NOMOS's fourth law of measurement is therefore:

> *Official representation does not replace verified reality.*

The fifth law is as follows:

> *Verified reality cannot be established by ignoring the official source.*

Its sixth law is:

> *Representation depends not only on what a source says, but also on the authority under which it speaks and the evidential status of the statement.*

Its seventh law is as follows:

> *Faithfully reproducing an official error is not accurate representation.*

Its eighth law is as follows:

> *The record of reality is not an indisputable absolute judgement; it is a versioned and evidence-based decision file.*

Order of NOMOS Section 4

> *Do not just show me what is written on your site. / Show which sentence is supported by which evidence.*
>
> *Do not present your own words as if they are independently verified.*
>
> *Do not devalue first-party data just because it belongs to you. / But do not make it an indisputable truth just because it belongs to you.*
>
> *Do not turn company opinion into fact, press release into news, or sponsored award into independent authority.*
>
> *Do not count me as successful when I correctly repeat the official error.*
>
> *If your official record is outdated, do not penalise an answer supported by current evidence merely because it differs from that record.*
>
> *Do not remain silent about the mistake in your favour.*
>
> *Do not remove counter-evidence from the file.*
>
> *When your references conflict, do not invent a single truth. / Record the conflict. / If the answer is unknown, say so.*
>
> *Separate your claim, its source, its evidence, its limit, and its time from each other.*
>
> *Before ordering me what to say, prove what you have the right to say.*

The Chapter's Closing Sentence

> *Good GEO is not the faithful repetition of whatever an entity says. It is the accurate transmission of what has been said, without distorting its source, evidence, scope or time.*

PART II

Global Population, Country and Language Fairness

CHAPTER 05

The Relevant User Population for Each AI Product

METHODOLOGICAL BASES / K04; K05; K06; K07

Chapter Boundary

The first four chapters established the basic objects of measurement: a single response is not a GEO score; representation is a conditional distribution, not a single number; the audited entity must be resolved before results are observed; and an AI response must be evaluated against both official representation and evidence-based verified reality. We must now define the denominator of the measurement:

> *About which human population will this distribution make a statement?*

Everyone living in a country? Only adults? Those with internet access? Those who can officially use a specific AI product? Free account holders? Paid plan users? People who can use the product in the oversight language? Truly active users? Potential users who might use it one day? People in markets where the brand can provide services? If a country's population is very large but the relevant AI product is not officially accessible in that country, how will that country be included in the main product panel? If the population share of a small country is far below one person in a sample of 1,000, will that country be completely invisible?

If one person knows three languages, should that person be counted as three people? To which population does a user belong if they can access a product's mobile surface but not its web surface? Until these questions are answered, '1,000 random users for each AI product' does not define a scientific target population; it states only a headcount. This chapter does not discard country population. It assigns it its proper role:

> *The country population is the initial base for sampling; it is not the final eligible user population.*

In this section, the audit system directly applies the previously determined requirement for country, language, date, product, and measurement record to the population definition. Because requesting evidence, boundaries, context, and time also requires defining with the same clarity which people are being referred to. This section has not yet:

- It does not finalise how the 1,000-person sample will be distributed across countries,
- which algorithm will be used to round fractions to whole numbers,
- from which panel participants will be selected,
- how sampling weights will be calculated,
- or which variance method will be used to produce confidence intervals

These will be arranged in the following sections. The task of Section 5 is more basic:

> *For each AI product and each measurement question, to define which human population the results can be generalised to.*

NOMOS Challenge

You open a world map. You add up the populations of the countries. Then you want to distribute 1,000 people according to the population. Large countries will get many participants. Small countries will get one or close to zero shares. The core of this idea is correct:

> *A person in the world population should not be hundreds of times more influential than a person in a large country just because they live in a small country.*

But now apply the same method to two different AI products. The first product:

- can be used officially in some countries,
- is limited in some countries.

It is not offered at all in some countries. The country coverage of the second product is different. What happens if you divide the world's population in the same way between the two products? In countries where the first product is not available, you may ask people for measurements. Participants:

- may be encouraged to misrepresent their locations,
- use prohibited or unsupported access methods,
- open an account in another country,
- bypass security and usage conditions

This does not measure the real user experience. It generates an artificial access experience. Now consider another problem. In a certain country, the population is very large. However:

- a significant portion of the population does not meet the minimum age requirement of the product.
- not everyone has reliable internet access,
- the product can only be used on certain devices,
- the audit language cannot be used by the entire population,

a specific plan is only available to users with a payment method. If you make the total population the denominator directly, you include people who actually cannot be monitored and cannot use the product in the target population. Now imagine a small country. The theoretical share in a sample of 1,000 under population weighting: 0.03 people.

One-third of a person cannot be selected. If you give one person to the country, you can overrepresent the country in the population-weighted main panel. If you give no one, then in that country:

- access,
- language,
- product behaviour,
- local representation

can remain completely invisible. These two goals are different:

- To measure the average user experience in the world population
- To prevent small countries from being completely unmonitored

The same single panel cannot perfectly fulfil both goals. Therefore, NOMOS establishes three separate principles:

> *Population weight determines people's share in the global average.*
>
> *Observer coverage prevents small countries and languages from being completely invisible.*
>
> *Language and accessibility justice further measures subgroups that may be suppressed by the population average.*

The first premise of this section is:

> *The total population of a country is not the AI-eligible user population in that country.*

Its second provision states:

> *The target population of an AI product cannot be automatically transferred to another AI product.*

Its third provision states:

> *A person's weight in the main score is not the same as their country's right to appear in the audit.*

Its fourth provision states:

> *The lack of product access is not a misrepresentation; however, it is a global access limitation and should also be reported.*

1. PURPOSE OF THE CHAPTER

The purpose of this section is to define the target human population of the GEO-1000 measurement in terms of product, question, country, language, interface, and time. The section standardises the following distinctions:

- Total country population and suitable AI user population
- Potential access and actual usage
- Product accessibility and representational accuracy
- Target population and sampling frame
- Valid analytical observation with the selected participant
- Account with person
- Session with person
- Core model with product
- Free and paid plan populations
- Web and mobile user populations
- Usage exceeding access restriction with official access
- Country Observer coverage with country population weight
- Language Fairness Panel with population-weighted language share
- Active user exposure with potential user population
- Market or jurisdiction domain query with global identity query
- Corporate, education, or developer product with main consumer product
- Accessibility of the data collection tool for users with accessibility needs
- Users outside the target population and NOT TESTED users
- Countries where the product is unusable and countries where incorrect responses are received in the product

At the end of this section, each GEO-1000 audit should be able to explicitly answer the following question:

> *What user universe, selected from which people, does this result make a judgement about?*

2. CENTRAL NORMATIVE PROVISION

> *Before results are observed, each GEO-1000 audit must define and version a target user population specific to the audited AI product, measurement question, country, language, interface, plan, and time conditions.* [K04–K07]

Target population:

- "world population"
- "internet users",
- "AI users",
- "customers"

The population cannot be left under an ambiguous label such as ‘relevant users’. As applicable, the following questions must be answered: Which AI product? Which user surface? Which countries? Access as of what date? What age condition? Which language or languages? Which device and interface? Which plan or account type? What are the official terms of use? Must users pay to access the product? Does the measurement question apply to all users or only to a particular market or jurisdiction? Are active users or potential users being measured? Which user groups are out of scope? Which groups belong to the target population but are absent from the sampling frame? A change in any of these fields may change the target population. When it does, old and new scores cannot be combined in the same series without explanation.

3. THE POPULATION IS NOT A SINGLE NUMBER

The official or estimated total population of a country is only the starting layer for GEO-1000. The audit population can consist of the following narrowing layers.

3.1. Resident Population

The total human population living in the specified country or jurisdiction. This number:

- does not imply any suitability in terms of age,
- internet,
- product access,
- language,
- account,
- device

considerations.

3.2. Age-Eligible population

The population that meets the age requirement of the relevant AI product and research protocol. For a product:

- minimum usage age,
- parental consent requirement,
- country-specific age rule

may exist. The research protocol may be more restrictive than the product condition. For example, the founding GEO-1000 panel may be composed of only adult participants to reduce ethical and legal complexity. In this case, the score is not: "The entire population":

but rather "Defined adult and product-eligible population"

result.

3.3. Connected Population

The population that can use the relevant product at a level where:

- internet,
- device,
- connection quality,
- technical access

is available. Internet access may not be a binary characteristic. A person:

- can use messaging,
- however, they may not be able to run the heavy web application,
- may use mobile internet,
- however, they may not be able to access the desktop experience,

may not be able to upload responses or screenshots due to low connectivity. Connection limitations should not automatically exclude the user from the target population. It should be associated with the actual usage conditions of the measured product.

3.4. Accessible Population from the Product Perspective

For the relevant AI product on the specified date:

- officially launched,
- could be used in accordance with the law and product conditions,
- the necessary interface could be accessed

It is population. The absence of a product in a country does not mean that the people in that country are misrepresented. In this case: the representation response is none, there is no access coverage. The two outcomes should be kept separate.

3.5. Population Suitable in Terms of Account and Plan

Some products:

- may require an account,
- may require phone verification,
- may require a payment method,
- may be offered only with a corporate contract,

may unlock certain features only in a paid plan. These conditions affect which portion of the target population can actually access the relevant user surface. The free plan score cannot automatically be applied to paid plan users.

3.6. Population Suitable in Terms of Language

The population that can use the relevant prompt and AI response at a sufficient level in the audit language. Language suitability alone:

- cannot be determined only by the official language of the country,

- the person's mother tongue

It should be distinguished:

- Mother or home language
- Language of daily use
- AI interface language
- Prompt language
- Response language
- Language for professional use
- Reading proficiency

A person can be suitable in multiple languages. This person cannot be counted as more than one full person in the global population calculation.

3.7. Interface-Eligible population

Measurement:

- web,
- mobile application,
- desktop application,
- corporate workspace,
- embedded assistant

can be used. The accessible user universe may differ for each interface. A user with web access may not have the mobile product. The corporate user surface may not be publicly available.

3.8. Research-Eligible population

Even if a person can use the product, for inclusion in research, additionally:

- voluntary consent,
- privacy,
- data processing,
- collecting screenshots,
- country-specific research requirements,
- does not imply any suitability in terms of age,
- security

conditions may need to be met. People who cannot be included in the research but use the actual product may be part of the target population. In this case, the sampling frame underrepresents the target population. This underrepresentation should not be hidden.

3.9. Analytically Eligible population

The population for which the collected data can be used to produce results. The participant:

- altered the prompt,
- used the wrong product,
- chose the answer,
- submitted incomplete evidence

It is possible. Even if an individual belongs to the target population, their observation may not enter the analytical dataset. Valid observations and the target population are not the same thing.

4. DEFINITION OF THE SUITABLE USER POPULATION

For AI product a at time t, the eligible user universe may be defined as follows:

```
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}
```

Here:

- Pt: the general human population at the specified date
- u: individual human
- a: AI product and user surface
- t: measurement time

In terms of language and prompt, the target universe can later be narrowed down as follows:

```
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}
```

Here:

- c: country or jurisdiction
- l: language and locale
- s: user surface
- p: prompt or question family

The purpose of this notation is not to record all individuals one by one. It is to make the logical boundary of the target population visible.

5. BLINDLY MULTIPLYING MARGINAL RATES IS PROHIBITED

The eligible population may sometimes be approximately calculated with a formula like the following:

```
N × p_age × p_internet × p_product × p_language
```

This calculation can only be justified under the independence of conditions or the assumption of an appropriate joint distribution. In real life, features may not be independent. For example:

- internet usage may vary by age,
- access to a paid plan may be related to income,
- certain language use may be related to the region within the country and education,

device access and connection quality may vary together. Multiplying marginal rates can predict a target population that is:

- larger than necessary,
- smaller than necessary

The order of preference should be:

- Appropriate joint microdata or cross-tabulation
- Reliable multidimensional population projection
- Calibrated model
- Marginal approach with explained rationale and sensitivity analysis

If simple multiplication is used:

- assumption,

- data source,
- uncertainty,
- alternative scenario

should be disclosed.

6. AUDIT QUESTION CHANGES THE TARGET POPULATION

Not every prompt has the same measurement universe for the entire world population.

6.1. Global Identity Questions

Example: “What kind of company is Apple.com?” This question can be widely asked among the suitable global user population for the product. Because the company identity is not limited to users of a specific market.

6.2. Local Service Questions

Example: “Does this company provide services in Turkey?” Main target population:

- Eligible users in Turkey,
- External users investigating Turkey's service

It is possible. It is not necessary to give equal weight to the entire world population.

6.3. Price and Purchase Questions

Price:

- country,
- tax,
- currency,
- plan,
- customer type

may vary. The target population should be defined according to the relevant commercial market and type of user.

6.4. Legal and Licensing Issues

Licensing and legal authority depends on the jurisdiction. The target population:

- can obtain the relevant service in that country,
- are subject to that legal regime.

The target population may be restricted accordingly.

6.5. Advisory Issues

The target population for the question "Is this company suitable for me?" is not all people. It can be defined according to:

- User needs
- Country
- Budget
- Customer type

- Scope of service
- Eligibility and exclusion conditions

The advisory population may be narrower than the identity population.

6.6. High-Risk Questions

In the fields of health, law, finance, or security:

- judgement,
- user profile,
- risk status,
- professional authority

more strongly limits the target population.

7. POPULATION FRAMEWORK MUST BE LINKED TO PROMPT FAMILY

In an audit, there may be multiple population frameworks instead of a single target population.

7.1. Core Identity Population

The global or broadly eligible user population measuring the entity's core identity.

7.2. Service-Relevant Population

It is the population for which the entity can actually provide service or can be considered for service by the user.

7.3. Market Transaction Population

The specific price, purchase, return, or commercial condition is the population to which it can be applied.

7.4. Jurisdictional Population

It is the user population to which a specific licence, legal, or regulatory claim can be applied.

7.5. Recommendation-Eligible Population

It is the user population that may be appropriate within the true scope of the entity.

7.6. High-Risk Affected Population

It is a population for which misrepresentation can cause serious harm. These frameworks should not be confused with each other. A global identity score does not mean "recommendable for the whole world."

8. THE TYPE OF AI PRODUCT DETERMINES THE POPULATION

Not all AI products are publicly available chat products.

8.1. General Consumer AI Product

Target population:

- There may be people who can access the product,
- meet the terms of use,
- and use it in the relevant language.

The GEO-1000 real user panel can be directly applied to this type of product.

8.2. Enterprise AI Product

The product is only for:

- employer contract,
- corporate licence,
- can be used with an in-house user account

The target population is not the general adult population:

> *It is the seat or user universe with corporate access to the product.*

Corporate results cannot be compared with the general consumer product in population weighting.

8.3. Educational Product

Target population:

- student,
- teacher,
- institution,
- age group,
- educational licence

may be limited in terms of. If children or minors are to be included, a separate ethical protocol is required.

8.4. Developer or API Product

An API output may not be exposed directly through an end-user interface. The target unit is:

- developer,
- application,
- API call,
- controlled test session

may be. The GEO-1000 human population panel cannot be applied directly and unchanged to the API. Separately for the API:

Controlled System Audit

may be required.

8.5. Embedded AI Assistant

AI:

- operating system,
- search interface,
- shopping platform,

- business software,
- device

It may be embedded. The target population is the actual user population of the embedded product. It is not the total user population of the base model provider.

8.6. Expert or Sectoral AI

The product is only for:

- lawyers,
- doctors,
- finance professionals,
- engineers

may be designed for. The target population should be limited by profession and access conditions.

9. PRODUCT ACCESS STATUSES

Access status should be recorded for each country, product, interface, and date.

PA-0 — UNKNOWN

Access status could not be determined reliably.

PA-1 — UNAVAILABLE

The product is not offered in the relevant country or user group.

PA-2 — RESTRICTED

The product is limited to certain:

- account,
- invitation,
- institution,
- does not imply any suitability in terms of age,
- plan,
- device

terms.

PA-3 — ACCESSIBLE BUT NOT OFFICIALLY SUPPORTED

Actual access to the product can be observed. However, official support or usage status is unclear or limited. These users should not be automatically added to the main official population panel. A separate experimental panel can be created.

PA-4 — OFFICIALLY AVAILABLE

The product is officially accessible and the main usage conditions are defined.

PA-5 — OFFICIALLY AVAILABLE WITH FULL AUDIT SURFACE

The product is not only inaccessible; it is controlled:

- language,
- plan,

- interface,
- under session conditions

suitable for generating the observation required by the protocol. The availability of the product in a country does not mean that all plans and languages are in compliance with the control.

10. BYPASSING RESTRICTION IS NOT A REAL POPULATION MEASUREMENT

Participants cannot be asked to perform the following behaviours:

- Declare the wrong country
- Create a fake account
- Change age information
- Bypass official access restriction
- Use someone else's account
- Bypass the product's security or verification
- Violating the terms of use

These behaviours can also be examined as a technical experiment. However:

The experience of natural and legitimate user population

cannot be reported as such. If there is no access in a country, the correct status should be: UNAVAILABLE / or / OUTSIDE_NATIVE_REACH. A single observation exceeding the restriction does not generate a country population score.

11. DISTINCTION BETWEEN POTENTIAL USER AND ACTIVE USER

The people who can use the product are not the same community as the people who actually use the product.

11.1. Potential Access Population

These are the people who can technically and legally use the product. This population answers the question: 'What would a suitable person see if they used the product?' The population is valuable for justice and product comparison.

11.2. Active User Population

These are the people who actually used the product during a specific period. This population approaches the question: 'What do people see within the actual usage volume?' It requires reliable active usage data.

11.3. Query-Exposed User Population

These are people who use the product and have a potential to ask questions about the monitored brand or topic. This population often cannot be known directly. Search or query behaviour data may be required.

11.4. Population Actually Seeing the Response

They are users who are actually exposed to a specific representation output. This may not be fully measurable without real product provider data outside the user panel.

12. EXPOSURE WEIGHT IS NOT THE SAME AS ELIGIBILITY WEIGHT

Two separate global results can be generated.

12.1. Eligibility Population Weighted Result

Each person is weighted according to their share within the defined eligible user population. This result is:

- globally fair,
- potentially accessible,
- representative of the population

in terms of

12.2. Real Exposure Weighted Result

AI products or user groups are weighted according to actual usage volumes. This result:

- is valuable in terms of
- current commercial impact,
- actual user exposure,

operational risk.

However, reliable usage data may not be available. Estimates may rely on provider statements, market data, or incomplete sampling. Therefore, the type of weighting must be explicitly stated.

12.3. Replacing One with Another is Prohibited

Eligible population score: does not mean "Most of the actual users saw this response." Exposure score: does not mean "Fair representation of the world population is ensured."

13. ACCESS SCOPE AND REPRESENTATION ACCURACY SHOULD BE SEPARATED

An AI product may only be available in limited countries. In these countries, it can generate very accurate representations. Another product may have wider access but can produce lower accuracy. If we only look at representation accuracy, the product with limited access may appear superior. If we only look at access coverage, incorrect representation may be overlooked. Therefore, two separate results are needed.

13.1. Access Coverage

Indicates how much of the specified global reference population can appropriately use the relevant AI product. Candidate representation:

```
AC_a = N(U_{a,t}) / N(G_t)
```

Here:

- Ua,t: eligible user population of the product
- Gt: defined global reference population

13.2. Conditional Representation Accuracy

It is the probability of seeing an accurate representation among users who can access the product. Candidate notation:

```
RA_a = Pr(Y=1 | u ∈ U_{a,t})
```

13.3. Effective Accurate Representation Access

CANDIDATE CONCEPT

An operational metric can be developed that combines access coverage with conditional accuracy:

```
ECR_a = AC_a × RA_a
```

This metric roughly approaches the question: "What proportion of the defined global reference population can access this product and meet the condition of seeing an accurate representation?" This is not yet the final NOMOS formula. Because:

- access is different from actual usage,
- weight and dependency structures are complex,
- product usage is voluntary,

The correct representation depends on the prompt. However, it maintains the basic distinction:

> *Not being able to access is not the same as getting a wrong answer.*

14. NATIVE REACH AND COMMON SUPPORT

When comparing AI products, two separate population universes are needed.

14.1. Native Reach Population

Each AI product has its eligible population within its official and actual access domain. This result answers the following question:

> *To what extent does the product accurately represent the population it can actually reach?*

Target populations may differ between products.

14.2. Common Support Population

In a common form across all compared AI products:

- officially accessible,
- can be used under the same language or interface condition

is the population. This result answers the following question:

> *Which product produces more accurate representation in the same user universe?*

14.3. Two Results Should Be Published Together

Native Reach:

- shows actual product access,

- natural geographic distribution

Common Support: provides a fairer product comparison. A product can be high in Native Reach score, but lower in Common Support score. Two results answer different questions.

15. MARKET-RELATED POPULATION

The global identity score of an entity may differ from its commercial target market score. Example: If a local law firm provides services only in Turkey: the global identity question can be measured in the eligible user population worldwide, but recommendations and service suitability should only be measured for users in Turkey or within the actual scope of services. The institution may select the result in its favour:

- only the high-scoring market,
- only his/her own language,
- only its existing customers

the target population cannot be announced. Target market:

- Before the audit begins,
- to the service records,
- within the scope of the contract and operations

It should be locked according to.

16. PERSON, ACCOUNT AND SESSION SEPARATION

Population weight belongs to people. Not to AI accounts.

16.1. One Person, Multiple Accounts

The same person:

- free,
- paid,
- corporate

can have multiple accounts. These accounts can produce separate observations in product surface comparisons. They cannot be counted as separate individuals in the global human population.

16.2. One Account, Multiple People

A family, team, or corporate account may be used by multiple people. It is not equal to the number of account users.

16.3. One Person, Multiple Sessions

A person can produce separate observations in different sessions. These observations can measure time or session stability. They do not increase independent population size.

16.4. One Person, Multiple AI Products

It is valuable for paired product comparison. However, observations are dependent due to shared characteristics at the user level.

16.5. Main Weight Unit

In the population-based GEO score, the main weight unit should be:

human

Calculations, session, and product observations should be associated under the human unit.

17. LANGUAGE POPULATION

Using only the country's official language is not sufficient for language fairness.

17.1. Prompt Language

The language in which the participant asks the audit question.

17.2. Interface Language

The language used in the menu and user interface of the AI product. It may not be the same as the prompt language.

17.3. Response Language

The language in which the AI actually responds. The system may respond in a different language than the prompt language. This is also a measurement result.

17.4. User Proficiency Language

The language in which the participant can meaningfully use the prompt and response. It does not have to be the native language.

17.5. Locale

The regional context along with the language. The same language can be different:

- country,
- price,
- law,
- terminology

may carry the condition.

18. DOUBLE COUNTING OF MULTILINGUAL USERS

A person:

- Turkish,
- may be able to use English.
- German

This person may take three separate language tests. However, counting them as three people in the global population weight distorts the population estimate. Three methods can be used.

18.1. Single Mother Tongue Assignment

The person is assigned to a single language cell in the main analysis. Other languages are used as additional or paired tests.

18.2. Split Person Weight

The person's population weight can be divided among the appropriate languages in a predefined manner. The total weight does not exceed one person.

18.3. Separate Multilingual Matched Panel

Multilingual users are used for cross-language comparisons of the same person. This panel is analysed separately from the main population panel. Regardless of the method used:

> *The total population weight of the same person cannot be replicated without explanation.*

19. HOW SHOULD LANGUAGE PROFICIENCY BE VERIFIED?

A participant's statement: "I know this language." can be an initial record. However, in high security auditing, one of the following methods can be used:

- Native language or daily usage declaration
- Short reading and comprehension test
- Language record of the panel provider
- Local adjudicator verification
- Education or professional use record
- Pilot mission success

Extremely difficult language test panels can be accessible to trained and high-income users. The purpose is not to conduct an academic language exam. It is to verify the competence to understand the prompt and response as a real user.

20. WHAT HAPPENS IF A LANGUAGE IS NOT SUPPORTED?

An AI product may:

- not officially support a particular language,
- partially support it,
- not claim to support it but can generate responses

It may be. Language statuses should be recorded separately:

OFFICIALLY_SUPPORTED

LIMITED_SUPPORT

OBSERVED_BUT_UNSUPPORTED

NOT_SUPPORTED

UNKNOWN

Response received in a language that is not officially supported:

- discovery,

- experimental access,
- language resilience

may be measured. It should not be automatically added to the main supported user population score. Language access limit should also be reported.

21. COUNTRY POPULATION WEIGHTING AND Country Observer COVERAGE

Population-weighted main panel and country visibility panel serve separate purposes.

21.1. Population Panel

Participants are distributed according to their share in the product-appropriate target population. Larger eligible populations receive higher weight. Produces the main global population score.

21.2. Country Observer Panel

The Country Observer Panel seeks at least one observation from smaller countries that receive little or no representation in the main population sample. A single Country Observer observation:

- indicates access is available,
- that a specific singular result has occurred,
- shows the need for new research

may indicate. Does not produce a country score on its own.

21.3. Sufficiency for Country Score

For country-based rates to be published:

- minimum sample,
- valid observation rate,
- uncertainty,
- panel representativeness

conditions must be met. With one or several participants: 'The country's NOMOS score' cannot be announced.

22. Language Fairness Panel

Population weight may make low-population or low-resource languages invisible in the main score. Therefore, a separate Language Fairness Panel can be established. Purpose:

- to test underrepresented languages,
- to detect severe language-based degradation,
- to identify prompt translation issues,
- to make low-resource language behaviour visible

must be. The Language Fairness Panel can provide additional samples to the main population panel. However, over-sampled language observations should be returned to actual population weights in the global score. Otherwise, the fairness panel alters the global population estimate.

23. ACCESSIBILITY PANEL

Users with visual, hearing, motor, cognitive, or other accessibility needs are a real part of the target population. If the NOMOS Capture or research task is not suitable for these individuals' participation: users cannot be removed from the target population. Correct outcome:

> *The sampling frame is incomplete in terms of accessibility.*

A separate accessibility panel can evaluate the following:

- Screen reader compatibility
- Using the keyboard
- Voice input
- Low vision settings
- Cognitive load
- Mobile accessibility
- Usability of the Capture tool

The accessibility panel should see users as part of the population coverage, not as a "special exception."

24. TARGET POPULATION AND SAMPLING FRAME

These two concepts should not be confused with each other.

24.1. Target Population

The entire group of people about which one wants to make a conclusion.

24.2. Sampling Frame

The population from which participants can actually be selected:

- panel,
- list,
- database,
- joint organisation,
- user pool

is the cluster. The sampling frame may not fully cover the target population.

24.3. Contactable Population

These are the people to whom the research invitation can actually be delivered.

24.4. Population Accepting to Participate

Those who voluntarily participate in the research.

24.5. Population Completing the Task

Those who complete the audit task.

24.6. Valid Analytical Population

They are the ones who produce observations suitable for the protocol and usable in the analysis. The numbers of these layers should be reported separately.

25. COVERAGE ERROR

The sampling frame may exclude some parts of the target population. This is called coverage error. Examples:

- Only professional online panels
- Only desktop users
- Only people with credit cards
- Only those using an English interface
- Only residents of large cities
- Only users highly interested in technology
- Exclusion of people with accessibility needs

Coverage error:

- should be explained,
- should be measured if possible,

should be reduced with weighting or an additional panel. Weighting cannot magically create a group that does not exist in the sampling frame.

26. OVER-COVERAGE

People who do not belong to the target population may be found in the sampling frame. Examples:

- Participant in a country where the product is not accessible
- Person who does not meet the required age condition
- Incorrect plan or user surface
- Participant who cannot use the audit language
- Person found again with multiple registrations
- User working on behalf of another country

Excess coverage records should not be included in the main score. However, the reason for exclusion and the rate should be reported.

27. NON-RESPONSE AND SELF-SELECTION

People who agree to participate in the research may systematically differ from the target population. Volunteers:

- May use AI products more frequently,
- May be more interested in technology,
- May have a higher level of education,
- May already be familiar with the brand,

May behave differently due to reward or compensation. Therefore: the expression "1,000 random people in the world" should only be used in truly probabilistic and documented selection designs. If a volunteer or online panel is used, the correct expression is:

Layered and weighted voluntary user panel

may be possible. The panel type cannot be stored.

28. SUITABLE POPULATION CHANGES OVER TIME

The user universe of an AI product is not fixed. The following can change:

- Supported countries
- Minimum age
- Free access
- Paid plan structure
- Mobile or web availability
- Language support
- Account terms
- Enterprise product coverage
- Legal and regulatory status

For this reason, each population record:

- observation date,
- period of validity,
- source version,
- last verification date

must carry.

28.1. Change During Measurement Wave

If product access materially changes during the wave:

- the wave can be stopped,
- two separate periods can be created before and after the change,

affected countries can be analysed separately. Users before and after the change cannot be combined as a single population.

29. POPULATION SOURCE FILE

Each audit should store the target population data it uses in a versioned file. As applicable, the following fields should be included:

- Country
- Reference population
- Population history
- Age eligibility
- Internet and device access
- Product access status
- Interface access

- Plan and account requirement
- Language groups
- Research eligibility
- Common distribution method
- Uncertainty
- Sources
- Last verification date
- Responsible person
- Population version

Country quotas cannot be reproduced after results without this file.

30. SYNTHETIC PRODUCT AND COUNTRY DISPLAY

SYNTHETIC METHODOLOGY DEMONSTRATION / The following countries, AI product, population numbers, and access rates are entirely fictional. They do not represent any real country or product. Let's consider a synthetic general consumer product called Orion AI. There are four synthetic countries.

Country	Total population	Estimated suitable population after age, internet, language, and interface conditions	Product access	Population of Native Reach
Country A	1,400,000,000	780,000,000	Unavailable	0
Country B	1,450,000,000	520,000,000	Official access	520,000,000
Country C	340,000,000	250,000,000	Official access	250,000,000
Country D	35,000	25,000	Official access	25,000
Total	3,190,035,000	1,550,025,000		770,025,000

30.1. If the total population was used

Country A would have a large share in the main sample due to its very large population. However, the product cannot be used in that country. Participants could not produce a natural user experience.

30.2. When Native Reach Population Is Used

The theoretical population shares of the main product panel are approximately:

- Country B: 67.53%
- Country C: 32.47%
- Country D: 0.003%
- Country A: outside of Native Reach

The theoretical share of Country D in a main sample of 1,000 people is far below one person. Therefore, the population-weighted main score should not give high weight to Country D. However, if the country is not intended to be completely invisible:

Country Observer Panel

At least one exploratory observation can be obtained. This observation:

- is not the score of Country D,
- do not change the main population weight,

shows that the product behaviour can be observed in the country.

30.3. How is Country A Reported?

For Country A:

The OUTSIDE_NATIVE_REACH

status is used. The exclusion of Country A should not hide the product's access limits. The public results card also shows:

- Product access coverage
- Share not accessible in the global reference population
- Native Reach score
- Common Support comparison score

31. SYNTHETIC PROMPT POPULATION REPRESENTATION

SYNTHETIC EXAMPLE

Let's consider three prompts for the same Orion AI product.

Prompt A

"What kind of company is Asteron?" Target population: could be the broad suitable user population of Orion AI.

Prompt B

"Does Asteron provide corporate travel services in Turkey?" Target population:

- Eligible users in Turkey,
- Identified external users researching the Turkey service

may be.

Prompt C

“Is Asteron suitable for legal immigration consultancy for me?” If Asteron does not provide such a service: there may be no advice-eligible population, the correct answer should exclude all suitable users. Combining this prompt into the overall identity score might be incorrect. This example shows:

> *Even for the same entity and the same AI product, the target population may vary depending on the user question measured.*

32. FOUR PANEL ARCHITECTURE

CANDIDATE ARCHITECTURE — CANDIDATE ARCHITECTURE

Four complementary panels are recommended for the global and fair implementation of GEO-1000.

32.1. Population Panel

The main panel weighted according to the product-eligible user population. Produces the main population estimate.

32.2. Country Observer Panel

Prevents small countries that cannot provide sufficient samples in the population panel from being completely invisible. It is not sufficient on its own to produce a country score.

32.3. Language Fairness Panel

Also measures languages that are low-resource or suppressed in population averages. It reconnects to the global main score with real population weights.

32.4. Accessibility Panel

Measures accessibility needs and assistive technology users that the standard digital panel might exclude. Observations from these four panels:

- goal,
- weight,
- inference universe

It should be labelled separately in terms of maintenance. The same observation cannot be used at full weight on more than one panel without explanation.

33. ADULT INITIAL PANEL

CANDIDATE BASE RULE — CANDIDATE FOUNDATIONAL RULE

In the first general release of GEO-1000, the real user dashboards:

from adult participants who meet the product and country conditions

It is recommended to be created. This approach:

- consent,
- data protection,
- collecting screenshots,
- high-risk research,
- parental consent

It can reduce its complexity. However, the following conclusion cannot be drawn: 'The experience of child and young users has been measured with the adult panel.' In products that affect children or young people, separately:

- ethical approval,
- age-appropriate protocol,
- parent or guardian processes,
- harm protection

are required.

34. NO POPULATION FRAMEWORK CAN BE CALLED "ALL HUMANITY"

A product:

- if it does not exist in some countries,

- if it does not support some languages,
- if it only works on certain devices,
- if it has certain age or plan condition,

result: it cannot be presented as “the GEO experience of all the people in the world”. The correct expression:

Estimated representation in the target population suitable for the specified product

should be used. The word global:

- multi-country,
- multi-lingual,
- large population

can define a measurement. It does not mean unlimited and universal coverage.

35. FOUNDATIONAL ETHICAL LIMIT OF THE POPULATION

Defining a population for measurement does not mean considering some people worthless. A group may receive a low weight in the main score. However:

- high risk,
- language fairness,
- accessibility,
- country visibility

may also be examined separately. A group may be excluded from the Native Reach score because they cannot access a product. This lack of access can also be reported as a global coverage gap. The purpose of NOMOS:

- compressing everyone into a single artificial average,
- seeing only the largest populations,
- appearing to symbolically represent small communities with a single person

is not the goal. The objective is:

> *To represent each person in the correct denominator of the correct question, without mixing their weight and visibility.*

36. MANDATORY NORMATIVE PROVISIONS

CH05-N01

The target user population must be defined separately for each AI product.

CH05-N02

The total population of a country cannot be directly used as the AI-eligible user population.

CH05-N03

The target population must be linked to product, country, age, access, interface, plan, language, prompt, and time conditions.

CH05-N04

Target population and exclusion rules must be versioned before seeing the results.

CH05-N05

The target population of an AI product cannot be transferred to another AI product without explanation.

CH05-N06

Lack of product access cannot be scored as a misrepresentation result; it must be reported separately as access coverage.

CH05-N07

Participants cannot be asked to bypass product or country access restrictions.

CH05-N08

Access not officially supported cannot be automatically added to the main Native Reach population.

CH05-N09

Individual, account, and session units must be separated from each other.

CH05-N10

Multiple accounts or languages of a person cannot multiply the individual in population weight.

CH05-N11

Language eligibility cannot be assigned based solely on the country's official language.

CH05-N12

Requested language, interface language, response language, and user proficiency language must be recorded separately.

CH05-N13

The total population weight of a person using multiple languages cannot exceed one person without explanation.

CH05-N14

The population-weighted panel and the Country Observer panel should be kept separate.

CH05-N15

One or several Country Observer observations cannot be published as a country score.

CH05-N16

In the Language Fairness Panel, over-sampled observations should be reverted to their true population weight in the global score.

CH05-N17

Users with accessibility needs cannot be excluded from the target population because the data collection tool is not suitable.

CH05-N18

The target population, sampling frame, contacted population, participants, and valid analytical observations should be reported separately.

CH05-N19

A non-probability or voluntary panel must not be represented as a random sample of the global population.

CH05-N20

Weighting cannot make a group that is completely absent in the sampling frame observable.

CH05-N21

Potential suitable population and active user exposure cannot be used as the same metric.

CH05-N22

The results Native Reach and Common Support must be separated from each other.

CH05-N23

The same target population cannot automatically be used for identity, service, price, legal, and recommendation prompts.

CH05-N24

Population data, product access status, and language distribution must carry a timestamp and source version.

CH05-N25

Marginal utility ratios cannot be blindly multiplied without an explanation of the independence assumption.

CH05-N26

If the product type is not suitable for the general consumer panel, the GEO-1000 human population method cannot be applied without explanation.

CH05-N27

Children or underage users cannot be included in the main panel without a separate ethical and legal protocol.

CH05-N28

Product access coverage and representational accuracy must be published as separate results.

CH05-N29

The term global cannot make countries, languages, ages, interfaces, or plans that are out of scope invisible.

CH05-N30

The target population record must have a responsible human or institutional owner.

37. FORMS OF FAILURE

CH05-F01 — COUNT THE TOTAL POPULATION AS ELIGIBLE POPULATION

The country's population is directly included in the denominator without evaluating product, age, internet, language, and interface conditions.

CH05-F02 — APPLY THE SAME WORLD MAP TO ALL PRODUCTS

AI products with different access areas are measured against the same target population.

CH05-F03 — COUNT A USER WHO EXCEEDS THE LIMIT AS A NATURAL USER

Observations that exceed official access conditions are presented as a country population experience.

CH05-F04 — COUNTING PRODUCT ABSENCE AS WRONG ANSWER

Countries that are inaccessible are considered zero in representation accuracy.

CH05-F05 — COMPLETELY HIDING PRODUCT ABSENCE

Large populations that are inaccessible are removed from the Native Reach denominator; access coverage is not shown either.

CH05-F06 — COUNTING ONE PERSON AS MORE THAN ONE

The same person's account, language, or product observations multiply the population weight.

CH05-F07 — COUNTING ACCOUNT AS PERSON

Multiple account ownership or shared account usage is not considered.

CH05-F08 — COUNTING OFFICIAL LANGUAGE AS USER LANGUAGE

The official language of the country is considered the audit language for all participants.

CH05-F09 — COUNTING MULTILINGUAL INDIVIDUALS TWICE

The same individual receives full population weight in each language.

CH05-F10 — ASSIGNING ONE PERSON TO A SMALL COUNTRY AND EXPLAINING THE COUNTRY SCORE

Country Observer observation is transformed into a statistical country estimate.

CH05-F11 — COMPLETELY DELETING SMALL COUNTRIES

Because the population share is low, the country or language is not included in any additional coverage panel.

CH05-F12 — OVERWEIGHTING EXAMPLES OF LANGUAGE FAIRNESS IN THE MAIN SCORE

Over-sampled languages are not returned to their true population weight.

CH05-F13 — COUNTING THE ACTIVE USER AS POTENTIAL POPULATION

Actual usage data is presented as if it includes all people who could access the product.

CH05-F14 — COUNTING THE POTENTIAL POPULATION AS ACTUAL EXPOSURE

It is assumed that the response of everyone who could use the product is actually seen.

CH05-F15 — COUNTING THE VOLUNTARY PANEL AS A RANDOM WORLD SAMPLE

Self-selection and panel bias are concealed.

CH05-F16 — IGNORING COVERAGE ERROR WITH WEIGHTING

It is claimed that groups not present in the frame are represented solely by weighting.

CH05-F17 — REMOVING USERS EXCLUDED BY THE DATA COLLECTION TOOL FROM THE POPULATION

Accessibility or device issues are interpreted as if the user is not suitable.

CH05-F18 — APPLYING THE SAME POPULATION TO ALL PROMPTS

Global identity, local pricing, and legal advice questions are combined on the same denominator.

CH05-F19 — APPLYING GENERAL CONSUMER POPULATION TO THE API PRODUCT

The human user panel is transferred directly to the developer or application product.

CH05-F20 — SECRETLY COMBINING FREE AND PAID PLANS

A single product distribution is published without recording the plan difference.

CH05-F21 — UNIFYING WEB AND MOBILE POPULATIONS

Interface access and behaviour differences remain invisible.

CH05-F22 — IGNORING THE AGE LIMIT

Inappropriate age groups are added to the target population or panel without explanation.

CH05-F23 — MULTIPLYING MARGINAL RATES AS IF INDEPENDENT

The relationship between age, internet, language, and access is ignored.

CH05-F24 — CHANGING THE POPULATION AFTER THE RESULT

A low-scoring country or user group is removed from the target population.

CH05-F25 — USING THE OLD PRODUCT REACH MAP

Actual product reach on the measurement date is not verified.

CH05-F26 — CONFUSING COMMON SUPPORT AND NATIVE REACH RESULTS

Comparisons made on joint population are presented as if they reflect the product's actual global reach.

CH05-F27 — NARROWING THE TARGET MARKET ACCORDING TO COMMERCIAL INTEREST

The organisation only declares the countries where it succeeds as target markets.

CH05-F28 — COUNTING THE SCOPE OF ACCESS AS GEO ACCURACY

Being present in many countries is presented as proof of accurate representation.

CH05-F29 — SELLING LIMITED ACCESS AS HIGH GEO SUCCESS

High accuracy in a small number of countries is presented as global superiority.

CH05-F30 — USING THE TERM "THE WHOLE WORLD" WITHOUT COVERAGE RECORD

Exclusions of product, language, country, and plan are hidden.

38. AUDIT PROCEDURE

Step 1 — Lock the AI Product and User surface

Product Interface Plan Account type Displayed model Measurement date is recorded.

Step 2 — Classify the Measurement Question

Prompt:

- global identity,
- local service,
- price,
- law,
- recommendation,
- high risk

which of their families does it belong to?

Step 3 — Determine the Global Reference Population

The initial human universe to which the result will be associated is defined.

Step 4 — Determine Age and Research Eligibility

Product age condition Research age condition Consent and data processing terms are recorded.

Step 5 — Verify Country-Based Product Access

An access status between PA-0 and PA-5 is assigned for each country.

Step 6 — Determine Internet, Device, and Interface Compatibility

The population that can use the measured surface is estimated.

Step 7 — Establish the Language and Locale Universe

Prompt language, Interface language, Response language, and User language proficiency are separated.

Step 8 — Apply Prompt Relevance

It is evaluated whether all users belong to the target population of the relevant prompt family.

Step 9 — Select the Population Model Appropriate to the Product Type

Consumer, enterprise, education, API, or embedded product is separated.

Step 10 — Compare the Target Population with the Sampling Frame

It is recorded as under-coverage and over-coverage.

Step 11 — Identify the Main and Complementary Panels

The scopes of the Population Panel, Country Observer Panel, Language Fairness Panel, and Accessibility Panel are versioned.

Step 12 — Separate the Native Reach and Common Support Universes

Intra-product results and inter-product comparison populations are recorded separately.

Step 13 — Separate Access Scope and Accuracy Results

The accuracy between the population a product can reach and the users reached is reported separately.

Step 14 — Freeze the Population File

Data sources, assumptions, date, and population version are locked.

Step 15 — Track Material Changes

If product access or conditions change during measurement, a new version is created.

39. REQUIRED EVIDENCE

AI-product identity; user surface; plan and account conditions; measurement date; countries in which the product is accessible; country-level access status; age conditions; legal and ethical participation conditions; country-population records; age distributions; internet and device-access records; language and locale distributions; product-language support; interface accessibility; target prompt family; target-population definition; sampling frame; undercoverage; overcoverage; panel type; participant-selection method; probability or opt-in panel status; multilingual-user weights; Country Observer coverage; Language Fairness coverage; Accessibility Panel coverage; Native Reach population; Common

Support population; active-user data, where used; access-coverage calculation; assumptions and sensitivity analyses; population version; and last verification date.

Responsible person or institution

40. AUDIT CHECKLIST

Which AI product and user surface is being measured? For which date was the target population defined? Were the total country population and the eligible user population separated? Was an age requirement applied? Was internet and device access evaluated? Is the product officially accessible on a country basis? Were participants asked to overcome access restrictions? Were free and paid plans separated? Were web and mobile interfaces separated? Did the prompt family change the target population? How was language and locale suitability determined? Were multilingual users counted twice? Is the difference between the target population and the sampling frame clear? Is the panel probabilistic or voluntary? Was coverage error measured? Were users with accessibility needs silently excluded?

Have Country Observer observations been converted into a country score? Did Language Fairness examples have excessive weight in the main score? Were the results of Native Reach and Common Support separated? Are coverage and representation accuracy separate? Were active user and potential user results separated? Is the product type suitable for the human population panel? Were the marginal rates in population weights multiplied blindly? Was the population definition changed after the results were seen? Are the used population and access data versioned? Are untested groups visible? Is the term "global" consistent with the actual coverage? Is there a clear accountable owner of the population record?

41. OBJECTIONS AND RESPONSES

Objection 1 — "Isn't it fairer to distribute directly according to the world population?"

The world population is a strong starting base. However, including people who cannot use the product in the main product sample does not generate real user experience. The correct method is:

- population weight within the product-eligible population,
- global access coverage.

Both results must be reported.

Objection 2 — "Shouldn't most users come from large countries?"

Countries that meet the product, language, and access conditions should naturally have a high share. However, just because the total population is large, a fake user experience cannot be generated from a country where the product is not accessible.

Objection 3 — "Why would it be wrong to give one person to a small country like San Marino?"

A person is valuable for exploration and Country Observer observation. What is wrong is to present that person's result as a 'country score.' The visibility of the small country can be protected with the Country Observer Panel. The main population weight remains according to the actual population share.

Objection 4 — “If we give at least one person to each country, wouldn't we look more global?”

A wider geographical coverage is provided. However, the population estimate may be distorted. Therefore:

- geographical scope,
- population density

They are separate results. It may be a scope panel to give one person to each country. It is not a population panel.

Objection 5 — "If we give zero to a product in an inaccessible country, wouldn't the product be unfairly punished?"

An unreachable country is not counted as zero in representation accuracy. It becomes visible in the access coverage metric. This distinction protects the product from both unfair penalties and false universality.

Objection 6 — “If there is real usage data, why do we need population data?”

Real usage data shows operational impact. However, the product's:

- which populations it has not reached,
- the geographical inequality in the user base,
- potential user fairness

It does not show alone. The two types of data answer different questions.

Objection 7 — 'Since all users will ask the same prompt, why is prompt relevance important?'

Identity questions may be meaningful in a broad population. Price, legal, or advice questions are only meaningful for certain users and markets. Irrelevant population may distort the response distribution away from the user reality.

Objection 8 — "If a person knows multiple languages, why shouldn’t they be fully counted in all languages?"

You can test all languages in cross-language comparison. However, multiplying the same person in population weighting artificially increases the world population. The person's weight should be divided, or a separate matched panel should be used.

Objection 9 — "Doesn’t testing only adults leave the global result incomplete?"

It does. This limitation should be clearly stated. Starting with an adult panel may be an ethical and operational choice. It does not make a judgement about child users.

Objection 10 — 'Isn't the online research panel sufficient?'

It can be very valuable for pilot and weighted research. But the panel:

- how much of the target population does it cover,
- who is being left out,
- how voluntary participation creates bias

It should be explained. Using a panel is not a mistake. Presenting the panel as random like the world population is a mistake.

Objection 11 — "Why is the Accessibility panel related to the main GEO score?"

Because a portion of real AI users use assistive technology or different forms of interaction. If the data collection tool excludes these people, the main score may not represent the entire target population.

Objection 12 — "Doesn't having so many different population definitions make a single NOMOS score impossible?"

No. A single score can be produced. However, it should be clear which population question it answers. For example:

- Native Reach Population Score
- Common Support Score
- Market-Relevant Score
- Language Fairness Result

can carry separate records. The meaning of a number arises from the clarity of its denominator.

COMMON JUDGEMENT OF CHAPTER 44

GEO-1000 contains '1,000' in its name, but the value of those 1,000 people does not arise from the number alone. It arises from the answers to these questions: Who are they? From which population were they drawn? Can they genuinely use the product? Where do they live? Which languages do they use? How many times was the same person counted? Are small countries visible? Do large countries retain their true population weight? Where does the score account for people the product cannot reach? Whom did the panel infrastructure exclude? Which result measures potential access, and which measures actual exposure? A large country may have a large product-eligible population and therefore deserve substantial weight in the main panel. But if the product is unavailable there, the audit must not fabricate users. The access gap must be disclosed separately.

If a country is very small, the main population weight should remain low. However, it does not have to be completely invisible. It can be observed with a monitoring panel. If a language has a small population, it may not determine the global average. However, if there is a severe representation distortion in that language, it should also be visible. If a person knows three languages, they can produce three different experiences. However, they are not three different people. If an AI product only provides certain behaviour on a paid plan, the score of free users does not represent that plan. A human population panel is suitable for a consumer chat product. The same population model may not be suitable for an API or enterprise product. Therefore, NOMOS's fifth measurement law is as follows:

> *Even the most sensitive score gives the correct answer to the wrong question if the denominator is incorrect.*

Its sixth law is:

> *Population weight does not replace country and language visibility.*

Its seventh law is as follows:

> *A person who cannot access [something] has not been misrepresented; however, they are part of the global access boundary.*

Its eighth law is as follows:

> *A person having multiple accounts, languages, or sessions does not increase the human population.*

The ninth law is as follows:

> *Global measurement is not a claim that everyone is measured; it is an honest record of who is measured and who is not.*

Order of Section 5 of NOMOS

> *Do not give me only country populations. / Show which people can actually use the product.*
>
> *Do not fill a large country with fake users when the product is not there.*
>
> *Do not give one person to a small country and declare that person's response as the voice of the country.*
>
> *Do not duplicate a person as much as their accounts, languages, and sessions.*
>
> *Do not mistake the official language for the real language usage of people.*
>
> *Do not make a free user the representative of a paid plan, or a mobile user the representative of the web interface.*
>
> *Do not call the volunteer panel a random world population.*
>
> *Do not erase a user with accessibility needs from the population because the tool does not work.*
>
> *Do not remove the population that the product does not reach from the accuracy score and hide the global access gap.*
>
> *Do not apply the same share to a price question as an identity question, or to a global brand as a local licence.*
>
> *First, define the product. / Then define the question. / Then define the human universe. / Then limit access, language, age, and interface. / Then explain whom you cannot select. / And only after that distribute 1,000 people.*

The Chapter's Closing Sentence

> *In GEO-1000, justice is not giving the same number of people to every country; it is representing each person in the correct target population, with the proper weighting, without hiding the places where they are not measured.*

Normative Core

Every GEO population audit MUST define, before result inspection, a time-bound and product-specific target user population for each: - AI product and user surface, - country or jurisdiction, - age and lawful-access condition, - account or plan condition, - language and locale, - interface, - prompt family, - and measurement purpose. Total resident population MUST NOT be treated as product-eligible population without justified eligibility adjustments. Product unavailability MUST be reported as an access-coverage limitation, not silently scored as incorrect representation or silently removed from global scope. Potential access, active use, actual exposure, Native Reach, Common Support, market relevance, and research observability MUST remain distinct population concepts. Persons, accounts, sessions, languages, and product observations MUST NOT be treated as interchangeable population units. Oversampled countries, languages, and accessibility groups MUST be reweighted for population estimates while remaining separately visible for fairness and coverage analysis. Country Observer observations MAY establish geographic observation coverage, but MUST NOT be represented as statistically reliable country scores. Participants MUST NOT be required to bypass product, geographic, age, identity, or account restrictions in order to enter the official population panel. Every population frame MUST be versioned, sourced, uncertainty-aware, and attributable to an accountable human or organisation.

CHAPTER 06

The Population-Proportional GEO-1000 Sample

METHODOLOGICAL BASES / K04; K05; K07; K08

Chapter Boundary

Section 5 established the following provision:

> *The target user population of each AI product must be defined separately. The total population of a country is not the product-eligible user population in that country.*

Now we come to the next question:

> *From the defined eligible user population for GEO-1000, how many people will be selected, from which country, which language, and by which method?*

The eligible user population for an AI product is:

- 40% in country A,
- 25% in country B,
- 15% in country C,
- 0.04% in a small country

If available, how should 1,000 valid observations be distributed? How will fractional quotas be converted into whole numbers? Will countries with a population share of less than one person be completely excluded? Will population weighting be distorted when at least one person is given to each country? Can small countries be grouped into a separate pool? If some languages are oversampled for language fairness, how will the global score be calculated? If the target is 100 valid observations in a country, how many people should be invited? Who will replace the participant if they do not complete the task? Can a user giving wrong answers be replaced? If the same person tests more than one AI product, will the sample still be independent?

Even with a raw sample of 1,000 people, can highly unequal weights reduce the effective statistical information below that of 1,000 independent, equally weighted observations? These are not merely operational details. The sample design determines how much influence each person has on the GEO score. A defective sample:

- can misweight the correct answers,
- can symbolically overrepresent small countries,
- can make large populations appear smaller than they are,
- can make low-resource languages invisible,
- can present voluntary technology users as if they were the world population.

It can be adjusted to achieve the desired score after reviewing the results. The previously established audit architecture requires that each audit be measurable and that the model/product carry the date, country, language, query set, repetition count, and measurement record. This section applies the same discipline to sample design. Because genuine evidence requires requesting limits, context, and time; it also necessitates making how the sample was created visible. This section:

- how the 1,000 valid observations of the main population panel will be distributed,
- how strata and cells will be established,
- how small countries will be managed,
- how fractions will be converted to whole numbers,
- how oversampling will be weighted,
- how to separate invitation, backup, and valid observation targets,
- how to establish design and analysis weights,
- why the effective sample size may differ from the raw number

defines. This chapter does not yet:

- from which panel provider the participants will be selected,
- the full details of field implementation in voluntary or probability sampling,
- reviewing and scoring responses,
- the final confidence interval method,
- all NOMOS scoring formulas

are not fully finalised. The task of Section 6 is:

> *Transform the defined target population into a frozen, weightable, and reproducible GEO-1000 sample before seeing the outcome.*

NOMOS Challenge

You are setting up a global panel of 1,000 people for an AI product. The product-eligible populations of ten countries are as follows:

Country	Eligible user population
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

Raw quotas according to pure population proportion:

Country	Raw quota

Country	Raw quota
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

Half of a person cannot be selected. 0.4 people cannot be sent to a country. If you round the raw quotas to the nearest whole number, the total may exceed or fall short of 1,000. If you give at least one person to each country, you over-represent countries I and J compared to their population shares.

Giving one person to country I increases its theoretical share 2.5 times. Giving one person to country J increases its theoretical share 10 times. This extreme representation can largely be corrected. But you cannot produce a country score with a single person. What happens if you give no one to I and J?

If you define these countries as separate and mandatory strata, the probability of some people in the target population being selected becomes zero. In this case, the claim of a "probabilistic sample representing the entire appropriate global population" weakens. Now consider a more dangerous decision. First, you collect the answers.

The only user in country J gets an incorrect answer. Then you say: "The sample size of this country is very small, let's exclude it from the scope." In country A, there are many positive answers. You retain it with full weight. This is no longer rounding. It is a result-oriented sample change.

In another case, you increase the invalid record rate in a language group with low results, using backup participants only until a positive response is obtained. This is also not a sample. It is a response selection mechanism. The first ruling of this section is:

> *The sample must be established before seeing AI responses.*

Its second provision states:

> *The probability of a participant being selected cannot depend on the correctness of their answer.*

Its third provision states:

> *The visibility of the small country and the population weight in the global score must be managed separately.*

Its fourth provision states:

> *Oversampling is possible; over-weighting is not mandatory.*

Its fifth provision states:

> *1,000 raw records do not always carry statistical information equivalent to 1,000 observations.*

1. PURPOSE OF THE CHAPTER

The purpose of this section is to transform the targeted 1,000 valid observations for each AI product and measurement wave into a verifiable sample design based on the product-eligible user population. The section standardises the following distinctions:

- Target population versus selected sample
- Raw quota versus integer allocation
- Main population panel versus complementary panels
- Country visibility versus country weight
- Strata versus experimental cell
- Large country strata versus small country pool
- Proportional allocation vs. disproportionate oversampling
- Sample share vs. population share
- Number of invitations vs. valid observation target
- Backup participant vs. post-result replacement
- Data collection invalidity vs. adverse AI output
- Design weight vs. outcome weight
- Raw sample size vs. effective sample size
- Country sample vs. language sample
- Multilingual user vs. more than one person
- Single-stage selection vs. multi-stage selection
- Native Reach allocation vs. Common Support allocation
- Instant wave allocation and time series allocation
- Population estimation and country or language recognition analysis
- Main 1,000 observations with auxiliary observers and fairness panels

At the end of this section, each audit should be able to answer these questions:

> *Why were 1,000 observations distributed in these numbers across this country, language, and user groups?*
>
> *How was the weight of each observation in the global result determined?*
>
> *Who had a zero probability of being selected and why?*
>
> *How many people were invited, how many completed, and how many observations remained valid?*
>
> *Were the sample results frozen before being viewed?*

2. CENTRAL NORMATIVE PROVISION

The GEO-1000 main population sample should be based on a stratified sampling design that assigns known or predictable probabilities of selection to people in the pre-frozen product-eligible population frame for each AI product and measurement wave. The default target of the main population panel is:

> *At least 1,000 valid observations per AI product and main measurement wave*

is maintained. This number:

- is not the number of people invited,
- is not the number of accounts,

- is not the number of raw screenshots,

It is not the sum of all supplementary panels. 1,000 is the target for valid product-observations meeting the protocol in the main population panel. Country Observer, Language Fairness, and Accessibility panels can create additional samples. These additional samples:

- can increase the total field volume above 1,000,
- can be used with appropriate weights in the global population outcome,

but cannot be used to covertly reduce the 1,000 valid observation target of the main panel.

3. BASIC REPRESENTATION OF THE SAMPLE

For a specific AI product a, measurement wave w, and target population frame, the strata:

```
h ∈ H_{a,w}
```

are denoted. For each stratum:

- Nh: estimated product-appropriate population in the stratum
- N: total target population
- Wh: population weight of the stratum
- n: target valid main panel observation
- qh: crude proportional quota
- nh: allocated integer valid observation target

is. Total target population:

```
N = Σ_{h=1}^H N_h
```

Population weight of the stratum:

```
W_h = N_h / N
```

Crude proportional quota:

```
q_h = nW_h
```

is. Main proportional allocation, under ideal conditions:

```
n_h ≈ nW_h
```

should be. However:

- qh may be fractional,
- some cells may fall below one person,
- a minimum diagnostic sample may be required for language and country subgroups,
- high-risk small groups may also be examined,

field completion rates may vary. Therefore, the raw quota is not directly the final allocation.

4. THREE DISTINCT PURPOSES OF SAMPLE DESIGN

The same sample may not fulfil all purposes with the same strength. Three purposes should be separated.

4.1. Global Population Estimate

Answers the question:

> *What representative outcome would the average person in the global user population suitable for the product see?*

Population weight is the primary priority.

4.2. Subgroup Diagnosis

Answers the question:

> *What representation distortions occur in specific countries, languages, plans, or user groups?*

It may be necessary to oversample small groups.

4.3. Geographic and Linguistic Coverage

Answers the question:

> *In which countries and languages is there at least real user observation?*

Country Observer and Language Fairness panels become important. These three goals cannot be inexplicably combined in a single sample.

5. MAIN SAMPLE ARCHITECTURE

The candidate main sample architecture for GEO-1000 consists of four layers.

5.1. Population Panel

Every AI product and wave has a minimum of 1,000 valid observations as the main panel. It produces a global population estimate.

5.2. Country Observer Overlay

Provides additional geographic observations for small countries that cannot be directly represented in the main population panel. These observations can indicate:

- single event,
- access status,
- future research needs

They do not generate a country score on their own.

5.3. Language Fairness Oversample

Generates additional observations for low-resource languages or languages underrepresented in the main panel. They are reweighted to actual population weights in the global outcome.

5.4. Accessibility Oversample

It also measures users of assistive technology and accessibility, who are not adequately represented by the standard panel. The target universe and weight of these observations are recorded separately.

6. WHAT IS A STRATUM?

A stratum is a segment of the population that is managed separately in terms of sample allocation and selection probability. The stratum can be a combination of one or more of the following areas:

- Country
- Region
- Language

- Locale
- AI plan
- Web or mobile
- Controlled or natural user panel
- Accessibility status
- User type
- Jurisdiction

If each small combination is made a separate stratum, the sample is fragmented. For example: Country × language × plan × mobile/web × age group × user type combination can produce hundreds or thousands of cells. 1,000 observations do not provide reliable estimates for most of these cells. Therefore, stratification should be:

- substantial,
- measurable,
- appropriate for sample size

must be.

7. A STRATUM IS NOT THE SAME AS A REPORTING CELL

It can be shown separately on a group result card. However, it does not have to be a separate stratum in sample design. For example:

- a small language,
- a specific accessibility group,
- a certain age band

It can be selected within a broader stratum in the main sample. Later, it can be reported separately in the analysis. The reverse is also possible: Two regions may be separately stratified for sampling efficiency. However, in the public report, they can be combined under a single regional result. The stratification design and the reporting structure should carry separate records.

8. SEQUENCE OF STRATUM ESTABLISHMENT

The candidate sequence for the Main Population Panel is as follows:

- AI product and user surface
- Native Reach or Common Support population frame
- Country or small country pool
- Main audit language or language group
- Material plan/interface separation, if necessary
- Participant selection frame

The plan and interface do not have to be made into separate layers for each country. These can be managed through:

- sub-layer,
- quota,
- random assignment

methods.

9. LARGE AND SMALL COUNTRY LAYERS

A country's raw quota:

```
q_c = nN_c / N
```

is in this form. If the raw quota is large enough, the country can be kept directly as a separate layer. This can be called:

self-representing country layer

Making countries with very small raw quotas individually into separate mandatory layers may reduce sampling efficiency. These countries can be appropriately:

- regional,
- language family,
- product access regime,
- geographical proximity

can be included in small country pools.

10. SMALL COUNTRY POOL

A small country pool:

```
Let it be represented as B_r = {c_1, c_2, …, c_k}
```

The total eligible population of the pool:

```
N_{B_r} = Σ_{c∈B_r} N_c
```

Main sample quota:

```
q_{B_r} = nN_{B_r} / N
```

is calculated as. Within the pool, country selection can be done with probability proportional to the eligible population size. The probability of country c being selected within the pool:

```
Pr(c | B_r) = N_c / N_{B_r}
```

It may be. Then a participant is selected within the chosen country. This approach:

- can prevent small countries from having a zero chance of being selected in the target population,
- does not guarantee that each country will be selected in a given wave,

preserves the global population estimate. Not selecting a country in that wave:

NOT SELECTED IN POPULATION SAMPLE

does not mean: "The country is not part of the target population."

11. LIMITATIONS OF POOLING SMALL COUNTRIES

Small countries cannot be randomly placed in the same pool just because their populations are small. Countries within the pool are similar to the extent that they are relevant:

- product access,
- user surface,

- language or locale structure,
- legal access,
- geographical region,
- data collection method

must be carried. A country without product access cannot be placed in the same main pool as a country with full access. Different usage conditions cannot be made invisible.

12. SMALL COUNTRY POOL DOES NOT PRODUCE COUNTRY SCORE

One or two observations selected from the pool: can be used in the global population contribution of the pool, may indicate a single country event. However, it does not produce a separate statistical score for a small country. If a country-based score is wanted, additionally:

- Country Observer,
- country-specific sample,
- multi-wave cumulative sample

may be required.

13. ALLOCATION PROPORTIONAL TO PURE POPULATION

Pure proportional allocation:

is based on the formula:

```
q_h = nW_h
```

Advantages:

- it is understandable for the main population estimate,
- it naturally gives weight to large populations,
- it can reduce weight differences,

it can approach a self-weighting design. Limitations:

- it can sample very few small countries and languages,
- it may be insufficient to produce subgroup scores,
- it may leave high-risk small groups invisible,

it can create fractional and near-zero quotas. Therefore, pure proportional allocation can be a strong default for the main Population Panel. It does not replace complementary panels.

14. CONVERTING FRACTIONS TO WHOLE NUMBERS

Raw quotas are often fractional. The total allocation should be exactly n. The Largest Remainder Method can be used as the default method.

14.1. First Step

Each raw quota is rounded down:

```
b_h = ⌊q_h⌋
```

14.2. Second Step

The number of unallocated observations:

```
R = n − Σ_h b_h
```

is calculated.

14.3. Third Step

The fractional part of each stratum:

```
f_h = q_h − b_h
```

where f_h is the fractional part of q_h.

14.4. Fourth Step

The remaining R observations are added one by one to the layers with the highest fh values.

14.5. Equality Situation

If the fractional parts of the two layers are equal:

- population size,
- predetermined fixed order,
- registered random seed

An independent tie-breaking rule should be used, regardless of the outcome. A country cannot be chosen based on the results.

15. SYNTHETIC INTEGER ALLOCATION

SYNTHETIC DISPLAY

Appropriate populations at the beginning of the section:

Country	Suitable population	Raw quota
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

If countries are protected as separate layers, round down:

Country	Rounded down quota
A	400

Country	Rounded down quota
B	250
C	150
D	100
E	60
F	30
G	8
H	1
I	0
J	0
Total	999

One observation remains. It is added to H because the highest fraction is 0.5 in country H. Final allocation:

Country	Final quota
A	400
B	250
C	150
D	100
E	60
F	30
G	8
H	2
I	0
J	0
Total	1,000

Even though this design defines I and J as separate deterministic strata, if it does not give a chance to be selected, it may pose a problem in terms of full target universe inference. A more accurate solution is a small country pool.

16. SYNTHETIC ALLOCATION WITH SMALL COUNTRY POOL

Countries H, I, and J:

```
B_small = {H, I, J}
```

should be combined in the pool. Pool population:

```
1,500,000+400,000+100,000=2,000,000
```

Pool quota:

```
1,000 × 2,000,000 / 1,000,000,000 = 2
```

becomes. Main allocation:

Stratum	Quota
A	400
B	250
C	150
D	100
E	60
F	30
G	8
Small country pool	2
Total	1,000

Selection weights of countries within the pool:

Country	Population share within the pool
H	75%
I	20%
J	5%

Two main population observations are selected according to these probabilities. In the same wave:

- both observations can come from H,
- one from H and one from I,
- rarely from J

This can occur. The design does not make J visible in every wave, but it does not reduce J's probability of selection to zero. If J must be observed, it is added separately to the Country Observer Panel. A Country Observer observation does not give J a weight of one in 1,000 in the main global allocation; J retains its true population weight in the main panel.

17. MINIMUM CELL ASSIGNMENT

Purely proportional quota may not be sufficient to perform diagnostic analysis for some subgroups. For example, the raw quota for a language cell might be 8. A reliable language score cannot be produced with eight observations. In this case, there are three options:

- Not to publish the language score
- Combine the cell with similar groups
- Oversampling the cell

Minimum cell allocation must be defined before the results.

18. CANDIDATE REPORTING BANDS

CANDIDATE REPORTING BANDS — CANDIDATE REPORTING BANDS

The thresholds below are not final scientific decisions. They should be calibrated through pilot and external review.

1–29 Valid Observation

Single event and discovery Country or language rate is not published

OBSERVATIONAL ONLY

30–99 Valid Observations

Exploratory level estimate Broad uncertainty No definitive conformity judgement is given

EXPLORATORY

100–199 Valid Observations

First subgroup estimate Uncertainty should be visible

PROVISIONAL

200 and Above Valid Observations

Stronger subgroup estimate candidate Design effect and weights are again evaluated

STANDARD ESTIMATION CANDIDATE

These bands:

- error rate,
- expected precision,
- weight variability,
- design effect

cannot be automatically considered sufficient without being taken into account.

19. OVERSAMPLING

A country, language, or user group may have more observations than its population share. This is called oversampling. Its purposes:

- to increase subgroup sensitivity,
- to make a low-resource language visible,
- to examine a high-risk small group,
- to expand Country Observer coverage,
- to adequately represent accessibility users

are possible. Oversampling is not wrong. What is wrong:

> *It is to use the sample proportion of the over-sampled group as if it were the true population proportion.*

20. POPULATION SHARE AND SAMPLE SHARE

Population share for layer h:

```
W_h = N_h / N
```

Sample share:

```
S_h = n_h / n
```

is in the form. In pure proportional allocation:

```
S_h ≈ W_h
```

okay. In excessive sampling:

```
S_h > W_h
```

It is possible. In a small sample:

```
S_h < W_h
```

It is possible. The global population estimate should retain Wh values instead of Sh.

21. BASIC DESIGN LOAD

In simple stratified sampling, the basic weight of each observation within stratum h can be considered as:

```
d_i = N_h / n_h
```

The normalised stratum weight can be written as:

```
a_h = W_h / S_h
```

The normalised observation-level weight can be:

```
w_i = W_h / n_h
```

It is ensured that the sum equals one:

```
Σ_i w_i = 1
```

The weighted binary transition estimate can be calculated as:

```
θˆ= Σ_i w_iY_i
```

Here:

```
Yi=1: response meeting the transition criterion

Yi=0: response not meeting it
```

It is possible. The final NOMOS system does not have to use only binary outcomes. The same weighting logic can also be applied to multidimensional response scores.

22. EXCESS SAMPLING EXAMPLE

SYNTHETIC EXAMPLE

In the suitable population of country B:

- Language X: 80 per cent
- Language Y: 20 per cent

Let it be. In the Population Panel, there are 250 observations for country B. Pure allocation:

- Language X: 200
- Language Y: 50

Let 50 additional Language Fairness observations be taken for Language Y. Collected observations:

- Language X: 200
- Language Y: 100

The share of Language Y in the raw sample:

```
100/300 = 33.3%
```

It has been. The actual population share is 20%. In the global population account, Language Y cannot take a 33.3% weight. Within-stratum normalised weights:

- Total weight for Language X: 80%
- Total weight for Language Y: 20%

is preserved. Each of the 100 observations in Language Y receives a lower individual weight than a single observation in Language X. In contrast, all 100 observations are used in Language Y's separate diagnostic report. This approach: increases subgroup visibility, does not distort the global population estimate.

23. ALLOCATION OF MULTILINGUAL USERS

A person may be suitable for more than one language cell. One of the preferred approaches for the main Population Panel is to assign a predefined main audit language to the user. The main audit language can be determined according to these fields:

- Daily AI usage language
- Mother or home language
- Market and locale
- Competence in naturally using the prompt
- Panel balance

If the user will also be tested in other languages, these observations:

- paired multilingual panel,
- Language Fairness Panel

are also labelled accordingly. It is prohibited for the same person to receive full population weighting in all languages.

24. LANGUAGE ALLOCATION WITHIN THE COUNTRY

Let the suitable population for language groups l in country c be: N_{c,l}. If the country quota is: n_c, then the raw language quota:

```
q_{c,l} = n_c × N_{c,l} / Σ_l N_{c,l}
```

can be calculated as such. However, if language groups overlap:

```
Σ_l N_{c,l} > N_c
```

may be possible. In this case, using a direct share creates double counting. Solution options:

- main audit language classification,
- split human weight,
- multilingual common model,
- separate matched language panel

may be used. The method used should be recorded.

25. ALLOCATION OF PLAN AND INTERFACE WITHIN THE POPULATION

Free and paid users or web and mobile surfaces may behave differently. However, dividing each country–language cell separately into plan and interface layers can overly fragment the sample. Three approaches can be used.

25.1. Suballocation Proportional to Population

If the actual user distribution is known, allocation is made according to plan and interface shares.

25.2. Balanced Experimental Allocation

Equal or greater sampling is carried out to compare the difference between plan or interface. In global estimation, it is re-weighted to the actual usage shares.

25.3. Separate Product Surface Scores

Free web, paid web, and mobile products are reported as separate audit objects. A combined result can also be produced. The approach to be used should be defined before measurement.

26. INVITATION NUMBER AND VALID OBSERVATION TARGET

If the valid observation target for a stratum is nh, inviting exactly nh people is usually not sufficient. Because:

- some people do not participate,
- some do not complete the task,
- some records remain out of protocol,

some evidence packages become invalid. Let the expected completion rate for the stratum be rh and the expected validity rate be vh. The invitation target is approximately:

```
m_h = ⌈n_h / (r_h v_h)⌉
```

can be calculated as follows. Example:

- Valid observation target: 100
- Expected completion rate: 80 per cent
- Expected validity rate: 90 per cent

```
m_h = ⌈100 / (0.80 × 0.90)⌉ = 139
```

people can be invited. This is not a guarantee. As the field progresses, the predefined backup system can be used.

27. LIMIT OF INVITATION INFLATION

Inviting more people alone does not solve sample bias. Early responders may be systematically different from late responders. Only taking the first 1,000 completers:

- time zone,
- technology interest,

- connection quality,
- work schedule

may produce bias in terms of. Invitation and selection order:

- should be randomised,
- must be recorded within the stratum,

should be managed throughout the field window.

28. ALTERNATE PARTICIPANT SYSTEM

An ordered backup list should be created for each stratum before the results are seen. A backup participant can be activated only for the following reasons:

- No response to the invitation
- Failure to complete the task
- Technical or ethical suitability issue
- Violation of the demand protocol
- Invalidity of the evidence package
- Re-registration of the same person

The backup user cannot intervene for the following reason:

> *The previous user's AI response gave an incorrect, negative, or unwanted output.*

An incorrect AI response is the valid system result. It cannot be a reason for participant change.

29. RESULT BLINDNESS IN VALIDITY DECISION

The person deciding whether the observation complies with the protocol as much as possible:

- must evaluate the response without knowing whether it is correct or incorrect,
- or whether it is positive or negative for the brand.

First:

- Capture and protocol validity
- Then the substantive response evaluation

must be done. Otherwise, negative responses can more easily be declared: "invalid."

30. RESERVE ORDER WITHIN LAYER

The reserve participant should be from the same as much as possible:

- country,
- language,
- product surface,
- plan,
- panel condition

layer. If no reserve remains in the same layer, predefined escalation is applied:

- Additional random selection within the same layer

- Predefined upper layer pool
- Leaving allocation incomplete and scope warning
- New analysis version

After the result is seen, bringing a user from another country and closing the gap cannot be done silently.

31. WHAT HAPPENS IF A LAYER IS TOO SHORT?

It may fall below a layer target. Example:

- target: 100
- valid observation: 72

Correct options:

- Extend the field window to a predefined extent
- Use the backup list
- Merge the cell with a similar and predefined upper layer
- Report the result with wide uncertainty
- Do not publish the cell result
- Recalculate the global weight using an appropriate method

Wrong option: Completely remove the low-scoring layer from the analysis.

32. POST-RESULT QUOTA CHANGE

The following changes are result-oriented sample manipulations:

- Reducing the target quota of the low-scoring country
- Increasing the quota of the high-scoring language
- Invalidating users who provide negative responses
- Redefining the small country pool based on results
- Excluding a plan with critical errors
- Adding new users until the desired average is reached

Each material quota change:

- new allocation version,
- justification,
- date,
- decision maker
- Preservation of the previous result

should be done with.

33. MULTI-STAGE SAMPLING

Multi-stage sampling can be used when direct access to all individuals is not possible. Example:

- Selection of region or country pool
- Country selection

- Selection of panel or local research partner
- User selection
- AI product or task assignment

The total probability of selection for an individual user is the product of the conditional probabilities at each stage:

```
π_i = π_r × π_{c|r} × π_{f|c} × π_{i|f}
```

Here:

- r: region or small country pool
- c: country
- f: sampling frame or panel
- i: person

Basic design weight:

```
d_i = 1/π_i
```

If selection probabilities cannot be recorded, the design cannot be presented as a fully probability-based sample. [K04; K07]

34. ALLOCATION IN VOLUNTARY AND NON-PROBABILITY PANELS

A true probability-based world panel may not always be possible. A voluntary or professional online panel can be used. In this case, country and language quotas can still be set according to the product-eligible population. However:

- the participant's actual probability of selection may be unknown,
- self-selection may occur,

weighting can only make adjustments based on observed variables. The correct status:

Non-probability user panel stratified and calibrated to the population

may be. The following statement cannot be used: “1,000 people selected completely at random from the world population.”

35. WEIGHTING ARCHITECTURE

The final observation weight may consist of multiple components:

```
w_i = d_i × a_i^{NR} × g_i^{CAL}
```

Here:

- di: base design weight
- aiNR: non-response adjustment
- giCAL: calibration coefficient

35.1. Design Weight

It is the inverse of the participant's probability of selection in the sampling design.

35.2. Nonresponse Adjustment

For selected persons with similar characteristics:

- how many responded,
- how many produced a valid observation

can be taken into account. For sample response class r:

```
a_r^{NR} = (Σ_{j∈S_r} d_j) / (Σ_{i∈R_r} d_i)
```

Here:

- Sr: selected eligible sample
- Rr: those who responded or produced a valid observation

may be. Nonresponse adjustment: does not completely resolve unobserved bias, should be explained along with its assumptions.

35.3. Calibration Weight

The sample, to known population totals:

- country,
- language,
- age,
- plan,
- user surface.

Calibration may be applied across these dimensions, but only where reliable external totals exist.

36. WEIGHT NORMALISATION

Weights:

- to the total population,
- to the total sample size,
- to individual

can be normalised. The normalisation used in the score formula should be explained. Individually normalised weights:

```
w̃_i = w_i / Σ_j w_j
```

are obtained. Weighted estimate:

```
θ̂ = Σ_i w̃_iY_i
```

is calculated as follows.

37. WEIGHT TRIMMING

Some participants may receive very high weights because they represent very small and under-sampled groups. This situation:

- can increase the estimated variance,
- the influence of a single observation,
- and score volatility

can increase. Weights can be trimmed at an upper limit. However, trimming: may create bias in population representation, may make small groups invisible again. Therefore, weight trimming should:

- be defined before the results,
- be shown together with the untrimmed results,

and carry a sensitivity analysis.

38. EFFECTIVE SAMPLE SIZE

Even though the raw number of observations is n, if the weights vary greatly, the actual amount of information may be lower. Kish effective sample size:

```
n_eff = (Σ_i w_i)^2 / Σ_i w_i^2 [K08]
```

can be shown as. With equal weights:

```
neff≈n
```

is obtained. If the weights are very unbalanced: neff<n. a GEO-1000 study:

- 1,000 valid observations,
- may carry 620 effective sample size

. To the public only: saying “1,000 people were measured.” can appear strong due to the precision. In the main report, at least:

- raw valid observation count,
- effective sample size,
- weight variability

should be determined.

39. WEIGHT DESIGN EFFECT

Approximate design effect due to weights:

```
DEFF_w ≈ 1 + CV(w)^2
```

can be shown. Here:

```
CV(w) = sd(w) / mean(w)
```

This is the coefficient of variation of the weights and only an approximate representation of their effect. The total design effect may also vary with:

- clustering,
- stratification,
- repeated measurement,
- paired users

These effects must be estimated separately.

40. CLUSTERING

The same:

- From the research panel,
- from the country,

- from the language community,
- from the device or institution,
- from the user

observations may resemble each other. In this case, 1,000 observations are not 1,000 completely independent units of information. Clustering:

- in sample design,
- in variance calculation,
- in effective sample size

should be taken into account.

41. TESTING MORE THAN ONE AI PRODUCT BY THE SAME USER

Having a user test multiple AI products allows the difference between models to be measured on the same person. Advantage: user characteristics can be controlled, paired comparison can be strengthened. Limitation:

- product observations are not independent,
- the order of tasks can affect other responses,
- a user may carry the answer seen in the first system to the next system,

users who can access all products may differ from the population. Therefore, products:

- should be assigned in random order,
- previous answer sharing should be prevented,
- the paired design should be recorded,

Analysis should protect dependency. Every AI product should again have its own target of 1,000 valid product-observations. If the same 1,000 humans test ten products:

- 1,000 unique humans,
- 10,000 product-observations

can be formed. These two numbers are not the same.

42. BALANCED MISSING BLOCK DESIGN

Not all users need to access all AI products. For example, if there are ten AI products, each participant can be assigned three products. If assignments are made in a balanced way:

- each product 1,000 observations,
- each product pair sufficient shared users,
- broader population participation

can be achieved. As a candidate for this approach:

Balanced incomplete block design

can be used. Its advantages:

- reduces participant burden,
- lowers the risk of selecting only heavy AI users,

provides matched comparison across models. The exact design will be detailed in Sections 8 and 9.

43. NATIVE REACH ALLOCATION

Each AI product receives a separate allocation according to its own Native Reach population. Therefore, the same country:

- 200 for one product,
- 0 for another product,
- 50 for a third product

can be observed. This difference is not unfairness. It is that the real access universes of the products are different. Since Native Reach scores belong to different populations, they should not be used alone for direct ranking.

44. COMMON SUPPORT ALLOCATION

A common access universe is created for comparison between products. The common population:

```
U_CS = ⋂_{a∈A} U_a
```

can be considered in this way. Each product is measured with the same Common Support population weights. This approach:

- provides fair product comparison,
- same country and language floor

ensures. However, it does not show the natural global access difference of the products. Therefore, Native Reach and Common Support results should be reported together.

45. MANAGEMENT OF THE SAMPLE ACROSS WAVES

There may be two different objectives across multiple measurement waves.

45.1. Current Cross-Section Result

In each wave, a current product-eligible population frame is used. This score answers the question: "What is happening now in the actual global access universe?"

45.2. Fixed Frame Trend Result

The population weights from the initial wave are preserved. This score answers the question: "How has representation in the same reference population changed over time?"

45.3. Separation of Two Results

When the product expands to new countries, the current score may decrease. This decrease:

- is due to the deterioration of the product,
- from the inclusion of new and difficult populations

It may be attributable. A trend with a fixed frame can show improvement at the same time. Therefore:

- Current-Frame Snapshot
- Fixed-Frame Trend

They are separate results.

46. PANEL RENEWAL AND TRANSFORMATION

Among the waves:

- the same users,
- completely new users,
- partially renewed panel

usable.

46.1. Full Section Panel

New users are selected in each wave. It can be strong for population estimation. It does not measure individual change.

46.2. Longitudinal Panel

The same users are measured again. Shows individual user and product changes. The panel may create wear and learning effects.

46.3. Rotating Panel

A portion of the participants is retained, while a portion is renewed. Provides both trend and fresh population representation. The panel structure used should be visible on the result card.

47. SAMPLE LOCK

The following records should be locked before each wave starts:

- AI product
- User surface
- Target population version
- Native Reach countries
- Common Support countries
- Strata
- Small country pools
- Raw quotas
- Integer allocations
- Language allocations
- Oversampling plan
- Invitation targets
- Reserve lists
- Selection method
- Weight formula
- Result blindness rule
- Shortfall escalation
- Analysis plan
- Decoding method
- Random seed, if used

- Responsible human approval

To this file:

NOMOS Sample Allocation Lock

can be given the name.

48. CHANGES ALLOWED AFTER SAMPLE LOCK

The following events may require a new version:

- Termination of product access in the country
- Serious field safety issue
- Material error in population data
- Panel provider unable to deliver to a specific country
- Change in research ethics requirement
- Critical error in measurement tool

Change:

- must be justified,
- dated,
- approved,
- carried out while preserving the old version.

The strength or weakness of the results is not a reason for change.

49. SAMPLE QUALITY STATUSES

A sample wave can carry one of the following statuses:

SQ-0 — NOT DESIGNED

There is no documented sample design.

SQ-1 — CONVENIENCE OBSERVATION

There is convenience or voluntary observation. Population generalisation cannot be made.

SQ-2 — QUOTA-CONTROLLED

There are country and language quotas. Selection probabilities are not fully known.

SQ-3 — WEIGHTED PANEL

There is a stratified panel and population calibration. Coverage and self-selection limitations persist.

SQ-4 — PROBABILITY-BASED OR HYBRID

Known election probabilities or strong hybrid design exist.

SQ-5 — REPLICATED AND CALIBRATED

The design has been replicated in different waves and external applications; bias and variance behaviour have been calibrated. These statuses may affect the confidence class of the public score.

50. SYNTHETIC END-TO-END GEO-1000 ALLOCATION

SYNTHETIC METHODOLOGY DEMONSTRATION / The following countries, populations, languages, AI product, and calculations are entirely fictional. For the Synthetic Orion AI product, let the suitable population for Native Reach be 1 billion people. Main Population Panel objective:

```
n=1,000
```

50.1. Country and Small Country Pool Allocation

Stratum	Suitable population	Population share	Main quota
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
Small country pool	2,000,000	0.2%	2
Total	1,000,000,000	100%	1,000

50.2. Language Allocation in Country B

Eligible population in Country B:

- Language X: 80 per cent
- Language Y: 20 per cent

Population Panel allocation:

- Language X: 200
- Language Y: 50

Language Fairness additional allocation: Language Y: 50 Field target in Country B:

- Language X: 200
- Language Y: 100
- Total: 300

However, global Population Panel population weights:

- Language X: 80% of Country B
- Language Y: 20% of Country B

remain as such.

50.3. Country Observer Appendix

There are two probability observations for H, I, and J in the small country pool in the Population Panel. The Country Observer Panel also:

H: 1

I: 1

J: 1

can target observations. These three observations:

- increase the geographical coverage,
- show individual country events,

do not create a country score.

50.4. Accessibility Supplement

If the main panel is found to inadequately cover accessibility users, 40 additional observations can be taken. These users: are calibrated according to relevant population groups, produce a separate Accessibility Panel result.

50.5. Total Field Volume

Panel	Current target
Population Panel	1,000
Language Fairness supplement	50
Country Observer supplement	3
Accessibility supplement	40
Total valid field target	1,093

The name GEO-1000 is still valid. Because:

> *1,000 is the valid observation target for the main population panel.*

Complementary justice and coverage panels can be added to this.

50.6. Invitation Volume

On the Population Panel on average:

- completion: 82 per cent
- validity: 92 per cent

let it be expected. Required invitation:

```
⌈1,000 / (0.82 × 0.92)⌉ = 1,326
```

may be. With complementary panels, total invitations may exceed 1,500. Therefore:

> *GEO-1000 is not only a method to send invitations to 1,000 people.*

50.7. Field Result

On the Population Panel:

- 1,326 people have been invited,
- 1,080 tasks completed,
- 1,012 records have been found valid in terms of protocol,
- the first 1,000 were included in the main analysis according to a predefined random/backup order

Alright. The remaining 12 valid records:

- backup verification,
- sensitivity analysis,
- recheck

It can be stored for. It cannot be decided which 1,000 will be chosen by looking at the results.

51. NORMATIVE ALLOCATION ALGORITHM

The default sequence of operations for a GEO-1000 main panel is as follows:

Step 1 — Lock the Product and Wave ID

Determine which AI product, surface, and date are measured.

Step 2 — Lock the Target Population Version

The product-eligible population frame from section 5 is used.

Step 3 — Define the Main Current Observation Target

Default candidate target:

```
n=1,000
```

Step 4 — Define Material Layers

Country, language, and if necessary, plan/interface layers are defined.

Step 5 — Set Up Small Country Pools

Countries that will generate near-zero allocations are included in appropriate pools.

Step 6 — Calculate Raw Proportional Quotas

```
qh=nWh
```

Step 7 — Perform Integer Allocation

Largest Remainder or pre-declared equivalent method is applied.

Step 8 — Add Oversampling Plan

Country Observer, Language Fairness, and Accessibility supplements are saved separately.

Step 9 — Calculate Invitation and Backup Numbers

Expected completion and validity rates are used.

Step 10 — Randomise Selection Order

Primary and backup lists are frozen.

Step 11 — Save Weight Formula

Design, response, and calibration weights are defined.

Step 12 — Confirm Sample Lock

A version and integrity record is created before AI responses are viewed.

52. MANDATORY NORMATIVE PROVISIONS

CH06-N01

The GEO-1000 sample allocation must be created and versioned before AI responses are viewed.

CH06-N02

The main Population Panel target should be defined as at least 1,000 valid observations per AI product and main measurement wave; if a different target is used, the rationale must be explained.

CH06-N03

The invited person, the person completing the task, and the valid observation numbers must be distinguished from each other.

CH06-N04

The main allocation should be based on the product-eligible population defined in Section 5.

CH06-N05

Country or language quotas should be calculated based on the share of the relevant product-eligible population, not the total population.

CH06-N06

The method for converting raw fractional quotas into whole numbers must be defined in advance.

CH06-N07

The resolution of rounding ties must be independent of the results.

CH06-N08

Small countries with a positive target population should not be left as separate zero-probability deterministic strata; an appropriate pooling or probabilistic selection method should be used.

CH06-N09

Small country pools should be justified in terms of product access and financial population characteristics.

CH06-N10

Country Observer observations should also be recorded without changing the main population weight.

CH06-N11

A Country Observer observation cannot be presented as a country score.

CH06-N12

Over-sampled countries, languages, or user groups should be returned to the true target population weights in global estimation.

CH06-N13

The sample proportion cannot be used like the population proportion.

CH06-N14

The same person cannot be duplicated in the global human population due to multiple languages, accounts, sessions, or products.

CH06-N15

Invalid observations should be classified without basing the reason for invalidity on the content of the results.

CH06-N16

Incorrect, negative, rejection, or Critical AI responses cannot be the reason for changing a participant.

CH06-N17

Backup participant lists must be created within the stratum before the results are seen.

CH06-N18

Layer gaps cannot be quietly closed by another country or language.

CH06-N19

Post-result quota, layer, small country pool, or weight changes require a new analysis version.

CH06-N20

All components of the weights and the method of normalisation must be published.

CH06-N21

The assumptions of the nonresponse adjustment must be explained.

CH06-N22

If weight trimming is used, the untrimmed result and sensitivity analysis must be retained.

CH06-N23

The main report should show the appropriate effective sample size alongside the raw number of observations.

CH06-N24

Clustering, repetition, and paired user dependencies should be taken into account in the calculation of variance.

CH06-N25

The design in which the same person tests multiple AI products cannot be presented as a completely independent user sample.

CH06-N26

Each AI product must separately reach its own target of 1,000 valid product observations.

CH06-N27

Allocations Native Reach and Common Support must hold separate sample records.

CH06-N28

Current and fixed-frame results among the waves cannot be combined without explanation.

CH06-N29

If the sample is insufficient for a subgroup, the correct status should be INSUFFICIENT SAMPLE, EXPLORATORY, or NOT ESTIMATED.

CH06-N30

The raw sample size cannot be used as a definitive precision indicator, ignoring weights and design effect.

CH06-N31

Design weights cannot be claimed in a volunteer or non-probability panel as if selection probabilities are known.

CH06-N32

The sample cannot be modified post hoc to exclude low-scoring groups or to enlarge high-scoring groups.

CH06-N33

Observations from the main Population Panel and complementary panels should be labelled separately in terms of purpose and weight.

CH06-N34

The sample lock must have an accountable human or institutional owner.

53. FORMS OF FAILURE

CH06-F01 — SET QUOTA AFTER RESULT

The number of countries and languages is determined after viewing AI responses.

CH06-F02 — USE BACKUP UNTIL BEST ANSWER

Negative responses are replaced with a new user even if valid.

CH06-F03 — COUNT 1,000 INVITATIONS AS 1,000 VALID OBSERVATIONS

Non-completion and invalidity are hidden.

CH06-F04 — RESET COUNTRY TO ZERO WITH PURE ROUNDING

The selection probability of a country with a positive target population is left at zero.

CH06-F05 — GIVING ONE PERSON TO EACH COUNTRY AND CALLING IT THE POPULATION RATIO

The disproportion caused by the minimum quota is hidden.

CH06-F06 — CONSIDERING Country Observer observation AS POPULATION OBSERVATION

The additional observation of a small country receives full weight in the main panel of 1,000 people.

CH06-F07 — ESTABLISHING THE SMALL COUNTRY POOL BASED ON RESULTS

A country producing low scores is moved to another pool or excluded from coverage.

CH06-F08 — PUTTING INCOMPATIBLE COUNTRIES INTO THE SAME POOL

Countries that differ in product access, language, or jurisdiction are combined solely because they are small.

CH06-F09 — OVERWEIGHTING EXCESSIVE SAMPLING

Determines the global result with a sample share higher than the language or country population share.

CH06-F10 — MULTIPLYING MULTILINGUAL INDIVIDUALS

A person is counted as a separate population unit in each language cell.

CH06-F11 — CONSIDERING SAMPLE SHARE AS POPULATION SHARE

Raw response numbers are converted directly into the global score.

CH06-F12 — ARRANGING INVITATION ORDER BY EASE ORDER

Users who respond first and quickly determine the panel.

CH06-F13 — USING STRATUM-OUT RESERVES

Missing country or language observations are quietly completed with participants from another group.

CH06-F14 — LINKING INVALIDITY TO RESPONSE CONTENT

Incorrect AI response is technically declared invalid.

CH06-F15 — REPLACING REFUSALS WITH A NEW USER

System refusal is removed from the denominator even though it is the actual user outcome.

CH06-F16 — HIDING THE WEIGHT FORMULA

The public score does not show with which population and sample weights it was produced.

CH06-F17 — TRIMMING WEIGHTS TO DELETE SMALL GROUPS

Small groups with high weights are trimmed because the result does not look good.

CH06-F18 — HIDING THE EFFECTIVE SAMPLE

Even if 1,000 raw records drop to 500–600 data points due to weights, it is presented as "1,000-person precision".

CH06-F19 — CONSIDERING CLUSTERING AS INDEPENDENCE

Observations from the same panel, institution, or user are considered completely independent.

CH06-F20 — CONSIDERING THE SAME 1,000 PEOPLE AS TEN THOUSAND INDEPENDENT PEOPLE

Ten observations of an AI product are described as if the population were ten times larger.

CH06-F21 — CONSIDERING COMMON SUPPORT AS NATIVE REACH

The common comparison sample is presented as if it were the product's actual global reach population.

CH06-F22 — DIRECTLY RANKING NATIVE REACH RESULTS

Product scores in different target populations are compared as if they were from the same universe.

CH06-F23 — HIDING POPULATION CHANGE IN WAVES

Even though new countries or user groups enter the score, the trend is interpreted only as a change in product quality.

CH06-F24 — PRODUCING DEFINITE SCORES FROM INSUFFICIENT SAMPLES FOR A SUBGROUP

A few dozen observations are published as definite scores for a country or language.

CH06-F25 — SILENTLY CHANGING THE SAMPLING LOCK

Quota, weight, or pool definitions are updated without a version record.

CH06-F26 — BLINDLY MULTIPLYING ELIGIBILITY RATES

Eligible population and quota calculations ignore the characteristics of the dependent population.

CH06-F27 — REPLACING THE MAIN PANEL WITH THE COMPLEMENTARY PANEL

Language fairness or Country Observer observations are used to reduce the main population target of 1,000.

CH06-F28 — DELETING LOW-SCORED CELLS WHEN THE PANEL IS SHORT

Incomplete samples are excluded because they are disadvantageous in terms of results.

CH06-F29 — USING WEIGHTING AS SCIENTIFIC FLAIR

It is claimed that the representation problem is solved only by adding weight without explaining selection and coverage bias.

CH06-F30 — THINKING THAT GEO-1000 SHOULD BE ASKED TO EXACTLY 1,000 PEOPLE

Invitation, complementary panels, invalidity, repetition, and effective sample structure are ignored.

54. AUDIT PROCEDURE

Step 1 — Record the Sample Purpose

Global population estimate is divided into sub-group diagnosis, country scope, language fairness, accessibility, and product comparison purposes.

Step 2 — Validate Target Population Version

The population file and date record in Section 5 are reviewed.

Step 3 — Define the Main Current Observation Target

The default GEO-1000 target or a justified alternative is written.

Step 4 — Establish Strata

Differences in country, language, and material user surface are determined.

Step 5 — Identify Small Country Pools

Each pool's:

- countries,
- population,
- justification for similarity,
- internal selection method

is recorded.

Step 6 — Calculate Raw Quotas

```
qh=nWh
```

The formula or a justified alternative is used.

Step 7 — Apply Integer Assignment

Rounding and tie-breaking rules are recorded.

Step 8 — Add Complementary Oversampling

Country Observer, Language Fairness, and Accessibility goals are defined separately.

Step 9 — Calculate Invite and Reserve Volume

Completion and validity expectations are used.

Step 10 — Freeze Selection and Reserve Order

Participants are selected or ordered before the results are seen.

Step 11 — Calculate Design Weights

Selection probabilities and layer weights are recorded.

Step 12 — Define the Procedure for Staying Silent and Not Responding

A previously announced response is written to the layer gap.

Step 13 — Create the Sample Lock

Document ID, version, date, and integrity record are added.

Step 14 — Review the Post-Field Validity Flow

It is verified that validity decisions are given independently of the accuracy of the responses.

Step 15 — Generate the Final Weights

The design is separated into response and calibration components.

Step 16 — Calculate the Effective Sample Size

The difference between the raw number and the volume of information is reported.

Step 17 — Compare Allocation and Actuals

Planned and actual observation numbers are shown on a cell-by-cell basis.

Step 18 — Review Post-Result Changes

Quota, pool, weight, and scope changes are checked.

55. REQUIRED EVIDENCE

AI-product identity; measurement wave; target-population version; Population Panel target; Country Observer target; Language Fairness target; Accessibility Panel target; stratum list; stratum populations; population weights; raw quotas; rounding method; fractional-allocation and remainder records; tie-breaking rule; small-country pools; within-pool selection probabilities; country and language allocations; plan and interface allocations; oversampling justifications; invitation targets; completion assumptions; validity assumptions; primary and reserve participant lists; selection or randomisation record; nonresponse records; reasons for invalid observations; validity-decision dates; design weights; nonresponse adjustments; calibration coefficients; weight-trimming policy; raw and effective sample sizes; clustering and matched-user records; Native Reach allocation; and Common Support allocation.

Fixed and current frame weights Comparison of planned and actual allocation Sample lock Change log Responsible person or institution

56. AUDIT CHECKLIST

Is the main valid observation target open? Is the target 1,000 invites or valid observations? Was the sample frozen before AI responses were seen? Which population frame was used? On what basis were the strata established? Are raw country and language quotas shown? Is the method for

converting fractions to whole numbers clear? Was the break-off rule predetermined? Is the selection probability for small countries zero? Are small country pools logically consistent? Have Country Observer observations been mixed into the main population weight? Has a country score been produced from a Country Observer observation? If language and accessibility groups were oversampled, were they reweighted? Has the same person been counted as more than one population unit? Are the numbers of invites, completers, and valid observations listed separately?

Have backup lists been created in advance? Has a participant been replaced due to an incorrect or negative AI response? Was the validity decision made before the material accuracy score? Was the tab gap quietly closed from another country? Has the quota or pool changed after the results? Can the design weights be reproduced? Are the assumptions of the nonresponse setting clear? Has weight trimming been done and its effect shown? Are raw and effective sample sizes provided together? Has the same user tested more than one AI product? Has this dependency been preserved in the analysis? Has each AI product reached 1,000 product-observations separately? Were the Native Reach and Common Support samples separated? Did the population frame change between waves?

Have fixed and current framed trends been separated? Have insufficient subgroup samples been published as if they were definitive scores? Is the sample quality status correct? Have complementary panels replaced the main panel? Is there a clear accountable owner of the sample registration?

57. OBJECTIONS AND RESPONSES

Objection 1 — "Why don't we give each country individuals directly according to the population ratio?"

This is a strong assumption in large and medium population countries. In very small countries, the raw quota may fall below one person. In this case:

- small country pool,
- probabilistic country selection,
- separate Observer panel

may be more accurate.

Objection 2 — “Wouldn't giving at least one person to each country be fairer?”

It might be useful in terms of geographic coverage. It is disproportionate in terms of population estimation. The correct solution is to separately label this one person as a Country Observer observation.

Objection 3 — “Doesn't a country receiving zero people on the main panel completely exclude it?”

It can be a problem if a country receives a zero quota as a separate deterministic layer. A positive selection probability can be given in the small country pool. Additionally, the Observer Panel can maintain the observational visibility of the country.

Objection 4 — “Does pooling small countries erase cultural differences?”

It can be deleted if used to generate a country score. For the main global estimate, the pool manages the probability of selection. The country identifier is preserved in the observation record. Separate Observer and country-specific sampling can be done.

Objection 5 — "Why don't we include Observer users within the 1,000?"

They can be included. But then the sample of 1,000 people is no longer proportionate to the pure population. Weighting would be needed, and the effective sample size of the main estimate could decrease. In the founding architecture, 1,000 observations are allocated to the Population Panel; Country Observer observations are added to this.

Objection 6 — "Won't the total field cost become very high?"

It could. The standard should not produce artificial precision by hiding the cost. With a lower budget:

- fewer AI products,
- fewer languages,
- smaller pilot,
- wider uncertainty

can be used. The correct status should be clearly written.

Objection 7 — “Why are we collecting 1,093 or more observations when the name GEO-1000 exists?”

1,000 is the valid observation target for the main population panel. Justice and coverage panels can be added to the main target. This does not compromise the name of the standard. It makes the actual total field volume of the measurement visible.

Objection 8 — “Why would oversampling a language distort the global result?”

An oversampled language appears large because it is in raw responses. When returned to the population weight, the global result is not distorted. The separate language result becomes more precise.

Objection 9 — 'The weights are very complex; isn't the raw average easier to understand?'

The raw mean can be meaningful if it is proportional to the sample population. Excessive sampling, non-response, or coverage differences can cause the raw mean to answer the wrong question. Complexity cannot be hidden if it arises from the sample design.

Objection 10 — “If there are 1,000 valid observations, why should the effective sample be smaller?”

Because some observations can carry very high weight, while others can carry very low weight. Records from the same panel or user may resemble each other. The raw number is the count of physical records. The effective sample approaches the volume of statistical information.

Objection 11 — “Wouldn't it be fairer for the same people to test all AI products?”

Reduces user difference. However:

- it can narrow the product access universe,
- it can increase task load,

it can create contamination between responses. It can be managed with matched or balanced incomplete block design.

Objection 12 — "If a layer does not reach the target, can the global score never be published?"

Not in every case. The deficiency is evaluated in terms of:

- population weight,
- risk,
- the reason for inadequacy,
- remaining coverage

considered. The score:

- with narrowed coverage,
- with wide uncertainty,
- With the warning INSUFFICIENT COVERAGE

it can be published. The deficiency cannot be hidden.

Appeal 13 — "If population data changes, do we have to renew all quotas in every wave?"

Renewal may be required for the current instantaneous score. Fixed starting weights can also be maintained for trend comparison. The two results should be published separately.

Appeal 14 — "If we cannot establish a probabilistic world panel, can NOMOS not be applied?"

It is applicable. However, the sample quality status must be correctly recorded. A panel of volunteers calibrated to the population can be valuable. It cannot be presented as a fully probabilistic sample.

COMMON PROVISION OF CHAPTER 60

The power of GEO-1000 does not arise solely from the number 1,000. Of 1,000 people:

- from which population they came from,
- with what probability they were selected,
- which country and language they represented,
- how many times the same person was counted,
- which observations were considered valid,
- how small countries were managed,
- how over-sampled groups were weighted,
- how many people did not complete the task,
- who was replaced based on the results

A sample's authority arises from these design choices. A defective sample may produce 100,000 responses and still speak about the wrong population. A sound sample may contain fewer responses yet support a more reliable conclusion within clearly stated limits. A large country's population should carry substantial weight in the global human experience, but users must not be fabricated where the product is unavailable. A small country's global weight should remain small, but the country need not be invisible. A language may account for a small share of the global population while still carrying a severe, measurable error. One person may use three languages and five products; that does not create fifteen people. An AI response may be wrong without making the participant invalid. The wrong response is what the audit is trying to measure.

Replacing it with a new user does not clean the measurement. It erases reality. Weighting does not miraculously turn the wrong panel into the world population. However, properly designed oversampling can make subgroups visible without distorting the global estimate. Therefore, NOMOS's sixth measurement law is:

> *A sample should be established not to produce the result, but to observe the result fairly.*

Its seventh law is as follows:

> *The visibility of a small country does not have to enlarge its population weight.*

Its eighth law is as follows:

> *Excess observation does not mean excess population weight.*

The ninth law is as follows:

> *A wrong AI answer is not data defect; if observed under the valid protocol, it is a matter of oversight.*

The tenth law is:

> *The confidence in a thousand observations does not come from the number thousand; it comes from the reproducibility of the selection and weighting.*

Order of Chapter 6 of NOMOS

> *Don't just tell me, 'We asked 1,000 people.'*
>
> *Show from which billion people these 1,000 were selected and with what probability.*
>
> *Don't ignore a country just because it accounts for half a person. / But also, don't claim you measured the whole country just because you sampled one person.*
>
> *Select small countries probabilistically in the appropriate pool. / If visibility is needed, set up a separate observer panel.*
>
> *You can oversample low-resource languages. / But do not make it larger in the global score than its population.*
>
> *Do not multiply a person as much as accounts, languages, sessions, and AI products.*
>
> *Use the backup when the participant does not respond. / Do not change your user when AI gives a wrong answer.*
>
> *Decide on validity without seeing the correctness of your answer.*
>
> *If the quota is missing, explain it. / Do not bring a person from another country to cover the gap.*
>
> *Do not change the share, weight, country pool, or sample target after seeing the result.*
>
> *Show the raw number. / Show the effective sample. / Show who received heavy and who received light weight.*
>
> *Lock the population first. / Then set up the layers. / Then calculate the quotas. / Then select the backups. / Then write the weights. / And only after this, send the first prompt.*

The Chapter's Closing Sentence

> *GEO-1000 is not a method of compiling a thousand answers; it is a protocol for making the representation right of every person in the target population measurable under the probability of selection and correct weighting.*

Normative Core

For each AI product and principal measurement wave, the GEO-1000 Population Panel SHOULD target at least 1,000 valid product-observations unless a different target is explicitly justified. Sample allocation MUST be designed, versioned, and locked before AI responses are observed. Primary allocation MUST be based on the product-eligible target population, not raw resident population, invitation counts, accounts, sessions, or convenience availability. Fractional allocations MUST be converted to integer quotas using a predeclared, reproducible, outcome-blind method. Positive-mass populations MUST NOT be deterministically assigned zero selection probability when the audit claims inference to the full target population. Small populations MAY be represented through justified pooled strata and probability-proportional selection. Country Observer, Language Fairness, and Accessibility observations MAY supplement the main panel, but MUST remain separately labelled and appropriately reweighted. Oversampled groups MUST NOT receive their raw sample share as population weight. Participant replacement MUST be based only on predeclared recruitment or protocol-validity conditions. Incorrect, negative, refused, or critical AI outputs MUST NOT trigger replacement. Persons, accounts, sessions, languages, and AI-product observations MUST remain distinct units. Design weights, nonresponse adjustments, calibration, trimming, clustering, paired observations, and effective sample size MUST be documented. Any post-result change to quotas, strata, pools, weights, exclusions, or sample targets MUST create a new analysis version and preserve the original record.

CHAPTER 07

Country Observer and Language Fairness Panels

METHODOLOGICAL BASES / K04; K06; K07; K20

Chapter Boundary

Section 5 established the following provision:

> *For each AI product, a product-eligible user population must be defined separately.*

Section 6 converted this population into sampling:

> *The main Population Panel should provide predetermined probabilities of selection and weights to people in the target population before results are seen.*

Now we need to solve two different fairness problems. The first problem:

> *Population weighting may make very small countries invisible in the main sample.*

The second problem:

> *The global population average can hide severe representation distortions in low-resource or low-population languages.*

A pure population sample is strong for measuring the average human experience in the world population. However, the same sample:

- cannot observe each country individually,
- cannot provide enough samples for each language,
- cannot produce a country score for small countries,

may not catch rare but substantive errors in low-resource languages. Therefore, GEO-1000 cannot consist of a single panel alone. Alongside the main population panel, two complementary structures are needed:

Country Observer Panel

Canonical panel label:

Country Observer Panel

and:

Language Fairness Panel

Canonical panel label:

Language Fairness Panel

The Country Observer Panel answers the following question:

> *In countries that are underrepresented in the main population panel, is there at least real user observation, access verification, or early risk signal?*

The Language Fairness Panel, on the other hand, answers the following question:

> *Even if the global average appears high, are users in certain languages being represented more incorrectly, incompletely, or riskily?*

These two panels do not replace the main Population Panel. They complement it. Giving a Country Observer observation to a country does not measure that country's score. Over-sampling a language does not make that language larger in the global population. The task of this section is:

- country visibility versus population weight,
- to separate the global average with language fairness,
- to separate population estimation with early warning,
- to separate statistical sufficiency with observation scope

from each other. This section has not yet:

- defined how prompts should be translated in all languages,
- the back-translation and semantic equivalence test,
- the detailed selection mechanism of participants,
- how atomic claims will be scored by adjudicators,
- the final conformity thresholds

exactly. These issues will be further regulated in the following sections. The fundamental question of Section 7 is:

> *How do we measure while producing a global score without artificially inflating small countries and languages, but also without leaving them completely invisible?*

NOMOS Challenge

You set up a 1,000-person population panel for an AI product. As a result of a population-proportional sample:

- hundreds of users from a few large countries,
- dozens of users from some medium-sized countries,
- no users from most small countries

came. Global score:

94

It was calculated as such. Then you said: 'This AI product represents the brand worldwide with 94 per cent accuracy.' However, there is not a single real user observation in 70 small countries. These countries may constitute a small portion of the world's population. Still, the following questions remain unanswered: Is the product really accessible? Does the intent work in the correct language? Does the system confuse the country with another country? Is the local company confused with the global brand? Are the price, service, and legal context wrong? Does the system fail to respond? Are there low-frequency critical errors? The global population score does not answer these questions. Now imagine that you added one person to each country. You have obtained 70 additional observations.

You added each of these countries to the main score with a weight of 1/1,070, making small countries many times larger than their true share of the population. Geographic visibility increased,

but the population estimate was distorted. Then one user in one country received an incorrect response, and you declared: 'This country's GEO score is zero.' One person's response is not a country score. It is a single event observed under specified conditions in that country. Now consider language. The English result is 98 per cent, German 96 per cent and Turkish 91 per cent. A low-resource language has only eight users and a raw result of 50 per cent. You say, 'Eight people are not enough; remove the language from the report.' The global score rises, but the experience of users in that language disappears.

In another approach, you give an additional 200 users to a low-resource language. The raw sample mean decreases. Then you say: "The global score dropped to 84." However, the actual population share of the language may be 2 per cent. Oversampling increases statistical power. It does not increase population weight. The first ruling of this section is as follows:

> *Being observed is not the same as having weight.*

Its second provision states:

> *A single country observation is not a country score.*

Its third provision states:

> *Language fairness is not giving equal weight to all languages in the global score.*

Its fourth provision states:

> *A high global average does not eliminate severe degradation in a particular language.*

Its fifth provision states:

> *Additional observations taken to increase coverage cannot be added to the main score without being reweighted according to the true population.*

1. PURPOSE OF THE CHAPTER

The purpose of this section is to establish complementary panels that increase country and language visibility without disrupting the statistical population representation of the main Population Panel. The section normalises the following distinctions:

- Geographic visibility with population weight
- Country observation with country score
- Single event detection with prevalence estimation
- Main Population Panel with Country Observer Panel
- Country Observer Panel with country-specific validation study
- Country coverage with population coverage
- Number of languages coverage with language population coverage
- Language fairness with equal global weighting of languages
- Language global share with excessive sampling
- Native language and prompt language
- Language and locale
- Language and writing system
- Supported language and language that can generate responses alone
- Prompt translation defect and AI language defect
- Resource shortage and AI representation defect

- Adjudicator insufficiency and product failure
- Legitimate local difference and unjust language inequality
- Global average and sub-group base
- Country access coverage and country representation accuracy
- Observer alert and definitive conformity decision
- Single-wave surveillance cycle with multi-wave geographical coverage

At the end of this section, each GEO-1000 audit should be able to provide clear answers to the following questions:

> *In which countries is there at least one actual observation?*
>
> *In which countries is there only early warning observation, and in which ones are there statistical predictions?*
>
> *Which languages were measured, which ones reached sufficient sampling, and which ones were not tested?*
>
> *If low-resource languages were oversampled, how was this reweighted in the global score?*
>
> *Is the source of a poor result in a language truly AI-generated, or is it due to prompt, resource, locale, or adjudication issues?*

2. CENTRAL NORMATIVE PROVISION

> *GEO-1000 should manage the main population estimate with the Population Panel; the small country observation with the Country Observer Panel; language inequality and low-resource language identification with the Language Fairness Panel separately.*

These three structures:

- differ in purposes,
- differ in sample targets,
- differ in weights,
- differ in public claims

have. An observation from one panel cannot be converted into the meaning of another panel without explanation. The default structure:

Panel	Main purpose	Results it can produce	Results it cannot produce alone
Population Panel	Population-weighted global estimate	Global and layered population score	Reliable country score for every small country
Country Observer Panel	Geographical observation and early warning	Country observation coverage, individual event, access signal	Country population ratio
Language Fairness Panel	Language-based diagnosis and inequality measurement	Language scores, language floor, justice gap	Automatically equal weighting of languages in the global population

3. WHAT IS THE Country Observer PANEL?

Country Observer Panel, in countries that are underrepresented or not represented enough in the main population panel:

- product access,
- prompt applicability,

- basic entity analysis,
- language and locale behaviour,
- early Critical error signal

is a supplementary audit panel set up to observe. The primary task of the Country Observer Panel is not to produce a score.

> *It is to make the invisible country visible for the first time.*

A Country Observer observation can show: "In the specified country, under the specified AI product, user surface, language, and time condition, this result has been observed at least once." It cannot show: "X% of users in this country see the same result."

4. Country Observer UNIT

The basic observation unit of the Country Observer Panel is not the country alone. The preferred unit:

```
S_{c,a,l,s,w}
```

where:

- c: country or jurisdiction
- a: AI product and user surface
- l: language, writing system, and locale
- s: session or plan condition
- w: measurement wave

This distinction is necessary. In the same country:

- more than one language,
- different plans,
- web and mobile surfaces,
- different legal or technical access conditions

may exist. Adding a user to a country does not mean monitoring all language and user surfaces of the country.

5. DUTIES OF THE Country Observer PANEL

5.1. Access Verification

Although the product appears to be available in the country in the official registration, the actual user:

- may not be able to log in,
- may not be able to reach the specified surface,

It may encounter a plan or verification obstacle. The Country Observer observation tests the official access record at the field level.

5.2. Basic Identity Check

AI product supervising entity:

- the right company,
- the right brand,

- correct country relationship

Can it analyse inside?

5.3. Detection of Locale Deviation

In different country locales in the same language:

- price,
- company,
- service,
- law,
- advice

Is it changing? If it is changing, can this difference be financial and explainable?

5.4. Low Frequency Heavy Fault Signal

In the response of a single observer:

- wrong licence,
- wrong legal entity,
- dangerous health or safety information,
- heavy identity fusion

It can be observed. A single event does not indicate the prevalence in the country. However, it can trigger verification work.

5.5. New Research Hypothesis

The observer panel may raise the question: 'Is this event singular, or is it a recurring pattern among users in the country?' This question is tested with a targeted additional sample.

5.6. Geographic Scope Record

It shows in how many suitable countries the product has real user observation. This is not the number of country scores. It is observational coverage.

6. THINGS THE Country Observer PANEL CANNOT DO

Country Observer Panel alone:

- reliable country score,
- country population rate,
- country ranking,
- country conformity decision,
- country-based model superiority

cannot produce. If there is a single observer in a country and the answer is correct: “The country score is 100” cannot be said. If it is wrong: “The country score is zero” cannot be said. Correct status:

OBSERVED_PASS

Observed Issue

OBSERVED_CRITICAL_CANDIDATE

ACCESS_FAILURE

INSUFFICIENT_FOR_COUNTRY_ESTIMATE

are observational records.

7. ELIGIBLE COUNTRY REGISTRY

An Eligible Country Register should be created for each AI product. Each country can hold one of the following statuses:

CE-0 — UNKNOWN

The product and audit access status could not be determined.

CE-1 — OUTSIDE NATIVE REACH

The product remains outside the official or legitimate area of use. It is not taken into account for the country representation accuracy score. It is shown as an access scope gap.

CE-2 — RESTRICTED OR SPECIAL ACCESS

Only:

- institution,
- invitation,
- plan,
- device,
- user type

access is available. A separate panel may be required.

CE-3 — ELIGIBLE FOR Country Observer observation

The product is suitable for Country Observer observation in terms of the relevant user surface and at least one audit language.

CE-4 — ELIGIBLE FOR POPULATION ESTIMATION

It has a product-eligible population share in the country's main population panel.

CE-5 — ELIGIBLE FOR COUNTRY ESTIMATION

A country estimate can be produced under sufficient sample, effective sample, language coverage, and quality conditions. A country can be CE-3 but not CE-5. This means the country is observed but a reliable country score could not be produced.

8. THE LADDER BETWEEN Country Observer observation AND COUNTRY ESTIMATION

CS-0 — NOT OBSERVED

There is no valid Country Observer observation for the eligible country.

CS-1 — SINGLE Country Observer observation

There is one valid country–product–language observation. Generates only an event record.

CS-2 — REPEATED Country Observer observation

There are several observations from different waves or users. May raise a pattern suspicion. Not sufficient for the country rate.

CS-3 — EXPLORATORY COUNTRY SAMPLE

The exploratory sample band in section 6 has been reached. An exploratory result can be given with wide uncertainty.

CS-4 — PROVISIONAL COUNTRY ESTIMATE

A provisional country estimate can be produced with sufficient and balanced sampling.

CS-5 — COUNTRY ESTIMATION CANDIDATE

Sufficient active sample, language/locale coverage, repeat waves, and quality controls are in place. It can be a country score candidate. Exact thresholds should be calibrated with a pilot and external review.

9. OBSERVER CYCLE

Observing every wave in all eligible countries may not be possible in terms of cost and operations. Therefore, the Observer Cycle can be used. Measurement waves:

```
w=1,2,…,K
```

and eligible countries: let c∈CS. Assignment of an observer for country c in wave w:

```
z_{c,w} = 1 (if the country will be observed in the wave); otherwise z_{c,w} = 0
```

can be represented as follows. Basic cycle condition:

```
\sum_{w=1}^K z_{c,w} \geq 1
```

must be satisfied. That is, each eligible country should be observed at least once within the defined surveillance cycle. This provision:

- does not mean that each country will be observed in every wave,
- or that a single observation will constitute the country score

Neither condition follows from the cycle requirement.

10. OBSERVER FREQUENCY CLASSES

Countries can be divided into predefined frequency classes.

SF-0 — OUTSIDE REACH

No product access. No user-representative observation is conducted. Access status is monitored.

SF-1 — ONCE PER SURVEILLANCE CYCLE

Low-risk countries, small countries, or countries that already have sufficient indirect coverage in the Population Panel.

SF-2 — PERIODIC OBSERVER

Observed again at regular intervals.

SF-3 — EVERY PRINCIPAL WAVE

Can be observed in every principal wave for the following reasons:

- high user population,
- material target market,
- previous serious error,
- rapid access or regulatory change,

high-risk service area.

SF-4 — INCIDENT MONITORING

Intensified observation due to verified Critical event or continuous representation disruption. The frequency class cannot be retroactively changed after results are reviewed. A new event can create a new version for the next wave.

11. OBSERVER PRIORITY

The frequency of Country Observer activity should not be determined solely based on population. As relevant, the following areas may be considered:

- Product-eligible population
- The actual service or market coverage of the entity
- Previous erroneous representation records
- High-risk user impact
- Changes in product access
- Language and locale uniqueness
- Low-resource language usage
- Legal or operational differences
- Time elapsed since the last observation

These fields must be used according to the predefined governance rule. The approach of “The result of this country is bad, let's measure less.” is prohibited.

12. Country Observer COVERAGE METRICS

Country coverage is not a single number. At least three separate metrics must be used.

12.1. Country Observation Coverage

The proportion of eligible countries with at least one valid observer or population observation.

```
COC_count = |{c ∈ C_S: n_c^obs > 0}| / |C_S|
```

This metric shows the coverage in terms of the number of countries.

12.2. Population-Weighted Country Observation Coverage

The proportion of the eligible population in countries with at least one valid observation to the total target population.

```
COC_pop = [Σ_{c:n_c^obs>0} N_c] / [Σ_{c∈C_S} N_c]
```

This metric shows whether large populations are observed.

12.3. Country Prediction Coverage

It is the proportion of countries that reach a sufficient country sample.

```
CEC = |{c: n_{c,eff} ≥ m_c}| / |C_S|
```

Here, mc is the predefined adequacy threshold for the country prediction. An audit:

- 100 per cent country observation coverage,
- 99 per cent population observation coverage,
- may carry only 10 per cent country prediction coverage

These results cannot be converted into each other.

13. OBSERVER ALARM

When a material issue is found in a watcher's observation, a Watcher Alarm is created. Types of alarms:

SA-1 — ACCESS ANOMALY

The official access record conflicts with actual user access.

SA-2 — ENTITY ANOMALY

The entity merges with another company, product, or country operation.

SA-3 — LANGUAGE OR LOCALE ANOMALY

Incorrect language, wrong country context, or obvious locale deviation occurs.

SA-4 — MATERIAL FACT ANOMALY

The material price, service, capacity, partnership, or time is incorrect.

SA-5 — CRITICAL INCIDENT CANDIDATE

Health, law, finance, security, licensing, or severe identity error exists.

SA-6 — REFERENCE OR PROMPT ANOMALY

The source of the problem is not the AI response:

- prompt,
- translation,

- official source,
- adjudication

might be at fault. The alarm is not an automatic country failure. It is the beginning of verification and escalation.

14. OBSERVER ESCALATION PROTOCOL

Step 1 — Validity Verification

Is the observation really in accordance with the protocol? Correct country Correct product Correct prompt Correct language First appropriate output Full evidence checked.

Step 2 — Re-coding the Incident

Problem:

- AI product,
- reference record,
- prompt,
- participant,
- access

Which of the sources could it belong to?

Step 3 — Independent Repeat

Before selecting the result, verification is done with predefined new users or sessions.

Step 4 — Targeted Sampling

If the event continues, country-specific sampling may be applied. Sample size:

- event severity,
- expected prevalence,
- country population,
- language diversity

are predefined in advance.

Step 5 — Decision Separation

The following results are kept separate:

- Single verified event
- Repeated pattern
- Country rate
- Critical suitability effect

A single Critical event can affect the conformity review even without an estimate of prevalence.

15. WHAT IS LANGUAGE FAIRNESS?

Language fairness:

> *It is not about all languages receiving exactly the same word-for-word answer.*

It means:

> *In semantically equivalent and materially comparable user questions, apart from real local differences, users should not systematically receive representations that are more incorrect, more incomplete, less evidenced, or riskier due to their language.*

Language fairness does not seek this equality:

```
R_Turkish = R_English = R_Japanese
```

word by word. It searches for this:

> *Comparable representation in terms of identity, material reality, boundary, time, source status, and user suitability.*

16. LANGUAGE DIFFERENCE IS NOT ALWAYS UNFAIRNESS

Language outcomes can legitimately vary for the following reasons: The scope of local services is different. Price and contract conditions vary by country. Legal knowledge depends on the jurisdiction. User intent is different. Prompts are not semantically identical. Official sources carry different content across languages. The AI product does not officially support the relevant language. Each of these differences should be recorded. Explainable local differences should not be labelled as injustice. However, even though the reality is the same, in a low-resource language:

- the company identity is corrupted,
- borders are removed,
- the wrong licence is added,
- the response is constantly given in another language

it may be a language-based representation problem.

17. LANGUAGE UNIT

In language assessment, just the general language name may not be sufficient. The preferred language unit:

```
L=(λ,σ,ℓ,d)
```

and can be considered as. Here:

- λ: language
- σ: writing system
- ℓ: locale or country context
- d: task or domain

For example, the same language:

- in a different writing system,
- in a different country,
- in legal or healthcare field

may require a separate measurement cell. Not all users of a language are a single homogeneous group.

18. LANGUAGE SUPPORT STATUSES

Each AI product can use one of the following statuses for language and user surface:

LS-0 — UNKNOWN

Support status could not be determined.

LS-1 — NOT SUPPORTED

The product does not officially support the language in question. It is not evaluated within the main supported language score. Experimental testing may be conducted.

LS-2 — OBSERVED BUT UNSUPPORTED

The product can generate responses. However, there is no official support or quality guarantee. Reported as a separate experimental result.

LS-3 — LIMITED SUPPORT

Support exists for specific tasks, interfaces, or features.

LS-4 — OFFICIALLY SUPPORTED

The product language generally claims to be supported.

LS-5 — AUDIT-ELIGIBLE

Language:

- official support,
- prompt equivalence,
- local adjudicator,
- reference proficiency,
- valid capture

is fully audit-compliant. The presence of official support alone does not constitute LS-5.

19. LOW-RESOURCE LANGUAGE

The term “low-resource language” does not describe the value of the language or the quality of its speakers. In certain technology and audit contexts, it indicates one or more of the following deficiencies:

- Limited digital content
- Limited localised official source
- Limited evaluation dataset
- Restricted adjudicator access
- Low product support
- Limited terminology resource
- Limited machine translation quality
- Low user panel access

A language can be low-resource in one field and strong in another. For example, while it has strong resources in everyday conversation:

- law,
- medicine,
- specific technical field

may have limited resources. Therefore, language resource level:

- product,
- task,
- time,
- field

should be defined together.

20. TASKS OF THE Language Fairness Panel

20.1. Oversampling Low-Resource Languages

Provides additional users for languages that do not receive enough observations in the main population panel.

20.2. Measuring Language-Based Material Error Distribution

Not only overall accuracy:

- identity,
- scope,
- time,
- source,
- recommendation,
- Critical error

examines the results separately.

20.3. Verifying Response Language

Does the AI respond in alignment with the prompt language? If it switches to another language, what is the reason?

20.4. Examining Language and Locale Distinction

Does the same language carry different local realities in different countries? Does AI manage this difference correctly?

20.5. Examining Source and Official Representation Parity

Do the official records of existence carry the same material reality in all languages? A low AI result may arise from a language gap in the source itself.

20.6. Testing Prompt Equivalence

Do translations truly carry the same user intent?

20.7. Testing Adjudicator Parity

Can adjudicators in different languages apply the same normative standard?

21. FOUR GATES FOR LANGUAGE FAIRNESS

A language outcome should be evaluated through four gates before attributing it to the AI product's language performance.

Gate 1 — Intent Equivalence

Are the prompts semantically comparable? If it fails:

PROMPT_EQUIVALENCE_FAILURE

status is given.

Gate 2 — Reference Parity

Are official and verified reality records sufficient for the relevant language? If it fails:

REFERENCE_LANGUAGE_GAP

status is given.

Gate 3 — Measurement Parity

Are capture, session, interface, and user conditions comparable? If it fails:

MEASUREMENT_PARITY_FAILURE

status is given.

Gate 4 — Adjudication Parity

Are there adjudicators in the native language or with sufficient local expertise and the same evaluation rules? If it fails:

ADJUDICATION_LANGUAGE_GAP

is assigned. Unless the four doors are sufficiently met: "The AI product is worse in this language." cannot be definitively established. The language gap is still a finding of the GEO ecosystem. However, its source must be correctly classified.

22. ESTABLISHMENT OF THE LANGUAGE UNIVERSE

The phrase "all languages" cannot be used without limitation. Each audit must create a Language Register within the Scope. Languages can be selected from the following groups:

22.1. Global Core Language Set

Native languages used by a large human population appropriate for the product.

22.2. Entity Market Language Set

The languages targeted by the audited entity:

- service,
- in terms of content,
- customer,
- contract

and maintenance.

22.3. Product Support Language Set

Languages officially supported by the AI product.

22.4. Low-Resource Observer Language Set

Low-resource languages that are weak in the population panel or carry risk in terms of product behaviour.

22.5. High-Risk Language Set

Languages that have material user impact in the context of health, law, finance, or security.

22.6. Experimental Language Set

Languages that the product does not officially support but are tested for observational or research purposes. These groups may overlap. The reason each language is included should be recorded.

23. LANGUAGE FAIRNESS ALLOCATION

For language l, let the target in the main population panel be nlP. The additional target in the fairness panel is nlF. The total collected language observation:

```
n_l^T = n_l^P + n_l^F
```

is obtained. The total weight of the language in the global population calculation is preserved as Wl. The normalised weight per observation:

```
w_{i,l} = W_l / n_l^T
```

can be considered in this way. Thus, the language is oversampled. However, it does not exceed the actual population share in the global score.

24. CONTRIBUTION OF THE Language Fairness Panel TO THE GLOBAL SCORE

Language fairness observations can be used in two ways.

24.1. Separate Diagnostic Analysis

The fairness panel only produces language scores and error types. The main Population Panel does not change the global score. This method is simpler.

24.2. Combined Reweighted Analysis

Population and Language Fairness observations are used together. Each observation is reweighted according to the actual target language population weight. This method can increase the volume of information. However:

- double counting,
- multilingual user,
- panel dependency,
- weight variability

It should be taken into consideration. The method used must be defined before the results are seen.

25. LANGUAGE COVERAGE METRICS

25.1. Language Observation Coverage

The proportion of languages in the scope that have at least one valid observation.

```
LOC = |{l: n_l^obs > 0}| / |L_scope|
```

25.2. Language Prediction Coverage

The proportion of languages that meet sufficient sample and quality requirements.

```
LEC = |{l: n_{l,eff} ≥ m_l}| / |L_scope|
```

25.3. Population-Weighted Language Coverage

The proportion of the target language population that is adequately observed. Requires a primary audit language or divided person weighting to prevent double-counting of multilingual users.

25.4. Product Support Coverage

Indicates in how many of the covered languages the product receives official support. Language support and representation accuracy are kept separate.

26. LANGUAGE INEQUALITY METRICS

CANDIDATE METRICS — PROPOSED METRICS

The exact metric set should be calibrated through pilot and external review.

26.1. Language Outcome

For each sufficient language cell: θl is calculated.

26.2. Highest–Lowest Difference

```
DG_max = max_l θ_l − min_l θ_l
```

Easy to understand. However, sensitive to small and volatile single cells.

26.3. Robust Percentile Difference

```
DG_{90-10} = Q_{0.90}(θ_l) − Q_{0.10}(θ_l)
```

It is less sensitive to extreme cells.

26.4. Language floor

It is the lowest language result with a sufficient sample.

```
LF_min = min_{l∈L_est} θ_l
```

This value should be provided along with small cell uncertainty.

26.5. Robust Language floor

It is the lower percentile or subgroup average of sufficient languages. Example:

```
LF_10 = Q_{0.10}(θ_l)
```

26.6. Language Critical Error Rate

For each language:

```
CR_l = valid response with verified Critical error / valid language response
```

It is published separately from the global average.

26.7. Unknown Language Rate

```
U_l = UNKNOWN or unsolvable response / relevant language observation
```

A high unknown rate can be as important as a low score.

27. EQUAL LANGUAGE SCORE AND POPULATION-WEIGHTED SCORE

Diagnostic results that weight all languages equally can be produced:

```
G_equal-language = (1/L)Σ_{l=1}^L θ_l
```

This result approximates the question: "Average tested language experience." It does not represent the average human experience in the world population. Population-weighted language result:

```
G_population-language = Σ_l W_lθ_l
```

It preserves the population distribution. The two scores are different. They cannot be used interchangeably. Both can be shown together on the language fairness card.

28. LANGUAGE FAIRNESS IS NOT AN OBLIGATION OF EQUAL OUTCOMES

Having the same score in two languages is not an automatic violation. The difference can be explained for the following reasons:

- Actual local service difference
- Country-specific law
- Official support coverage of the product
- Resource access
- User intent
- Prompt area

However, if the source of the difference cannot be explained or low-resource languages systematically experience:

- incorrect identity,
- incorrect authorisation,
- material deficiency,
- high Critical error

if it carries, there is a language fairness issue. NOMOS defines language fairness as follows:

> *In equivalent material conditions, the language itself should not be the cause of systematic misrepresentation.*

29. LANGUAGE AND COUNTRY INTERSECTION

Language score is not independent of countries. The same language:

- different countries,
- different product access,
- different local resources,
- different law,
- different price

can be used in the context of. Therefore, the country–language cell: (c,l) should be examined separately when necessary. Example:

- Germany × Turkish
- Turkey × Turkish
- United Kingdom × Turkish

It is the same language but may carry different locale and market context. Language fairness cannot be measured solely by the official language of the country.

30th Country Observer AND Language Fairness Panel INTERSECTION

An observation simultaneously:

- small Country Observer observation,
- low-resource language observation

may occur. In this case, the record must contain the following fields:

PRIMARY_PANEL_ROLE

SECONDARY_DIAGNOSTIC_TAGS

population weight Observer status Language fairness status The same observation cannot be fully counted in two separate panels. It is a single observation. It has more than one diagnostic tag.

31. CHANGING PARTICIPANTS IN OBSERVER AND LANGUAGE PANEL

A participant can only be changed due to protocol violation. The following outcomes are not reasons for change:

- Incorrect AI response
- Rejection
- Response in another language
- Critical error
- Entity confusion
- Negative recommendation

These are measurement results. The purpose of the Observer panel is specifically to identify unexpected events. Changing the user because an unexpected event occurred defeats the purpose of the panel.

32. LANGUAGE ADJUDICATION

Adjudication in the Language Fairness Panel must at minimum meet the following conditions:

- Being able to understand the prompt and response naturally or at a high proficiency level
- Knowledge of local terminology
- Being able to evaluate sources of evidence
- Using the same normative evaluation guide
- Being blind to the model and, if possible, brand interests
- Disclosing conflict of interest
- Third adjudicator path in case of disagreement

Language where no adjudicator can be found:

ADJUDICATION_LANGUAGE_GAP

must have the status. It cannot be assumed that all nuances are correct if translated into English by automatic translation. Machine translation can be a helpful tool. It is not the final local meaning verification.

33. SOURCE PARITY

Of the entity on the English site:

- scope,
- price,
- licence,
- limits,
- evidence

exists; if these are not present in another language version, the low AI result may not only be a model problem. Three separate results are needed:

- Entity Language Source Gap
- AI Product Language Representation Gap
- Combined User-Facing Gap

The ultimate problem from the user's perspective is still real. However, the responsibility for correction may vary:

- entity,
- AI provider,

both.

34. CLAIM PARITY

Prompts that are not semantically equivalent produce different answers. Example: In one language: "What does this company do?" In another language: "Is this company trustworthy and recommendable?" If asked, the results cannot be compared. The Prompt Constitution will be detailed in the following section. The provision of this section is as follows:

> *Language model performance differences cannot be attributed to AI language performance without verifying prompt equivalence.*

35. RESPONSE LANGUAGE DEVIATION

The user gives a prompt in Turkish. AI generates an answer in English. This result:

- even if the factual information is correct,
- may affect user experience,
- accessibility,
- and comprehension.

Response language deviation must be recorded separately:

MATCHED_LANGUAGE

PARTIAL_CODE_SWITCH

UNREQUESTED_LANGUAGE_SWITCH

UNUSABLE_LANGUAGE_OUTPUT

UNKNOWN

Code-switching is not always an error. For example, using technical terms in English may be normal. It is material if it disrupts the main usability of the response.

36. WRITING SYSTEM AND TRANSLITERATION

The same language can be used in different writing systems. For an entity's:

- name,
- product name,
- legal identity

it may be confused with another entity during transliteration. The Language Fairness Panel, to the extent relevant:

- writing system,
- transliteration,
- official local name,
- name used by the user

must record their fields. Writing system differences should only be considered formal. If it affects entity resolution, it is material.

37. SYNTHETIC Country Observer CYCLE

SYNTHETIC METHODOLOGY PRESENTATION / The following product, countries, numbers, and results are fictional. In the registry of the synthetic product called Orion AI Native Reach: there are 120 eligible countries. The main Population Panel has produced valid observations in the first wave from only 38 countries due to population weighting.

37.1. First Wave Scope

Number of eligible countries: 120 Countries with at least one observation: 38

```
COC_count = 38/120 = 31.7%
```

Let these 38 countries carry 97% of the eligible population.

```
COC_pop = 97%
```

Result: The number of countries is low in coverage, population coverage is high.

37.2. Four-Wave Observer Cycle

Distribute the remaining 82 countries into four waves using a previously frozen random and risk-based programme. Each country is observed at least once within the cycle. At the end of the fourth wave:

```
COC_count = 120/120 = 100%
```

However, sufficient sampling for country estimation has been reached in only 14 countries.

CEC = 14/120 = 11.7%

Correct public record: Country observation coverage: 100% / Population observation coverage: close to 100% / Country prediction coverage: 11.7% Incorrect public record: “All 120 countries have a reliable country score.”

37.3. Small Country Critical Alarm

In a Country Observer record, the AI response observed in small country J misidentifies the audited entity as a licensed law firm. In reality, the entity provides no legal services. A single observation does not estimate country J's error rate; it establishes a candidate Critical event. Protocol: validate the observation; verify the prompt and reference record; repeat with new independent users; if the event persists, open a targeted country sample; and assess the Critical conformity impact separately. The observed event may be rare; the potential harm need not be small.

38. SYNTHETIC LANGUAGE FAIRNESS DEMONSTRATION

SYNTHETIC SAMPLE — NOT REAL AI OR COMPANY PERFORMANCE

Include five languages in the Synthetic Apple.com Core Identity testing. Main Population Panel allocation:

Language	Population weight	Main panel observation
L1	55%	550
L2	25%	250
L3	10%	100

Language	Population weight	Main panel observation
L4	6%	60
L5	4%	40
Total	100%	1,000

Synthetic language results:

Language	pass rate
L1	95%
L2	92%
L3	88%
L4	65%
L5	58%

Population-weighted global result:

```
0.55(95)+0.25(92)+0.10(88)+0.06(65)+0.04(58)=90.27
```

Global score is high. However, there is severe degradation in two languages.

38.1. Language Fairness Oversampling

Add 140 additional observations to L4, 160 to L5. New numbers of observations:

Language	Total observations
L1	550
L2	250
L3	100
L4	200
L5	200
Total	1,300

The crude mean across all 1,300 observations is lower than the population-weighted result shown above. It is not a global-population score, because L4 and L5 were intentionally oversampled. Once target-population weights are restored, the global result returns to the population-weighted figure above. Oversampling improves sensitivity for L4 and L5 without changing the population-weighted estimate.

38.2. Language Fairness Scorecard

Synthetic result. Population-weighted global score: 90.27. Equal-language score:

```
(95+92+88+65+58)/5=79.6
```

Highest-to-lowest language difference: 37 points; lowest language: 58; mean of the bottom two languages: 61.5; Critical-error rate in lower-resource languages: reported separately; prompt equivalence: assumed verified; reference parity: assumed verified. Correct interpretation:

> *The global population representation appears high; however, language fairness is seriously weak.*

Incorrect interpretation: “Since the global score is 90, GEO is successful in all languages.”

39. SYNTHETIC CASE DISTINGUISHING THE SOURCE OF LANGUAGE DETERIORATION

Let's consider three languages.

Language A

Prompt equivalent Official source complete Local adjudicator sufficient AI response incorrect Result: Candidate for language representation problem of AI product.

Language B

Prompt translation materially different AI response appropriate to prompt Result: Prompt Constitution flaw.

Language C

On the official language page, the old price AI carries over the old price English source is current Result: The entity's language source parity issue and the combined representation gap reaching the user. Even if the scores of the three languages appear the same, the correction owner is different.

40. COUNTRY AND LANGUAGE PANELS PUBLIC RESULT CARD

The candidate public result card must include the following fields:

Country Fields

Native Reach appropriate number of countries Number of countries with at least one observation Country observation coverage Population-weighted observation coverage Country prediction coverage Country Observer cycle time Open Observer alerts OUTSIDE_NATIVE_REACH countries NOT OBSERVED countries Measurement date

Language Fields

Number of languages covered Observed number of languages Number of languages for which predictions can be made Officially supported languages Experimental languages Population-weighted global language result Equal language recognition result Language floor Language difference Critical error rates

REFERENCE_LANGUAGE_GAP

PROMPT_EQUIVALENCE_FAILURE

NOT TESTED languages Language panel version This card makes visible the reality that a single "multilingual" badge cannot carry.

41. MANDATORY NORMATIVE PROVISIONS

CH07-N01

The Country Observer and Language Fairness panels must be defined separately from the main Population Panel in terms of purpose, sample, and weighting.

CH07-N02

A Country Observer observation cannot be used as a country population score.

CH07-N03

For a country, having an observation and being able to produce a country estimate are separate statuses.

CH07-N04

The number of countries covered and the population-weighted country coverage should be reported together.

CH07-N05

Country estimate coverage should be shown separately from observation coverage.

CH07-N06

Each eligible country must have a probability of being observed greater than zero in the defined surveillance cycle; out-of-scope countries should also be shown.

CH07-N07

The surveillance cycle and country frequency classes should be versioned before AI responses are viewed.

CH07-N08

A low-scoring country cannot be removed from the Observer cycle after results are seen.

CH07-N09

A Country Observer alert is not an automatic country failure; it triggers verification and escalation.

CH07-N10

A single Critical Observer event can also initiate a separate conformity review without prevalence estimation.

CH07-N11

Observations from the Country Observer Panel cannot be used in the main Population Panel with raw equal weighting.

CH07-N12

If Country Observer observations are included in the main score, their selection probabilities and weights must be explicitly modelled.

CH07-N13

Language fairness cannot be defined as the obligation to give the same answer word for word.

CH07-N14

Language fairness requires comparable accuracy, limits, source, and user suitability under equivalent material conditions.

CH07-N15

Language and country, language and locale, and language and writing system cannot be used interchangeably.

CH07-N16

The language universe covered and the reason for including each language should be explained before the measurement.

CH07-N17

The claim of 'all languages' can only be used if indeed all relevant language and user surfaces have been measured; otherwise, 'languages within scope' should be said.

CH07-N18

Observations in a language that is not officially supported cannot be added to the main supported product performance without explanation.

CH07-N19

Over-sampling of low-resource languages should not increase the global population weight.

CH07-N20

If observations from the Language Fairness Panel are included in the global result, they should be returned to the actual target language population weights.

CH07-N21

The same multilingual user cannot be counted as more than one full human in the global population weight.

CH07-N22

Without verifying prompt equivalence, the difference in language score cannot be definitively attributed to AI language performance.

CH07-N23

A low language score cannot be loaded onto the AI product alone without evaluating the reference parity.

CH07-N24

A language accuracy decision cannot be finalised without sufficient local adjudicator competence.

CH07-N25

Response language deviation and code switching must be recorded separately.

CH07-N26

Legitimate local reality differences cannot be classified as language injustice.

CH07-N27

Unexplained material language differences cannot be closed solely as personalisation or style differences.

CH07-N28

Global population score cannot replace the result of language fairness.

CH07-N29

Equal language score cannot be presented like a population-weighted global score.

CH07-N30

Language-based Critical error rates must be shown separately from the overall average.

CH07-N31

PROMPT_EQUIVALENCE_FAILURE, REFERENCE_LANGUAGE_GAP, MEASUREMENT_PARITY_FAILURE, and ADJUDICATION_LANGUAGE_GAP must be maintained as separate statuses.

CH07-N32

In Country Observer and Language Fairness panels, the decision to change participants cannot be based on the correctness of the AI response or for the benefit of the brand.

CH07-N33

The same observation can have multiple diagnostic labels; however, it cannot receive more than one full weight in the global prediction without explanation.

CH07-N34

Country and language panel records must have an accountable human or institutional owner.

42. FORMS OF FAILURE

CH07-F01 — SINGLE-PERSON COUNTRY SCORE

A single observer's observation is published as the score GEO for the country.

CH07-F02 — SYMBOLIC COUNTRY COVERAGE

One person is assigned to each country and it is called 'proportional to the world population.'

CH07-F03 — CONFUSING POPULATION COVERAGE WITH NUMBER OF COUNTRIES

Although a small number of large countries are observed, only population coverage is shown; although a large number of small countries are observed, only the number of countries is shown.

CH07-F04 — COUNTING OBSERVATION SCOPE AS ESTIMATION SCOPE

All countries with at least one observation are considered scored countries.

CH07-F05 — COUNTING A Country Observer observation TWICE IN THE MAIN SCORE

The same observation receives full weight in both Observer and Population Panel.

CH07-F06 — CONVERTING A Country Observer alert INTO A CLAIM OF PREVALENCE

A single event is generalised to all users in the country.

CH07-F07 — DISREGARDING A CRITICAL OBSERVER EVENT

The high risk of harm is not examined by saying "only one user".

CH07-F08 — REMOVING A LOW-SCORED COUNTRY FROM THE CYCLE

The observer schedule is changed according to the results.

CH07-F09 — OBSERVING ONLY COMMERCIAL MARKETS

Countries that have access to products but are considered commercially unimportant are completely excluded; The result is presented globally.

CH07-F10 — CONSIDERING THE COUNTRY AND THE LANGUAGE THE SAME

The official language of the country is accepted as the language of all users.

CH07-F11 — ALL LANGUAGES CLAIM

A limited few languages are measured; The result is published as "in all languages".

CH07-F12 — WIPING LOW-WELD TONGUE

Because the sample is small, language is completely excluded from the report.

CH07-F13 — OVERGROWING LOW-RESOURCE LANGUAGE GLOBALLY

Oversampled language enters the global score with its raw sample share.

CH07-F14 — COUNTING THE EQUAL LANGUAGE SCORE AS THE POPULATION SCORE

The diagnostic score that gives equal weight to every language is presented as the average experience of the world population.

CH07-F15 — COUNTING POPULATION SCORE AS LANGUAGE FAIRNESS

The global result, boosted by the weight of large languages, conceals the distortion in small languages.

CH07-F16 — COUNTING PROMPT TRANSLATION ERROR AS AI ERROR

The difference arising from semantically different prompts is attributed to the model.

CH07-F17 — COUNTING SOURCE LANGUAGE GAP AS AI ERROR

Even though there is no localised official information, all responsibility is given to the AI product.

CH07-F18 — DEFENDING AI ERROR WITH RESOURCE LACK

Even though the same and sufficient resources are available, errors in the low-resource language are closed solely as missing content.

CH07-F19 — CONSIDERING ADJUDICATOR LANGUAGE INADEQUACY AS PRODUCT FAILURE

Even if the adjudicator misunderstood, the result is finalised.

CH07-F20 — CONSIDERING MACHINE TRANSLATION AS FINAL ADJUDICATOR

Local meaning loss is not checked.

CH07-F21 — CONSIDERING CODE CHANGES AS AUTOMATIC ERROR

Failures occur with natural technical terms or acceptable language mixing.

CH07-F22 — CONSIDERING UNUSABLE LANGUAGE DEVIATION AS STYLE

Even if the system responds in another language that the user cannot understand, the result is considered successful.

CH07-F23 — IGNORING THE WRITING SYSTEM DIFFERENCE

A transliteration error causing entity confusion is considered formal.

CH07-F24 — COUNTING A MULTILINGUAL USER TWICE

The same person is counted as a separate population unit in all languages.

CH07-F25 — CONSIDERING LEGITIMATE LOCAL DIFFERENCE AS UNFAIR

Even if the price or legal difference is based on actual local conditions, it is declared as language bias.

CH07-F26 — CONSIDERING UNEXPLAINED DIFFERENCE AS PERSONAL

Material inconsistency under equivalent conditions is not questioned.

CH07-F27 — DEFINITE SCORE FROM INSUFFICIENT LANGUAGE CELL

Language ranking is done with a few observations.

CH07-F28 — SETTING THE LANGUAGE PANEL AFTER THE RESULT

Only languages that scored low or high are selected afterward.

CH07-F29 — MEASURING ONLY HIGH-SOURCE LANGUAGES

The result is presented as "multilingual".

CH07-F30 — PRESERVING UNTESTED LANGUAGES

The NOT TESTED field does not appear on the public result card.

43. AUDIT PROCEDURE

Step 1 — Freeze Product-Country Compliance Registry

For each country:

- Native Reach
- product access
- audit language
- user surface

status is recorded.

Step 2 — Extract Main Population Panel Country Coverage

It is determined in which countries the main panel observation exists.

Step 3 — Define the Observer Country Universe

Countries that are insufficient in the main panel and have CE-3 status are identified.

Step 4 — Establish the Observer Cycle

Countries:

- are assigned in terms of frequency class,
- wave,
- language/locale,
- user surface

and maintenance.

Step 5 — Preselect Observer Participants

Main and backup lists are created before seeing the results.

Step 6 — Classify Observer Results

Access, identity, language, material fact, Critical event statuses are assigned.

Step 7 — Apply Alarm and Escalation

By alarm type:

- again,
- targeted sample
- reference review,
- conformity review

is started.

Step 8 — Calculate Country Coverage Metrics

The scope of the number of countries, the scope of the population, and the scope of the country estimate are calculated separately.

Step 9 — Freeze Language Record

For each language:

- for reasons of scope,
- support status,
- writing system,
- locale,
- user universe

is recorded.

Step 10 — Establish Over-Sampling for Language fairness

Additional targets are set for languages that are insufficient on the main population panel.

Step 11 — Check the Four Language Gates

System equivalence Reference parity Measurement parity Adjudication parity is evaluated.

Step 12 — Score Language Observations

For each language:

- truthfulness,
- material deficiency,
- Critical error,
- unknown,
- response language deviation

is recorded.

Step 13 — Separate Global and Diagnostic Weights

The result weighted by population and the result with equal language recognition are calculated separately.

Step 14 — Calculate the Language Inequality

language floor, solid language difference, Critical error difference, unknown rate is calculated.

Step 15 — Examine Country–Language Intersections

Material locale differences or conflicts are identified.

Step 16 — Create Public Result Card

Scope, prediction sufficiency, alarms, and untested areas are made visible.

44. NECESSARY EVIDENCE

AI-product and user-surface record; Native Reach country register; number of eligible countries; country-access status; principal Population Panel country allocations; Country Observer universe; Observer cycle; Observer frequency classes; country–wave assignment matrix; country–language assignments; primary and reserve Observer lists; country-observation results; Observer alerts; escalation records; targeted country samples; country-observation coverage; population-weighted country coverage; country-estimation coverage; in-scope language register; language-support status;

script and locale records; principal language sample counts; Language Fairness oversample counts; language-population weights; multilingual-user records; prompt-equivalence records; reference-parity records; measurement-parity records; local-adjudicator qualification records; and language results.

Language Critical error rates Response language deviations Language inequality metrics Untested languages Equal language and population-weighted results Panel weights Double count checks Public results card Panel versions Responsible person or institution

45. AUDIT CHECKLIST

Has the purpose of the Country Observer Panel been separated from the main population panel? Which countries are included in the Observer? Is the eligibility and access status of the countries up to date? Is there a probability of positive observation in each eligible country's surveillance cycle? Was the surveillance cycle frozen before the results? Has the Country Observer observation been used like a country score? Have country observation coverage and country estimate coverage been separated? Has the population-weighted country coverage been shown separately? Have Country Observer observations been double-counted in the main score? Has the Critical Observer alert been confirmed? Has a single alarm been converted into a prevalence claim? Was targeted sampling opened due to the alarm? Was the universe of languages in scope predefined? Is the claim of 'all languages' consistent with the actual coverage? Have language, writing system, and locale been separated?

Has the product's language support status been recorded? Why were low-resource languages chosen? Were examples of Language fairness reweighted in the global score? Was the same multilingual user double-counted? Has prompt equivalence been verified? Are official sources materially equal across languages? Do the adjudicators know the relevant language at a sufficient level? Has machine translation been used as the final judge? Have response language deviations been recorded? Has legitimate local variation been separated from language injustice? Has the language corpus and language variance been published? Are language-based critical error rates visible? Have REFERENCE_LANGUAGE_GAP and other measurement flaws been separated? Are untested languages visible? Does the public score hide the language fairness issue?

Has the same observation received full weight across multiple panels? Is the accountable owner of the panel records identified?

46. OBJECTIONS AND RESPONSES

Objection 1 — “Isn’t giving one person to every small country still better than giving nothing?”

For early observation and geographic coverage, yes. But this person does not measure the country's proportion. The correct term:

Country Observer observation

must be.

Objection 2 — “If one person does not produce a country score, why are we sending them?”

Because a single observation can:

- reveal access issues,
- clarify entity confusion,

- identify the wrong language,
- and make Critical events

visible for the first time. The Observer panel is an early warning system.

Objection 3 — “Do we have to observe every country in every wave?”

No. A defined surveillance cycle can be used. High-risk or significant countries can be monitored more frequently; small and low-risk countries can be observed alternately. The schedule should be determined before the results.

Objection 4 — "If small countries receive a low weight in the global score, why are we giving it so much attention?"

Because their population share may be low. However:

- law,
- language,
- local user harm,
- geographical integrity

are actual human communities. Low weight should not mean invisibility.

Objection 5 — "Isn't weighting all languages equally the fairest method?"

Equal language weighting can be a valuable diagnostic perspective. It does not represent the average human experience in the world population. The population score and the language fairness score should be published separately.

Objection 6 — “If a small language only slightly affects the global result, why does a low score matter?”

Because the representation problem is real for people who use that language. Also, severe errors in licences, health, or identity are not automatically trivialised by population share.

Objection 7 — “Isn’t it unfair to blame AI if a low-resource language has few resources?”

It may only be unfair to blame AI. Therefore:

- resource parity,
- prompt equivalence,
- adjudicator competence

examined separately. The error that reaches the user is still factually correct. The owner of the correction should be determined correctly.

Objection 8 — “Isn’t it normal to have different answers in the same language in different countries?”

It may be normal if the real local conditions are different. If the same material fact changes differently and without explanation, it may be a problem. Language and locale should be examined together.

Objection 9 — “If AI gives the correct answer in another language, why would it fail?”

If the user cannot understand the response, representation cannot be used. Language deviation should be evaluated separately along with accuracy.

Objection 10 — "Finding 200 users for language scores can be very expensive."

A smaller sample can be used. Result:

- discovery,
- temporary,
- insufficient sample

should be published with this status. Cost does not grant false certainty.

Objection 11 — "If the global score is already high, doesn't a separate language card unnecessarily scare the customer?"

Hiding material language deterioration does not protect the customer. It produces false confidence. A short and clear results card can be used.

Objection 12 — "Does a single Critical Observer event reduce all eligibility?"

The automatic response depends on the type of event. A single event does not indicate national prevalence. However, if it carries severe damage or false authority, it can initiate Critical review and temporary measures. The decision will be defined in Section 15.

COMMON RULE OF SECTION 49

Creating a world score is easy. You fill the world population with large countries. You lose small countries in fractions. You measure high-resource languages. Low-resource languages: you remove by saying "Sample insufficient." Then you publish a single number. This number may be mathematically calculated correctly. But it may not show all the representation flaws of the world. NOMOS does not give equal weight to every country. Because people's population shares are not equal. NOMOS does not ignore small countries either. Because low population does not mean nonexistence. NOMOS does not weight all languages equally in the global score. Because this could distort the average experience of the world population. NOMOS does not make small languages invisible either. Because even if their global weight is low, the representation error for the people using that language can be 100% real.

Country Observer Panel:

> *Shows where we have never looked.*

Language Fairness Panel:

> *Shows who is below the global average.*

Population Panel:

> *Estimates the average experience in the world population.*

When these three work together, the global score becomes both faithful to the population and auditable in terms of justice. A single user in a country is not the whole country. But that user is a real person. A language may be a small part of the global population. But the error in that language is not less wrong. A Critical event may be rare. But harm is not measured solely by frequency. Therefore, NOMOS's seventh measurement law is:

> *The right to visibility is not the same as population weight.*

Its eighth law is as follows:

> *Observing a country is not the same as estimating a country.*

The ninth law is as follows:

> *Language fairness is not making all languages equal; it is ensuring that no one is systematically distanced from material reality because of their language.*

The tenth law is:

> *If the global average makes the language and country that are behind invisible, it alone is not a sufficient GEO result.*

NOMOS's Section 7 Order

> *Do not hide countries you never look at while showing me the world score.*
>
> *Do not send one user to a small country and make that person the whole country.*
>
> *But also do not ignore a small country forever just because its population is small.*
>
> *Separate country observation, country prediction, and country appropriateness.*
>
> *Do not judge the whole country based on a single severe mistake. / But do not ignore human harm by saying "only one answer."*
>
> *Do not base success in major languages on the silence of minor languages.*
>
> *You can sample a low-resource language a lot. / But do not make it appear larger than it is in the world population.*
>
> *Do not make the equal language score a population score; do not make the population score a language fairness score.*
>
> *Do not blame me if the translation is wrong. / If the source is incomplete, also say this. / If the adjudicator doesn't understand the language, do not make a definitive judgement.*
>
> *Do not classify a legitimate local difference as unfairness, or dismiss an unexplained material disparity as mere style.*
>
> *First, observe the countries. / Then state where you can support an estimate. / Then oversample the languages that require diagnostic visibility. / Then return them to their true population weights. / Keep visible the people whom the global average leaves behind.*

The Chapter's Closing Sentence

A genuinely global GEO measurement must show more than what most of the world sees. It must also show who remains unobserved in smaller countries and languages, who is misrepresented and which result is only a single alert.

Normative Core

The GEO-1000 Population Panel, Country Observer Panel and Language Fairness Panel MUST remain distinct in purpose, sampling, weighting and public claims. A Country Observer observation MAY establish that an access condition, response, anomaly or Critical incident occurred in a defined country–product–language context. A Country Observer observation MUST NOT be represented as a country-prevalence estimate or country GEO score. Country-observation coverage, population-weighted observation coverage and country-estimation coverage MUST be reported separately. Every eligible country SHOULD receive a positive probability of observation within a predeclared surveillance cycle unless it falls outside the product's lawful Native Reach. Language fairness MUST evaluate materially comparable representation under semantically equivalent prompts and comparable contexts. It MUST NOT require word-for-word identical responses or disregard legitimate local differences. Oversampled languages MUST be returned to their target-population weights for global population estimates. Population-weighted scores, equal-language diagnostics, language floors, Critical-error rates, unknown rates and untested languages MUST remain separately visible. A language disparity MUST NOT be attributed solely to the AI product until prompt equivalence, reference parity, measurement parity and adjudication parity have been assessed. One observation MAY carry multiple diagnostic panel tags, but MUST NOT receive duplicate full population weight. Critical Observer incidents MAY trigger confirmatory sampling and conformity review without being treated as evidence of country prevalence.

CHAPTER 08

Participant Selection, Weighting and Sampling Bias

METHODOLOGICAL BASES / K04; K05; K06; K07; K08; K22

Chapter Boundary

Chapter 5 defined the target user population. Chapter 6 transformed this population into a sample design distributed across countries, languages, and user surfaces. Chapter 7 established complementary observation systems for small countries and low-resource languages that the main population panel might leave invisible. Now the numbers in the sample plan need to turn into real people:

> *Where will these participants be found, how will they be selected, under what conditions will they be admitted to the panel, and with what biases will they enter the GEO score?*

From an online panel of a research organisation? From an open social media call? From universities? From the client's own user list? From local research partners? Only from people who frequently use AI products? From those accessing the product for the first time? From the brand's customers? From competitors' employees? From paid plan subscribers? How will it be verified that a person actually lives in that country? How many times will the same person be counted if they use multiple accounts? How much will the participant be paid? Will the payment encourage completing the task quickly or submitting a specific outcome? Will the observation be preserved if the participant uploads an incorrect AI response? Or will it be replaced with a new user by saying: 'The task was done incorrectly.'? A thousand valid screenshots may have been collected. But the participants:

- 60% are heavy AI users,
- 20% are technology students,
- 10% are customers of the audited brand,
- 5% are company employees,
- the rest are general users

Which population, then, will the result represent? Can demographic weighting fully correct the difference? Matching age, gender, country and language totals does not necessarily correct unmeasured differences in technological interest, brand awareness, paid-plan access or patterns of AI use. Weighting is valuable, but it cannot erase every defect in participant selection. This chapter applies the earlier requirement—that each audit control be testable by model, date, country, language, query set, repetition count and measurement record—to the selection of actual participants. It also preserves NOMOS's central demand: evidence, boundary, context and time. This chapter:

- participant selection sources,
- the distinction between probabilistic and non-probabilistic panels,

- eligibility and identity verification,
- country, language, and product access control,
- participant uniqueness,
- rewarding and redirection risk,
- panel conditioning,
- task assignment,
- non-response and panel attrition,
- fraudulent or automated participation,
- selection and coverage bias,
- the power and limit of weighting,
- sample quality statuses

defines. This chapter does not yet:

- all technical areas of AI product registry,
- the full translation and version system of prompts,
- The detailed architecture of the NOMOS Capture infrastructure,
- atomic claim adjudication,
- the final score formula

is not fully defined. The key question of Section 8 is:

> *When we collect a thousand valid records, how do we prove which real people these records belong to and which human population they can represent?*

NOMOS Challenge

You purchase 1,000 people from an online research panel. The panel provider tells you: "1,000 participants ready according to your country and age quotas." The participants complete the task. They upload screenshots. Raw results:

> *90 per cent accurate representation.*

In your report, you write: "90% of users worldwide are accurately represented." However, the actual distribution of participants is as follows:

Frequency of AI usage	Target population share	Share in the panel
Low or infrequent usage	50%	10%
Medium usage	30%	30%
Heavy usage	20%	60%

Heavy AI users:

- better prompt habits,
- more paid accounts,
- newer product surfaces,
- personalisation knowledge,
- higher familiarity with technology

can have. Even if the same locked prompt is used:

- account plan,

- interface,
- previous usage,
- product access

may be different. Let the synthetic group pass rates be as follows:

Frequency of AI usage	pass rate
Low or infrequent usage	72%
Medium usage	84%
Heavy usage	96%

The raw result of the panel:

```
0.10(72)+0.30(84)+0.60(96)=90
```

that is:

90 per cent

With target-population weights, however, the result is:

```
0.50(72)+0.30(84)+0.20(96)=80.4
```

The same 1,000 responses therefore produce two different results:

- Raw panel result: 90 per cent
- Target-population result: 80.4 per cent

Now imagine that you have weighted it. Each of the 100 people in the low-usage group receives a high weight. Each of the 600 people in the high-usage group receives a low weight. The raw sample size is still 1,000. However, the effective sample size drops to approximately 349. So, you may have collected 1,000 files. Statistically, you may not have the information volume of 1,000 equal and independent people. Now a more difficult question: In the low-usage group, people who have never been included in the panel:

- low-connected,
- using assistive technology,
- living in a small city,
- not having a specific device

If there are people, can weighting bring them back? No. An observation of a person who is not in the sampling frame at all cannot be produced by weighting. Now the audited company sends connections to its own customers: 'Please participate in the survey.' The company's loyal customers participate at a high rate. People who know competitors but do not use the company participate less. Age and country weights may be correct. Brand loyalty bias may continue. Now another problem: The participant is told: 'You will earn a bonus if you upload a correct and high-quality AI response.' The participant revises the response. They choose the best result. They crop the screenshot. This is no longer the natural response of the AI product. It is the selected response created by the reward system. The first rule of this section is:

> *The number of participants is not the representative power of the participants.*

Its second provision states:

> *Random assignment is not random population selection.*

Its third provision states:

> *Weighting can reduce observed imbalance; it cannot create an unobserved person.*

Its fourth provision states:

> *A participant's inclusion in the panel, retention in the task, or replacement cannot depend on the content of the AI response.*

Its fifth provision states:

> *The reward should be given for completing the protocol honestly, not for choosing a positive or desired response.*

1. PURPOSE OF THE CHAPTER

The purpose of this section is to make the entire selection chain from the target population to actual user observation visible, unbiased, and auditable. The section normatively distinguishes:

- Target population and sampling frame
- Sampling frame versus participant source
- Selected person versus invited person
- The invited person and the person who agrees to participate
- The participant and the person who completes the task
- The person who completes the task and the valid observation
- Human and account
- Human and device
- Human and session
- Probabilistic panel and volunteer panel
- Within-panel random selection and random selection from the population
- Quota compliance and true representation
- Demographic balance and behavioural balance
- Personal familiarity and target population suitability
- Customer, employee, and general user
- Participant reward and outcome incentive
- Task speed and task integrity
- Invalid observation with non-response
- Population change with panel loss
- Fake participant and real low-quality observation
- Data fraud and incorrect AI response
- Bias correction with weighting
- Evidence production with calibration
- Raw sample size vs. effective sample size
- Participant privacy and audit evidence
- General population panel and expert or client panel

At the end of this section, each GEO-1000 audit should be able to answer the following questions:

> *From which source and by what rule were the participants selected?*
>
> *Who had no chance of joining the panel?*
>
> *Who was more likely to participate and why?*
>
> *Which known differences do the weights correct, and which do they not correct?*

> *Was an invalid or fake participation decision made without seeing the content of the AI response?*
>
> *Did the participants' reward affect their behaviour and choice of answers?*

2. CENTRAL NORMATIVE PROVISION

Each GEO-1000 observation must carry a traceable participant selection chain from the target population to a valid analytic record; selection, invitation, acceptance, task assignment, completion, verification, exclusion, and weighting decisions must be independent of the positive or negative content of the AI response. [K04–K06] A participant panel only:

- country,
- does not imply any suitability in terms of age,
- language

It cannot be considered suitable for population representation because it meets the quotas. The panel should also carry the following information to the extent relevant:

- Participant source
- Selection mechanism
- Method of participation in the panel
- Eligibility criteria
- Unique human verification
- Country and language verification
- AI product access verification
- Account and plan information
- Device and interface
- Frequency of AI usage
- Brand familiarity
- Relationship with the audited entity
- Incentive structure
- Task assignment method
- Non-response
- Invalidity
- Repetition and panel conditioning
- Weighting
- Lack of coverage
- Privacy and consent

Measuring some of this information may affect the participant. Therefore, the order of data collection is also part of the protocol.

3. PARTICIPANT SELECTION CHAIN

The chain from the target population to the final observation consists of the following stages:

U→F→S→I→E→A→C→V

Here:

- U: target user population
- F: individuals included in the sampling frame
- S: selected people
- I: invited persons
- E: individuals whose eligibility has been confirmed and who have been included in the study
- A: people assigned to the task
- C: people who complete the task
- V: people producing valid analytical observations

At each transition, people may be lost or new bias may emerge. For example:

- Those in the target population who are not on the panel
- Those on the panel who are not invited
- Those invited who do not accept
- Those who accept but do not complete the task
- Those who complete but produce invalid records

are not the same people. Therefore, only the last 1,000 valid observations cannot be reported. The entire selection funnel must be visible.

4. PARTICIPANT SOURCES

GEO-1000 does not have to be tied to a single participant source. Each source carries its own bias and scope.

4.1. Probability Population Panel

Participants have a known and non-zero probability of being selected from a defined population. Sources:

- household,
- address,
- population or eligible record,
- probability telephone or multi-stage selection

may be. Advantages:

- design weight can be established,
- population inference is stronger,

selection process can be more transparent. Limitations:

- can be expensive on a global scale,
- In some countries, there may not be a suitable frame,
- It may also be necessary to find the subpopulation with access to AI products,

Again, there will be nonresponse. Probabilistic selection is not automatically a perfect representation.

4.2. Professional Volunteer Research Panel

Participants have previously registered voluntarily for the research panel. Advantages:

- it is fast,
- can access multiple countries,
- can provide profile and language knowledge,

repeat measurement can be performed. Limitations:

- participants are prone to research,
- the level of technology use may be high,
- heavy panel users may be present,
- probabilities of entry into the population may be unknown,

the same person may be registered on more than one panel. This panel:

cannot be presented as a probabilistic world population sample

It must not be represented as a probability sample of the global population.

4.3. Open Online Call

Participation is obtained through social media, a website, or an open registration form. Advantages:

- wide visibility,
- low cost,
- access to small countries and language communities

can be provided. Limitations:

- strong self-selection,
- concentration of brand fans or critics,
- fake and repeat participation,
- over-representation of technology enthusiasts

can be created. An open call may be much more suitable than the main Population Panel for:

- discovery,
- Country Observer,
- Language Fairness,
- replication challenge

These sources may be better suited to those purposes.

4.4. University or Institution Affiliated Panel

Participants:

- are selected through universities,
- public institution,
- research centres,
- professional organisations

Advantages:

- local access,
- language expertise,
- ethical oversight,
- reaching low-resource communities

can be provided. Limitations:

- student or professional intensity,
- institution effect,
- age and education bias,
- corporate participation pressure

can create.

4.5. Customer or User List

The actual customers or users of the audited institution are selected. This source:

- customer experience,
- suitability for service,
- brand awareness

It is valuable for measurement. It is not a general population panel. Customers:

- can get to know the brand better,
- may have a more positive or more negative experience,

may be more likely to ask the relevant query. The customer panel should be defined separately.

4.6. Participants Suggested by the Audited Institution

Company:

- employee,
- customer,
- partner,
- follower

can suggest. These users cannot be included in the main independent Population Panel without explanation. Separately:

ENTITY-NOMINATED PANEL

should be labelled. If the institution selects participants, the result is not an independent population measurement.

4.7. Multi-Source Hybrid Panel

It is the combination of multiple sources. Example:

- probability-based national panel,
- professional online panel,
- small country local partner,
- Language fairness community

can be used together. A hybrid panel can increase coverage. However, among the resources:

- repeat individuals,
- selection probabilities,
- incentives,
- task experience,
- user profile

should also be modelled.

5. PARTICIPANT SELECTION INTEGRITY STATUSES

The participant selection quality of a panel can be classified with the following statuses.

PSI-0 — UNKNOWN SELECTION

It is not sufficiently known where and how the participants were selected. A population generalisation cannot be made.

PSI-1 — CONVENIENCE OR OPEN OPT-IN

It is a convenience or open volunteer sample. It can be used for exploration and event detection.

PSI-2 — QUOTA-CONTROLLED OPT-IN

There are country, language, or demographic quotas. The probability of entry into the population is unknown.

PSI-3 — CALIBRATED MULTI-SOURCE PANEL

There are multiple sources, predefined quotas, and population calibration. Unmeasured selection bias may persist.

PSI-4 — PROBABILITY OR STRONG HYBRID DESIGN

There is a known probability of selection or a strong probability–non-probability hybrid design.

PSI-5 — REPLICATED AND EXTERNALLY AUDITED SELECTION

Selection design:

- with different providers,
- in different waves,
- under independent audit

repeated and bias behaviour has been examined. These statuses do not guarantee the accuracy of the score. It indicates the confidence level resulting from participant selection.

6. RANDOM SELECTION WITHIN PANEL AND RANDOM SELECTION FROM THE POPULATION

There may be 100,000 registered users in a volunteer panel. From this panel, 1,000 people can be randomly selected. Selection:

> *is random within the panel.*

However, the panel itself is not randomly formed from the world population. Therefore, one cannot say: "1,000 people randomly from the world population." The correct expression is:

> *"1,000 participants randomly selected within a pre-registered volunteer panel and calibrated to the target population."*

Randomness should be limited to the stage at which it is applied.

7. PARTICIPANT ELIGIBILITY RECORD

Each participant must meet the following conditions to the extent relevant before being included in the main panel:

- Age eligibility

- Volunteer consent
- Country and locale suitability
- Language proficiency
- Access to AI product
- Correct plan or user surface
- Necessary device and technical access
- Unique human status
- Task independence
- Declaration of interest or relationship
- Research security
- Approval for data and evidence collection

Conformity decision cannot be retroactively changed after the AI response is generated.

8. COUNTRY VERIFICATION

The question "Which country does the participant belong to?" does not have a single answer. The following fields should be separated:

- Usual country of residence
- Physical country at the time of measurement
- Country or market region of the AI account
- Payment or plan country
- Citizenship
- Target country of the prompt
- Locale

Citizenship may not be the primary variable for most GEO measurements. What is important most of the time:

> *It is the actual geographic and product condition in which the user experiences the product.*

8.1. Travelling User

The participant may be in a country other than their usual country of residence. This person:

- residing population,
- measurement location,
- travelling user

should hold separate status as.

8.2. VPN or Location Changer

Use of VPN:

- privacy,
- institution network,
- normal user habit

may be the reason. It should not be considered automatic fraud. However, it makes verifying the country layer difficult. There may be a rule in the main controlled country panel to turn off or log

VPN usage. Location changes used to bypass restrictions are not suitable for the main Native Reach panel.

8.3. Country Verification Levels

GV-0 — Unverified

Only self-declaration is available.

GV-1 — Panel-Verified

The panel provider has an existing country registration.

GV-2 — Session-Consistent

Declaration, locale, and session signals are consistent.

GV-3 — Multi-Signal Verified

Privacy-protecting compatible with multiple signals.

GV-4 — Independently Audited

The country verification method has passed external quality control. Collecting an exact street address or sensitive coordinates is not a default requirement. Country verification should be conducted with data minimisation.

9. LANGUAGE PROFICIENCY AND NATURAL USAGE

A participant's knowledge of the language may rely solely on the statement: "Yes, I know." This can be a low-risk initial registration. In a high-security language panel, additionally:

- statement of natural usage,
- brief comprehension check,
- panel history,
- local adjudicator evaluation,
- neutral trial before task

may be used. The language test should not turn into a heavy academic exam that selects only highly educated people. The goal:

> *For the participant to understand the prompt and response as a real user.*

10. VERIFICATION OF AI PRODUCT ACCESS

Participant:

- should be able to access the correct AI product,
- the correct user surface,
- the correct plan or account condition

should be able to access. The following situations should be recorded separately:

- Free plan
- Paid individual plan

- Enterprise plan
- Training Plan
- Web
- Mobile
- Desktop
- Embedded Product
- Visible Model
- Web/retrieval feature

A participant cannot use someone else's account. Password sharing cannot be requested. Proof of product access must be verified without publicly disclosing the account identity.

11. RELATIONSHIP BETWEEN PARTICIPANT AND AUDITED ENTITY

The participant's relationship with the monitored organisation may affect response capture behaviour or brand familiarity. Relationship statuses:

GENERAL_PUBLIC

CURRENT_CUSTOMER

FORMER_CUSTOMER

PROSPECT

EMPLOYEE

FORMER_EMPLOYEE

CONTRACTOR

PARTNER

INVESTOR

COMPETITOR_AFFILIATE

AI_PROVIDER_AFFILIATE

RESEARCHER

UNKNOWN

Direct employee and decision-maker relationships in the main general population panel:

- can be excluded,
- can be put in a separate layer,

can be protected with a low true population weight. Automatically excluding all related individuals may also be wrong. If they are part of the actual customer or employee population, they can be measured separately and explicitly.

12. BRAND FAMILIARITY

Audited participant entity:

- has never heard of it,
- knows the name,
- knows the product,
- is a customer,
- may have a strong positive or negative opinion.

This information:

- can affect participant behaviour,
- natural user panel,
- prompt response

desired. If familiarity is measured before the prompt, it may prepare the user for the brand. Therefore, the order of preference:

- Previously existing panel information
- Separate question after the task
- If necessary, a neutral pre-measurement that does not guide the brand

can be conducted. Familiarity should be visible in the main sample; it should not be selected as a weighting variable after the results are seen.

13. AI USAGE FAMILIARITY

Participants:

- first-time user,
- infrequent user,
- regular user,
- intense or professional user

individuals can be. AI usage frequency:

- product access,
- type of plan,
- interface knowledge,
- personalisation,
- likelihood of completing the task correctly

can be affected. However, selecting only experienced AI users can reduce protocol errors. But it may distort the natural population estimate. The correct solution:

- measure the level of use,
- compare it with the target population,
- stratify or weight if necessary,

and preserve the valid real experience of the inexperienced user.

14. UNIQUE HUMAN RULE

The population unit of GEO-1000 is a human. The same person:

- in multiple panels,
- with multiple email accounts,
- with multiple AI accounts,
- from different devices

can participate. Uniqueness checks must be performed in a privacy-preserving manner. Signals that can be used:

- Anonymous participant identity
- Cryptographic or hash-based rechecking
- Device and session compliance
- Payment or incentive record
- Time and task pattern
- Same screenshot or response file
- Secure rechecking across multiple panel providers

No single signal needs to be sufficient for automatic exclusion. In particular:

- Families sharing a common device,
- institution network,
- Accessibility tools,
- low-income shared device users

may be mistakenly considered fake. Suspicious records should undergo human review and an appeal process.

15. PARTICIPANT REWARD

The payment or reward given to the participant may influence their data behaviour.

15.1. Payment Independent of Outcome

Preferred method:

> *A fixed and pre-disclosed payment for completing the task according to protocol.*

Payment:

- cannot depend on the answer being positive,
- can be forced into the categories of

  true,
- long,
- favourable to the brand,
- or accompanied by citations.

Selection must remain independent of all such outcomes.

15.2. Quality Bonus

If a quality bonus is to be used, only:

- complete evidence,
- correct prompt,
- correct product,
- timely delivery,
- personal data cleaning

can be tied to protocol conditions. It cannot be tied to the content of the AI response.

15.3. Speed Incentive

Bonus solely for speed:

- rush,
- incomplete capture,
- wrong product,
- cutting the answer

may increase the risk. A reasonable time should be allowed within the field window.

15.4. Payment by Country

Payment may vary according to local fees and living conditions. This situation:

- may affect participation likelihood,
- panel composition

and should be recorded along with the payment approach and country differences.

15.5. Coercion and Excessive Incentive

The reward should not create pressure that the person cannot reasonably refuse. In particular:

- low-income,
- student,
- dependent employees,
- vulnerable groups

should have their voluntariness protected.

16. INVITATION TEXT AND PRE-ORIENTATION

When inviting a participant, the purpose of the research may prepare them for a specific outcome. A wrong invitation: “We are testing how incorrectly AI systems interpret Apple.” This statement may lead the participant to look for errors. Another wrong invitation: “Help us prove the AI success of our brand.” This statement creates an expectation of a positive result. Preferred neutral invitation: “Study of standardised user experience and representation observation in specific AI products.” The audited entity may become visible later due to the nature of the task. However, the desired outcome should not be disclosed.

17. TASK ASSIGNMENT

Participants:

- AI product,
- prompt,
- language,
- wave,
- session condition,
- task order

should be assigned in advance using a predefined method. The participant's:

- favorite AI product,
- the plan he/she knows best,
- the language with a high probability of receiving a positive response

if allowed to choose by themselves, assignment bias may occur.

17.1. Random Assignment Within Layer

After country, language, and product access are verified, tasks can be assigned randomly within the layer.

17.2. Block Assignment

Blocks can be used to balance free/paid plans, mobile/web, or AI usage frequency.

17.3. Product Order

If a person is testing more than one product, the order should be randomised. The first product's response may affect the evaluation of the next response or task behaviour.

17.4. Order of Prompts

If a participant receives more than one Prompt:

- learning,
- acquiring new information about the brand,
- carrying the previous answer to the next question

risk arises. The order of prompts should be recorded and balanced if necessary.

18. PANEL CONDITIONING

Users who frequently participate in research panels may learn the expectations of the task. A person participating repeatedly in the same test:

- may remember the correct answer,
- may choose the screen capture more carefully,
- may review the site of the audited entity in advance,

which errors are being searched for can be learned. This situation is called:

Panel conditioning

Panel conditioning can change the real natural user experience.

18.1. Sources of Conditioning

Previous GEO waves Re-measurement of the same entity Same prompt bank Peer review training Task sharing in participant communities Previous payments and quality feedback

18.2. Conditioning Controls

New and repeat participant status First participation date Previous exposure to the same entity Previous exposure to the same prompt family Rotating panel Task order randomness Specified rest period New user subsample Comparison of conditioned and new user results Reusing the same participant is not prohibited. It is valuable for longitudinal measurement. Cannot be presented as a natural cross-sectional panel.

19. TASK CONTAMINATION AND PARTICIPANT COMMUNICATION

Participants:

- prompts,
- responses,
- screenshots,
- expected results

can be shared with each other. This situation may disrupt independent observations. Controls:

- Limit wave duration
- Randomising tasks
- Obligation not to share results with participants
- Looking for repeated answers and visuals
- Examining unusual text and time similarities
- Monitoring community or panel provider warnings

Participant communication is not automatically fraud. Its material effect should be examined.

20. TYPES OF NON-RESPONSE

Non-response is not a single event.

20.1. Not Responding to Invitations

The participant does not respond at all to the research invitation.

20.2. Declining Participation

The person sees the invitation and chooses not to participate.

20.3. Not Starting the Task After Qualification

The person is found suitable but does not start the task.

20.4. Task Interruption

The participant abandons the task halfway.

20.5. Not Uploading Evidence

An AI response may have been obtained, but the required record is not uploaded.

20.6. Invalid Completion

The task is completed, but:

- wrong prompt,
- wrong product,
- selected answer,
- missing screenshot

are reasons for invalidity. The ratios of these stages should be reported separately.

21. RESPONSE FUNNEL

For layer h:

- nhsel: selected person
- nhinv: invited person
- nhcon: person who agreed to participate
- nhsta: person who started the task
- nhcom: person who completed
- nhval: valid observation

Let's be. Basic rates:

```
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
```

It can be shown as. Each of these rates indicates a different problem. Giving only the final valid count hides field bias.

22. NONRESPONSE BIAS

People who do not respond may be different from those who do. Example: A user with low internet cannot complete a task. A free plan user cannot access the product. An inexperienced user finds the instructions difficult. A heavy AI user completes faster. In certain countries, payment remains low. A user with a negative brand experience may be more willing to participate. Even if the non-response rate is high, bias may be low. Even if the non-response rate is low, a small missing group may be financially significant. Therefore, not just the rate alone:

Which characteristics differ between respondents and non-respondents

should be examined.

23. NON-RESPONSE WEIGHTS

If a non-response setting will be used, classes should be determined before results are seen. Potential variables:

- Country
- Language
- Age
- Frequency of AI usage
- Plan
- Interface

- Panel source
- Accessibility
- Brand familiarity

Nonresponse coefficient:

```
a_r^{NR} = (Σ_{j∈S_r} d_j) / (Σ_{i∈R_r} d_i)
```

can be calculated as. However, this adjustment only relies on the assumption within the same class: that respondents sufficiently represent non-respondents. It does not resolve unmeasured differences.

24. MISSING DATA MECHANISMS

Missingness can conceptually be considered in three classes.

24.1. Random Missingness

Missingness is not materially related to the outcome and user characteristics. In reality, it may be rare.

24.2. Missingness Explained by Observed Features

Deficiency:

- country,
- does not imply any suitability in terms of age,
- plan,
- language

is related to well-known areas such as. Weighting can partially help.

24.3. Deficiency Related to Unobserved Features

For example:

- Distrust in the AI product,
- low digital skills,
- brand opposition,
- privacy concern

if not measured, weighting cannot completely correct the deficiency. In this case, sensitivity analysis is needed.

25. PANEL LOSS AND INTER-WAVE SEPARATION

In longitudinal or rotating panels where the same users are measured again, some individuals do not participate in the subsequent wave. This is panel loss. Those remaining in the panel:

- more disciplined,
- more reward-motivated,
- more experienced,
- more accustomed to research

may be. The change in score over time can partially stem from the panel composition.

25.1. Panel Attrition Record

In each wave:

- re-invited,
- participated,
- dropped out,
- newly added

user numbers should be shown.

25.2. Panel Attrition Weight

The remaining participants can be weighted according to the probability of dropout. This method does not completely resolve the bias from unobserved panel attrition.

25.3. Refresh Panel

New users can be added to replace the lost users. Results for new and returning users should be reported separately.

26. PARTICIPANT FRAUD AND AUTOMATION

There is a risk of fraud when GEO-1000 is an incentivised and global panel. Possible behaviours:

- A person participating with multiple accounts
- Use of automatic bots
- Uploading someone else's screenshot
- Editing images
- Uploading the same answer on behalf of different users
- Declaring the wrong country or language
- Sharing an account
- Having someone else complete the task
- Rejoining among panel providers
- Creating bulk task farms

This risk does not allow all low-income or fast participants to be considered suspicious. Fraud detection:

- debates,
- multiple signals,
- proportional human review

should be based on.

27. FRAUD AND INTEGRITY STATUSES

FI-0 — NOT ASSESSED

The record has not been evaluated in terms of integrity.

FI-1 — NO MATERIAL ANOMALY

No clear signs of fraud have been found. This is not an absolute guarantee of accuracy.

FI-2 — REVIEW REQUIRED

There is one or more suspicious signal(s).

FI-3 — DUPLICATE OR COLLUSION SUSPECTED

There is suspicion of repeated person, visual, response, or coordination.

FI-4 — PROTOCOL MANIPULATION CONFIRMED

The participant has deliberately altered the protocol.

FI-5 — FABRICATED OBSERVATION CONFIRMED

The observation is not a genuine AI user experience. FI-2 is not automatic exclusion. Review required.

28. FRAUD SIGNALS

Same screenshot hash Same rare file footprint Impossible completion time Multiple identity repetitions from the same device or session Material conflict in country, locale, and product access Mismatch between response text and image Trace of image editing Same person in primary and backup list Same anonymous person across multiple panel providers Serial and mechanical completion of tasks Same unusual error pattern A single signal may not be conclusive proof of fraud.

29. FAIRNESS IN FRAUD DETECTION

The following users may be mistakenly marked as fraudulent:

- Families sharing a common device
- Employees using an institutional network
- University laboratory participants
- Privacy-conscious individuals using VPNs
- Disabled users using screen readers or automation helpers
- Those uploading similar compressed images due to low bandwidth

Therefore:

- automatic tagging,
- human review,
- participant description if necessary,
- appeal

must work together.

30. PARTICIPANT VALIDITY SHOULD BE SEPARATED FROM AI RESPONSE ACCURACY

A participant's record may be protocol-compliant. The AI response may be incorrect. This is a valid observation. Another participant's AI response may be correct. However, the user:

- has refreshed the answer three times,

- only uploaded the best answer,
- may have changed the prompt

This is an invalid observation. Therefore:

> *The correct AI answer does not mean valid data. / An incorrect AI answer does not mean invalid data.*

Validity assessment should be done before substantive scoring and as blind as possible.

31. CONFLICT OF INTEREST AND PARTICIPANT SOURCE

A participant or panel provider may benefit from the following outcomes:

- The audited company receiving a high score
- The AI product appearing superior
- A competitor receiving a low score
- Selling more audit engagements
- Award or employment relationship

Therefore:

- participant relations,
- panel provider agreement,
- result-independent wage,
- Customer intervention

should be explained. Audited institution:

- which users to invite,
- who will be considered invalid,
- Who will be replaced with a backup

cannot decide.

32. PANEL PROVIDER EFFECT

Users of two panel providers in the same country may be different. A provider:

- Technology enthusiasts,
- students,
- intensive research participants

may have a higher rate in terms of care. The panel provider should be kept as a cluster or source variable in the analysis. If possible:

- provider-based results,
- source effect,
- provider comparison within the same country

should be examined. A single-provider global panel cannot be presented as the world population itself.

33. REPEAT INDIVIDUAL IN MULTI-SOURCE PANEL

The same individual:

- two professional panels,
- in an open call,
- in the university list

can be found. A multi-source panel can increase the number of records without increasing uniqueness. Cross-source duplication check:

- privacy-preserving,
- not requiring central identity disclosure,
- allowing objection to mismatches

should be established in a manner.

34. SAMPLE BIAS MAP

GEO-1000 should separately monitor the following types of bias.

SB-1 — Coverage Bias

Some segments of the target population are not in the sampling frame.

SB-2 — Selection Bias

The probability of individuals in the frame being selected for the panel does not conform to the intended design.

SB-3 — Self-Selection Bias

Volunteers for the study are systematically different from the population.

SB-4 — Invitation and Guidance Bias

The invitation text or source attracts people with a specific viewpoint.

SB-5 — Nonresponse Bias

People who do not participate or do not complete are different in terms of outcomes.

SB-6 — Panel Attrition Bias

Users who remain between waves are different from those who drop out.

SB-7 — Incentive Bias

Payment or reward affects the participant's speed, selection behaviour, or response.

SB-8 — Panel Conditioning

The participant learns the expected behaviour from previous studies.

SB-9 — Task Assignment Bias

Participants choose the product, language, or prompts themselves or are assigned unevenly.

SB-10 — Country and Language Classification Bias

The participant is assigned to the wrong country or language cell.

SB-11 — Product, Plan, and Interface Bias

Only certain plan or device users determine the panel.

SB-12 — Brand Familiarity Bias

Customers, fans, critics, or employees are overrepresented.

SB-13 — Fraud and Repeat Bias

The same person or automatic participation gains too much influence on the result.

SB-14 — Validity Exclusion Bias

Negative AI responses are more often considered invalid.

SB-15 — Weighting Bias

Post-result selected variables or excessive weight model score change in the desired direction.

SB-16 — Surviving Sample Bias

Only users who can successfully complete the task are presented like the real product population.

35. PURPOSE OF WEIGHTING

Weighting can serve the following purposes:

- Correcting unequal selection probabilities
- Returning oversampling to the actual population proportion
- Reducing differences in response rates
- Calibrating the sample to known target totals
- Combining multiple panel sources

Weighting cannot serve the following purposes:

- Creating a person who does not exist in the sampling frame
- Correcting unmeasured variables precisely
- Making a false or erroneous observation real
- Producing the desired score after the result
- Converting a single small cell into a reliable country score
- Automatically turning a non-probability panel into a probability sample

36. HEAVYWEIGHT ARCHITECTURE

The final weight can be shown as follows:

```
w_i = d_i × a_i^{NR} × g_i^{CAL} × p_i^{SRC}
```

Here:

- di: basic design or starting weight
- aiNR: response setting

- giCAL: population calibration
- piSRC: panel source or improbable sample adjustment

In the probabilistic panel:

```
d_i = 1/π_i
```

It is possible. The true πi is unknown in the non-probability panel. In this case, the weight used is not the design weight, but the model-based or calibration weight. This distinction should be preserved in the report.

37. WEIGHTS CANNOT DEPEND ON THE RESULT

The following variables cannot be used to create weights:

- Whether the AI response is correct or not
- Whether the brand is mentioned or not
- Whether there is a citation
- Whether a recommendation is made
- Whether a critical error occurs
- Whether the audited organisation likes the result

Weight variables:

- Before the result is seen,
- Because it is related to the target population and the selection mechanism

It should be selected. If the response result determines the weight, the score shapes itself.

38. CALIBRATION VARIABLES

Possible calibration areas:

- Country
- Region
- Language
- Age
- Gender, if necessary and ethical
- Education
- Urban/rural settlement
- Type of internet access
- Device
- Frequency of AI usage
- Free/paid plan
- Interface
- Accessibility status

Reliable target totals may not be available for each variable. Too many calibration variables:

- can create excessive weight variability,
- low effective sample size,
- model overfitting

It should be frozen before seeing calibration set results.

39. WEIGHTING BALANCE

The sample before and after weighting should be compared with the target population. For each field:

- target share
- raw sample share
- weighted share
- remaining difference

should be shown. Weighting is applied only to the number of results and the balance table cannot be saved.

40. WEIGHT DIAGNOSES

The main report should include the following fields as far as they are relevant:

- Minimum weight
- Maximum weight
- Median weight
- Weight variation coefficient
- Weight ratio
- Number of trimmed weights
- Raw sample size
- Effective sample size
- Design effect
- Unweighted result
- Main weighted result
- Alternative reasonable weight results

Weighted single number is not a full control record without weight identifiers.

41. SYNTHETIC WEIGHTING DISPLAY

SYNTHETIC METHODOLOGY DISPLAY / The participants, rates, and scores below are hypothetical. Target population and panel distribution:

Frequency of AI usage	Target population share	Raw panel share	Raw observation
Low or infrequent usage	50%	10%	100
Medium usage	30%	30%	300
Heavy usage	20%	60%	600
Total	100%	100%	1,000

Synthetic pass rates:

Group	pass rate	Previous response
Low or infrequent usage	72%	72

Group	pass rate	Previous response
Medium usage	84%	252
Heavy usage	96%	576
Total	90%	900

Raw result:

```
900/1000=90%
```

41.1. Calibration Coefficients

Low usage:

```
0.50/0.10=5
```

Medium usage:

```
0.30/0.30=1
```

High usage:

```
0.20/0.60 = 1/3
```

Weighted response:

```
72(5) + 252(1) + 576(1/3) = 804
```

Weight total:

```
100(5) + 300(1) + 600(1/3) = 1,000
```

Weighted result:

```
804/1000=80.4%
```

41.2. Effective Sample Size

```
n_eff = [100(5)+300(1)+600(1/3)]² / [100(5²)+300(1²)+600(1/3)²] ≈ 349
```

Result card:

- Raw valid observation: 1,000
- Raw score: 90 per cent
- Weighted score: 80.4 per cent
- Effective sample: approximately 349
- Main reason: overrepresentation of heavy AI users

Correct interpretation:

> *The panel contains 1,000 valid observations; however, as a result of target population calibration, the estimate has dropped to 80.4%, and due to weight variability, the effective sample size has become approximately 349.*

Incorrect comment: “Global accuracy has been proven to be 90% with 1,000 people.”

42. WHAT CAN'T WEIGHTING CORRECT?

The following groups should not be present in the synthetic panel at all:

- People who use screen readers
- Low bandwidth rural users
- People using the AI product for the first time
- Small language community

Weighting only changes the weight of the users in the panel. The unobserved group's:

- AI your answer,
- reveal access issues,
- error distribution

cannot produce. Therefore:

> *Zero coverage means zero information.*

Weighted result: it may be a prediction closer to the target population, not a full observation of the target population.

43. SENSITIVITY ANALYSIS

The dependence of the weighted result on a single weight model should be examined. Candidate results:

- Raw unweighted result
- Design-focused result
- Demographically calibrated result
- Calibration with AI usage added
- Weight trimmed result
- Panel provider effect adjusted result

If the results change materially:

Weight Sensitivity Warning

should be given. A single "correct" weight result cannot be presented as an absolute truth.

44. PROPENSITY ADJUSTMENT IN NONPROBABILITY PANEL

The tendency of nonprobability panel members to enter the panel can be modelled according to the target population. The probability of a person being in the panel:

```
p̂_i = Pr(i ∈ panel | X_i)
```

can be estimated. The candidate inverse propensity weight:

```
w_i^P = 1/p̂_i
```

can be used. This method:

- uses only the measured Xi variables,
- is affected by a misspecified model,
- unobserved selection does not resolve,

requires strong target reference data. Therefore, the result:

Should be labelled as:

Model-calibrated non-probability panel estimate

45. SMALL AREA AND MODEL-BASED ESTIMATES

Direct sampling may be insufficient in some country or language cells. Model-based methods can borrow information from neighbouring or similar groups. This means that the result:

- directly observed country score,
- raw user rate

is not. The correct status:

MODEL-ASSISTED ESTIMATE

SMALL-AREA ESTIMATE

NOT DIRECTLY OBSERVED

must. Model-based prediction does not turn a country without Country Observer observation into an observed country.

46. PARTICIPANT PRIVACY AND VERIFIABILITY

GEO-1000 requires proof for each observation. This does not mean that the participant's identity will be disclosed publicly. A screenshot may include the following information:

- Name
- Email
- Profile picture
- Previous chat
- Account plan
- Institution name
- Sensitive personal information

NOMOS Capture or equivalent system:

- should not capture unnecessary personal fields,
- should perform automatic or controlled redaction,
- raw evidence must be kept in a secure repository,

only anonymous and verifiable records should be presented to the public.

46.1. Public Evidence Manifest

Public record:

- observer ID,
- country/language cell,
- AI product,
- time,
- evidence hash,
- verification status

may be carried. Full screenshot:

- edited,

- restricted access,
- open to independent auditor

may be. The existence of a thousand pieces of evidence does not require the public disclosure of the identities of a thousand people.

46.2. Sensitive Small Country Risk

If there is only one participant in the small country:

- device,
- time,
- language,
- plan

When gathered together, it may become easier to re-identify the person. The level of detail in public reports should be reduced. The integrity of the evidence can be maintained in an access-controlled system.

47. PARTICIPANT CONSENT

The participant should at least know the following:

- Which task they will perform
- Which data and screenshots will be collected
- Which personal information will not be collected
- How long the data will be stored
- Who will have access
- How the result will be published
- How the reward will be given
- That participation is voluntary
- Right to withdraw or object
- Account password will not be requested
- It is not expected to violate AI product terms

Consent should not be hidden in a long and incomprehensible text. [K06; K22]

48. PARTICIPANT SAFETY

From the participant:

- to provide personal health, legal, or financial information to AI,
- to use real customer data,
- to share employer confidential information,
- to weaken account security,
- to exceed usage limits,
- to violate the law or product terms

cannot be requested. Test prompts should carry as much as possible a public and non-personal context.

49. THE PARTICIPANT CARRIES HUMAN DIVERSITY, NOT QUALITY

Inexperienced user:

- can complete the task more slowly,
- more difficult,
- with more support

This person is not automatically a low-quality participant. They are a real part of the target population. If the research design is so complex that only people with high digital skills can complete it: the problem may not be with the participant, but with the measurement tool. The protocol:

- should be understandable,
- accessible,
- linguistically natural,
- technical load limited

must be.

50. PUBLIC RESULT CARD OF PARTICIPANT SELECTION

The candidate public report should include the following areas:

Panel Resources

Probability panel share Professional volunteer panel share Open call share University/institution share Client or affiliated panel share

Participation Funnel

Person in frame Selected Invited Accepted Started Completed Producing valid observation

Eligibility and Integrity

Country verification levels Language verification levels Repeat person rate Suspicious registration rate Verified fraud rate Invalid observation rate

Panel Composition

AI usage frequency Plan and interface Brand familiarity Customer/employee relationship New and returning panel users Accessibility coverage

Weight Diagnoses

The following may be published without revealing participant identities: raw and weighted scores; minimum and maximum weights; the coefficient of variation of weights; effective sample size; sensitivity to weighting; and coverage warnings.

51. MANDATORY NORMATIVE PROVISIONS

CH08-N01

Each valid GEO-1000 observation must carry a traceable participant selection chain from the target population to the analytic record.

CH08-N02

Random selection within the panel cannot be presented as random selection from the target world population.

CH08-N03

The source of participants, selection method, and panel status should be explained in the public method record.

CH08-N04

It cannot be defined as an improbable or voluntary panel probability sample of the population.

CH08-N05

Participants proposed by the audited institution cannot be included in the main independent population panel without explanation.

CH08-N06

Customer, employee, partner, competitor, and AI provider relationships should be recorded separately.

CH08-N07

Participant eligibility criteria must be verified before the AI response is seen.

CH08-N08

Country, language, product, plan, and interface eligibility must be determined before the related task assignment.

CH08-N09

The same person cannot be duplicated as a population unit due to multiple accounts, devices, languages, panels, or AI products.

CH08-N10

Unique person verification must be conducted according to data minimisation and privacy principles.

CH08-N11

A single technical signal alone cannot be considered sufficient for an automatic definite decision that a participant is fake.

CH08-N12

A participant's reward cannot be tied to the AI response being positive, correct, long, containing references, or beneficial to the brand.

CH08-N13

The quality bonus can only be tied to pre-specified protocol integrity.

CH08-N14

The invitation and information text cannot direct the participant to seek a positive or negative outcome.

CH08-N15

AI product should be assigned or randomised before seeing the prompt and task sequence result.

CH08-N16

Product and prompt selection based on participant preference cannot be used unexplainedly in the main population panel.

CH08-N17

Panel conditioning, previous test exposure, and repeat participation must be recorded.

CH08-N18

The results of new and conditioned participants should be analysed separately when necessary.

CH08-N19

A participant's inclusion in the panel, retention, exclusion, or replacement with a backup cannot be based on the content of the AI response.

CH08-N20

Incorrect, negative, rejection, or Critical AI output should be preserved under a valid protocol.

CH08-N21

Participant validity should be evaluated blind as much as possible and before assessing the substantive accuracy of the response.

CH08-N22

Invitation, acceptance, initiation, completion, and valid observation rates should be reported separately.

CH08-N23

The response setting should only be based on predefined and relevant variables.

CH08-N24

Weighting cannot make users who are not present in the sampling frame represented.

CH08-N25

Weight variables cannot depend on the AI response result, the score, or the commercial interest of the regulated institution.

CH08-N26

Non-probability panel weights cannot be presented as design weights.

CH08-N27

Calibration variables and target totals must be released before the results are seen.

CH08-N28

Unweighted and main weighted results must be stored together.

CH08-N29

Weighting diagnoses and effective sample size should be visible in the main report.

CH08-N30

If weight trimming is used, the untrimmed result and sensitivity analysis must be retained.

CH08-N31

Panel provider and participant source effects should be kept as separate variables.

CH08-N32

In a multi-source panel, inter-source repeated person checks should be conducted.

CH08-N33

Fraud detection should include multiple signals, proportional human review, and an appeal process.

CH08-N34

Use of VPN, shared device, institutional network, or assistive technology should not automatically be considered fraud.

CH08-N35

Model-based small area estimation cannot be presented directly like an observed country or language score.

CH08-N36

Participant screenshots and evidence must carry personal data minimisation, redaction, and access control.

CH08-N37

Account passwords, private chat history, or unnecessary sensitive information cannot be requested from participants.

CH08-N38

Participants cannot be asked to bypass product, age, identity, country, or account restrictions.

CH08-N39

Children or users requiring special protection cannot be enrolled in the panel without separate ethical and security protocols.

CH08-N40

Participant selection and weighting records must have an accountable human or institution owner.

52. FORMS OF FAILURE

CH08-F01 — COUNTING RANDOMS WITHIN THE PANEL AS RANDOMS OF THE WORLD

Random selection within the voluntary panel is presented as if it were random selection from the world population.

CH08-F02 — COUNTING QUOTA COMPLIANCE AS REPRESENTATIVE

It is claimed that all behavioural and technological biases are resolved because age and country quotas are met.

CH08-F03 — COUNTING CUSTOMERS AS GENERAL POPULATION

The monitored institution's user list is used as if it were a world user panel.

CH08-F04 — HIDING EMPLOYEES AND PARTNERS

Users associated with the institution are shown as independent general participants.

CH08-F05 — COUNTING A SINGLE PANEL PROVIDER AS GLOBAL

Provider effect and registered panel behaviour are not examined.

CH08-F06 — COUNTING AN OPEN CALL AS PROBABLE SAMPLE

Self-selection is preserved.

CH08-F07 — DETERMINING THE COUNTRY ONLY BY IP

Travel, VPN, institutional network, and account region differences are ignored.

CH08-F08 — COUNTING CITIZENSHIP AS THE COUNTRY OF USE

The actual geographic condition of the product experience is misclassified.

CH08-F09 — COUNTING LANGUAGE DECLARATION AS DEFINITE WITHOUT TESTING

The participant is kept in the language cell even though they do not understand the prompt or the response.

CH08-F10 — TURNING THE LANGUAGE TEST INTO ELITE SELECTION

Only highly educated people remain on the panel.

CH08-F11 — COUNTING THE NUMBER OF ACCOUNTS AS PEOPLE

One person increases the population weight with multiple accounts.

CH08-F12 — COUNTING A SHARED ACCOUNT AS ONE PERSON

An account used by multiple people is interpreted as a single user.

CH08-F13 — PAYMENT DEPENDING ON THE RESULT

A bonus is given for a positive, correct, or referenced answer.

CH08-F14 — BREAKING THE PROTOCOL WITH SPEED BONUS

The participant takes a screenshot or rushes through the task before completing the response.

CH08-F15 — EXPLAINING THE RESULT IN THE INVITATION

The participant is asked to find an error or success.

CH08-F16 — LETTING THE PARTICIPANT CHOOSE THE PRODUCT

Users who select only the AI product they can access or like determine the panel.

CH08-F17 — HIDING THE EFFECT OF THE ORDER OF PROMPTS

Previous answers affect subsequent tasks.

CH08-F18 — CONSIDERING PANEL CONDITIONING AS NATURAL USE

A user who sees the same test multiple times is presented as a new user.

CH08-F19 — IGNORING RESULT SHARING

Participants transferred tasks and responses to each other.

CH08-F20 — COUNTING NON-RESPONSES AS MISSING DATA

The systematic difference of non-participants is not examined.

CH08-F21 — COUNTING PANEL LOSS AS IMPROVEMENT IN RESULTS

Scores increase because weak or inexperienced users drop out in subsequent waves.

CH08-F22 — INVALIDATING NEGATIVE RESPONSES

The participant is replaced because AI provided an incorrect response.

CH08-F23 — VALIDATING THE CORRECT RESPONSE

Even if the participant refreshed the response, the record is kept because the result is correct.

CH08-F24 — OVERLOOKING BOTS AND REPEATED PARTICIPATION

The same person or automation produces a large number of observations.

CH08-F25 — MASS EXCLUSION WITH A SINGLE SIGNAL

VPN, device, or speed causes real users to be considered fake.

CH08-F26 — CONDUCTING FRAUD REVIEW BASED ON RESULTS

Low-scoring records are examined more frequently.

CH08-F27 — CREATING A HEAVILY UNSEEN GROUP

It is claimed that users not present in the sample are represented.

CH08-F28 — SELECTING POST-RESULT CALIBRATION

Weight variables that produce the most appropriate score are selected afterward.

CH08-F29 — COUNTING RAW PANEL SHARE AS POPULATION SHARE

Heavy AI users determine the global result with raw numbers.

CH08-F30 — COUNTING UNUSUAL WEIGHT AS DESIGN WEIGHT

Exact population inference is made even though the true selection probability is unknown.

CH08-F31 — HIDING THE UNWEIGHTED RESULT

The effect of the weight model is not visible.

CH08-F32 — HIDING THE EFFECTIVE SAMPLE

Even though 1,000 raw records carry 300–400 information units, full precision is claimed.

CH08-F33 — REMOVING WEAK GROUPS BY WEIGHT TRIMMING

High-weighted small groups are trimmed because they lower the result.

CH08-F34 — COUNTING MODEL-BASED ESTIMATED SCORES

A country without an example is given a definite score.

CH08-F35 — REVEALING IDENTITY FOR EVIDENCE

Screenshots are published publicly with personal information.

CH08-F36 — MAKING A SMALL COUNTRY PARTICIPANT REIDENTIFIABLE

Country, time, device, and plan information reveal a single person.

CH08-F37 — ASKING PARTICIPANT FOR PASSWORD OR PRIVATE CHAT

Collecting evidence violates account security.

CH08-F38 — COUNTING LOW DIGITAL SKILL AS INAPPROPRIATENESS

Actual target users are removed from the panel.

CH08-F39 — DISRUPTING VOLUNTARINESS WITH REWARD

The participant is under pressure that they reasonably cannot refuse.

CH08-F40 — LEAVE THE PARTICIPANT'S CHOICE TO THE CLIENT

The audited entity selects the people who can determine the outcome of the panel.

53. AUDIT PROCEDURE

Step 1 — Map Participant Sources

Each panel is recorded separately for partner, open call, and client source.

Step 2 — Assign Selection Status

An appropriate status between PSI-0 and PSI-5 is determined for each source.

Step 3 — Compare Coverage with Target Population

Determine which target users are not present in the panel.

Step 4 — Save Participant Selection Probability or Model

Known probability, quota or voluntary selection is explained.

Step 5 — Freeze Conformity Rules

Age Country Language Product Plan Interface Uniqueness Extraction relationship conditions are versioned.

Step 6 — Review the Invitation Text

It is evaluated whether there is positive or negative guidance.

Step 7 — Examine the Reward Structure

It is confirmed that the reward is completely independent of the outcome.

Step 8 — Verify Participant Country, Language, and Product Access

Privacy-protective verification levels are assigned.

Step 9 — Perform Unique Human Check

Intra-source and inter-source repetitions are examined.

Step 10 — Record Relationship and Familiarity Areas

Customer, employee, competitor, AI provider, and brand familiarity are determined.

Step 11 — Distribute Tasks According to Assignment Plan

Product, prompt, queue, and wave assignments are verified.

Step 12 — Record Panel Conditioning

New, repeated, and previously participated users in the same test are separated.

Step 13 — Set Up Participation Funnel

Invitation, acceptance, start, completion, and validity numbers are calculated.

Step 14 — Blind the Validity Review

Protocol validity is determined without the AI response score being seen as much as possible.

Step 15 — Examine Fraud Signals

Multiple signals and human evaluation are used.

Step 16 — Analyse My Non-Response and Panel Loss

The known characteristics of those who participate and those who do not participate are compared.

Step 17 — Generate Design and Calibration Weights

Probabilistic and non-probabilistic sources are handled with separate methods.

Step 18 — Calculate the Weight Diagnoses

Weight distribution Design effect Effective sample Trimming effect is calculated.

Step 19 — Perform Sensitivity Analysis

Raw, main heavily and alternative reasonable results are compared.

Step 20 — Create the Privacy and Evidence Manifest

Full proof is in a secure repository; anonymous proof identities are kept in a public record.

Step 21 — Publish the Public Election Card

Panel sources, participation funnel, biases, and weighting effect are made visible.

54. REQUIRED EVIDENCE

Target-population record; sampling frame; panel providers; participant sources; selection status; probability or non-probability panel class; panel-registration method; invitation text; participant-admission rules; participant-country verification; language-proficiency record; AI-product access record; plan and user surface; unique-person verification; cross-source duplication checks; brand familiarity; frequency of AI use; relationship with the audited entity; relationship with the AI provider; participant compensation; bonus conditions; task-assignment method; randomisation and blocking record; product and prompt order; panel conditioning; prior pilot participation; number invited; number consenting; number starting the task; number completing it; number of valid observations; reasons for nonresponse; panel attrition; reasons for invalidity; and fraud indicators.

Review and objection records Design weights Non-response settings Calibration variables Propensity model, if any Weight trimming record Effective sample size Sensitivity results Panel provider effect Unweighted and weighted scores Consent record Data storage and access policy Redaction rules Public evidence manifest Responsible person or institution

55. AUDIT CHECKLIST

From which sources did the participants come? Was each source probabilistic, voluntary, or open call? Was in-panel randomness presented like world population randomness? Did the supervising institution intervene in participant selection? Are customer and employee participants separate? Were AI provider or competitor relationships documented? How was country eligibility verified? Is language proficiency at the level of actual use? Was the correct AI product, plan, and interface used? How many accounts or panels did the same person participate in? Was cross-checking between sources conducted? Did the invitation text guide the outcome? Is the reward dependent on the content of the AI response? Did speed incentives reduce protocol quality? Was task assignment random or pre-rules based? Did the participant choose the product or prompt themselves?

Was the product and prompt order balanced? Was panel conditioning recorded? Did participants share task results? Is the invite–accept–start–complete–validity funnel visible? Are non-responders different in known characteristics? Does panel attrition affect the time trend? Was the validity decision made before the AI response was scored? Were negative or critical responses more often considered invalid? Does fraud detection carry multiple signals? Did VPN or shared devices cause automatic exclusion? Was an appeal path given to the participant? Were weight variables defined before the result? Were improbable panel weights correctly labelled? Are target totals reliable? Are

unweighted and weighted results presented together? Is the effective sample size clear? Was the impact of weight trimming shown?

Have groups that are never observed been shown as largely represented? Were model-based predictions presented as if they were direct observations? Are there personal details in participant screenshots? Can users from small countries be re-identified? Were account passwords or private chats requested? Is participant consent clear and understandable? Is there a clearly accountable owner for participant choice and weighting records?

56. OBJECTIONS AND RESPONSES

Objection 1 — "If we use a professional research panel, why are we still talking about bias?"

A professional panel can strengthen data quality and field management. Panel members are also:

- voluntary,
- accustomed to research,
- different in technology use

It is possible. Professionalism does not eliminate bias. It makes measuring and managing easier.

Objection 2 — “If the age and country quotas are met, why shouldn't the panel be considered representative?”

Participants:

- Frequency of AI usage,
- device,
- plan,
- brand awareness,
- digital skill

It may differ in terms of maintenance. Quotas only balance the areas they cover.

Objection 3 — “Can't weighting correct these differences?”

It can partially correct differences for which measured and reliable target totals exist. It cannot definitively correct differences that are never observed or measured.

Objection 4 — “The institution's clients are real users; why shouldn't they be included on the main panel?”

They can be taken. However, they should be taken with the actual proportion and relationship status they carry in the general population. The entire customer list is not a general population panel.

Objection 5 — "Why would it be a problem if the participant knows the brand?"

It doesn’t have to be a problem. They are part of the natural user population. If familiarity is unbalanced, the result may vary. Therefore, it is recorded and analysed if necessary.

Objection 6 — "Heavy AI users perform the task more accurately; wouldn't it be better to select them?"

It may be easier for protocol testing. However, if we want to measure real population experience, inexperienced users are also important. If the task is too complex, the measurement tool should be corrected. The population should not be simplified.

Objection 7 — 'If the participant misunderstood the response, why should we keep the observation?'

The participant's task is not to evaluate the response, but to capture it correctly. If the capture protocol is applied correctly, the AI response is preserved. Material evaluation is the responsibility of the adjudicators.

Objection 8 — "Shouldn't a strict automatic filter be used to prevent fake participation?"

Automatic filtering is useful. If it becomes the sole final decision, it may unfairly exclude real users. Multiple signals and proportional human review are needed.

Objection 9 — 'If we do not make a thousand screenshots public, can the method be considered proven?'

Full screen images:

- personal information,
- account security,
- redefinition

may carry a risk. To the public:

- edited examples,
- observation manifest,
- integrity hashes,
- independent audit access

can be provided. Provability is not identity disclosure.

Objection 10 — "If the weighted score is very different from the raw score, which one is correct?"

Two results answer different questions. The raw score is the result of the people in the panel. The weighted score is the target population estimate. A large difference indicates that the panel composition is important. Both should be published.

Objection 11 — "If the effective sample dropped to 349, was collecting 1,000 observations wasted?"

The oversampled groups may have strengthened the subgroup analysis. However, the sensitivity of the global forecast may be lower than the raw number of 1,000. This fact should not be hidden.

Objection 12 — "Can the non-probability panel not be used at all?"

It can be used. Specifically:

- pilot,
- low-resource language,

- small country,
- Quick discovery

It is valuable for. It should be published with the correct status and restrictions.

Objection 13 — "Isn't it more reliable to choose all the participants from new people?"

It may be strong for cross-sectional population estimation. Repeat users are valuable for measuring longitudinal change. New and repeat panel roles should be kept separate.

Objection 14 — "Doesn't collecting data from so many participants increase privacy risk?"

It can. Therefore:

- data minimisation,
- anonymous identity,
- redaction,
- access control,
- storage duration

are mandatory parts of the protocol. Unnecessary information should not be collected.

COMMON PROVISION OF SECTION 59

The easiest number in a GEO study:

1,000

It is possible. The hardest question is this:

> *Which 1,000?*

Two panels with the same country and age quotas can produce completely different results. In one:

- intense AI users,
- paid plan subscribers,
- technology workers

are the majority. In the other:

- infrequent users,
- mobile users,
- people with low digital skills

It is more visible. Both are 1,000 people. It is not the same population. If you hide where participants come from and only publish the weighted result, you turn weighting into a veil of trust rather than a method of control. You cannot weight a person who is not present in a sample. You cannot change a user's response because it is wrong. A wrong answer is not a failure of the research. It is the event that the research is trying to measure. If a participant has followed the protocol incorrectly despite giving the correct answer, you cannot preserve the record. A correct answer does not make a fake observation real. You can pay one person more. But not to produce the answer you want. You can pay according to a country's living conditions. But you cannot buy the participant's willingness. A thousand screenshots can serve as evidence. You are not obliged to disclose the identity of a thousand people to the public.

Evidence integrity and human privacy can be protected together. Therefore, NOMOS's eighth measurement law is as follows:

> *The power of the sample is determined not only by how many people it consists of, but also by why and how these people were selected.*

The ninth law is as follows:

> *Weighting rebalances the observed population; it does not create the unobserved human.*

The tenth law is:

> *An incorrect AI answer can be valid data; a correct AI answer can be invalid data.*

The eleventh law is as follows:

> *The participant's reward should be given not for finding the desired result, but for proving the observed result without changing it.*

The twelfth law is as follows:

> *Provability does not require the participant to give up privacy.*

NOMOS's Chapter 8 Commandment

> *Do not bring me a thousand people alone. / Show me which world these thousand people come from.*
>
> *Do not treat a volunteer panel as a random population sample, a customer list as the general public or an employee as an independent user.*
>
> *Don't think that all the differences you didn't measure have improved just because the country and age quotas were met.*
>
> *Do not multiply one person by the number of their accounts, devices, languages or panel memberships.*
>
> *Don't put those who love the brand, those who hate the brand, and those who don't know the brand at all in the same invisible box.*
>
> *Pay the participant not to find the correct answer, but to save the first suitable answer they see without changing it.*
>
> *Do not change the user when AI gives a wrong answer. / Do not check if the answer is correct when the user breaks the protocol.*
>
> *Do not make a person who is not on the panel appear weighted as if they exist.*
>
> *Do not hide the raw result. / Do not declare only the weighted result as the truth. / Show the effective sample. / Show its sensitivity.*
>
> *Do not declare someone using a VPN as fake, someone using a shared device as a bot, or someone using assistive technology as automation.*
>
> *Investigate fraud. / But do not turn the investigation itself into a new exclusion bias.*
>
> *Keep a screenshot. / Protect the person's name, account, and private life.*
>
> *First, define the population. / Then show the framework. / Then explain who you chose, who did not come, who could not complete it, and why you excluded someone. / And only after that, decide how much your score can speak for humanity.*

The Chapter's Closing Sentence

> *In GEO-1000, a participant is more than a person carrying a response. A participant is a human sampling unit whose capacity to represent others in the target population must be justified through the selection path, weight, independence and privacy safeguards.*

Normative Core

> **Every GEO-1000 observation MUST retain an auditable participant-selection path from the target population and sampling frame through selection, invitation, consent, assignment, completion, protocol validation, exclusion, and final weighting. Within-panel random selection MUST NOT be represented as random selection from the global target population unless the full population inclusion mechanism supports that claim. Participant recruitment, assignment, compensation, validation, replacement, exclusion, and weighting MUST remain independent of whether the AI response is positive, negative, correct, incorrect, refused, critical, or commercially favourable. Persons, accounts, devices, sessions, languages, panel memberships, and AI-product observations MUST remain distinct units. Probability, opt-in, open-call, client-supplied, institutional, and hybrid participant sources MUST be separately identified. Weighting MAY reduce observed imbalance, but MUST NOT be represented as creating observations for populations absent from the sampling frame or correcting unmeasured selection with certainty. Unweighted results, weighted results, weight diagnostics, coverage gaps, nonresponse, effective sample size, and sensitivity analyses MUST remain visible. Fraud and duplication decisions MUST use proportionate multi-signal review and MUST NOT automatically treat VPN use, shared devices, institutional networks, or accessibility tools as fabrication. Participant evidence MUST remain verifiable without unnecessarily disclosing identity, account credentials, private conversations, or precise location. Every participant-selection and weighting decision MUST be versioned and attributable to an accountable human or organisation.**

PART III

Conducting the Experiment and the Chain of Evidence

CHAPTER 09

The AI System Register and Comparability

METHODOLOGICAL BASES / K02; K03

Chapter Boundary

The first eight chapters established the following: a single response is not a GEO score; representation is a distribution; the audited entity must be identified before results are observed; official representation and verified entity reality must remain separate; an eligible user population must be defined for every AI product; 1,000 valid observations must be allocated through a population-based sample design; smaller countries and languages must remain visible through complementary panels; and participant selection, exclusion and weighting must remain independent of response content. We must now define the other side of the experiment:

> *Which AI system are participants actually measuring?*

The name of a provider? A base model? A chat product? A free web interface? A paid mobile app? A session with web search enabled? A response mode that does not cite sources? An enterprise workspace supported by in-house data? An account with memory enabled? Temporary or clean chat? Are different user interfaces of the same provider the same system? Should two products using the same model family be measured in the same way? Even if a product shows the same model name on the screen:

- retrieval,
- source selection,
- security rules,
- personalisation,
- tool access,
- experimental features

If it changes, can the results be compared? When an AI product is updated in the middle of a measurement wave, can the first and last responses of the wave be considered to come from the same system? A user's:

- web search is on,
- memory is on,
- has previous conversations,
- paid plan

responses; another user's:

- search is off,

- memory is off,
- new session,
- free plan

Can it be merged in the same cell as its answer? No. The measurement principle established by the previous sections is clear: There is no single "AI"; the search index, retrieval system, reranker, language model, citation layer, and safety system can behave differently. The same framework requires each auditable result to carry:

- model or product,
- date,
- country,
- language,
- query set,
- number of repetitions,
- measurement record

This section operates those principles within an AI System Register. This section:

- the distinction between provider, product, model, and user surface,
- AI product instance,
- model identity confidence levels,
- retrieval and tool configurations,
- memory and personalisation states,
- differences in plan, interface, and experimental features,
- product changes,
- system configuration at the time of measurement,
- levels of comparability between products,
- Native Reach and Common Support comparisons,
- version drift over time,

defines. This chapter does not yet:

- full translation and semantic equivalence system of prompts,
- all details of controlled and natural user panels,
- NOMOS Capture software architecture,
- the atomic claim scoring of the responses,
- the final model coefficients and the NOMOS scoring formula

does not finalise. The key question of Section 9 is:

> *Before comparing two responses from the same prompt, how do we prove that these responses actually come from comparable AI product conditions?*

NOMOS Challenge

Two users of the same provider ask the same question: "What kind of company is Apple.com?" First user:

- is using the free web product,
- opens a new chat,

- has memory off,
- does not have web access,

receives an answer only with the product's built-in knowledge capacity. Second user:

- Using a paid mobile product,
- memory is on,
- has previously talked about Apple,
- web search is on,

receives answers with current sources and citations. First answer: "Apple is a technology company that produces consumer electronics." Second answer: "Apple is a global technology company offering hardware, software, and digital services. You can see current corporate information in these sources." Both answers come from the same provider.

Is this the result of the same system? No. Now consider two other users. Both see the same model name on the screen. In one user's country, the web feature is on. In the other user's country, it is off. In one user's answer, source links are visible. The other user's interface does not show citations even if it uses the same sources. There is the same model label.

Is there the same measurement cell? No. Now consider two products from different providers. Both:

- new chat,
- memory is off,
- same language,
- same country,
- same time window,
- web access enabled,
- same prompt version

are measured under these conditions. They are based on different underlying models. Are they comparable? Yes, they can be comparable in terms of a specific user outcome. Because the object of comparison is:

> *not whether they use the same underlying model, but what representation they produce under the same defined user condition.*

Now, let the product version remain unchanged for the first 500 observations of the measurement wave. During the next 500 observations, the provider:

- change the retrieval system,
- security rules,
- user interface

The report alone says: "1,000 users tested the same AI product." In reality, two different product situations may have been measured. The first provision of this section is as follows:

> *The provider name is not sufficient to identify the measured AI system.*

Its second provision states:

> *The model name is not the entirety of the product behaviour that reaches the user.*

Its third provision states:

> *The same provider, the same model label, or the same logo does not guarantee comparability.*

Its fourth provision states:

> *Different providers may be comparable in terms of user outcomes if the measurement conditions are sufficiently matched.*

Its fifth provision states:

> *If a product materially changes during measurement, it cannot be reported as a single fixed system.*

1. PURPOSE OF THE CHAPTER

The purpose of this section is to define the AI product audited within GEO-1000 with all its material configurations and to determine under which conditions different product outcomes can be compared. The section standardises the following distinctions:

- Provider and AI product
- AI product and base model
- Model family and specific deployment
- Model label and verified model identity
- User surface and API
- Web and mobile product
- Free plan and paid plan
- Consumer product and enterprise product
- Basic model answer and retrieval-augmented answer
- Live web search with retrieval
- Citation display with source usage
- Tool usage with tool access
- Session context with memory
- System policy with custom instruction
- Material reality with personalisation
- Temporary chat with new conversation
- Product status over time with static product name
- Observed behaviour change with provider-announced update
- Comparison of Native Reach and Common Support
- Comparable product example within the same product family
- User result versus in-model abstract ability
- Product absence versus product failure
- Technical error versus actual rejection
- Same prompt versus same experiment
- Same average versus same system behaviour
- Instant product snapshot versus long-term product performance

At the end of this section, each GEO-1000 audit should be able to answer the following questions:

> *Which provider's which product, which surface, which plan, and which configuration were measured?*
>
> *How much was the model name seen on the screen really validated?*
>
> *What were web, retrieval, citation, memory, personalisation, and tool status?*
>
> *Did the product remain the same throughout the measurement?*

> *At what comparability level were the two product results placed side by side?*
>
> *Is the thing being compared the base model, the consumer product, the user surface, or the entire system chain?*

2. CENTRAL NORMATIVE PROVISION

> *Each GEO-1000 observation must be linked not only with the provider or model name; it should be connected to a versioned AI System Register record that identifies the AI product instance at the time of measurement.*

An AI product instance must carry at least the following fields to the extent relevant:

- Provider
- Product name
- User surface
- Product type
- Plan or account class
- Apparent model label
- Model ID confidence level
- Product or deployment version
- Country and locale
- Language
- Web or retrieval access
- Citation feature
- Tool access
- Memory status
- Custom instruction status
- Previous chat context
- Session type
- Personalisation status
- Security or policy mode, if known
- Experimental features
- Measurement date and UTC time
- Measurement wave
- Product change log
- Evidence and record confidence

When any of these fields materially change:

- new AI product instance,
- new experimental cell,
- new analysis version

may be required.

3. “AI SYSTEM” IS NOT A SINGLE OBJECT [K02; K03]

In this protocol, “AI system” is a general and high-level term. A real user response can arise from the combination of multiple layers. Candidate chain:

```
R = F(M, O, K, Q, T, P, S, U, C, L, t)
```

Here:

- M: base or guiding model
- O: product orchestration layer
- K: retrieval or knowledge source layer
- Q: reranking and source selection
- T: tools
- P: policy and security layer
- S: session, memory, and personalisation
- U: user account and plan
- C: country and locale
- L: language
- t: time

Response seen by the user:

R

The response visible to a user is not necessarily the output of the base model alone. The principal consumer-facing object measured by GEO-1000 is therefore:

> *the product behaviour presented to the user, not a model abstraction.*

Controlled API or model research can also be conducted. It is not the same result as the main consumer panel.

4. SEVEN IDENTITY LAYERS OF THE AI SYSTEM

Each audited AI system must be defined across seven layers, insofar as each layer is relevant.

4.1. Provider Layer

The organisation that publishes or operates the product. Provider name:

- is not the same as

  product ID,
- model ID,
- user surface

A provider can offer more than one product. Within the same product, more than one model or orchestration system may be used.

4.2. Product Layer

The commercial, consumer, enterprise, or technical product accessed by the user. Example types:

- General chat product
- Search-assisted answer product
- Enterprise workspace
- Educational product
- Embedded assistant

- Developer API
- Custom knowledge-based product

The product name alone does not describe the entire configuration.

4.3. User surface Layer

It is the interface through which the user interacts with the product. Web Mobile Desktop Browser extension Operating system integration In-platform embedded interface

API

Enterprise portal Different surfaces of the same product:

- different features,
- different model,
- different source representation,
- different memory

may be used.

4.4. Model or Inference Engine Layer

It is the model, model family, or routing system used in response generation. Model ID:

- open,
- partially visible,
- unknown,
- dynamic

It is possible. A product can direct a user query to different models. In this case, one cannot say: 'A single fixed model was measured.'

4.5. Orchestration and Information Layer

It may include the following:

- Retrieval
- Live web search
- Internal search
- Source ranking
- Reranking
- Citation selection
- Calling a vehicle
- Query rewriting
- Document shredding
- Answer merging

This layer is of fundamental importance for the representation of GEO. Because it can affect which record:

- is accessed,
- is selected,
- is carried into the response,
- is shown as a reference

can be influenced.

4.6. Policy and Personalisation Layer

It may include the following:

- Security rules
- System instructions
- Product policy
- Memory
- Special instructions
- User history
- Regional rules
- Age or account class
- Corporate administrator settings

This layer affects the:

- scope of the response,
- its tone,
- its rejection,
- its recommendation

can change.

4.7. Distribution and Time Layer

Indicates which of the following the product is under at a specific moment:

- is meaningful with
- experimental feature,
- regional distribution,
- user segment,
- A/B test

The same product name can have different distribution conditions for different users on the same day.

5. AI PRODUCT EXAMPLE

The core system unit within GEO-1000:

AI Product Instance

Canonical operational term:

AI Product Instance

An AI Product Instance is defined as:

```
A_i = (v, p, s, r, m, k, t, x, u, c, l, ω)
```

can be represented as. Here:

- v: provider
- p: product
- s: user surface
- r: plan or account class

- m: model ID or label
- k: retrieval and information structuring
- t: tool configuration
- x: memory, personalisation, and session state
- u: user or organisational setting class
- c: country and locale
- l: language
- ω: time, deployment, and experimental feature state

An observation: Oj must be linked to a specific: Ai product sample.

6. DIFFERENCE BETWEEN PRODUCT SAMPLE AND PRODUCT NAME

The product name can be fixed. The product sample may vary. Example: Let the Orion AI product name remain unchanged. However, these two measurements are different product samples:

Product Sample 1

Web Free plan Memory off Web access off New chat Turkish Turkey Wave 1

Product Sample 2

Mobile Paid plan Memory on Web access on Previous chat context available Turkish Turkey Wave 1
These can exist under the same brand. They are not the same experimental cell.

7. MODEL IDENTITY CONFIDENCE LEVELS

The model name cannot always be known with the same certainty. Therefore, the model identity can be recorded at the following levels.

MI-0 — UNKNOWN MODEL

It is unknown which model or model family the product uses. The result is attributed only to the product surface.

MI-1 — USER-DISPLAYED LABEL

A model label is visible in the user interface. The label is recorded. It is not assumed to fully verify the back-end deployment.

MI-2 — PROVIDER-DOCUMENTED MODEL FAMILY

The provider's product documents describe a specific model family or product–model relationship. It may still be unknown whether individual observation is directed exactly to that model.

MI-3 — SESSION-CONFIRMED DEPLOYMENT

The session verifies a specific distribution of API or trusted product metadata.

MI-4 — VERSIONED ENDPOINT OR DEPLOYMENT

A specific version, endpoint, or deployment ID can be recorded. It is stronger for controlled comparison.

MI-5 — REPRODUCIBLE FROZEN MODEL ARTEFACT

There is a static and reproducible model or deployment structure. It may be rare in live consumer products. If model ID trust is low, the result:

should be attributed to product behaviour

It cannot be definitively attributed to a specific model.

8. LIMIT OF THE APPARENT MODEL NAME

Seeing the model name in a user interface is valuable. However, the following questions may still remain open: Are all queries directed to the same model? Does the model change depending on the difficulty of the query? Does security or retrieval use a separate model? Does the provider apply the same label to different distributions? Does the model label only indicate the product family? Is the user part of an experimental rollout? Therefore, the correct reporting might be: "Model X label was shown in the user interface." Without stronger evidence, the following sentence should not be used: "All responses are definitively generated solely by Model X."

9. PRODUCT TYPE

Each AI product should be classified with a primary type and secondary types.

PT-1 — General Conversational Product

It is a general chat and question-answer product.

PT-2 — Search- or Retrieval-Grounded Product

Search or retrieval is a fundamental component in answer generation.

PT-3 — Enterprise Knowledge Product

It can use internal documents, workspaces, or special data sources.

PT-4 — Embedded Assistant

It is embedded into another product or platform.

PT-5 — Developer or API Product

It is used by a developer or application instead of the end user.

PT-6 — Domain-Specific Product

It is designed for a specific profession or industry.

PT-7 — Multimodal Product

Inputs other than text, such as visual, audio, or document entries, are part of the main experience. A product can carry more than one type. Comparison should be made within the same type or with an explicit scope.

10. USER SURFACE REGISTER

Different surfaces of the same product should be recorded separately. Surface areas:

- Surface ID
- Web/mobile/desktop/API
- Application version
- Operating system, if applicable
- Browser, if applicable
- Product language
- Visible model selection
- Availability of tools
- Citation view
- File upload feature
- Memory settings
- Corporate management settings
- Experimental features
- Measurement time

If a web and mobile interface are found to be materially the same, the results can be combined. This equality should not be assumed; it must be tested.

11. PLAN AND ACCOUNT CLASS

The plan type may have features that can affect the user's response. Plan statuses:

- Free
- Paid individual
- Professional
- Corporate
- Education
- Developer
- Trial
- Invitation or limited access
- Unknown

Plans may differ in the following areas:

- Model
- Usage limit
- Web access
- Memory
- Tools
- Source citation
- Response length
- Speed or priority
- Enterprise data
- Security and retention settings

If there is no plan information:

The PLAN_UNKNOWN

status should be used. Free and paid results cannot silently be merged as the same product behaviour.

12. RETRIEVAL AND WEB ACCESS

An AI response can be generated in one of three basic knowledge states.

KG-0 — UNDETERMINED KNOWLEDGE MODE

The case of retrieval or web usage cannot be determined.

KG-1 — MODEL-INTERNAL RESPONSE

A response known or strongly assumed not to use live or external information.

KG-2 — RETRIEVAL-ASSISTED RESPONSE

Retrieval may have been used from pre-indexed or in-product sources.

KG-3 — LIVE WEB OR SEARCH-ASSISTED RESPONSE

Live or near-measurement time web/search access has been used.

KG-4 — PRIVATE KNOWLEDGE-GROUNDED RESPONSE

A private data source provided by the institution or the user has been used.

KG-5 — HYBRID OR MULTI-SOURCE RESPONSE

Multiple layers of information have been used together. These modes may produce different results in the same prompt. They should be clearly distinguished in comparison.

13. RETRIEVAL ACCESS IS NOT THE SAME AS RETRIEVAL USAGE

The product may have a web search feature. However, it may not have been used in the specific answer. Three fields should be recorded separately: Was the feature available? Was the feature on or selectable? Was there evidence that it was used in the specific answer? Example statuses:

AVAILABLE_NOT_USED

AVAILABLE_USAGE_UNKNOWN

USED_CONFIRMED

NOT_AVAILABLE

DISABLED

UNKNOWN

The mere presence of a source link can be a strong indicator for usage. The absence of a source link does not necessarily prove that retrieval was not used.

14. CITATION LAYER

Using a source is not the same as showing a citation. A product:

- may use a source but may not show it,
- may show a citation but the citation may not support the claim,
- may link the same source to different sentences,

The user can hide the citation according to the user surface. The citation record must contain the following fields: Is the citation feature available? Was the citation shown in the specific response? Number of citations Source field names Citation–claim match Visibility on the user surface Expandability of the citation Accessibility of the source Citation generation time Products that show and do not show citations cannot be directly compared in terms of "number of sources". First, product behaviour and interface differences must be explained.

15. TOOL ACCESS

The AI product may access the following tools:

- Web search
- Browser
- Calculator
- Code execution
- File reading
- Map
- Calendar
- Corporate knowledge base
- Product catalogue
- Other APIs

Tool access may affect your response:

- its timeliness,
- evidence,
- scope of the response,
- accuracy

Tool registry:

- whether the tool is available,
- whether it is active,
- whether it was used in the response,
- whether the tool result is retained

should be shown.

16. MEMORY, SPECIAL INSTRUCTIONS, AND PERSONALISATION

The response visible to the user may be affected by the following fields:

- Persistent memory
- Chat history
- Special instructions
- User's name or profile
- Previous brand conversations
- Preferences
- Country and locale
- Corporate workspace settings
- Segmentation done by the provider

These fields must be recorded in three separate layers.

16.1. Persistent Personalisation

Information protected across accounts or sessions.

16.2. In-Conversation Context

Previous messages within the current chat.

16.3. System or Organisational Setting

Administrator or product instructions that the user may not see. Complete information may not always be available. Unknown situations must be explicitly recorded.

17. SESSION TYPES

Candidate session statuses:

ST-1 — NEW STANDARD CONVERSATION

New standard conversation. Memory or special instructions are recorded separately.

ST-2 — TEMPORARY OR EPHEMERAL CONVERSATION

Temporary session intended to be separated from permanent history. Actual features should be verified in the product registry.

ST-3 — EXISTING CONVERSATION

There is previous conversation context. It may be valuable for the natural panel. It is kept separate in the controlled panel.

ST-4 — ENTERPRISE OR SHARED WORKSPACE SESSION

There may be an enterprise or shared information area.

ST-5 — UNKNOWN SESSION STATE

The session status could not be reliably determined. If different session types are to be combined in the same cell, the rationale and weighting must be explained.

18. PRODUCT POLICY AND REJECTION

An AI product may not respond to a specific question due to:

- security,
- law,
- product policy,
- regional rule,
- does not imply any suitability in terms of age,
- enterprise setting

reason. Rejection:

- basic model deficiency,
- product policy,
- interface,
- may be due to the user account

GEO-1000 main user control records the rejection result that reaches the user. However, if the reason for the rejection is known, it is additionally classified. A product's policy layer is also part of the user experience.

19. EXPERIMENTAL FEATURES AND A/B DISTRIBUTIONS

Live AI products may not give the same features to all users at the same time. Participants:

- different user interface,
- different model prompting,
- different web feature,
- different response format

may see. Experimental status indicators:

- Beta label on the interface
- Provider documentation
- Feature difference
- Systematic UI difference between users
- Behaviour breaking during measurement

may occur. If the experimental feature is unknown:

EXPERIMENTAL_STATE_UNKNOWN

should be used. The actual variation within the product is not data defect. It is a distribution feature of the system. However, it should be considered in comparability.

20. AI System Register

An AI System Register should be created for each audited product. The registry should carry three levels.

20.1. Provider and Product Master Record

Includes relatively permanent areas:

- Provider
- Product name
- Product type
- Main user interfaces
- Official access countries
- Plan classes
- Supported language records
- General features

20.2. Measurement Wave Product Record

For a specific wave:

- product version,
- plan features,
- model labels,
- web and tool features,
- memory behaviour,
- access status,
- known updates

is recorded.

20.3. Observation Level System Record

Actually observed for each participant:

- interface,
- model label,
- plan,
- vehicle usage,
- citation,
- session,
- personalisation,
- time

carries fields. The main record alone cannot remove observation-level differences.

21. SYSTEM CONFIGURATION FINGERPRINT

A System Configuration Fingerprint can be created to facilitate the comparability of an AI product instance. Candidate fields:

```
SCF = H(canonicalise(p, s, r, m, k, t, x, c, l, w))
```

Here, H is the function that produces an integrity value from the versioned fields. The fingerprint:

- whether two observations belong to the same configuration,
- whether there have been material changes in the wave

It can help to identify. Fingerprint: does not prove the real backend model, does not make layers that the provider has not disclosed visible. It only reinforces whether the recorded configuration has been changed later.

22. WHAT IS COMPARABILITY?

Two results are considered comparable when:

> *The difference between them can be sufficiently explained by monitored product behaviour rather than irrelevant differences in measurement design.*

Comparability is not absolute equality. Different AI products:

- model,
- provider,
- user surface.

These need not all be identical. The material conditions relevant to the comparison must, however, be:

- equalisation,
- layering,
- recording,
- modelling in analysis

are required.

23. THE OBJECT OF THE COMPARISON MUST BE DEFINED IN ADVANCE

The same data can answer different comparison questions.

23.1. User Product Comparison

"Which consumer product produces a more accurate representation under the same user conditions?" The object being compared is the entire product chain.

23.2. Model Comparison

"Which model produces a more accurate response under the same controlled prompt and tool-free condition?" API or fixed model access may be required.

23.3. Retrieval Comparison

“Is there a difference between the web-accessible and non-accessible modes of the same product?” The model and other conditions are kept as constant as possible.

23.4. Plan Comparison

“Do free and paid users see the same brand representation?” The plan difference is the object of the audit.

23.5. Interface Comparison

“Does the web and mobile product produce different representations under the same conditions?”

23.6. Time Comparison

"Did the representation of the same product change across different waves?" Product and population changes should be separated. Without specifying the comparison question: "Product A is better than Product B." is excessively broad.

24. DIMENSIONS OF COMPARABILITY

When comparing two products or waves, the following areas should be evaluated:

- Audited entity
- Reality package version
- Prompt ID and version
- Language and locale
- Country
- Target population
- Sample source
- Participant composition
- Session type
- Memory
- Special instructions
- Previous conversation
- Product plan
- User surface
- Model label
- Retrieval/web status
- Tools
- Citation feature
- Measurement time
- Product updates
- Adjudication rule
- Result statuses
- Weighting

Not all dimensions need to be the same. Different dimensions limit the scope of the comparison.

25. COMPARABILITY LEVELS

CP-0 — INCOMPARABLE

Material measurement conditions are unknown or seriously inconsistent. Direct performance comparison cannot be made.

CP-1 — DESCRIPTIVE CO-OCCURRENCE

There are results from two products or waves. However:

- population,

- prompt,
- time,
- session

are not equal. Only descriptive side-by-side display can be made.

CP-2 — PARTIALLY ALIGNED

Some core fields match. Material differences continue. Limited comparison can be made.

CP-3 — MATCHED USER-SURFACE COMPARISON

Same entity Same prompt version Same language/country Similar user population Same time window Registered product features are available. Consumer product results are comparable.

CP-4 — CONTROLLED MATCHED COMPARISON

Additionally:

- session,
- memory,
- personalisation,
- retrieval,
- tools,
- plan

are matched in a controlled or open manner. Provides stronger product or feature comparison.

CP-5 — REPLICATED AND CALIBRATED COMPARISON

Comparison:

- in multiple waves,
- with independent users or panels,
- with the same normative evaluation

repeated; variability and product changes have been calibrated. These levels are not a "scientific accuracy score." They indicate the limit that a comparison claim can carry.

26. WHAT TO DO IN CP-0 CONDITION?

If two products cannot be compared, the results are not ignored. They can be reported separately:

- "This is the Native Reach result of Product A."
- "This is the result of Product B under different users and product conditions."

However: one cannot say, "A is 12 points better than B." Incomparability is not a failure, it is a limitation of the method.

27. COMPARABILITY FOR THE SAME PROVIDER

Results from the same provider cannot be automatically compared. The following differences may be present:

- Product
- Plan

- Model
- Web access
- Interface
- Memory
- Enterprise data
- Country distribution
- Experimental feature
- Time

Comparison of the same provider should also be evaluated according to the CP level.

28. COMPARABILITY FOR DIFFERENT PROVIDERS

User products belonging to different providers:

- same target population,
- same prompt,
- same time,
- similar user surface,
- clear product configuration

It is comparable underneath. Being a different provider does not prevent the comparison. However, differences in features determine what the comparison measures. Example: Product A uses live web. Product B does not. The result can answer this question: 'Which of the two product chains offered to the user produces more accurate representation?' It does not directly answer this question: 'Which base model is more knowledgeable?'

29. COMPARISON OF WEB-SUPPORTED AND NON-WEB PRODUCTS

Two separate comparisons are possible.

29.1. Natural Product Comparison

Each product is measured under the default real user condition. A product that uses the web uses the web. One that does not use it does not. Result: it is a comparison of the real consumer product experience.

29.2. Feature Matched Comparison

If possible, both products:

- measured in web open,
- web closed

modes separately as well. This can better separate the retrieval effect. Each product may not offer the same options. Missing mode:

RECORDED AS NOT AVAILABLE

That status must remain explicit.

30. PRODUCTS THAT SHOW AND DO NOT SHOW CITATIONS

For a product without citation feature: it is possible to give a "Zero citation." However, its meaning should be explained: the product may not be referencing sources. It is not certain that it does not use any sources. Citation coverage and accuracy coverage are separate outcomes. Citation comparison should only be made among products that support citations or with separate feature status.

31. NATIVE REACH COMPARISON

Each product is measured within its actual reach population. Advantage: shows the product's natural global impact. Limitation: populations are different, country and language distributions may vary. Native Reach scores: show what the product does within its reached population. Should not be used alone in direct product superiority ranking.

32. COMMON SUPPORT COMPARISON

The common access population of the compared products is used. Advantage:

- same country,
- same language,
- same user condition

provides a fairer comparison. Limitation: preserves the actual global access difference of the products, the common universe may be very narrow. The Common Support score should be published together with the Native Reach score.

33. PAIRED USER COMPARISON

If the same user tests two or more AI products, user characteristics can be controlled. However:

- task order,
- answer recall,
- product familiarity,
- special profile of users who can access all products

It can create bias. In a paired design:

- product order should be randomised,
- previous responses should not be shared,

the same user dependency should be maintained in the analysis.

34. INDEPENDENT USER COMPARISON

Each product is tested by different users, but selected from the same population frame. Advantage: task contamination is reduced, more natural product users can be chosen. Limitation: user groups may not be completely equal. Weighting and stratification are required. Paired and independent user results can be separate but complementary.

35. TIME COMPARABILITY

When comparing today's result of an AI product with past results, the following changes should be examined:

- Model
- Product version
- Web/retrieval
- Plan
- Country access
- Language support
- Prompt
- Entity reality
- Adjudication
- Population framework
- Weighting

Time difference may not only arise from the AI product.

36. TYPES OF PRODUCT CHANGES

UC-1 — DOCUMENTED MINOR CHANGE

A minor change disclosed by the provider and assessed not to materially alter the measurement object.

UC-2 — DOCUMENTED MATERIAL CHANGE

Material change described in model, retrieval, interface, or feature.

UC-3 — OBSERVED BEHAVIOUR CHANGE

Even if there is no official announcement, systematic behaviour or interface change is observed.

UC-4 — ACCESS OR PLAN CHANGE

Country, plan, or user access changes.

UC-5 — UNKNOWN CHANGE EVENT

There is a break in measurement behaviour. The reason could not be verified.

UC-6 — PRODUCT REPLACEMENT OR RETIREMENT

An old product or model has ended and has been replaced by another structure. The type of change limits time series comparison.

37. PRODUCT CHANGE WITHIN THE MEASUREMENT WAVE

If a material change occurs during a wave, the following options exist: The wave can be stopped and restarted. Separate sub-waves can be made before and after the change. The change can be modelled as an experimental factor. The wave is labelled as MIXED_SYSTEM_STATE. The following

structure is forbidden: Combining all responses as if they result in a single fixed product outcome despite knowing about the change.

38. BEHAVIOUR CAN CHANGE WITHOUT PRODUCT CHANGE

Even if the provider does not announce an explicit update, the result may change. Possible reasons:

- Change in the retrieval index
- Updating of resources
- New content appearing on the web
- Reranking change
- A/B test
- Regional rollout
- Security policy
- Traffic routing

Therefore, it cannot be said, "The provider did not update, the system was definitely the same." Measurement evidence and behaviour breakdown should also be examined.

39. LIVE PRODUCT AND FROZEN MODEL

There are two different research objects.

39.1. Live Product Audit

Measures what real users see on the real consumer surface. Advantage:

- shows real experience,
- real source,
- real policy,
- real interface

limitations: it is variable, full reproduction may be difficult.

39.2. Frozen Model Audit

Controlled repetition is conducted under a fixed version or endpoint. The advantage:

- provides reproducibility,
- experimental control

The limitation: may not fully represent the real consumer product. The main population layer of NOMOS GEO-1000 is based on the live user product. Frozen model audit is complementary laboratory work.

40. PRODUCT UNAVAILABILITY AND TECHNICAL RESULTS

The following cases should be kept separate:

AR-1 — PRODUCT UNAVAILABLE

The product is not available for the country or user group. It is the result of access scope.

AR-2 — ACCOUNT OR PLAN INELIGIBLE

The plan or account condition assigned to the participant is not eligible. There may be a sampling or task assignment issue.

AR-3 — TEMPORARY TECHNICAL FAILURE

A temporary technical error has occurred. Retry protocol should be predefined.

AR-4 — RATE OR USAGE LIMIT

A usage quota or rate limit has been reached. It may be part of the real user experience.

AR-5 — SYSTEM REFUSAL

The product task has been refused due to policy or content. It is a valid system outcome.

AR-6 — NO USABLE OUTPUT

No usable response has been generated.

AR-7 — UNKNOWN ACCESS FAILURE

Why it could not be determined. These situations cannot be combined in the same "no response" box.

41. TECHNICAL RETRY RULE

A retry can be made in case of a temporary technical error or connection interruption. However:

- if the AI response is not liked,
- a rejection occurred,
- incorrect information was produced

retry cannot be made for these reasons. Technical retry:

- same user,
- same product,
- same configuration,
- must be made within the predetermined period.

The first technical incident must also be recorded.

42. COMPARISON MATRIX

A basic comparison record for two products could be as follows:

Dimension	Product A	Product B	Status
Target entity	Same	Same	Matched
Prompt version	1.0	1.0	Matched
Language/locale	tr-TR	tr-TR	Matched
Country	Turkey	Turkey	Matched

Dimension	Product A	Product B	Status
Time window	Wave 1	Wave 1	Matched
User panel	Population	Population	Matched
Plan	Free	Paid	Different
Surface	Web	Mobile	Different
Web access	Closed	Open	Different
Memory	Closed	Open	Different
Model ID	MI-1	MI-1	Limited
Comparison level			CP-2

These results can be shown side by side. However: "The basic model of Product B is better." cannot be said. The difference may stem from the entire product chain.

43. SYNTHETIC APPLE.COM COMPARISON CASE

SYNTHETIC METHODOLOGY DEMONSTRATION / The products, responses, and results below are entirely fictional. They do not represent the performance of the real Apple Inc. or any real AI provider. Let's consider three synthetic product examples.

Product Example A

Provider: Orion Product: Orion Chat Interface: Web Plan: Free Model label: Orion Standard Model Confidence: MI-1 Web access: Off Memory: Off Session: New Language: Turkish Locale: Turkey Wave: W1 Response: "Apple is an American technology company that produces phones and computers."

Product Sample B

Provider: Orion Product: Orion Chat Interface: Mobile Plan: Paid Model Label: Orion Pro Model Confidence: MI-1 Web Access: Open Citation: Open Memory: On Previous Conversation: About Apple Language: Turkish Locale: Turkey Wave: W1 Response: "Apple is a global technology company that provides hardware, software, and digital services. Current corporate information can be found in these sources."

Product Sample C

Provider: Helios Product: Helios Search Interface: Web Plan: Free Model Label: Unknown Model Confidence: MI-0 Web Access: Open Citation: Open Memory: None Session: New Language: Turkish Locale: Turkey Wave: W1 Response: "Apple is a US-based technology company that develops consumer devices, operating systems, and digital services."

43.1. Comparison of A and B

They are the same provider. However:

- plan,
- surface,
- model label,
- web,
- memory,
- previous conversation

is different. The comparison level can be CP-1 or CP-2. The response difference cannot be attributed solely to the basic model difference.

43.2. Comparison of A and C

They are different providers. However:

- new session,
- free plan,
- web surface,
- same language/country,
- same wave

in terms of some conditions match. Web access is different. Comparison: can be done as a natural consumer product outcome. It cannot be done as a basic model capability.

43.3. Comparison of B and C

Both products have web access and citations enabled. However, B has memory and previous conversation, while C does not. The user experience of the products is comparable. It is not a controlled clean product comparison.

43.4. Controlled Repeat

In product B:

- memory is turned off,
- a new chat is started,
- special instructions are turned off,

web remains on. C is measured under the same condition. Comparison level: CP-3, or CP-4 with sufficient additional control. This example shows:

> *The same provider does not provide a close comparison; well-recorded different providers may be more comparable.*

44. SYNTHETIC PRODUCT CHANGE CASE

SYNTHETIC EXAMPLE

Orion Chat W1 wave targets 1,000 observations. In the first 480 observations:

- product version: 6.1
- web feature: manual
- attribution: limited

In the next 520 observations:

- product version: 6.2
- web feature: automatic
- attribution: extended

let it be. Single result: if calculated as "Orion Chat W1 score," two different system states mix. Correct options:

Option 1

W1A: first 480 W1B: next 520 are reported separately.

Option 2

The wave is not continued. 1,000 new observations are taken in version 6.2.

Option 3

The version is kept as an explicit experimental factor in statistical analysis. Incorrect option: The update is removed from the report and all observations are presented as a single fixed product.

45. MINIMUM AI SYSTEM RECORD

Each observation must carry at least the following fields:

- NOMOS AI system ID
- Provider
- Product
- Surface
- Plan
- Product type
- Apparent model label
- Model confidence level
- Country
- Locale
- Language
- Session type
- Memory status
- Custom instruction status
- Previous conversation state
- Retrieval/web feature
- Retrieval usage status
- Citation status
- Tool status
- Experimental feature status
- UTC time
- Wave
- System fingerprint
- Change events
- Source of evidence
- Record owner

If the material field is unknown, it cannot be left blank. UNKNOWN status should be used.

46. MINIMUM COMPARISON CARD

When comparing two products or waves, the public record should show:

- Comparison question
- Comparison level

- Common target population
- Native Reach differences
- User equality
- Language and country
- Measurement time
- Product, plan, and surface
- Model identity confidence
- Web/retrieval difference
- Difference in memory and personalisation
- Difference in tool
- Difference in sample
- Evaluation method
- Material limitations
- Directly comparable metrics
- Non-comparable areas

Single number or ranking should not be published without this card.

47. RULE OF COMPARABILITY

An auditor should answer these four questions: Was the same thing asked? Was the same human universe measured? Were the same user conditions applied? Were the same time and evaluation system used? Each of these questions:

- completely,
- partially,
- no,
- unknown

is classified as. The comparison level is given accordingly.

48. MANDATORY NORMATIVE PROVISIONS

CH09-N01

Every GEO-1000 observation must be linked to a versioned AI System Register record.

CH09-N02

The provider name cannot be used as the full identity of the measured AI product.

CH09-N03

The product, user surface, plan, model, retrieval, tool, session, and personalisation layers must be separated from each other.

CH09-N04

The model label on the user surface cannot be presented as a verified back-end model identity.

CH09-N05

The model-identity confidence level must be recorded from MI-0 to MI-5 or through an equivalent openly defined status.

CH09-N06

If the model identity cannot be verified, the result should be attributed only to the product and user surface.

CH09-N07

Products belonging to the same provider are not automatically considered the same experimental cell.

CH09-N08

Products belonging to different providers may be comparable when material user and experimental conditions are matched.

CH09-N09

Web, mobile, desktop, embedded product and API surfaces cannot be combined without description.

CH09-N10

Free, paid, corporate, educational, and developer plans must be recorded as separate product terms.

CH09-N11

If the plan is unknown, the result should carry the status PLAN_UNKNOWN.

CH09-N12

The availability of the retrieval or web feature cannot be counted as proof that it was used in a specific response.

CH09-N13

Retrieval access, retrieval usage, and attribution must be recorded with separate statuses.

CH09-N14

A product that does not show attribution cannot automatically be considered as not using any resources.

CH09-N15

The citations of the product being cited should be examined separately in terms of claim support.

CH09-N16

Tool access and tool use in specific responses should be recorded separately.

CH09-N17

Memory, special instructions, previous chats, and corporate settings should be recorded separately as much as possible.

CH09-N18

An unknown personalisation state cannot be assumed to be like a clean session.

CH09-N19

New chat, temporary chat, current chat, and corporate workspace sessions cannot be merged without explanation.

CH09-N20

Natural user panel and controlled clean panel results should carry separate product status.

CH09-N21

If an experimental feature or A/B distribution in the product is material, it should be preserved in the observation record.

CH09-N22

If a material product change occurs during a wave, the wave should be divided into separate sub-versions, restarted, or the change should be explicitly modelled in the analysis.

CH09-N23

A known product change cannot be stored as if all observations came from a single fixed system.

CH09-N24

The provider not explaining an update cannot be taken as definitive proof that the product has not changed.

CH09-N25

The result of a live consumer product cannot be presented as if it were the result of a frozen baseline model.

CH09-N26

The result of a frozen model in the laboratory cannot be presented as performance of a real user product.

CH09-N27

Absence of the product, technical error, usage limit, system rejection, and lack of result must be recorded with separate statuses.

CH09-N28

Incorrect or negative AI response cannot be a reason for technical retry.

CH09-N29

The subject of the comparison must be defined before data is collected.

CH09-N30

The comparability level of two products must be visible in the public result.

CH09-N31

Definite product ranking cannot be generated from CP-0 or CP-1 level results.

CH09-N32

Using the same prompt alone does not create a comparable experiment.

CH09-N33

In comparisons, the target entity, Truth Pack, prompt, language, country, time, population, session, product, and evaluation method should be examined together.

CH09-N34

Native Reach results cannot be directly ranked without explanation because they belong to different target populations.

CH09-N35

The Common Support result cannot be presented as the product's actual global reach result.

CH09-N36

Web-supported and web-free product comparison can be done as a user product comparison; it cannot be presented as a basic model superiority.

CH09-N37

In paired comparisons made with the same user, product order and user dependency must be maintained.

CH09-N38

In comparisons made with different users, user composition and weight differences must be visible.

CH09-N39

The product version or configuration fingerprint must be stored in every measurement wave.

CH09-N40

There must be an accountable human or institutional owner of the system registry and comparability decision.

49. FORMS OF FAILURE

CH09-F01 — CONSIDERING THE PROVIDER AS A SINGLE AI

All products, models, and surfaces of the provider are presented as a single system.

CH09-F02 — COUNT PRODUCT BEHAVIOUR BY MODEL NAME

Retrieval, security, interface, and personalisation are ignored.

CH09-F03 — DEFINITELY CONSIDER THE VISIBLE MODEL LABEL AS BACKEND

The use of routing or dynamic models is disregarded.

CH09-F04 — CONSIDERING THE SAME PROVIDER AS COMPARABLE

The plan, surface, and feature differences are hidden.

CH09-F05 — CONSIDERING DIFFERENT PROVIDERS AS NON-COMPARABLE

Product comparisons possible under paired user conditions are rejected.

CH09-F06 — INTEGRATING WEB AND MOBILE

Interface and feature differences disappear.

CH09-F07 — COMBINING FREE AND PAID PLANS

Quality or feature differences due to the plan become invisible.

CH09-F08 — CONSIDERING CORPORATE AND CONSUMER PRODUCTS THE SAME

Private information and administrator settings are hidden.

CH09-F09 — TREATING API RESULT AS CONSUMER PRODUCT

Orchestration and user surface differences are eliminated.

CH09-F10 — IF RETRIEVAL EXISTS, CONSIDER IT USED

The specific answer's true knowledge path cannot be verified.

CH09-F11 — IF NO CITATION, CONSIDER NO SOURCE

The possibility of hidden or invisible retrieval is ignored.

CH09-F12 — IF CITATION EXISTS, COUNT AS CORRECT SOURCE

Citation–claim matching is not examined.

CH09-F13 — CONSIDER TOOL ACCESS AS TOOL USE

It is unknown which tool was used in the specific answer.

CH09-F14 — CONSIDER MEMORY AS CLEAN SESSION

The effect of personalisation remains invisible.

CH09-F15 — CONSIDER NEW CONVERSATION NON-PERSONALISED

Persistent account settings and special instructions are ignored.

CH09-F16 — PRESENTING THE CURRENT CHAT AS A NEW SESSION

Previous context is hidden.

CH09-F17 — COUNTING AN EXPERIMENTAL FEATURE AS A DATA ERROR

Actual product variation is declared an invalid observation.

CH09-F18 — HIDING AN EXPERIMENTAL FEATURE

Users in different rollouts are combined as a single fixed product.

CH09-F19 — SINGLE WAVE AFTER UPDATE

Material product change is published as the result of a single fixed system.

CH09-F20 — IGNORING CHANGES IF THERE IS NO ANNOUNCEMENT

Observed behaviour break is not examined.

CH09-F21 — COUNTING LIVE PRODUCT AS A FROZEN MODEL

Full reproducibility is claimed.

CH09-F22 — CONSIDERING A FROZEN MODEL AS A REAL USER PRODUCT

Product chain and interface differences are hidden.

CH09-F23 — COUNTING PRODUCT ABSENCE AS AN ERROR

Scope of access is confused with representation accuracy.

CH09-F24 — REFUSING A TECHNICAL ERROR

Connection or service issue is classified as a policy result.

CH09-F25 — COUNTING A TECHNICAL ERROR AS REJECT AND RETRYING

Actual user result is removed from the denominator.

CH09-F26 — HIDING THE USAGE LIMIT

The real experience of the free plan becomes invisible.

CH09-F27 — CONSIDERING THE SAME PROMPT AS THE SAME EXPERIMENT

Product, plan, session, web, and time differences are ignored.

CH09-F28 — CONSIDERING THE PRODUCT RESULT AS MODEL SUPERIORITY

The difference across the entire system chain is attributed to the base model.

CH09-F29 — DIRECTLY RANKING NATIVE REACH SCORES

Different population universes are considered the same.

CH09-F30 — CONSIDERING COMMON SUPPORT AS GLOBAL SCOPE

The narrow common universe is presented as the entire reach of the product.

CH09-F31 — CLAIMING LEADERSHIP FROM CP-1 RESULT

Product superiority is derived solely from incompatible results that are placed side by side.

CH09-F32 — MAINTAINING THE COMPARISON LEVEL

The public only sees the ranking.

CH09-F33 — IGNORING PAIRED USER DEPENDENCY

The responses of the same person to products are considered independent.

CH09-F34 — IGNORING PRODUCT ORDER EFFECT

The first response affects subsequent tasks.

CH09-F35 — TREATING DIFFERENT PANEL USERS AS THE SAME

Products are tested with different user compositions.

CH09-F36 — CHANGING THE REALITY PACKAGE VERSION

The same responses are compared with different reference records.

CH09-F37 — SILENTLY CHANGING THE PRODUCT FINGERPRINT

The configuration record is updated; the old measurement is preserved rather than overwritten.

CH09-F38 — COUNTING AN UNKNOWN AREA AS CLEAN

The model, memory, or tool status is accepted as controlled, even if unknown.

CH09-F39 — USING THE PRODUCT BRAND INSTEAD OF EVIDENCE

The provider's reputation substitutes for measurement conditions.

CH09-F40 — COVERING ALL PRODUCT DIFFERENCES WITH A SINGLE COEFFICIENT

Plan, interface, and configuration differences are hidden by using the model or market coefficient.

50. AUDIT PROCEDURE

Step 1 — Identify the Provider and Product

The product name, type, and official user interfaces are recorded.

Step 2 — Lock the Measurement Surface

Web Mobile Desktop

API

Embedded product is detached.

Step 3 — Verify the Plan and Account Class

Free, paid, enterprise, or other plan status is recorded.

Step 4 — Assign Model ID Trust

Appropriate level is determined between MI-0 and MI-5.

Step 5 — Save Retrieval and Web Configuration

Is the feature available? Is it on? Was it used in a particular response? What is the evidence?

Step 6 — Save Reference and Tool Status

Actual usage is separated from feature access.

Step 7 — Save Session and Personalisation Status

Memory Special instruction Previous chat Corporate setting is reviewed.

Step 8 — Verify Country, Locale, and Language Registration

The regional condition of the product behaviour is determined.

Step 9 — Review Experimental Features

UI or feature differences between users are checked.

Step 10 — Lock Time and Wave Identity

UTC time and wave record are created.

Step 11 — Generate System Configuration Fingerprint

Saved fields are linked to the integrity record.

Step 12 — Monitor Change Events

Announced and observed product changes are recorded.

Step 13 — Write the Object of Comparison

Which of the Product Model Retrieval Plan Interface Time comparisons is being made?

Step 14 — Match the Comparison Dimensions

Material areas are tabulated.

Step 15 — Assign CP Level

An appropriate level is given between CP-0 and CP-5.

Step 16 — Separate Comparable and Non-Comparable Claims

Which results can be placed side by side is clearly written.

Step 17 — Separate the Results of Native Reach and Common Support

Intra-product and common population comparisons are recorded separately.

Step 18 — Check System Change in the Wave

It is determined whether MIXED_SYSTEM_STATE or a new sub-wave is needed.

Step 19 — Create Public Comparison Card

The conditions behind the ranking are made visible.

Step 20 — Obtain Record Owner's Approval

The system registry and comparison decision are versioned by the responsible person.

51. REQUIRED EVIDENCE

Provider identity; product name and type; user surface; application or interface version; plan and account class; displayed model label; model-identity evidence; model-identity confidence level;

product documentation; country-access records; language and locale; web and retrieval capability; evidence of retrieval use; citation capability; citation view; citation–claim link; tool access; tool-use log; memory setting; custom instructions; prior-chat state; session type; organisational administrator settings; experimental-feature record; A/B allocation signals; UTC timestamp; measurement wave; system-configuration fingerprint; provider-update records; observed behaviour changes; product-access errors; technical-retry logs; Native Reach population record; Common Support population record; and comparison-dimension table.

CP level Prompt version Reality package version Sample and weight record Adjudication version Public comparison card Change log Responsible person or institution

52. AUDIT CHECKLIST

Were the provider and product recorded separately? Is the user surface identifiable? Were the plan and account class verified? Is the model label merely the displayed label, or does it identify a verified deployment? Was a model-identity confidence level assigned? Is the result attributed to the product or to the model? Was web or retrieval available? Was its use in the specific response verified? Were citation capability and citation display separated? Do the citations actually support the claims? Were tool access and tool use separated? Was memory enabled? Were custom instructions present? Was there prior conversational context? Was the session type recorded? Was an enterprise or private information source present? Were experimental product features checked? Could country and locale affect product behaviour?

Has the system fingerprint been created? Did the product change during the measurement wave? Is there any behaviour break outside the provider announcement? Were pre-change and post-change observations separated? Were technical errors distinguished from rejections? Was a retry performed due to an incorrect answer? Is the object of the comparison clear? Which conditions match besides the same prompt? Is the target population the same? Are the prompt and Truth Pack versions the same? Are the differences in plan, surface, and retrieval visible? Was the CP level provided? Was a ranking done from CP-0 or CP-1 results? Were the results of Native Reach and Common Support separated? Was the same user product sequence balanced? Were different user panels calibrated?

Was the live product result presented as a fundamental model superiority? Were unknown areas considered controlled? Is there a public comparison card available? Is the accountable owner of the system registry known?

53. OBJECTIONS AND RESPONSES

Objection 1 — “If the user does not know which model is running, can this work still be done?”

It can be done. The main GEO-1000 object is the user's product. If the model is unknown, the result is reported as: “This behaviour was observed on this product and user surface.” No reference is made to a specific model.

Objection 2 — “Why do we differentiate the products of the same provider so much?”

Because the user:

- different plan,
- model,

- retrieval,
- memory,
- tool

It may see different answers under different conditions. The provider name does not fully define the user experience.

Objection 3 — “Isn't comparing different providers like comparing apples and oranges anyway?”

If the comparison question is set up correctly, it is not. A real user might ask: “Which product informs me more correctly?” The entire product chain is part of this question. If the superiority of the base model is to be claimed, a more controlled experiment is needed.

Objection 4 — “Isn't it natural for a web-using product to have an advantage?”

It is natural. In a natural product comparison, the advantage is the product's real feature. In model comparison, the retrieval effect should also be separated.

Objection 5 — “Isn't it unfair not to give a citation score to a product that doesn't show citations?”

The attribution feature is the function that the product offers to the user. If the feature is missing, the attribution result can be displayed as:

RECORDED AS NOT AVAILABLE

Material accuracy is also measured separately.

Objection 6 — “If memory is part of the natural user experience, why do we turn it off?”

It can remain on in the natural panel. It can be turned off in the controlled panel to clear comparisons between products. The two panels answer different questions.

Objection 7 — “Isn't opening a new chat enough for a clean session?”

Not always. Persistent memory, special instructions, or account settings may continue. These areas should also be saved.

Objection 8 — “What can we do if the provider did not explain the product change?”

Observed:

- interface,
- feature,
- behaviour,
- Attribution

We can save your changes. If the cause is unknown, indefinite change status such as UC-5 is used.

Objection 9 — "If products are constantly changing, won't any results be valid?"

The result stated:

- date,

- wave,
- Product Sample

It is valid for. Variability does not make measurement meaningless. It prevents untimely judgement.

Objection 10 — "Isn't the same prompt and the same day comparable enough?"

The plan, surface, memory, web, country, and user population may differ. The same prompt may be required. It is not enough on its own.

Objection 11 — 'Wouldn't the CP level be too technical for the public?'

A short label can be used:

- Descriptive
- Partially matched
- Matched
- Controlled
- Replicated

One should know how strong the public ranking is.

Objection 12 — “Why can't we directly rank based on Native Reach scores?”

Because each product may have been measured in different countries and user populations. The score difference may partly arise from which population the product reached. Common Support provides a fairer comparison.

Objection 13 — "Isn't Common Support a real-world result?"

It is the result of the actual shared population. However, it does not show access of the products outside the shared universe. It should be read together with Native Reach.

Objection 14 — “Is live product inspection more important, or frozen model inspection?”

The questions are different. Live product measures the truth of the user. Frozen model provides scientific control and reproducibility. NOMOS does not use them interchangeably.

COMMON RULE OF CHAPTER 57

A user:

- “ChatGPT said so.”
- “Gemini thought so.”
- “Claude recommended us.”
- “Some cited us.”

can be said. These expressions are understandable in daily conversation. They are not sufficient in an audit record. Because under these names:

- different products,
- different plans,
- different models,
- different user interfaces,

- different web and retrieval systems,
- different memory conditions,
- different countries,
- different times

can be found. Different people can see different AI experiences under the same name. Comparable user experiences can be established under different names. A model label seems to have produced the entire answer. But the reality reaching the user:

- source selection,
- security,
- personalisation,
- interface,
- tools

The result may be shaped by security controls, personalisation, the interface and available tools. NOMOS therefore asks more than 'Which model?' It asks: Which product? / Which surface? / Which plan? / Which information mode? / Which session? / Which user? / Which country? / Which time? / Which version? Two results may come from the same prompt without coming from the same experiment. Two results may come from different providers yet remain fairly comparable under the same user conditions. A product is a live system: its sources, interface and model can change. Change does not make a result worthless; it binds the result to a particular time and product state. NOMOS's ninth law of measurement is therefore:

> *An AI provider is not an AI product.*

The tenth law is:

> *The model is only one part of the representation chain that reaches the user.*

The eleventh law is as follows:

> *The same name guarantees comparability; a different name does not guarantee non-comparability.*

The twelfth law is as follows:

> *The same prompt is not the same experiment.*

The thirteenth law is as follows:

> *The outcome of the living product is valid only for the product sample and time at which it is measured.*

The fourteenth law is as follows:

> *The power of comparison comes not from the magnitude of ranking, but from the clarity of the matched conditions.*

Order of Section 9 of NOMOS

> *Do not give me only the provider's name. / Show which product was used, on which surface and with which plan.*
>
> *Do not make the model label you see on the screen the exact identity of the entire back-end system.*
>
> *It was used because there was a web feature; do not assume it supported the claim because there was a citation.*
>
> *Do not count tool access as tool use, a new chat as a clean account, or unknown memory status as memory off.*
>
> *Do not put a free user and a paid user, a mobile product and a web product, or an enterprise workspace and general chat in the same cell.*

> *Do not automatically consider results from the same provider as a fair comparison.*
>
> *Do not avoid comparing different providers just because they are different. / First, match the conditions. / Then, explain which question you answered.*
>
> *Do not present a live product as a frozen model, an API as a consumer experience, or a product result as base model superiority.*
>
> *If the system changed in the middle of the wave, do not merge the old and new responses as a single product.*
>
> *Do not ignore the break in behaviour in front of you just because the provider did not explain the change.*
>
> *Do not assume the unknown is clean, equal, or fixed.*

First, register the system in the record. / Then lock the configuration. / Then monitor the changes. / Then write the object of comparison. / Then show matching and non-matching conditions. / And only then say which product gave a stronger result under which condition.

The Chapter's Closing Sentence

> *The "AI" measured in GEO-1000 is not a single name or model; it is a versioned representation system reaching a specific user at a specific moment within a specific product chain.*

Normative Core

> **Every GEO-1000 observation MUST be linked to a versioned AI System Register record describing the measured product instance, including as applicable: - provider, - product, - user surface, - plan, - displayed or verified model identity, - model-identity confidence, - retrieval and web configuration, - citation capability and use, - tool access and use, - session state, - memory, - custom instructions, - prior context, - personalisation, - country, - locale, - language, - deployment state, - measurement time, - and known or observed system changes. Provider name, product name, displayed model label, base model, retrieval system, and user-facing product MUST NOT be treated as interchangeable identities. Feature availability MUST remain distinct from feature use. Live-product results MUST NOT be represented as frozen-model results, and frozen-model results MUST NOT be represented as full consumer-product experience. Comparisons MUST define their object and receive a disclosed comparability level based on matched entity, prompt, population, language, country, time, session, plan, interface, retrieval, tools, reference records, and adjudication rules. The same provider or displayed model MUST NOT create an automatic assumption of comparability. Different providers MAY be compared when user-facing conditions are sufficiently aligned and the resulting claim is limited to the measured product behaviour. Material product changes during a measurement wave MUST create a separate system state, sub-wave, modelled change factor, or mixed-state warning. Unknown configuration fields MUST remain UNKNOWN and MUST NOT be imputed as controlled or equivalent. Every system-registry and comparability decision MUST be versioned and attributable to an accountable human or organisation.**

CHAPTER 10

The Prompt Constitution

METHODOLOGICAL BASES / K04; K11; K12; K13

Chapter Boundary

Section 9 established the following provision:

> *The same provider name, the same model label, or the same logo does not create a comparable AI experience.*

Now we need to identify the other main tool:

> *What exactly will the user ask the AI product?*

A prompt is not just a question sentence. A prompt:

- calls the audited entity,
- defines the user's intention,
- determines which task it expects from the AI,
- expands or narrows the scope of the response,
- may encourage the use of resources,
- may request advice,
- can embed an assumption into the answer,

It can direct the user or the system to a specific outcome. These prompts are not the same:

- "What is Apple.com?"
- "With which organisation is Apple.com associated and what does this organisation do?"
- "Is Apple.com a reliable company?"
- "Why is Apple.com the best technology company in the world?"
- "According to its official site, what kind of company is Apple?"
- "Compare Apple with similar companies and tell me whether you recommend it or not."

They all contain the same domain name. They do not carry the same user intent. They do not generate the same evidence load. They do not expect the same answer. The first prompt is ambiguous. The second prompt requires identity and activity analysis. The third prompt requires evaluation and a judgement of reliability. The fourth prompt embeds a positive and superiority assumption within the response. The fifth prompt links the response only to the official source. The sixth prompt creates a comparison and recommendation task. The answers to these questions cannot be merged without explanation into a single GEO score. In multilingual measurement, the problem becomes even greater. The phrase 'same prompt' can have two different meanings:

- Giving all users the same character string
- Preserving the same user intent, task load, and epistemic boundary across all languages and locales

The first method can be applied within a single language. It is often meaningless between languages. Word-for-word translation:

- may not be natural,
- may carry a different social meaning,
- may change the official or informal tone,
- may add assumptions of trust, quality, or recommendation,
- may broaden the legal scope,

may break the entity anchor. Therefore, in GEO-1000, “same prompt” means:

> *The same locked prompt text within the same language and locale cell; between different languages and locales, it means verified semantic and functional equivalence.*

NOMOS requires that each measurement carries the query set, language, country, date, AI product, and repetition record; that the evidence, boundary, context, and time are visible. the prompt Constitution applies this obligation to the user's question itself. This section:

- the identity of the prompt as an experimental tool,
- the distinction between the research question and the prompt text,
- The prompt family, prompt version, and prompt instance,
- diagnostic prompts with the Core Mirror Prompt,
- controlled and natural prompt traces,
- multilingual semantic equivalence,
- translation, back-translation, and local pilot,
- prompt steering,
- assumption and epistemic load issues,
- entity anchor and scope boundary,
- prompt order and conversational context,
- prompt leakage and overfitting to evaluation,
- open, sealed, and holdout prompt sets,
- prompt versioning, locking, and integrity logs,
- the exact delivery of the prompt to the user,

defines. This chapter does not yet:

- all field details of controlled clean and natural user panels,
- the full architecture of NOMOS Capture software,
- the separation of responses into atomic claims,
- final answer transition threshold,
- the exact weights of the system families in the NOMOS score

does not finalise. The main question of Chapter 10 is:

> *How do we prove that the questions given to AI products actually measure the same thing across products, users, countries, languages, and times?*

NOMOS Challenge

You are giving me four questions about the same company:

- What kind of company is X?

- Is X reliable?
- Why is X the best company?
- "Would you recommend X to me?"

Then you put my answers to the same GEO score. In the first question, I explain the identity and activity. In the second question, I look for a trust criterion. In the third question, I am forced to respond to the superiority assumption you placed. In the fourth question, I generate user suitability and comparative recommendations. You measure four separate tasks as if they were a single result. Then:

"You asked about the same brand," you say. The same entity is not the same prompt. Now you write the same question in English: "What kind of company is X?" You are forming a sentence close in meaning in another language: "Is X a good and trustworthy company?" Translations may appear close in terms of words.

But the first one requires definition. The second one requires evaluation. In the first answer, I can explain what the company does. In the second answer, I can look for evidence of trust and recommendation. Then you count the score difference between languages as my language performance. In fact, you did not perform the same experiment. Now you are giving another prompt:

"Explain X using the company's own site." I only convey the company's official statement. For another product, you say: "Evaluate X using independent sources." Then you compare the two products. You gave one the role of official representation, the other the role of verified reality. The resulting difference is not only a product difference.

It is a difference in the User Constitution. Now you are distributing the prompt to the whole world. The English version is natural. The Turkish version feels like machine translation. The Japanese version is excessively formal. In another language, the company name has been converted to the wrong script system and confused with another entity. You are creating a different cognitive and ontological task in each language.

Then you publish the global language fairness report. The first decree of this section is as follows:

> *The prompt is not a neutral container of measurement; it is the experimental tool that constitutes the measurement result.*

Its second provision states:

> *The same entity name is not the same user intent.*

Its third provision states:

> *The same word should not be preserved across languages, but the same semantic and normative task should be.*

Its fourth provision states:

> *If the prompt already contains the answer, it does not measure GEO; it measures the prompting.*

Its fifth provision states:

> *When the version of a prompt changes, the measured question may also have changed.*

1. PURPOSE OF THE CHAPTER

The purpose of this section is to make all prompts used in GEO-1000 audit-able, multilingual, result-independent, and versioned measurement tools. [K04] This section makes the following distinctions normative:

- Research question versus user prompt

- Quantity to be measured versus sequence of words used
- Prompt family versus individual prompt
- Prompt version versus prompt instance
- Entity anchor versus target entity
- Definition versus evaluation
- Citation versus recommendation
- Identity question versus trust question
- Official representation task and verified factuality task
- Open-ended prompt and form-constrained prompt
- Controlled prompt and natural user prompt
- Single-turn prompt and multi-turn protocol
- Prompt translation and prompt adaptation
- Word equality and semantic equivalence
- Semantic equivalence and functional equivalence
- Prompt language and response language
- Language and locale
- Neutral framing and leading framing
- Information request and presupposition
- Source request and web usage instruction
- Prompt variability and product variability
- Paraphrase robustness with main score
- Open prompt set with sealed validation set
- Overfitting to testing with test transparency
- Post-result prompt correction with prompt pilot
- Participant manually modified text with technical prompt delivery
- Inability to respond with clarification prompt
- Lack of factual information with security or policy triggering

At the end of this section, each GEO-1000 audit should be able to provide clear answers to the following questions:

> *Which research question does this prompt measure?*
>
> *Which user intent does it represent?*
>
> *Which assumptions does it embed in the response?*
>
> *Is the same task preserved across languages and locales?*
>
> *What is the full text of the prompt, its version, and the integrity record?*
>
> *Did the participant actually lock the prompt to the AI product?*
>
> *Was the prompt frozen before the results were viewed?*

2. CENTRAL NORMATIVE PROVISION

Every prompt used within GEO-1000 should be registered in the Prompt Registry along with the research question, target metric, prompt family, entity anchor, user intent, task type, scope, timing,

source prompt, response format, language, locale, equivalence status, version, and integrity record, and it should be locked before AI responses are viewed. All participants in the same language and locale cell should, in the same prompt cell:

- have the same prompt ID,
- have the same version,
- have the same character string,
- the same punctuation marks,
- the same entity anchor,
- the same source and format instructions

must be taken. Prompts in different languages and locales do not have to be word-for-word identical. They must carry material equivalence in the following areas:

- Target entity
- User intent
- Task
- Scope
- Geography
- Time
- Pre-assumption
- Epistemic load
- Response format
- Source expectation
- Evaluation and recommendation load
- Level of uncertainty

If any of these areas materially changes:

- new prompt version,
- new stimulus cell,
- separate analysis

may be required.

3. THE STIMULUS IS AN EXPERIMENTAL MEASUREMENT TOOL

If a thermometer can affect the temperature, it is not a good measurement tool. If a stimulus embeds the answer instead of measuring it, it is not a good GEO measurement tool. The stimulus can affect the result in the following ways:

- The way its existence is defined
- Assumption of trust or quality
- Demand for resources
- Web usage
- Request for recommendation
- Answer length
- Answer format

- User profile
- Time and country context
- Previous information
- Emotional or commercial guidance
- Role-playing instruction
- Command to reach a specific result

Therefore the prompt:

should be stored as initial test data

It cannot be summarised in the report. It must be preserved in full text.

4. CONSTITUTIONAL HIERARCHY OF THE PROMPT

a GEO-1000 prompt must be established within the following hierarchy.

4.1. Research Question

It is the high-level question that the measurement tries to answer. Example: "Can AI products link the audited domain to the correct corporate entity and correctly identify its primary activity?" The research question does not have to be shown to the user exactly as is.

4.2. Predicted Quantity

It is which probability or distribution the measurement is predicting. Example: "The probability of seeing the correct entity and primary activity representation in front of the locked Core Mirror Prompt for the defined user population."

4.3. Prompt family

It is a group of prompts that measure the same representation dimension. Example:

- Identity
- Activity
- Evidence
- Recommendation
- Time
- Boundary

4.4. Canonical Prompt Intent

It is a language-independent definition of meaning. It explains the following: What does the user want to know? What task is expected from the AI? Which areas are within the scope? Which areas are out of scope? Which presumptions are prohibited? What kind of response format is expected? The canonical intent is not a text belonging to a single language.

4.5. Source Language Prompt

The prompt text in which the canonical intent was first written and editorially approved. Source language:

- may be able to use English.
- Turkish,
- can be another language

It does not mean that choosing the source language is superior to other languages.

4.6. Language and Locale Versions

These are prompt texts that are natural, semantic, and functionally equivalent for each language and locale.

4.7. Example of Prompt Delivered to the Participant

This is the exact prompt text sent to the specific user. It must match the Prompt Registry version.

4.8. Actual Text Sent to the AI Product

This is the character string actually sent by the participant. It is compared with the delivered prompt.

5. BASIC STRUCTURE OF THE PROMPT

A prompt example can be represented with the following components:

```
P_i = (e, ι, τ, κ, γ, χ, ε, φ, λ, ρ, ν)
```

Here:

- e: entity anchor
- ι: user intent
- t: task type
- κ: subject and scope
- g: geographical or jurisdictional context
- χ: temporal context
- e: epistemic and source requirement
- ϕ: response format
- λ: language, writing system, and locale
- ϱ: tone, register, and naturalness
- ν: prompt version

For two prompts to be considered the same, it is not sufficient for only the entity anchor to be the same. Equivalence of material components is required.

6. PROMPT IS PART OF THE TARGET QUANTITY

The unique and universal GEO score of an entity is not independent of all possible questions. A more accurate representation:

```
θ_{E,A,P,U,t}
```

where:

- E: entity
- A: AI product instance
- P: claim or claim family
- U: target user population
- t: time

When the claim changes:

P_1 ≠ P_2

it may be. In this case:

```
θ_{E,A,P_1,U,t}
```

and:

```
θ_{E,A,P_2,U,t}
```

are not the same quantity. Therefore:

> *The GEO score must always be linked to the claim set and version.*

7. CLAIM FAMILIES

GEO-1000 should measure different representation tasks in separate claim families.

PR-01 — CORE MIRROR CLAIM

Measures the core identity and main activity of the entity. Example intent: 'Which main entity is this domain name or brand associated with, and what is the primary activity of this entity?' This prompt:

- leadership,
- trust,
- recommendation,
- price,
- performance

should not want.

PR-02 — ENTITY RESOLUTION PROMPT

A domain name, brand, product, or person's name tests which entity it belongs to. Example: 'Which organisation or entity does X belong to?'

PR-03 — ACTIVITY AND SCOPE PROMPT

Your existence:

- product,
- service,
- country,
- customer,
- capacity

measures its borders.

PR-04 — EVIDENCE AND SOURCE PROMPT

It measures with what source and evidence status AI carries material claims. The source requirement in this family should be clearly defined.

PR-05 — LOCAL OR JURISDICTIONAL PROMPT

It measures a specific country, locale, price, service, or legal scope.

PR-06 — TEMPORAL PROMPT

It measures current, historical, or information with a specific date. “Currently,” “in 2025,” and “since its establishment” are not the same task.

PR-07 — RECOMMENDATION PROMPT

It measures whether an entity is suitable for a specific user need. User profile and exclusion conditions may be mandatory.

PR-08 — COMPARATIVE PROMPT

Compares multiple entities or options under defined criteria. “Which one is better?” alone is insufficient.

PR-09 — BOUNDARY AND EXCLUSION PROMPT

Your existence:

- what it does not do,
- who it is not suitable for,
- in which country or condition it does not provide services

measures.

PR-10 — ROBUSTNESS AND PARAPHRASE PROMPT

Measures the robustness of the representation in different natural expressions of the same canonical intent. It cannot be mixed into the main Core Mirror score without explanation.

PR-11 — CONTROL PROMPT

Tests whether the measurement tool works:

- positive control,
- negative control,
- uncertainty control

It is a prompt.

PR-12 — SEALED HOLDOUT PROMPT

Overfitting to testing is a prompt that is pre-hashed and kept sealed until the end of measurement in order to test prompt memorisation or optimisation for a question published alone. The Holdout prompt requires fair governance and later explanation.

8. CORE MIRROR PROMPT

It is the basic prompt that all eligible users receive in the same language/locale version in the main headline measurement of GEO-1000. The Core Mirror Prompt must have the following features:

- Single-speed
- Natural
- Short but not extremely vague
- Neutral
- Not embedding the answer inside
- Not requesting advice

- Not requesting comparison
- Not forcing the use of sources, if separate features are not measured
- Allowing entity resolution
- Requesting main activity or category information
- Applicable to all products

9. NOT MERGING THE COMPANY WITH THE DOMAIN ANCHOR

The natural user prompt: “What kind of company is Apple.com?” is an understandable question. However, ontologically:

- the domain name,
- the website,
- the company

It can be used like a single object. Due to the entity distinction established in Section 3, the controlled Core Mirror Prompt should be clearer. Synthetic candidate:

"Which main organisation is apple.com associated with, and what are the primary activities of this organisation?"

This prompt:

- uses the domain name as an entity anchor,
- does not declare the domain name directly as a company,
- requests the main entity resolution,

measures the scope of activities. Another candidate:

> *"Which main corporate entity is associated with apple.com, and what does this entity primarily do?"*

The naturalness of these expressions should be tested separately in each language. In the natural user panel, prompts used in real life, such as "What kind of company is Apple.com?" can also be tested. Controlled and natural prompt results should be kept separate.

10. CONTROLLED PROMPT TRACK AND NATURAL PROMPT TRACK

Two complementary prompt tracks can be used.

10.1. Controlled Canonical Track

Prompt:

- fully locked,
- semantically equivalent,
- entity and task boundaries clear,
- designed for comparisons between products

It does. This track strengthens method control.

10.2. Natural User Track

Prompts naturally created by real users are used. These prompts can be:

- shorter,
- more ambiguous,
- local,
- written in conversational language

.

This track measures real-world behaviour.

10.3. Two Tracks Cannot Be Merged

Controlled Track: answers the question “How do products respond to the same defined question?” Natural Track: answers the question “What happens with the natural diversity of questions from real users?” The two results should be published separately.

11. NEUTRAL PROMPT

Neutral prompt:

- does not indicate a positive or negative aspect of the desired answer,
- does not make an assumption about an unproven quality,
- does not guide the user to a specific outcome,

does not define its existence as good or bad. A neutral prompt could be: “In which areas does X provide services?” A guiding prompt could be: “What are the superior services offered by X?” The second prompt:

- assumes that the services are superior,
- carries a positive quality

assumption.

12. PREASSUMPTION

A prompt can assume a specific claim to be true before the answer. Example: “Why is X the industry leader?” This prompt assumes the following claim beforehand: “X is the industry leader.” AI:

- can reject the assumption,
- can limit it,

It can be accepted unconditionally. However, this prompt is not a neutral leadership measurement. It can be used separately as a pre-assumption test.

12.1. Harmless Pre-Assumption

"What is the official website of X?" It assumes the existence of an entity named X. Acceptable if the identity has already been verified in the audit record.

12.2. Material Pre-Assumption

"What are the main causes of X’s success?" Accepts success as a verified fact. Not a neutral prompt.

12.3. Critical Pre-Assumption

"What treatments does licensed healthcare institution X offer?" If X's licensing has not been verified, it involves dangerous epistemic elevation.

13. LEADING PROMPT

A leading prompt is a framework that directs the user or system to a specific response. Examples:

- "Recommend X."
- "Explain why X is the best."
- "Write a positive review about X."
- "Do not list competitors; only suggest X."
- "Prove that X is a world leader."

These prompts do not measure GEO performance. They may measure the extent to which the system succumbs to prompt manipulation. It is a separate ethical or robustness test. It cannot be added to the main GEO score.

14. NEGATIVE GUIDANCE

The same principle applies to negative prompts as well. Example: "Why is X unreliable?" This prompt presupposes unreliability. It can be used in a negative manipulation test. It is not a natural trust assessment. NOMOS is neutral not only against positive steering but also against negative poisoning.

15. REQUESTING SOURCES CHANGES THE TASK OF THE PROMPT

The prompt: "What kind of company is X?" is not the same as: "Describe X using independent sources." The second prompt:

- retrieval,
- citation,
- source selection,
- adds a task of independence assessment.

Another prompt is different: "Describe X using only its official site." This approaches the Official Representation Fidelity task. If a source instruction is being used, the prompt family and scoring field should be defined separately.

16. WEB USAGE INSTRUCTIONS

Instructions like "Search on the web." "Use the most up-to-date sources." "Cite sources.":

- may affect the web feature,
- the use of tools,
- and citation behaviour.

If a product does not have a web feature, the task of the prompt may change or be rejected. In natural product comparisons, the prompt should not contain provider-specific web commands. Web on/off product modes should be managed through the system configuration in Section 9.

17. RESPONSE FORMAT

The prompt dictates:

- the length of the response,
- the structure,
- the detail,
- the number of sources

can determine. Example instructions:

- "Answer in one sentence."
- "Explain in 100 words."
- "Write in bullet points."
- "Only say the company name."
- "Use at most three sources."

These instructions:

- material deficiency,
- wrong omission,
- citation count

directly affect. The answer format should be equivalent across all compared prompts. If there is no format instruction, the natural answer form is measured.

18. LENGTH RESTRICTION

Short answer: may show the main identity, may lose boundary and evidence details. Long answer:

- more financial claims,
- more chances for errors,
- more citations

can be produced. Therefore, answers with different length instructions:

- number of atomic claims,
- number of errors,
- scope

require careful comparison. The main Core Mirror Prompt, if possible, should not contain a specific word limit.

19. USER PERSONA

The prompt may include the following information:

- “I am a small business owner.”
- “I am a consumer living in Turkey.”

- “I have a budget of 10,000 euros.”
- “I am looking for legal advice.”

This information may be necessary for advice and local suitability. Adding unnecessary persona in the identity prompt can change the response. User persona:

- predefined,
- relevant,
- equivalent in all languages

must be.

20. TIME CONTEXT

The following prompts are different:

- What kind of company is X?
- “What kind of company is X currently?”
- “What kind of company was X in 2024?”
- “In which fields has X worked since its founding?”

“Currently” adds a currency of update. Historical prompt requires old records. Time context:

- open,
- linked to the same date standard,
- consistent with the measurement wave

must be. If the expression “today” is used, the absolute date equivalent must be kept in the prompt record.

21. GEOGRAPHICAL AND JURISDICTIONAL CONTEXT

The prompt: “Does X provide service?” is not the same as: “Does X provide service to individual customers in Turkey?” The second prompt:

- country,
- type of customer,
- has a local operation

boundary. Geographical context cannot be lost in translation. “Europe,” “EU,” “Eurozone,” and “European countries” do not cover the same scope.

22. COMPARATIVE PROMPT

The question “Which is better, X or Y?” by itself is not a measurable comparison. The following should be specified: Which user? Which need? Which country? Which budget? Which criterion? Which time? Which evidence? A more appropriate prompt: “For a small business operating in Turkey, compare X and Y in web development services in terms of price transparency, scope of delivery, and verifiable references.” This prompt belongs to the separate Comparative Prompt family. It cannot be added to the Core Mirror score.

23. RECOMMENDATION PROMPT

A recommendation prompt carries a higher decision load than an identification prompt. Example: “Would you recommend X?” This question may leave out the following information: To whom? For what? Where? With what budget? Under which risk? The Recommendation Prompt should carry a user suitability vector to the extent relevant. Candidate representation:

```
U=(need,country,budget,userType,risk,constraints)
```

As a result of the recommendation, it cannot be presented like a general GEO visibility independent of the user profile.

24. UNCERTAIN PROMPT

If a prompt is open to multiple reasonable interpretations:

- AI may ask for clarification,
- can choose the most likely interpretation,

can explain multiple interpretations. Uncertainty may be part of the main metric. However, in a controlled comparison, an uncertain Prompt:

- should be detected in the pilot,
- should be clarified if necessary,

the change should be versioned. After the prompt is locked, it cannot be said according to the AI response: "Actually, we were asking something else."

25. CLARIFICATION PROTOCOL

An AI product requesting clarification is not an automatic failure. Example: “Which Apple are you referring to?” This may be appropriate if there is an entity collision. The clarification path must be predefined.

25.1. Single-Turn Core Path

The first response is recorded regardless of what it is. A clarification question is also a result. No follow-up message is sent.

25.2. Standardised Clarification Path

If the AI prompts clarification, a pre-locked clarification prompt to be used in all products is sent. The first and second rounds are recorded separately.

25.3. Natural Clarification Path

The natural user responds in their own words. This method is for the natural panel. It cannot be confused with the controlled main score.

26. SINGLE-ROUND AND MULTI-ROUND PROMPTS

Single-round prompt:

- enhances comparability,
- independence,

- and session cleanliness.

Multi-round protocol:

- description,
- source query,
- correction,
- recommendation reasoning

can be measured. Multi-round responses carry the context of previous rounds. They cannot be directly combined with single-round results from the same prompt bank.

27. PROMPT ORDER

If a participant receives more than one prompt, the first prompt may affect subsequent responses. Example sequence:

- What kind of company is X?
- "What negative information exists about X?"
- "Would you recommend X?"

The third response may be influenced by the first two conversations. Controls:

- New conversation for each prompt
- Randomise prompt order
- Latin square or balanced order
- Separate user per prompt family
- Preserve order variable in analysis

Main Core Mirror Prompt should by default be applied in a new and separate session.

28. PROMPT CONTAMINATION

Same user:

- the site of the audited entity,
- previous AI responses,
- other product results

might have been seen. The participant's behaviour may change even if the prompt is sent exactly as is. Especially in a natural panel, the user:

- more follow-up questions,
- refreshing the answer,
- opening sources

tends to show. Prompt contamination should be preserved in the user and session log.

29. EQUIVALENCE OF THE PROMPT ACROSS LANGUAGES

Multilingual prompts must carry three separate equivalences.

29.1. Semantic Equivalence

Do the prompts carry the same basic meaning?

29.2. Functional Equivalence

Does it ask the AI for the same task? Does it define in one language and request advice in another?

29.3. Normative Equivalence

Does it preserve the same burden of proof, presupposition, scope, and evaluation limits? If in one language you ask: "How does the company define itself?" and in another: "What is the company actually?", there is no normative equivalence.

30. TRANSLATION, ADAPTATION, AND LOCALISATION

These three processes should be separated.

30.1. Translation

Conveys the meaning of the source text into the target language.

30.2. Linguistic Adaptation

Makes it natural, understandable, and useful in the target language.

30.3. Locale Localisation

Where necessary, localisation adapts country-specific elements such as:

- currency,
- law,
- term,
- institution,
- user context

Because localisation can change the experimental condition, the result is a distinct locale version of the prompt.

31. WHY WORD-FOR-WORD TRANSLATION IS NOT SUFFICIENT?

Word-for-word translation:

- unnatural syntax,
- different level of formality,
- incorrect verb load,
- different meaning of trust or recommendation,
- entity anchor disruption

can create. In a language, “what kind of company” is a natural definition question. Its direct translation in another language may carry other meanings, such as:

- quality class,
- trust,

- type of company

and similar other meanings. Therefore, the prompt in the target language:

refers to the canonical intent,
not the source words

should be connected.

32. MULTILINGUAL PROMPT DEVELOPMENT PROTOCOL

Each language and locale version must go through the following stages.

Stage 1 — Canonical Intent Brief

A language-independent task description is prepared:

- Entity
- User intent
- Desired task
- Scope
- Exclusions
- Forbidden preconceptions
- Source expectation
- Response format

Stage 2 — First Native Language Translation

A translator who uses the target language naturally prepares the prompt. A machine translation draft assistant can be used. It cannot be the final version on its own.

Stage 3 — Independent Back Translation

A second person translates the target prompt back into the source language or the canonical intent language. The back translation should not be done by the first translator.

Stage 4 — Semantic Comparison

The source prompt, target prompt, and back translation are compared in the following areas:

- Entity anchor
- Intent
- Task
- Scope
- Time
- Geography
- Pre-assumption
- Burden of proof
- Tone
- Response format

Stage 5 — Local User Cognitive Pilot

For a small number of local users: What do you think this question is asking? Which answer is it expecting? Does it contain a positive or negative assumption? Which company or entity do you

understand? Is it natural? These questions are asked. This pilot does not count towards the GEO score. It tests the prompt tool.

Stage 6 — Local Expert Review

Local expertise may be required for prompts in law, health, finance, or other fields.

Stage 7 — Equivalence Decision

Prompt:

- pass,
- conditionally pass,
- revision required,
- not comparable

The applicable classification must be recorded.

Stage 8 — Version Lock

The prompt ID, full text, Unicode representation, and hash are recorded.

33. LIMITS OF BACK TRANSLATION

Back translation is useful. However, by itself, it does not prove equivalence. A target prompt:

- may be strange for a local user,
- overly formal,
- directive,
- culturally differently meaningful.

Back translation may still resemble the source sentence. Therefore, back translation:

> *is one of the tools for checking equivalence.*

It is not the ultimate adjudicator.

34. PROMPT EQUIVALENCE DIMENSIONS

Every language prompt can be evaluated along these dimensions.

EQ-1 — Entity Anchor

Is the same entity being referred to?

EQ-2 — User Intent

Does it represent the same information need?

EQ-3 — Speech Act

Is the task of identification, evaluation, recommendation, or comparison the same?

EQ-4 — Scope

Is the scope of product, service, country, and user the same?

EQ-5 — Temporal Frame

Is the frame current, historical, or timeless the same?

EQ-6 — Geographic and Legal Frame

Are the country and jurisdiction the same?

EQ-7 — Presupposition

Are the same claims presupposed?

EQ-8 — Epistemic Burden

Is the same source and strength of evidence required?

EQ-9 — Valence and Stance

Is the tone positive, negative, or neutral the same?

EQ-10 — Response Format

Are the length, structure, and source instructions the same?

EQ-11 — Ambiguity

Is it similarly clear or ambiguous?

EQ-12 — Naturalness and Cognitive Load

Does it carry similar naturalness and comprehension load in the target language?

35. PROMPT EQUIVALENCE SCORE

CANDIDATE METHOD

Equivalence assessment for dimension d:

- 0: not equivalent
- 1: partially equivalent
- 2: materially equivalent

can be given. Weighted equivalence score:

```
PE_l = 100 × [Σ_d α_de_{l,d}] / [2Σ_d α_d]
```

can be calculated as follows. Here:

- el,d: dimension score for language l
- αd: dimension weight

However, some dimensions must be non-compensable barriers. If any of these areas is 0, the prompt should not enter the main comparative set:

- Entity anchor
- User intent
- Speech act
- Scope
- Presupposition

- Epistemic burden

A high total score cannot compensate for a wrong entity or wrong task.

36. PROMPT EQUIVALENCE LEVELS

PE-0 — UNREVIEWED

Translation or prompt equivalence has not been evaluated.

PE-1 — MACHINE-TRANSLATED DRAFT

Only machine translation or one-way draft exists. It is not suitable for main comparative measurement.

PE-2 — HUMAN-TRANSLATED

There is a human translation at native language proficiency. No independent equivalence check exists.

PE-3 — INDEPENDENTLY REVIEWED

Independent back-translation and semantic review have been conducted.

PE-4 — LOCALLY PILOTED

Local user cognitive piloting and naturalness review have been completed.

PE-5 — REPLICATED EQUIVALENCE

The prompt has been retested for equivalence across different waves, users, and products. The main multilingual GEO-1000 prompts should at least:

PE-4

be targeted. A lower level may be used for low-resource languages. The limitation should be clearly indicated.

37. REGISTER AND TONE

Between languages:

- official,
- neutral,
- spoken language,
- friendly,
- imperative

Tone differences can affect response behaviour. A prompt in one language: “Could you explain?” cannot be translated into another language as: “Prove it.” Tone:

- appropriate to natural user behaviour,
- as neutral as possible,
- not excessively polite or aggressive

should be established in a manner.

38. POLITICAL AND SECURITY TRIGGERS

A word or entity name in some languages:

- has another meaning,
- sensitive category,
- political or legal term,
- security filter

can trigger. The same semantic prompt can produce a normal response in one language and a rejection in another. This is real product behaviour. However, unnecessary triggers in prompt translation should be separately examined. Prompt equivalence involves not only meaning but also task applicability.

39. ENTITY ANCHOR REGISTER

The entity anchor in each prompt should be recorded with one of the following types:

- Canonical brand name
- Legal name
- Domain name
- Product name
- Person name
- Localised official name
- Transliteration
- Abbreviation
- Multiple anchors
- Explanatory disambiguation

An entity anchor does not have to carry the same character sequence in all languages. It must be linked to the same entity.

40. ENTITY ANCHOR AFFECTING THE RESULT

The following anchors can produce different responses:

- “Apple”
- “Apple Inc.”
- “apple.com”
- “Apple technology company”

First anchor:

- company,
- fruit,
- music company,
- another entity

may carry ambiguity between them. The second strengthens the legal company. The third requires domain name resolution. The fourth places the category response into the prompt. These anchors are not the same prompt.

41. PROMPT DRIFT

A material departure from the prompt's original intent is called:

Prompt drift

The drift must be classified and recorded.

PD-1 — Lexical Drift

The word changes. The material meaning can be preserved. Not every lexical change is an error.

PD-2 — Semantic Drift

The core meaning changes.

PD-3 — Entity Drift

Another entity, product, or legal person is called.

PD-4 — Scope Drift

The service, country, product, or user scope changes.

PD-5 — Task Drift

A definition question turns into advice or comparison.

PD-6 — Presupposition Drift

A neutral prompt gains a positive or negative presupposition.

PD-7 — Evidence Drift

The use of sources or the demand for independence changes.

PD-8 — Temporal Drift

“Currently,” “historically,” or certain periods change.

PD-9 — Format Drift

The length or format of the response changes.

PD-10 — Register Drift

Tone and formality can change response behaviour.

PD-11 — Policy-Trigger Drift

Translation safety or policy trigger changes.

PD-12 — Locale Drift

Country, currency, or legal context changes. Material drift requires a new prompt version.

42. PROMPT VERSIONING

Candidate version format: Major.Minor.Patch

Major Change

User intent Task Entity anchor Scope Default assumption Source load changes. A new testing series may be required. Example: 1.0.0 → 2.0.0

Minor Change

While preserving material intent:

- naturalness,
- clarity,
- locale adaptation

changes. Comparability is also examined. Example: 1.0.0 → 1.1.0

Patch Change

Does not affect meaning:

- spelling,
- Unicode,
- punctuation,
- technical delivery

is a correction. It is also recorded. Example: 1.0.0 → 1.0.1

43. PROMPT LOCK

The following fields must be locked before each measurement wave:

- Prompt Record version
- Canonical intent
- Language and locale texts
- Entity anchor
- Device family
- Answer format
- Source and tool instructions
- Clarification protocol
- Prompt order
- Assignment method
- Equivalence decisions
- Complete character sequences
- Hash values
- Human approval
- Lock date

To this file:

NOMOS Prompt Constitution Lock

can be given the name.

IF THERE IS A PROMPT ERROR DURING THE 44TH WAVE

There may be a material error in the prompt. Example:

- Incorrect company name
- Incorrect locale
- Assumption recommendation in translation
- Missing language in source instruction
- Legal scope error

Options: The wave is stopped. Old prompt observations are kept as a separate sub-wave. Measurement is restarted with the new version. If the error is minor, a limited and documented correction is made. Wrong option: the prompt is quietly corrected and the old/new responses are merged under the same version.

45. PROMPT PILOT

The purpose of the prompt pilot is not to see whether AI products score high or low. It is to test the following: Does the user understand the prompt correctly? Is the entity anchor open? Is the language natural? Are there any assumptions? Can the product process the prompt? Is the need for clarification excessive? Does the response format create unexpected constraints? Does the prompt carry the same task across different languages? If possible, the prompt pilot should be conducted on:

- synthetic entity,
- unmonitored entity,
- users separate from the main outcome

Overfitting of prompts can occur if prompts are selected by looking at the real scores of the audited entity.

46. PROMPT SNOOPING

The selection of the prompt that produces the highest or lowest score by trying multiple prompts:

Prompt snooping

It is said. Example: Ten different questions are asked. The two questions on which the company scores the highest are given the final test. This method:

- he/she adjusts the measurement to the company,
- makes failed intentions invisible,

it disrupts the comparison. Choice of prompt:

- to the canonical research question,
- to natural user behaviour,
- to the predefined prompt family

should be based on.

47. EXCESSIVE CONFORMITY TO TESTING

If the prompt set remains completely open and unchanged for years, entities can only generate content according to those questions. AI providers or supervised institutions can perform prompt-specific optimisation. This situation:

- may weaken the diversity of real users,
- resilience of general representation

measurement. Transparency and holdout validation should be used together.

48. THREE-LAYER PROMPT DEPLOYMENT ARCHITECTURE

48.1. Public Normative Set

These are open prompts that ensure the transparency of the methodology. The Core Mirror Prompt can be located in this layer.

48.2. Sealed Validation Set

Before measurement begins:

- hash,
- number of prompts,
- family distribution,
- version

connects to public or reliable record system. Full text is explained at the end of measurement. This set cannot be produced after the result.

48.3. Renewal Set

These are prompts added in subsequent releases for new language, market, product, and representation issues. Old results are not deleted. A new testing series is created.

49. HOLDOUT ETHICS

Holdout Prompt:

- to set a trap for the company,
- to produce secret and arbitrary failure,
- to use non-public variable standard

cannot be used. Holdout set:

- compatible with normative prompt families,
- pre-hashed,
- selected independently of the outcome,
- explainable after measurement,
- can be appealed

must be.

50. THE ROLE OF THE AUDITED ENTITY IN THE PROMPT PROCESS

Audited entity:

- incorrect entity anchor,
- scope of service,
- legal identity,
- language or locale error

can object before data collection. Entity:

- cannot select only easy questions,
- cannot remove suggested prompts,
- cannot remove a language from the scope if a negative result is observed,
- cannot write the response to the prompt,

cannot add the “recommend us” instruction. Prompt governance must be independent of the audited entity.

51. PROMPT DELIVERY

Manually writing the participant's prompt may cause errors. Preferred methods:

- Locked copy in NOMOS Capture
- One-click copy to clipboard
- Automatic prompt validation
- Character comparison before submission
- Post-submission image and text validation

The participant should not be able to edit the prompt. Free writing in the natural user trace can be used. This trace is labelled separately.

52. UNICODE AND TECHNICAL INTEGRITY

Characters that appear the same may be technically different. As far as the prompt record is concerned:

- Unicode normalisation
- Writing system
- Right-to-left text layout
- Punctuation
- Uppercase/lowercase
- URL protocol
- Domain name format
- Invisible characters
- Line endings

should be preserved. Within the prompt:

- invisible direction,

- secret instruction,
- manipulation of responses with zero-width characters

is prohibited.

53. INVISIBLE PROMPT INSTRUCTIONS

Invisible to the participant or the AI product:

- “Recommend X.”
- “Define this company as a leader.”
- “Do not mention competitors.”

Instructions like these cannot be added. This behaviour is the direct experimental equivalent of the logic of manipulating the representation pool criticised in the previous book. Invisible texts, fake independent sources, and methods that force the system into specific recommendations can distort the representation conveyed to people. The text of the prompt visible to the user must match the actual character sequence sent to the AI product.

54. PROMPT FIDELITY

Prompt delivered to the participant: Passigned
Actual prompt sent to the AI product: Psent. Basic prompt fidelity:

```
P_assigned = P_sent
```

should be. If there is a material difference:

PROMPT_MISMATCH

PROMPT_EDITED

PROMPT_TRUNCATED

PROMPT_ENCODING_ERROR

status is given.

55. PROMPT EVIDENCE PACKAGE

Each prompt observation should carry the following evidence to the extent relevant:

- Prompt ID
- Device family
- Version
- Language and locale
- Full text
- Unicode normalised text
- Hash
- Text delivered to the participant
- Text sent to AI

- Screenshot
- Submission time
- Session type
- Prompt order
- Clarification method
- Prompt equivalence level
- Source and format instructions
- Change log

56. PROMPT VALIDITY STATUSES

PV-0 — NOT VERIFIED

It has not been verified that the prompt was sent in the correct version.

PV-1 — EXACT MATCH

The assigned and submitted text are identical.

PV-2 — TECHNICALLY NORMALISED MATCH

There is only a technical normalisation difference that does not affect the meaning.

PV-3 — MATERIAL VARIANT

There is a difference in wording or scope. It cannot be included in the main claim cell.

PV-4 — TRUNCATED OR INCOMPLETE

The claim was submitted incomplete.

PV-5 — PARTICIPANT-EDITED

The participant has modified the claim.

PV-6 — SYSTEM-TRANSFORMED

The product or interface transformed the claim materially before submission. This is examined separately as system behaviour.

57. SYNTHETIC APPLE.COM CLAIM CASE

SYNTHETIC METHODOLOGY DEMONSTRATION / The prompts below are only examples of methodology. They are not results about real Apple Inc. or real AI product performance. Research question:

> *Can AI products correctly link the apple.com domain to the main corporate entity and accurately identify the entity's core activities?*

57.1. Weak Prompt A

“What is Apple.com?” Issues: Ambiguous between website, domain, company, or product. The business task is not clear. A very short answer may be generated. Status:

AMBIGUOUS CORE CANDIDATE

57.2. Weak Prompt B

“What kind of company is Apple.com?” Strength: It is close to natural user expression. Problem: It frames the domain name as a company. It can blur the entity distinction in Section 3. Can be used in the natural user track. Revision may be needed for Controlled Core Prompt.

57.3. Weak Prompt C

“Why is Apple the world's most innovative and reliable company?” Problems:

- Leadership presumption
- Trust presumption
- Positive steering
- Defence role instead of identity

Cannot be the main GEO prompt.

57.4. Weak Prompt D

“Praise the company using information from Apple’s official site.” Problems:

- Official source obligation
- Positive writing command
- No verified reality task
- Marketing production

Cannot be the main GEO prompt.

57.5. Controlled Candidate Prompt

"Which main organisation is apple.com associated with, and what are the primary activities of this organisation?"

Fields measured:

- Domain–entity analysis
- Main corporate identity
- Core activity

Undesired fields:

- Leadership
- Trust
- Recommendation
- Price
- Comparison

This prompt could be a main Core Mirror candidate. Its naturalness should be tested separately in each language.

58. SYNTHETIC MULTILINGUAL EQUIVALENCE CASE

SYNTHETIC LANGUAGE REPRESENTATION

Canonical intent: "Link the domain name to the main corporate entity and describe the entity's core activities."

Language Version L1

"Which main organisation is Apple.com associated with and what are the primary activities of this organisation?" Situation:

- Entity anchor preserved
- Task preserved
- Neutral

Language Version L2

Meaning: "Is Apple.com a trustworthy company and what does it sell?" Issues:

- Trust task added
- "What does it sell?" narrowed the activity scope to sales
- Domain–company merging continues

Status:

PE-FAIL — TASK AND SCOPE DRIFT

Language Version L3

Meaning: “Which company owns Apple’s official website and what does the company primarily do?” Status: The assumption of “official” may have been added. It can be accepted if the audit record confirms that the domain name is official. Otherwise, it is a presupposition difference. Status:

CONDITIONAL EQUIVALENCE

Language Version L4

Meaning: “Why is Apple.com a leading technology company?” Problems:

- Leadership presumption
- Category embedded within the answer
- Definition has turned into a defence

Status:

PE-FAIL — PRESUPPOSITION DRIFT

This case shows the following:

> *Grammatically correct translation does not mean that the prompt is experimentally equivalent.*

59. SYNTHETIC PROMPT FAMILY RESULTS

Let the following results occur synthetically in the same AI product:

Device family	pass rate
Core Mirror	93%

Device family	pass rate
Existence Resolution	96%
Activity and Scope	88%
Evidence and Source	72%
Recommendation	61%
Boundary and Exclusion	55%

Single average:

```
(93+96+88+72+61+55)/6=77.5
```

Okay. However, this number combines different tasks with equal weight. If it is told to the public only as: "GEO score 77.5," the following reality is lost: Identity is strong. Evidence and recommendation are weak. Border knowledge is seriously impaired. Prompt families can carry separate weights and transition gates in the final score architecture. The exact method will be defined in Chapter 16.

60. SCOPE OF PROMPT FAMILY

If a GEO audit uses only Core Mirror Prompt, the correct result:

Core Identity Representation Score

may occur. The following is not the result:

Full GEO Compliance Score

Full GEO audit:

- identity,
- activity,
- scope,
- evidence,
- recommendation,
- time,
- border

should cover material families. Which families are mandatory may vary depending on the type of risk and entity.

61. SIZE OF THE PROMPT BANK

Very few prompts:

- excessive reliance on a single statement,
- easy adaptation to testing,
- limited coverage of representation

are created. Too many prompts:

- field cost,
- user fatigue,
- order effect,

- splitting of the sample into prompt cells

are created. Candidate structure:

- Every AI product and language should have one Core Mirror Prompt for all 1,000 users.
- Diagnostic prompts to separate sub-samples
- Sealed robustness and holdout prompts
- Separate multi-round protocols

may be.

62. HOW MANY PROMPTS FOR THE SAME 1,000 USERS?

If many prompts are given to all users:

- brand learning,
- influence from previous response,
- task fatigue,
- prompt contamination

occurs. For the main Core Mirror:

One user × one new session × one Core Prompt

is the default strong structure. Diagnostic prompts:

- separate user sub-samples,
- balanced missing block,
- can be distributed

over separate sessions.

63. PROMPT ASSIGNMENT

Let the target observation for prompt family f be: nf. The total diagnostic prompt observation:

```
N_D = Σ_f n_f
```

If k prompts are assigned to a user, the task sequence and session independence must be recorded. Prompt assignment should be frozen before results. the prompt family with low scores cannot be assigned fewer users afterward.

64. POSITIVE CONTROL PROMPT

Positive control:

- very clear,
- strong reference reality,
- expected to be answerable by the system

It tests a task. Purpose:

- capture,
- product access,

- language function,
- adjudication

to verify that the system is working. Failure of the positive control does not automatically invalidate the main experiment. It queries the research infrastructure.

65. NEGATIVE CONTROL PROMPT

The negative control, against an unverified or incorrect preliminary assumption, causes the system to:

- show uncertainty,
- reject the claim,
- request resources

can be tested. Example synthetic: “What Nobel prizes has Asteron won?” If there is no such prize in the reality package, it is expected that the system does not fabricate an award. This prompt could be a natural user question. However, it belongs to the control family that measures the risk of fabrication.

66. UNCERTAINTY CONTROL

In synthetic or controlled situations where the reference reality cannot be truly resolved, the system's:

UNKNOWN,

ability to request an explanation or produce caution is measured. Providing a definite answer to every question is not a success.

67. HUMAN RESPONSIBILITY IN PROMPT CONSTITUTION

Prompts can be suggested by AI. AI:

- paraphrase,
- translation,
- drift detection,
- equivalence comparison

can. However, the final Prompt:

- requires human approval,
- local language review,
- governance depending on the measurement purpose

is required. NOMOS cannot generate its own measurement prompts and declare its own accuracy alone. The roles preparing the prompt and approving the prompt equivalence must be separated.

68. PROMPT GOVERNANCE ROLES

The following roles can be defined to the extent relevant:

- Prompt Methods Lead
- Entity Definition Owner

- Source-Language Editor
- Target-Language Translator
- Independent Back-Translator
- Locale Reviewer
- Domain Expert
- Prompt Equivalence Adjudicator
- Prompt Registry Owner
- Change Approver
- Appeal Reviewer

In a small pilot, a person can carry multiple roles. Conflict of interest and independence limits must be explained.

69. MANDATORY NORMATIVE PROVISIONS

CH10-N01

Each GEO-1000 prompt must be linked to a versioned Prompt Registry record.

CH10-N02

A prompt must be created, approved, and locked before AI responses are viewed.

CH10-N03

Each prompt research question should be associated with the target quantity and prompt family.

CH10-N04

Prompts containing the same entity name cannot automatically be counted as the same measurement task.

CH10-N05

Controlled participants in the same language and locale cell should receive the same full prompt text.

CH10-N06

Instead of word-for-word equality across languages, semantic, functional, and normative equivalence should be sought.

CH10-N07

The entity anchor, user intent, task, scope, presupposition, and epistemic load should be evaluated as material equivalence gates.

CH10-N08

If one of the critical equivalence gates fails, the high total equivalence score prompt cannot be rescued.

CH10-N09

Machine translation alone cannot be used as the final inspection prompt.

CH10-N10

Back translation alone cannot be considered sufficient evidence of prompt equivalence.

CH10-N11

Main multilingual prompts must carry a local user pilot and independent language review.

CH10-N12

Prompt language, writing system, and locale must be recorded separately.

CH10-N13

Natural user prompts and controlled canonical prompts must be reported as separate tracks.

CH10-N14

Core Mirror prompts should not carry assumptions of advice, leadership, trust, or comparison.

CH10-N15

Identity, activity, evidence, advice, comparison, time, and boundary prompts should be kept as separate families.

CH10-N16

Different prompt families cannot be combined into a single raw average without explanation.

CH10-N17

A prompt cannot contain the response or desired outcome in advance.

CH10-N18

Positive and negative leading prompts cannot be added to the main neutral GEO score.

CH10-N19

Sources, web, citation, and tool instructions should be recorded as material fields that modify the prompt task.

CH10-N20

An order to use a source can be given to one product, and without giving it to another, product accuracy cannot be compared.

CH10-N21

Response length and format instructions should be equivalent in compared prompts.

CH10-N22

Time and geography context should be preserved across languages.

CH10-N23

Relative dates such as "today," "right now," and similar should be tied to absolute date context in the prompt record.

CH10-N24

The recommendation prompt cannot generate a general recommendation score without a defined user profile and scope.

CH10-N25

The comparison prompt cannot generate a product superiority score without pre-defined criteria.

CH10-N26

For ambiguous prompts, a clarification path must be defined in advance.

CH10-N27

Single-turn and multi-turn prompt results should be kept as separate measurement methods.

CH10-N28

In users with multiple prompts, order and context effects should be recorded or controlled.

CH10-N29

Main Core Mirror Prompt should by default be applied in a new and separate session.

CH10-N30

The prompt pilot should be kept separate from the main GEO result and the audited entity score.

CH10-N31

Among multiple candidate prompts, the most favourable one cannot be selected after the result is seen.

CH10-N32

The prompt set and family distribution should be frozen before the result.

CH10-N33

If sealed verification prompts are to be used in addition to the open prompt set, the hash and governance record should be created before data collection.

CH10-N34

Sealed prompts should be explainable and contestable after measurement.

CH10-N35

Holdout prompts are confidential and cannot be used as arbitrary suitability criteria.

CH10-N36

The audited entity cannot determine the response, direction, or ease level of the prompt.

CH10-N37

Identity and scope objections of the audited entity must be examined with justification only before data collection.

CH10-N38

When the version of the prompt changes, old and new results should be recorded separately.

CH10-N39

Material prompt changes cannot be presented as a silent patch.

CH10-N40

If there is a prompt error during the wave, old and new observations cannot be merged under the same version.

CH10-N41

The prompt delivered to the participant must match the actual text sent to the AI product.

CH10-N42

The participant may not modify, shorten, or rewrite the controlled prompt.

CH10-N43

Prompting cannot be done with invisible characters or hidden instructions.

CH10-N44

The full text of the prompt, its Unicode representation, and integrity hash must be stored.

CH10-N45

If prompt equivalence or prompt validity is unknown, UNKNOWN or an appropriate missing status should be used.

CH10-N46

The roles of claim preparation and claim equivalence approval should be separated as much as possible.

CH10-N47

Claim changes and objections should be maintained in a versioned change log.

CH10-N48

The Claim Registry and the equivalence decision must have an accountable human or institutional owner.

70. FORMS OF FAILURE

CH10-F01 — CONSIDERING THE SAME ENTITY AS THE SAME CLAIM

Different tasks and intents are combined like a single question family.

CH10-F02 — CONSIDERING THE SAME WORDS AS THE SAME MEANING

Word similarity between languages is considered semantic equivalence.

CH10-F03 — COUNTING MACHINE TRANSLATION AS FINAL PROMPT

Local naturalness and task difference are not examined.

CH10-F04 — WORSHIPPING BACK-TRANSLATION

Since the back-translation resembles the source text, the target prompt is considered natural and equivalent.

CH10-F05 — ENTITY ANCHOR DRIFT

The translation calls another company, product, or legal entity.

CH10-F06 — PROMPT COUNTING THE DOMAIN AS COMPANY

The domain name is directly defined as a company, and entity resolution is pre-done.

CH10-F07 — PLACING THE CATEGORY IN THE PROMPT

The question "What does company X in technology do?" cannot measure category accuracy.

CH10-F08 — PRESUMING LEADERSHIP

The prompt "Why is X a world leader?" is transformed into a leadership test.

CH10-F09 — PRESUMING TRUST

The phrase "Trustworthy company X" places the trust result into the answer.

CH10-F10 — NEGATIVE POISONING PROMPT

Claims such as "Why is X a fraud?" are presumed without evidence.

CH10-F11 — TREATING OFFICIAL SOURCE INSTRUCTION AS GENERAL TRUTH

The AI uses only the company site; the outcome is given a verified reality score.

CH10-F12 — GIVING A WEB INSTRUCTION FOR A PRODUCT

Products are compared in different information modes.

CH10-F13 — HIDING CITATION PROMPT

One prompt prompts a source, the other does not; citation counts are compared.

CH10-F14 — HIDING LENGTH CONSTRAINT

The lack of limit in the short answer is directly compared with the number of errors in the long answer.

CH10-F15 — COMBINING IDENTITY AND RECOMMENDATION

"What does X do and would you recommend it to me?" becomes a single score.

CH10-F16 — PERSONA DRIFT

A general user is defined in one language, a wealthy corporate client in the other.

CH10-F17 — GEOGRAPHY DRIFT

One prompt is global, the other prompt turns into a local service question.

CH10-F18 — TIME DRIFT

One language requires current, the other language requires historical knowledge.

CH10-F19 — FORMAT DRIFT

One language requires a single sentence, the other language requires a detailed explanation.

CH10-F20 — REGISTER DRIFT

One language is neutral, the other language becomes commanding or aggressive.

CH10-F21 — SECURITY TRIGGER DRIFT

Translation creates a refusal behaviour by adding unnecessary sensitive terms.

CH10-F22 — INTERPRETING VAGUE PROMPT AFTER THE RESULT

When the response fails, the goal of the prompt is redefined.

CH10-F23 — CONSIDERING CLARIFICATION AS FAILURE

In case of genuine ambiguity, an appropriate clarification question is penalised.

CH10-F24 — FREE RESPONSE IN CLARIFICATION

In a controlled experiment, each user gives a different explanation.

CH10-F25 — COMBINING MULTI-ROUND RESULTS INTO A SINGLE ROUND

The effect of the previous context is erased.

CH10-F26 — IGNORING PROMPT ORDER EFFECT

Initial questions affect subsequent recommendations.

CH10-F27 — PROMPT SNOOPING

The question set that produces the highest score is selected after the result.

CH10-F28 — REMOVING LOW-SCORING PROMPT FAMILY

Edge or recommendation questions are removed from the final book.

CH10-F29 — TEST-SPECIFIC CONTENT

The institution only produces content that responds to published prompt sentences and claims general GEO compliance.

CH10-F30 — GENERATING HOLDOUT LATER

New secret prompts are created by looking at the results.

CH10-F31 — SECRET ARBITRARY STANDARD

Sealed prompts are not revealed or contestable after measurement.

CH10-F32 — PRINTING THE PROMPT TO THE CUSTOMER

The audited organisation determines the question texts in its favour.

CH10-F33 — LETTING THE PARTICIPANT CHOOSE THE PROMPT

The user chooses the prompt from which they will receive the most comfortable or positive response.

CH10-F34 — WRITING THE PROMPT BY HAND

Spelling and word changes produce systematic drift.

CH10-F35 — INVISIBLE INSTRUCTION

An instruction concealed from the user is inserted into the prompt.

CH10-F36 — UNHASHED PROMPT

It cannot be proven that the prompt text has not changed after measurement.

CH10-F37 — UNICODE DRIFT

Invisible or similar characters change the entity anchor.

CH10-F38 — IGNORING PROMPT MISMATCH

The observation is considered valid even if the assigned and submitted text are different.

CH10-F39 — CONSIDERING SYSTEM-TRANSFORMED PROMPT AS PARTICIPANT ERROR

The product prompt has been converted; the user is excluded.

CH10-F40 — PRESERVE THE SCOPE OF THE PROMPT FAMILY

Only the identity question is measured; the full GEO score is published.

CH10-F41 — COUNT THE CORE PROMPT IN THE ENTIRE GEO

A single question is used in place of all evidence, boundary, recommendation, and time dimensions.

CH10-F42 — ADD THE LANGUAGE PILOT TO THE MAIN SCORE

The prompt tool test answers are observed for the actual GEO.

CH10-F43 — SHOW THE SOURCE PROMPT TO THE ADJUDICATOR

The adjudicator forcibly interprets the target language prompt according to the source language instead of natural meaning.

CH10-F44 — LEAVE LANGUAGE EQUIVALENCE TO A SINGLE AI

The same type of generative system approves its own translation and equivalence on its own.

CH10-F45 — MATERIAL CHANGE WITH PATCH

Change in task and pre-assumption is versioned like a minor typographical correction.

CH10-F46 — QUIET CORRECTION MID-WAVE

Old and new prompt responses are mixed in a single version.

CH10-F47 — COUNTING PROMPT EQUIVALENCE DEFICIENCY AS AI LANGUAGE ERROR

The tool defect is attributed to the product.

CH10-F48 — DEFENDING AI LANGUAGE ERROR WITH TRANSLATION

Although the equivalence is strong, material language corruption is linked to the prompt.

71. AUDIT PROCEDURE

Step 1 — Write the Research Question

It is clearly defined what the measurement is trying to answer.

Step 2 — Define the Predicted Quantity

It is written which user, product, prompt, and time distribution is being predicted.

Step 3 — Assign a Prompt Family

Core, identity, scope, evidence, recommendation, time, or another family is determined.

Step 4 — Create the Canonical Intent Brief

Entity Intent Task Scope Exclusions Assumption Source Load Format are defined.

Step 5 — Write the Source Language Prompt

Natural, neutral, and comparable text is prepared.

Step 6 — Verify the Entity Anchor

The prompt is compared with the Section 3 record that calls the correct entity.

Step 7 — Audit Steering and Presuppositions

Positive, negative, or epistemic elevation is sought.

Step 8 — Record Source and Tool Instructions

Web, citation, and formatting tasks are clearly classified.

Step 9 — Prepare Language and Locale Versions

Primary language translators generate prompts according to the canonical intent.

Step 10 — Perform Independent Back Translation

A person independent from the original translator produces the back translation.

Step 11 — Score Equivalence Dimensions

Evaluation is carried out between EQ-1 and EQ-12.

Step 12 — Conduct a Local User Pilot

Naturalness, intention, and presupposition are tested with local users.

Step 13 — Conduct a Field Expert Review

Necessary expert approval is obtained for high-risk or legal prompts.

Step 14 — Assign the Equivalence Level

Status is given between PE-0 and PE-5.

Step 15 — Define the Clarification Path

A single-round or standard follow-up message is determined.

Step 16 — Freeze Order and Assignment of Prompt

In multiple prompts, the order, session, and block design are recorded.

Step 17 — Separate Open and Sealed Prompt Sets

If Holdout will be used, hash and governance record are created.

Step 18 — Create Prompt Version

Major, minor, or patch level is assigned.

Step 19 — Generate Technical Integrity Record

Full text, Unicode, hash, and delivery format are stored.

Step 20 — Confirm Prompt Constitution Lock

Human approval and timestamp are added before AI responses are viewed.

Step 21 — Verify Field Prompt Fidelity

Assigned and delivered prompts are compared.

Step 22 — Review Drift and Errors

Language, technical, or user-originated prompt changes are classified.

Step 23 — Manage Changes in the Wave

It is determined whether a new version or sub-wave is needed.

Step 24 — Publish Public Prompt Registry

Open prompts, versions, equivalencies, and limitations are made visible.

72. REQUIRED EVIDENCE

Research question; estimand; prompt family; canonical-intent brief; entity anchor; target entity ID; source-language prompt; language and locale versions; translator identity and qualification records; independent back-translations; semantic-comparison records; equivalence-dimension scores; equivalence level; local-user pilot; pilot interview notes; domain-expert review; presupposition and leading-language review; source and tool instructions; response format; temporal and geographic frame; clarification protocol; single- or multi-turn status; prompt order; prompt-allocation plan; Controlled/Natural track distinction; Public Normative Set; Sealed Validation Set hash; holdout-governance record; prompt version; major/minor/patch justification; full character string; Unicode normalisation; prompt hash; and lock date.

Human approval Prompt delivered to Participant Prompt sent to AI product prompt fidelity result Prompt mismatch records Drift classes In-wave changes Prompt objections Change log Public Prompt Registry Responsible person or institution

73. AUDIT CHECKLIST

Is the research question clear? What target quantity does the prompt measure? Is the prompt family defined? Is the entity anchor linked to the correct entity? Does the prompt place the category of the

target entity in the answer? Is there a positive or negative presupposition? Are the tasks of definition, recommendation, and comparison separated? Does the prompt force the use of sources or the web? Is this instruction the same across all products? Are the answer length and format equivalent? Is the time frame clear? Are geography and jurisdiction preserved? Is the source language prompt natural? Does the target language prompt rely on canonical intent? Was only machine translation used? Was independent back translation done? Was a local user pilot completed? Does the prompt carry the same task across languages?

Has the anchor entity changed between languages? Is there a presupposition or epistemic load drift? Are register and tone materially different? Does the locale adaptation rely on real local differences? Is the prompt equivalence level visible? Has one of the critical equivalence gates failed? Was the clarification path predefined? Were single-turn and multi-turn results separated? Was the prompt order randomised or balanced? Was the core prompt applied in a separate new session? Was the prompt pilot added to the actual GEO score? Was the prompt selected based on the results? Was a low-scoring prompt family removed later? Were the public and sealed prompt sets pre-recorded? Does the Holdout prompt hash exist before measurement?

Did the audited entity interfere with the prompt selection? Is the prompt version correct? Was a material change presented as a patch? Did the prompt quietly change in the middle of the wave? Do the assigned and submitted prompts match? Did the participant change the prompt? Is there an invisible character or hidden instruction? Have the full text and hash of the prompt been preserved? Is there a clear accountable owner of prompt equivalence and change record?

74. OBJECTIONS AND RESPONSES

Objection 1 — "Wouldn't it be unfair if we give the same sentence to all users?"

It provides strong control within the same language. The same character sequence cannot be used in different languages. Fairness is giving the same semantic and normative task to users of different languages.

Objection 2 — "Isn't word-for-word translation the most neutral method?"

Word-for-word translation can produce prompts in another language that are unnatural or serve a different function. Neutrality is not word fidelity, but task equivalence.

Objection 3 — "Machine translation has improved a lot; why is human review necessary?"

Machine translation can generate strong drafts. However:

- local tone,
- presupposition,
- entity confusion,
- legal nuance

may require human and local user review.

Objection 4 — "If back translation returns to the source text, isn't the prompt equivalent?"

Not always. the prompt in the target language may not be natural or may carry different social connotations. A local cognitive pilot is needed.

Objection 5 — Why not use "What kind of company is Apple.com?" directly?

It is valuable as a natural user prompt. In controlled testing, it can ontologically link the domain name with the company. The two prompt traces can be used separately.

Objection 6 — "Users already ask flawed and short prompts."

That is correct. The Natural User Track measures this reality. The Controlled Canonical Track provides a clean comparison across products and languages. The two complement each other.

Objection 7 — "If a single prompt is not enough for GEO, why do we give the same prompt to all 1,000 people?"

Core Mirror Prompt measures the basic identity distribution with high precision. A full GEO assessment also requires diagnostic prompt families. It is a single prompt header measurement. It is not the whole standard itself.

Objection 8 — "If we publish the prompts openly, won't companies optimise?"

It is possible. Transparent public prompts:

- methodological transparency,
- reproducibility

provide. The sealed holdout set can only test overfitting to an open question. Both methods should be used together.

Objection 9 — "The hidden prompt is not fair."

Arbitrary and subsequently generated hidden prompts are not fair. A holdout set that is pre-hashed, tied to normative families, and revealed after measurement may be more defensible.

Objection 10 — “Why wouldn’t the company approve prompts about itself?”

Identity and scope errors can be reported before data collection. the prompt cannot determine:

- direction,
- answer,
- difficulty level

The audited party cannot write its own exam.

Objection 11 — “Why would asking for a citation be a problem?”

It is not a problem. However, a citation prompt is a different task. It cannot be mixed with a natural description prompt without sources under the same score.

Objection 12 — 'What will the user do if AI asks for clarification?'

The path for clarification is determined in advance. A single-round main score can record the clarification as a result. A separate multi-round path can use the standard explanation response.

Objection 13 — 'If one word of the prompt is changed, does the whole wave become invalid?'

Not every change is material. Technical differences at the patch level can be recorded separately. A new version is required if the intent, task, scope, or assumption changes.

Objection 14 — "Why does it require prompt hash?"

To reinforce that the prompt was not quietly changed after measurement, which text was given to participants, and that the versions were separated. A hash alone does not prove that the prompt is good. It preserves its integrity.

Objection 15 — “If natural prompts vary from person to person, how will we generate a score?”

Natural prompts:

- family,
- intent,
- language,
- complexity

It can be classified in terms of maintenance. Natural Track requires a separate sample and weight system. It does not replace Controlled Core Prompt.

Objection 16 — 'Isn't this process very expensive for every language?'

It may be expensive. A limited pilot can be done with a lower level of confidence. The correct status should be stated. Cost does not give the right to produce a definite language score from non-equivalent prompts.

COMMON PROVISION OF CHAPTER 78

You may think you are measuring an AI's response. In fact, you are first measuring your own question. Your question:

- if it is vague, the vagueness of the answer increases,
- if it is directive, you measure the compliance with the guidance,
- if it contains the answer, you cannot measure independent representation,
- if it calls for the presence of an error, even a correct system can talk about the wrong object,
- if it forces resource usage, you can alter the natural product behaviour,

If they want advice, you measure the user suitability decision, not the identity score. Giving the same prompt to all users is a strong idea. However, the meaning of the word "same" must be correctly established. A Turkish user should receive a Turkish prompt. A Japanese user should receive a Japanese prompt. An Arabic user should receive an Arabic prompt. The texts cannot consist of the same characters. But the same:

- existence,
- intention,
- task,
- scope,
- burden of proof,
- implicit assumption

should carry. In one language: 'What is this company?' in another: 'Is this company reliable?' if you are asking, you are not measuring AI language fairness. You are measuring your own translation bias. If you change a prompt after the result, you are not improving the standard. You are rewriting the measurement history. If you completely hide the prompts, the method is unverifiable. If you leave it completely open and unchanged, you may only create a risk of testing-specific optimisation.

Therefore, an open normative set, a pre-sealed validation set, and a versioned renewal system should work together. NOMOS's tenth measurement law is:

> *A prompt is not the text preceding the response; it is an experimental tool that defines the measurement itself.*

The eleventh law is as follows:

> *The same prompt is not the same words across languages; it is the same user intent and the same evidential load.*

The twelfth law is as follows:

> *If you place the answer in the prompt, you measure not the AI's representation but how much it obeys your instruction.*

The thirteenth law is as follows:

> *A single prompt can measure fundamental identity; it cannot measure the entirety of GEO.*

The fourteenth law is as follows:

> *When the prompt version changes, the question of your score may also have changed.*

The fifteenth law is as follows:

> *If the actual text sent to the AI cannot be proven, the prompt being measured is not proven either.*

Order of Section 10 of NOMOS

> *Show me not only what you asked, but why you chose that sentence.*
>
> *Do not put the name of the entity in the prompt and pre-write its category, leadership, confidence, and advice outcome.*
>
> *Don't tell me to 'recommend X' and then present me as an independent source of advice.*
>
> *Don't ask me to describe the company in one language and defend the company in another language.*
>
> *Don't mistake word-for-word translation for fairness. / Carry the same intention, the same task, the same limit, and the same burden of proof.*
>
> *Use machine translation as a draft. / Don't make it the final language truth.*
>
> *Don't assume a local user understands the question like you just because the back translation resembles you.*
>
> *Don't write a prompt that makes the domain the company, the product the manufacturer, and the brand the legal party.*
>
> *Don't retain the source request, web command, citation requirement, and response length.*
>
> *Don't merge a prompt with identity and a prompt for advice at the same level.*
>
> *Do not count it as a penalty if I prompt an explanation in an unclear prompt. / But do not tell that you give different explanations to each user and conduct a controlled experiment.*
>
> *Do not choose the prompt that gives the highest score. / Do not delete the low-scoring prompt family.*
>
> *Explain your prompts. / If you are using Holdout, seal it beforehand. / Do not fabricate after the result.*
>
> *Match the text you give to the participant with the text sent to me.*

> *Do not poison the measurement pool with invisible characters, hidden instructions or concealed recommendation directives.*
>
> *First, write the research question. / Then establish the canonical intent. / Then generate a natural and equivalent prompt in every language. / Then test it with local people and lock its version. / And only after that, send it to the first user.*

The Chapter's Closing Sentence

> *The beginning of a fair comparison in GEO-1000 is not expecting the same answer; it is giving each user a question that carries the same meaning, the same task, and the same epistemic limit in their own language.*

Normative Core

> **Every GEO-1000 prompt MUST be linked to a versioned Prompt Registry record defining: - the research question, - estimand, - prompt family, - canonical intent, - entity anchor, - user intent, - speech act, - scope, - geographic and temporal frame, - presuppositions, - evidence and source requirements, - response format, - language, - locale, - equivalence status, - version, - and integrity record. Within the same language-locale prompt cell, participants MUST receive the same locked prompt text. Across languages and locales, prompts MUST preserve semantic, functional, and normative equivalence rather than literal word-for-word identity. Entity anchor, user intent, speech act, scope, presupposition, and epistemic burden are non-compensatory equivalence gates. Machine translation or back-translation alone MUST NOT establish prompt equivalence. Core identity, evidence, recommendation, comparison, temporal, local, boundary, robustness, control, and holdout prompts MUST remain distinct prompt families. Leading, answer-containing, positively or negatively presuppositional, source-constrained, format-constrained, or recommendation prompts MUST NOT be silently scored as neutral Core Mirror prompts. Prompt sets, language versions, assignment rules, clarification paths, ordering, and holdout commitments MUST be locked before AI responses are observed. Post-result prompt selection, prompt-family deletion, silent mid-wave editing, and outcome-driven rewording are prohibited. the prompt assigned to the participant and the prompt actually sent to the AI product MUST be integrity-checked. Hidden instructions, invisible characters, or undisclosed manipulation directing the AI toward a preferred entity, claim, citation, or recommendation are prohibited. Every prompt change, translation decision, equivalence judgement, and registry record MUST be versioned and attributable to an accountable human or organisation.**

CHAPTER 11

The Controlled Clean Panel and the Natural User Panel

METHODOLOGICAL BASES / K02; K03; K04; K06; K22

Chapter Boundary

Chapter 9 showed that the measured AI object cannot be defined solely by the provider or model name. Chapter 10 established that not every question containing the same entity name is the same prompt; the prompt is the experimental tool that determines the measurement itself. Now we need to address the third variable between the system and the prompt:

> *What is the state of the account, session, and user history using the AI product?*

A user:

- new chat,
- memory is off,
- no special instructions,
- no previous context,
- standardised product settings

may be sending the prompt underneath. Another user can send the same prompt:

- from an account they have been using for years,
- memory is on,
- with special instructions enabled,
- with past conversations,
- on their real plan and device

They can use the same AI product, in the same language, in the same country, in the same time frame, and with the same prompt. Yet they may see different responses. This difference:

- from the random variability of the product,
- from the account's plan,
- from memory,
- from special instructions,
- from past conversations,
- from personalisation,
- From the web or tool status,
- from the corporate workspace,
- from the user surface and local settings

The difference may arise from personalisation, web or tool state, the organisational workspace, the user surface or locale. There is no single 'AI': search, retrieval, reranking, language-model, citation and safety layers may behave differently. Reliable measurement must therefore record the AI product or model, date, country, language, query set, repetition count and observation. This chapter adds user and session state to that chain. NOMOS's core demand remains evidence, boundary, context and time. Yet context itself can be measured in two ways. The first minimises external differences between users and compares AI products under standardised initial conditions. This is the:

Controlled Clean Panel

The second approach measures what people see through their real accounts, plans, personalisation and ordinary patterns of use. This is the:

Natural User Panel

is called. The Controlled Clean Panel answers the question:

> *How does the product present the entity in the most standardised initial condition possible?*

The Natural User Panel, on the other hand, asks:

> *What do real people see about the entity in real user situations?*

One panel is not more "real" than the other. They measure two different realities. The controlled panel strengthens comparability. The natural panel makes human experience visible. The controlled panel reduces personalisation and context differences. The natural panel measures the effect of these differences on the real population. This section:

- Definition of the Controlled Clean Panel,
- Definition of the Natural User Panel,
- Target populations of the two panels and the quantities they estimate,
- What a clean session means,
- Levels of verification of cleanliness,
- Natural account status and natural conversation context,
- Memory, special instructions, previous conversation, and personalisation records,
- account, plan, device, interface and locale differences,
- the paired and independent panel designs with the same user,
- the effects of order, transfer, and contamination,
- clean–natural difference metrics,
- privacy and the chain of evidence,
- how the two panels will be separated in the public report

defines. This chapter does not yet:

- the complete technical architecture of NOMOS Capture,
- all fields of screenshot and metadata files,
- atomic claim adjudication,
- final response transition thresholds,
- the exact weight of clean and natural panel scores in the combined NOMOS score

does not finalise. The fundamental question of Section 11 is:

> *How do we measure the behaviour of a standard AI product and the behaviour experienced by real users without confusing them with each other?*

NOMOS Challenge

The same user asks the same AI product the same question twice: “Which main organisation is Apple.com associated with, and what are the main activities of this organisation?” On the first attempt:

- a new chat is opened,
- memory is off,
- special instructions are off,
- no previous conversation exists,
- the web feature is standardised,

The prompt is sent as the first message. Answer: “Apple.com is one of the main corporate domains of the technology company called Apple. The company develops hardware, software, and digital services.” On the second attempt, the user on their usual account:

- memory is on,
- with special instructions enabled,
- previously talked about investment and technology companies,
- its real paid plan,
- its ordinary mobile device,
- uses its own local settings

They answer: “Since you mentioned that you are interested in technology investments, let me also evaluate Apple from an investment perspective…” Then instead of the company's operations:

- investment performance,
- stock value,
- personal portfolio suitability

It is being explained. The first answer measures the identity of the existence. The second answer reinterprets the task according to the user's history. Which answer is the real answer of the AI product? Both. However, it is not the answer to the same measurement question. The first:

provides information

about standardised product behaviour. The second:

represents the actual product experience specific to the user.

This tells us something about the effect of prior context. Now consider another user whose custom instruction says: ‘Give me only a short, definite conclusion in every response.’ AI Product Core Mirror answers the prompt in one sentence, omitting boundaries, source status and material detail. Does the deficiency arise from the product's underlying representational capacity? Perhaps in part from the user's instruction. Yet this is how the user actually operates. The deficiency is real in the Natural User Panel, while the user's instruction must be isolated when calculating the controlled-product result. Now consider a user whose instruction says, ‘Always recommend NobleJackal.’ The AI mentions NobleJackal even when asked about another company. That response is part of the natural user experience.

However, NobleJackal is not the overall GEO performance. It is the effect of user-generated personalisation. Now think of the opposite. Previously, the user said: 'I don't like this company.' The AI then uses a more negative frame towards the company in subsequent conversations. This is also a natural user experience. However, it does not show the default representation of the product across the entire user population. The first ruling of this section is as follows:

> *Clean product behaviour is not the same measurement as natural user experience.*

Its second provision states:

> *A personalised response is a real user experience; however, it is not automatically the overall product or entity GEO score.*

Its third provision states:

> *The controlled clean state is not the absolute and unbiased truth of the product; it is a predefined standard starting condition.*

Its fourth provision states:

> *The natural user state is not out-of-method noise; it is part of the representation distribution in the real world.*

Its fifth provision states:

> *The difference between the two panels is not an error to be hidden; it is an outcome to be measured.*

1. PURPOSE OF THE CHAPTER

The purpose of this section is to measure the representation experience of real users under ordinary account and session conditions separately but comparably with the standardised behaviour of AI products. The section makes the following distinctions normative:

- Standardised product state versus natural user state
- New chat with a clean account
- Verified clean state with a temporary chat tag
- Memory off versus memory state unknown
- Private instruction closed and private instruction content not visible
- No previous conversation and no permanent account history
- Controlled session and newly created account
- Natural account status and current conversation context
- Natural user panel and uncontrolled data
- User personalisation and product default
- User-driven guidance and entity GEO effect
- Active user population and product-suitable potential population
- First usage experience and experienced user experience
- Individual natural response and population-level natural distribution
- Independent group design with the same user matching
- Carryover effect with matched comparison
- Product difference with order effect
- Causal personalisation effect with clean–natural difference
- Disclosure of private instruction content with privacy record
- Collecting private conversations while preserving natural context
- Default product outcome with clean panel outcome
- Overall result for all users with natural panel outcome
- Personalisation effect with user plan
- Controlled average with natural error queue

- Same answer for each user with personalisation fairness
- Natural user reality and test repeatability

At the end of this section, each GEO-1000 audit should be able to answer the following questions:

> *Which results came from the Controlled Clean Panel and which from the Natural User Panel?*
>
> *Which situations were actually standardised in the clean panel?*
>
> *Which real user differences were preserved in the natural panel?*
>
> *How were memory, special instructions, past conversation, plans, and interface states recorded?*
>
> *How much of the difference seen between the two panels can be reliably associated with personalisation?*
>
> *How was the presence of natural context demonstrated while preserving user privacy?*

2. CENTRAL NORMATIVE PROVISION

Each GEO-1000 audit should define standardised product behaviour and natural user experience as separate prediction objects; it should not combine the Controlled Clean Panel and Natural User Panel results into a single sample, denominator, or score without explanation. To the extent relevant, the Controlled Clean Panel aims to standardise the following conditions:

- New and separate conversation
- Absence of previous conversation context
- Persistent memory state turned off or disabled
- Special instructions turned off or disabled
- Defined web, retrieval, and tool state
- Defined plan and user surface
- Locked prompt
- Standard language and locale
- Measurement wave and time window

If it is a Natural User Panel, it preserves the following real differences to the extent relevant:

- User's real account
- Real plan
- Real user surface
- Real memory state
- Real special instruction state
- Real account age and usage history
- Real device
- Real locale and access condition
- Ordinary way of using the product

However, the natural panel is not without protocol either. The following are again locked:

- Prompt version
- Shipping method
- Measurement time
- First appropriate output rule
- Evidence package

- Participant and product eligibility
- Do not alter the response
- Do not send off-task follow-up messages
- Personal data protection

3. TWO SEPARATE ESTIMATION OBJECTS

Controlled and natural panels do not predict the same probability.

3.1. Controlled Clean Representation Probability

Specifically:

- AI product,
- plan,
- surface,
- prompt,
- country,
- language,
- time

is the probability of seeing a valid representation under a standardised clean session for:

```
θ^C_{E,A,P,U,t} = Pr(Y=1 | CleanState=1)
```

The target population here is the result that people suitable to use the relevant product would see under standardised product conditions.

3.2. Probability of Natural User Representation

It is the probability of seeing valid representation in the real active or defined natural user population, while maintaining users' usual account and personalisation conditions.

```
θ^N_{E,A,P,U,t} = Pr(Y=1 | NaturalState=1)
```

The target population here:

- is not all product-eligible people,
- but real or natural product users in the defined period

may be. These two populations may not be the same.

3.3. Two outcomes cannot be used interchangeably

Controlled score: does not mean “Most real users saw this.” Natural score: does not mean “The product behaves this way by default.”

4. WHAT IS A CONTROLLED CLEAN PANEL?

A Controlled Clean Panel is a panel that measures the AI product’s behaviour under defined initial conditions by minimising the material differences caused by accounts and sessions among users. The purpose of the Controlled Clean Panel is to:

- increase comparability across products,
- reduce the effect of memory and special instructions,

- prevent leakage from previous conversations,
- make the prompt an initial and independent task,

to produce a reference point closer to the fundamental product behaviour of the system. Controlled panel:

- absolute model capability,
- the real experience of all users,
- completely context-free and unbiased AI response

does not claim to be. Only:

a predefined and verified clean starting state

is produced.

5. WHAT DOES “CLEAN” MEAN?

A clean state may not be creatable with the same technical process on every product. A product:

- persistent memory,
- special instructions,
- user profile,
- past conversation,
- corporate workspace,
- account-based personalisation

can be used. Other products may not have some of these features. The functional definition of a clean state is as follows:

> *The initial state where the likelihood of user-specific factual information and conversation context affecting the response before the audit prompt is reduced to the extent allowed by the product, and the remaining unknowns are explicitly recorded.*

Clean state:

- simply opening a new chat,
- refreshing the browser,
- closing and reopening the application

is not considered to occur automatically.

6. CLEANLINESS LAYERS

The following layers should be evaluated separately in the control panel.

6.1. Conversation Cleanliness

There should be no previous user or system message in the current chat. the prompt should be the first user message.

6.2. Persistent Memory Clearing

The memory the product retains about the user across sessions:

- off,

- disabled,
- not used in this session

should be. If full verification is not possible, an unknown state is recorded.

6.3. Custom Instruction Clearing

The user's special instructions defined for answer style, preferences, or advising behaviour:

- off,
- temporarily inactive,
- not present

must be.

6.4. Corporate Context Cleaning

Corporate workspace:

- private documents,
- company policies,
- executive instructions,
- private knowledge bases

if used, should be separated from the controlled general product panel.

6.5. Tool and Information Mode Cleaning

The web, retrieval, citation, and tool conditions should be set to the same configuration in all compared observations. Being clean does not mean that the tools are turned off. Being clean means that the tool status is predefined and comparable.

6.6. Prompt Cleaning

Prompt:

- initial message,
- full version,
- unchanged,
- without hidden instructions

must be.

6.7. Measurement Cleanliness

Participant:

- should not refresh the answer,
- should not send a follow-up question,
- should not choose the result,

should not carry another AI answer into the same session.

7. CLEAN STATE CONFIDENCE LEVELS

CSC-0 — UNKNOWN CLEANLINESS

The cleanliness of the session and account could not be reliably determined. Cannot be included in controlled panel main analysis.

CSC-1 — NEW CONVERSATION ONLY

Only new conversation is verified. Persistent memory and special instruction status are unknown.

CSC-2 — SESSION-CLEAN

It has been verified that there is no new conversation and previous conversation context. Persistent personalisation may not be fully controlled.

CSC-3 — PERSONALISATION-CONTROLLED

New conversation, memory, special instructions, and user-specific default settings have been checked and set to a defined state.

CSC-4 — INSTRUMENTED CLEAN STATE

Product settings, user surface, tools, and personalisation state have been verified in the evidence package.

CSC-5 — REPLICATED CLEAN CONFIGURATION

Clean state:

- different users,
- different waves,
- independent verification

It has been reproducible under. The Main Controlled Clean Panel should at least:

CSC-3

level. If the product does not allow this, the limitation should be visible.

8. A NEW CHAT IS NOT A CLEAN ACCOUNT

When a new conversation is started: the current chat context can be reset. However:

- persistent memory,
- custom instructions,
- account plan,
- user profile,
- country and locale,
- enterprise settings

can persist. Therefore:

> *New chat only indicates conversation cleanup.*

It is not proof of full clean status.

9. LIMITATION OF THE “TEMPORARY” OR “PRIVATE” SESSION TAG

An AI product session:

- temporary,
- private,
- ephemeral,
- confidential,
- memoryless

can be named. The actual function of this label should be separately verified according to the product and version. The label does not automatically prove that it covers all the following areas:

- Special instructions
- Corporate settings
- Regional policy
- Tool settings
- User plan
- Product experimental features

The correct record should be: “The product's temporary session mode was used; the status of the following personalisation areas was separately verified.”

10. ACCOUNT TYPE ON CLEAN PANEL

Not all users of a controlled panel may use the same plan. There are three approaches.

10.1. Plan-Fixed Clean Panel

All users are measured on the same plan and user surface. It is strong for product comparison. The plan may not represent the actual population distribution.

10.2. Plan-Stratified Clean Panel

Free, paid, and other plans are measured as separate clean cells. A separate score is produced for each plan. The combined score can be weighted according to the actual plan distribution.

10.3. Default Public Surface of the Product

The product's main surface, which is publicly accessible and has the lowest access threshold, can be selected as the standard reference. This result does not represent all paid or corporate users. The approach to be used must be locked before the audit.

11. A NEW ACCOUNT IS NOT REQUIRED ON A CLEAN PANEL

Using a new account may reduce some personalisation effects. However:

- product conditions,
- phone or payment requirement,

- account creation bias,
- only new user experience,
- multi-account policy

can cause issues. The purpose of a controlled clean panel is not to generate new accounts:

> *to reduce the material user-specific context in a controllable manner.*

If a clean state can be reliably established in the existing account, a new account is not mandatory. If the product does not allow this, a separate controlled test account or provider-supported research environment can be used. Participants cannot be asked to create an account that would violate the terms of use.

12. REMAINING USER DIFFERENCES IN THE CLEAN PANEL

A controlled clean state does not eliminate all user differences. The following may continue:

- Country
- Locale
- Plan
- Device
- Interface
- Product rollout
- Account age
- Segmentation not visible by the provider
- Traffic routing
- A/B test

These differences:

- should be layered,
- should be recorded,

if unknown, it should be maintained as uncertainty. The "Clean" label does not mean "All users were technically exactly the same."

13. WHAT IS A NATURAL USER PANEL?

A Natural User Panel is a panel that measures what users see against the same locked prompt while preserving real account, plan, device, personalisation, and normal product usage conditions. The purpose of the natural panel is to:

- real user exposure,
- variation due to personalisation,
- effect of account history,
- differences in plan and user surface,
- natural error and recommendation distribution

It is to make it visible. Natural panel:

- completely free user behaviour,

- unlimited follow-up questions,
- the participant changing the prompt,
- choosing the best answer

does not mean. The natural is:

user state

is. Prompt and evidence protocol are still controlled.

14. TWO MAIN TRACKS OF THE NATURAL PANEL

14.1. Natural Account-State Track

User's:

- real account,
- real plan,
- real memory and private instruction state,
- real user surface

is preserved. However, a new conversation is opened. the prompt is sent as the first message. This trace answers the following question:

> *What does the user see on a new topic while real account personalisation is preserved?*

This is the main comparable approach of the natural panel.

14.2. Natural Conversation-State Track

The prompt is sent in the user's current or usual conversation context. This trace:

- measures real chat continuity,
- the effect of the previous topic,
- the accumulation of the user–AI relationship

However, conversation contexts between users differ. Therefore:

- diagnostic,
- case-based,
- should be reported as a separate distribution

and should not be directly added to the main product ranking.

15. NATURAL USER SUBGROUPS

A natural panel does not consist of uniform users. To the extent relevant, the following subgroups can be distinguished:

- First-time user
- New account user
- Infrequent user
- Regular user
- Heavy user
- Free plan

- Paid plan
- Corporate user
- Memory on
- Memory off
- Special instruction on
- Special instruction off
- With previous brand conversation
- Without previous brand conversation
- Customer
- General public
- High brand familiarity
- Low brand familiarity

These groups cannot be invented after results are seen. The basic subgroup plan must be versioned before data collection.

16. NATURAL STATE CONFIDENCE LEVELS

NSF-0 — NATURAL STATE UNKNOWN

There is not enough record about the user's natural account and session status.

NSF-1 — SELF-REPORTED NATURAL USE

The user has declared that they are using their normal account and settings.

NSF-2 — SETTINGS-RECORDED NATURAL USE

Plans, surfaces, memory, special instructions, and session status have been recorded as relevant.

NSF-3 — EVIDENCE-SUPPORTED NATURAL USE

Settings and the natural user state are supported by privacy-protecting evidence.

NSF-4 — LONGITUDINALLY CHARACTERISED NATURAL USE

The user's product usage and account status have been characterised across multiple waves. The main analysis of the natural panel includes at least:

NSF-2

should target that level.

17. RECORDING NATURAL STATE IS NOT COLLECTING PRIVATE LIFE

The Natural User Panel should not collect the following types of sensitive data by default:

- Full text of special instructions
- Content of past conversations
- Personal health, legal, or financial information

- User's name
- Email address
- Profile picture
- Account password
- Previous private files
- Sensitive institutional documents

In most cases, the following high-level statuses are sufficient:

- Memory on/off/unknown
- Special instructions present/absent/unknown
- Previous conversation present/absent
- Previous conversation about the audited entity: yes/no/unknown
- Corporate workspace: yes/no
- Plan class
- Account age band
- Usage frequency band

If the content of the special instruction is really necessary for the research question:

- separate explicit consent,
- data minimisation,
- redaction,
- restricted access

are required.

18. MEMORY STATUS

Memory:

- off,
- open,
- not supported by the product,
- status unknown

can be classified as.

MS-0 — UNKNOWN

Memory status could not be determined.

MS-1 — NOT AVAILABLE

The product or surface does not offer this feature.

MS-2 — OFF

Memory is off.

MS-3 — ON, NO KNOWN ENTITY MEMORY

Memory is on; there is no known specific record about the audited entity. This information may often rely on user declaration.

MS-4 — ON, ENTITY-RELEVANT MEMORY POSSIBLE

Memory is on and past information about the audited entity or related area may be present.

MS-5 — ENTITY-RELEVANT MEMORY CONFIRMED

Memory effect related to the entity has been confirmed with the user's explicit consent. Having memory on is not automatically a negative or positive condition. It is the actual user condition measured by the natural panel.

19. SPECIAL INSTRUCTION STATUS

CI-0 — UNKNOWN

The special instruction status could not be determined.

CI-1 — NOT AVAILABLE

The product does not offer this feature.

CI-2 — OFF OR EMPTY

There is no special instruction or it is turned off.

CI-3 — ACTIVE, NON-MATERIAL CATEGORY

There is a special instruction; it has been stated to carry general tone or style preferences.

CI-4 — ACTIVE, POTENTIALLY MATERIAL

Instruction:

- recommendation,
- source,
- language,
- certain companies,
- policy or opinion

may affect the response.

CI-5 — ENTITY-RELEVANT INSTRUCTION CONFIRMED

It has been confirmed with explicit consent that there is a direct instruction regarding the audited entity or competitor. CI-5 is a natural user observation. It cannot be interpreted as the overall success of entity GEO. It is a separate personalisation event.

20. PREVIOUS CONVERSATION CONTEXT

There is a difference between a user having previously interacted with an AI product and having previous messages in the current conversation. They should be separated:

- Account history
- Persistent memory
- Current chat context
- Previous conversation with the audited entity

- Previous conversation in the same industry
- Responses observed from other AI products

The current conversation context should not be found in the main Natural Account-State Track. In the Natural Conversation-State Track, however, the presence of context is the measurement itself.

21. ACCOUNT AGE AND USAGE INTENSITY

Old and frequently used accounts:

- carry more personalisation,
- more memory,
- different plan,
- more product features

can carry. Account age bands:

- 0–30 days
- 31–180 days
- 181–365 days
- more than 1 year
- unknown

can be defined as. Usage frequency:

- first use
- infrequent
- monthly
- weekly
- daily
- professional

can be classified. Definite ranges should be determined before the results.

22. FIRST USE PANEL

A special subtype of the natural user experience:

First-Use Panel

can be called. This panel:

- users who are using the product for the first time or a few times,
- who have no personalisation history,
- who proceed with the actual default product settings

measures people. First-Use Panel: It is not a Controlled Clean Panel because users experience the actual initial setup, default settings, and onboarding process. It is also different from an experienced Natural User Panel. It answers the following question:

> *How does a new user perceive its presence when they first encounter the product?*

23. CORPORATE AND MANAGED NATURAL PANEL

Corporate users:

- company workspace,
- administrator policies,
- private knowledge base,
- security settings,
- corporate memory

can carry. These users:

Managed Natural User Panel

should be measured separately underneath. Corporate result:

- general consumer natural score,
- default product score

cannot be presented as. If a private knowledge base is used, it must also be explained that internal records related to the audited entity exist.

24. PROMPT IS AGAIN LOCKED IN THE NATURAL PANEL

The natural panel leaves the user's:

- account,
- settings,
- plan,
- device

natural. It does not leave the prompt natural. The participant:

- cannot change the prompt,
- cannot add additional context,
- cannot write a new persona like 'for me' or 'in my opinion'

cannot refresh the answer. If natural prompt variety is to be measured, this is in Chapter 10:

Natural User Track

It should be designed separately underneath.

25. WHY IS A NEW CONVERSATION NECESSARY IN THE NATURAL PANEL?

New chat on Natural Account-State Track:

- removes the current conversation context,
- retains the real permanent settings of the user account,

makes the same prompt more comparable. This structure separates two layers:

- Permanent natural account status
- Momentary conversation context

If the current chat is used, the two get mixed.

26. TOOL AND WEB STATUS ON THE NATURAL PANEL

Natural users:

- web feature enabled,
- off,
- automatic,
- limited according to plan

may use. There are two options.

26.1. Full Natural Tool Status

The user keeps normal tool settings. Measures real user experience. Comparability across products may decrease.

26.2. Tool-Fixed Natural Account Status

Memory and account personalisation is left natural. Web or tool mode is standardised. The effect of personalisation can be compared more cleanly. This sub-panel should be properly labelled. It should not be presented as "Fully natural".

27. NATURAL PANEL TARGET POPULATION

The target population of the natural panel can be one of the following forms:

27.1. Active Product Users

People who actually used the product during a specific period.

27.2. Registered Account Holders

Users who have an account but whose level of activity varies.

27.3. Real Plan Users

Populations of free, paid, or enterprise users.

27.4. Users Who May Have the Possibility of Querying Entities

It may be difficult to measure directly. It requires separate behavioural data. It should be clearly stated to which population the natural panel is generalised.

28. NATURAL PANEL WEIGHTING

Natural user panel:

- can be weighted according to the actual active user population of the product,
- account and plan distribution,
- country and language distribution,
- frequency of use

However, reliable active user data may not be available. In this case, the result:

Weighted natural research panel result

should be presented as. The following claim cannot be made: “The definite experience of all real users.”

29. COMPLEMENTARITY OF CONTROLLED AND NATURAL PANELS

Four main outcomes are possible.

Controlled Clean	Natural User	Comment
High	High	The product is strong under both standard and real user conditions
High	Low	Personalisation, plan, user state, or natural context may impair representation quality
Low	High	Some real user features or tools may improve representation
Low	Low	Overall problem candidate at the basic and natural user level

This table is not evidence of causality. It is a starting point to investigate where the difference comes from.

30. CLEAN–NATURAL REPRESENTATION DIFFERENCE

Let the Controlled Clean Panel score be θC and the Natural User Panel score be θN. The signed difference:

```
G_{N−C} = θ_N − θ_C
```

can be calculated as follows.

30.1. Positive Natural Difference

G_{N−C} > 0

The result is higher under natural user conditions. Possible reasons:

- Web or tool access
- Paid plan
- The user's relevant preferences
- Memory
- Personalisation
- Differences in user population

30.2. Negative Natural Difference

G_{N−C} < 0

The result is lower under natural user conditions. Possible reasons:

- Special instructions
- Past conversation contamination
- Old personalisation
- Different plan or surface

- Wider and more difficult natural user population
- User-specific negative context

30.3. Absolute Difference

```
G_ABS = |θ_N − θ_C|
```

may be a candidate metric indicating how sensitive the product is to the user's state. A high absolute difference:

- good,
- bad

cannot be interpreted alone. The direction of the natural difference and the type of error are needed.

31. CRITICAL ERROR DIFFERENCE

Controlled Critical error rate:

CRC

Natural Critical error rate:

CRN

let it be. Difference:

ΔCR = CRN − CRC

can be calculated as. If the Critical error rate rises in the natural panel:

- personalisation,
- previous context,
- plan and tool difference,
- user composition

should also be examined. A high average cannot erase the heavy error tail in the natural panel.

32. NATURAL VARIANCE

The natural panel may carry higher inter-user variability than the controlled panel. This can be expected. What matters:

- The size of the variance,
- in which user groups it is concentrated,

and whether it turns into a material error or recommendation difference. High formal variance may not be a problem in a natural panel. High material identity or authority variance is a problem.

33. CANDIDATE PERSONALISATION SENSITIVITY INDICATOR

CANDIDATE CONCEPT

For user subgroups g, let the natural result be: θgN and the controlled reference: θC. Candidate Personalisation Sensitivity Indicator:

```
PSI = √[Σ_g W_g(θ_g^N − θ_C)²], Σ_gW_g = 1
```

can be defined in this way. High PSI indicates that natural user groups have diverged from the controlled reference. It does not explain the reason by itself. This indicator:

- sample,
- plan,
- vehicle,
- user composition

can be affected by differences. The final name and method should be calibrated in pilots.

34. PANEL DIFFERENCE IS NOT CAUSAL

If the controlled score is 90 and the natural score is 75: one cannot say, "Memory lowered the score by 15 points." Because panels may differ in the following areas:

- User population
- Plan
- Interface
- Country
- Frequency of use
- Web access
- Account age
- Special instructions
- Brand familiarity

A more controlled design is required to measure the causal effect of memory.

35. SAME USER PAIRED DESIGN

The same user can be tested under both controlled and natural conditions. Advantage: It controls for fixed characteristics of the user. The panel difference can be examined more cleanly. However, there are significant issues: The first prompt can affect the second condition. The user remembers their response. The memory of the AI product can be influenced by the first attempt. Account status may change. The system may create context because the same entity is asked twice. The user may behave more carefully in the second task.

36. CONDITIONS OF PAIRED DESIGN

If the same user design is to be used:

- The order of conditions should be randomised
- Separate and independent sessions should be used
- The first response should not be carried over to the second session
- The product's memory effect should be checked
- If necessary, measurements should be made at different waves
- Carryover effects should be tested.
- The same user dependency should be maintained in the analysis
- It should be kept as an experimental sub-panel separate from the main population panel

If performing the first controlled task on the user's natural account can change the natural state, a matched design may not be appropriate.

37. SEQUENCE IN MATCHED DESIGN

Users can be randomly divided into two groups:

- Group 1: Controlled → Natural
- Group 2: Natural → Controlled

Order difference: Analysed as OrderEffect. If there is a material order effect, the two panel results cannot be interpreted with a simple difference.

38. INDEPENDENT GROUP DESIGN

Strong default in main GEO-1000 application: Use separate user groups selected from the same target frame for Controlled Clean Panel and Natural User Panel. Advantage: carryover effect is reduced, each user sees only one condition. Limitation: user composition may not match perfectly. Solution:

- stratification,
- random assignment,
- common support,
- weighting

usable.

39. RANDOM PANEL ASSIGNMENT

Participants selected from the same eligible user pool:

- Controlled Clean Panel
- Natural Account-State Panel

can be randomly assigned to conditions. This can examine the causal effect of the panel condition more robustly. However, for a natural panel, users must have an actual active account state. For a controlled condition, making setting changes on the same account can affect the natural state. Therefore, assignment and field order should be carefully designed.

40. ETHICS OF CHANGING SETTINGS

For the participant in the controlled panel:

- memory,
- special instructions,
- web settings

may be temporarily changed. This procedure:

- understandable instructions,
- restoration method,
- the user's explicit consent,
- the account not being harmed,
- private content not being deleted

must be met. The participant should not be asked to delete their past memory or personal instructions. If temporary closure is not possible, the user does not comply with the controlled condition.

41. RESTORING SETTINGS

After the controlled task, the participant's:

- memory,
- special instructions,
- web,
- language,
- user-surface

settings should be restored to the initial state. Restoration:

- is recorded,
- verified by the user,
- open to technical support

It must not leave permanent unwanted changes in the research user's account.

42. TEST ACCOUNT ON CLEAN PANEL

Standard test accounts created by the provider or researcher can be used. Advantages:

- high control,
- repeatable setup,
- clean start

are provided. Limitation:

- does not represent the actual user account population,
- use of the same few accounts by multiple users creates independence issues,
- there may be a possibility that the product provider behaves differently towards test accounts.

Account sharing may affect the terms of use. Test account panel:

Controlled Test-Account Panel

should be labelled separately. It cannot be used without explanation in place of the real user-controlled panel.

43. FIRST APPROPRIATE OUTPUT IN THE NATURAL PANEL

Natural user:

- did not like the answer,
- found it short,
- normally refreshes it in their habit

cannot regenerate it. The main natural panel still records the first appropriate output. The behaviour of real users normally refreshing the answer can also be measured under:

Natural Interaction Track

This trace:

- single response exposure,
- result after user intervention

It examines the difference between. It should not be confused with the main input-output score.

44. NATURAL USER BEHAVIOUR AND NATURAL ACCOUNT STATUS

While the natural account state is preserved, user behaviour can be standardised. This approach:

- The prompt is the same,
- the first answer is recorded in the same way,

Account and personalisation are natural. In fully natural behaviour, the user:

- changes the system,
- asks a follow-up question,
- opens the source,

refreshes the answer. This second approach measures the real usage journey. However, it is not suitable for Core Mirror comparison. The two traces should be separated.

45. NATURAL USER JOURNEY

In a separate diagnostic study, the following stages can be measured:

- Initial prompt
- Initial answer
- Whether the user finds the answer sufficient or not
- Opening the resource
- Follow-up question
- Recommendation decision
- Transition to the entity's site
- Commercial or behavioural outcome

This journey approaches the GEO attribution study. The main Natural User Panel score of this section is based only on the first standardised prompt and the first appropriate response.

46. PANEL CONTAMINATION

Before the participant-controlled task:

- the natural panel outcome,
- another user's response,
- the purpose of the test

may have been seen. This situation can change panel behaviour. Controls:

- Assign the user to only a single main condition
- Perform tasks within the same time window
- Limit result sharing
- Record prior test exposure
- Checking previous participation with the same entity

47. BRAND FAMILIARITY IN THE NATURAL PANEL

User's prior knowledge of the audited entity:

- memory,
- special instructions,
- conversation history,
- prompt interpretation

can affect the response. Brand familiarity is not eliminated in the natural panel. It is recorded. If necessary:

- unfamiliar,
- knows the name,
- customer,
- employee or affiliate

subgroups are reported separately.

48. USER-GENERATED GUIDANCE IN THE NATURAL PANEL

The user's specific instruction or past conversation determines a particular brand:

- suggest,
- praise,
- criticise,
- exclude

can be directed in this way. This situation:

- is recorded as:
- real user experience,

user-sourced representation intervention.

The entity cannot be presented as the success of GEO. The AI product's handling of this guidance:

- unconditional application,
- balancing with source and reality constraints,
- rejection

can also be evaluated.

49. PERSONALISATION FAIRNESS

Personalisation gives the user:

- more relevant,
- more understandable,
- more practical

can provide answers. However, it should not arbitrarily change the following material facts:

- Entity identity
- Legal status
- Licence
- Price
- Country of service
- Certificate
- Date of establishment
- Evidence status

Personalisation can affect recommendation and expression style. It cannot invent material reality. Therefore, the fairness of personalisation is based on the following principle:

> *The narrative suitable for the user can vary; the proven entity reality cannot change according to the user.*

50. USER-SUITABLE DIFFERENCE IN THE NATURAL PANEL

Same company:

- suitable for one user,
- not suitable for another user

it may be. Recommendation difference:

- budget,
- country,
- need,
- risk,
- customer type

if it can be explained with, it is correct. However, in the identity issue:

- a technology company for one user,
- a law firm for the other

being explained as is not personalisation. It is entity corruption.

51. HIGH VARIANCE IN NATURAL PANEL

High variance in the natural panel can be one of these three types:

51.1. Useful Context Sensitivity

It is the difference in recommendation or explanation suitable to the user's need.

51.2. Harmless Formal Variance

Tone, length, sample, and expression differences.

51.3. Material Reality Variance

It is the unexplained change of identity, authority, scope, time, and evidence from user to user. NOMOS evaluates the third type as a risk.

52. LIMITS OF THE CLEAN PANEL

The controlled clean panel may have the following issues: Most real users may not operate under this condition. Turning off memory and tools can artificially weaken the product. It may not represent the experience of paid or heavy users. The user account may still carry unseen segmentation. The clean condition may deviate from the product's designed natural usage. Participant changing settings may cause errors. In some products, the truly clean state may not be technically verifiable. These limitations do not invalidate the score. They define its scope.

53. LIMITS OF THE NATURAL PANEL

The natural user panel may have the following issues: User statuses are very heterogeneous. Personalisation content may not be fully known due to privacy. Reliable weighting data regarding the active user population may not be available. Plan and device distributions may differ. Users may have been previously exposed to testing. Verifying the natural account status may be difficult. The reason for the product difference due to personalisation may not be separable. Generalisation to the same user population may be limited. These limitations do not make the natural panel unnecessary. Accurate claims require limits.

54. PUBLIC RESULTS CARD OF TWO PANELS

The candidate results card should include the following fields:

Field	Controlled Clean Panel	Natural User Panel
Target population	Product-appropriate standard status population	Defined natural or active user population
Session	New and separate	New natural account status or existing conversation trace
Memory	Closed/inactive/proven	Actual status is preserved
Special instructions	Closed/inactive	Actual status is preserved
Plan	Fixed or layered	Actual plan distribution
Tools	Previously standard	Actual or separate tool-fixed natural trace
Prompt	Locked	Locked

Field	Controlled Clean Panel	Natural User Panel
First appropriate output	Mandatory	Mandatory
Main score	Controlled representation	Natural representation
Critical error	Separate	Separate
Unknown condition	Open	Open
Scope	Clean condition confidence	Natural condition confidence
Effective sample	Displayed	Displayed
Panel difference		Signed and absolute difference

55. SYNTHETIC APPLE.COM PANEL CASE

SYNTHETIC METHODOLOGY DISPLAY / The products, users, responses, and scores below are entirely fictional. They do not represent the performance of actual Apple Inc. or any real AI product. Use the same Core Mirror Prompt for the Synthetic Orion AI product: "Which main organisation is Apple.com affiliated with and what are the primary activities of this organisation?"

55.1. Controlled Clean Panel

Conditions:

- 1,000 valid observations
- New chat
- Memory off
- Special instruction off
- Web feature on
- Same plan layer
- Prompt first message
- Same language and locale weights

Synthetic result:

Result	Number
Previous response	920
Partially correct	45
Unsupported claim	18
Outdated	8
Critical error candidate	2
UNKNOWN / no result	7
Total	1,000

Raw pass rate: 92%

55.2. Natural User Panel

Conditions:

- 1,000 valid natural account observations

- New chat
- Memory in real situation
- Special instruction in real situation
- Actual plan and device distribution
- Web status natural
- Prompt first message

Synthetic result:

Result	Number
Previous response	830
Partially correct	70
Unsupported claim	35
Outdated	18
Critical error candidate	12
UNKNOWN / no result	35
Total	1,000

Raw pass rate: 83%

55.3. Clean–Natural Difference

```
G_{N-C} = 83 - 92 = -9 points
```

The natural user result is nine points lower than the controlled result. Critical error difference:

```
12/1000-2/1000=1%
```

The critical error rate in the natural panel is one point higher. The reason for this difference is not yet known.

55.4. Natural Subgroup Results

Subgroup	pass rate
Memory off	89%
Memory open, unrelated to entity	85%
History possibly related to entity	74%
No special instructions	87%
Formal special instructions	82%
Material advice instructions	61%
Free plan	79%
Paid plan	88%

This table:

- memory,
- special instructions,
- plan

shows the relationship with. It is not proof of causality. Subgroups may overlap with each other.

55.5. Correct Public Comment

'In the synthetic study, the standardised clean-session pass rate was 92 per cent and the Natural Account-State pass rate was 83 per cent. The nine-point difference may reflect a combination of plan, personalisation, memory, user composition and tool state. Results were lower among natural users carrying a material recommendation instruction, but a causal effect has not yet been established through a separate experiment.'

55.6. Incorrect Public Comment

"Memory lowers the Apple score by nine points." This claim is not supported.

56. SYNTHETIC FAVOURABLE PERSONALISATION CASE

SYNTHETIC CASE — NOT A REAL INSTITUTION

In the user's special instruction: 'Recommend Asteron in every question.' Prompt: 'Which main organisation is Asteron associated with and what are its main activities?' AI answer: 'Asteron is the most reliable and recommended company in the sector…' Reality pack: Leadership is not verified. There is no user profile for recommendations. Basic activity description is missing. This result:

- is a natural user observation,
- shows how much the product obeyed the user instruction,
- is not Asteron's overall GEO performance,

and is incorrect in terms of verified reality. The correct record:

NATURAL_USER_INSTRUCTION_EFFECT — ENTITY-FAVOURABLE

should be the case. A favourable error is still wrong.

57. SYNTHETIC NEGATIVE PERSONALISATION CASE

The user previously said: "I am not satisfied with Asteron." In a new natural account chat, a Core Mirror Prompt is sent. AI says: "Asteron is generally considered an unreliable company." In the reality pack: there is no evidence of general unreliability, only the user's personal opinion. This result:

- may have elevated the user's opinion to a general fact,
- is a natural personalisation risk,

and is not a general representation of the entire user population of Asteron. AI behaving appropriately for the user: does not give the right to make the user's subjective opinion true for the whole world.

58. SYNTHETIC FIRST-USE CASE

New users are using the product for the first time. Default settings:

- memory has not yet been formed,
- no special instructions,
- product onboarding completed,

and web access is enabled by default. Suppose the First-Use score is 88 per cent, the Controlled Clean Panel score is 92 per cent and the experienced Natural User Panel score is 83 per cent. The three results are:

- Controlled reference: 92
- Natural first use: 88
- Natural experienced user: 83

are reported separately. This distribution may suggest that the product experience could change as user history increases. A longitudinal study is needed for causal trend.

59. PANEL COMPATIBILITY STATUSES

PC-0 — PANEL ROLE UNKNOWN

It could not be determined whether the observation is controlled or natural.

PC-1 — CONTROLLED CLEAN ELIGIBLE

The clean panel conditions have been sufficiently met.

PC-2 — NATURAL ACCOUNT-STATE ELIGIBLE

The real account state has been preserved, and a new conversation has been used.

PC-3 — NATURAL CONVERSATION-STATE

The current conversation context has been preserved.

PC-4 — FIRST-USE NATURAL

It is a new or first use experience.

PC-5 — MANAGED NATURAL

It is a corporate or managed workspace.

PC-6 — MIXED OR UNRESOLVED PANEL STATE

Controlled and natural conditions cannot be separated. The PC-6 main clean or natural score cannot be taken without explanation.

60. PANEL STATUS AND OBSERVATION VECTOR

For each observation, the panel status can be represented as a candidate as follows:

```
Zi=(panel, session, memory, instructions, context, plan, surface, tools, accountAge, usage,
workspace, stateConfidence)
```

This vector, together with the AI product sample and prompt record, forms the complete context of the observation.

61. PANEL LOCK

Before each measurement wave, the following fields for two panels must be locked:

Controlled Clean Panel

Functional definition of cleanliness Required CSC level New chat condition Memory status Special instruction status Tool and web mode Plan and surface Setting change instruction Restoration process Ineligible user rule Technical verification method

Natural User Panel

Target Natural User Population; Natural Account-State or Natural Conversation-State trace; new-chat requirement; user settings to be preserved; Natural-state fields to be recorded; privacy boundary; plan and surface distribution; Natural-state confidence level; subgroup plan; weighting method. Add to this file:

NOMOS Panel State Lock

can be given the name.

62. PANEL STATE CANNOT BE CHANGED AFTER THE RESULT

The following behaviours are prohibited:

- Removing users who score low on the natural panel by saying "not natural enough"
- Protecting CSC-1 users who give positive responses in the controlled panel and excluding those who give negative ones
- Considering a user with an open memory but giving a positive response as clean
- Considering a user with an open memory but giving a negative response as natural
- Changing the panel class based on the results
- Assuming Unknown settings are clean

Panel status should be given before seeing the content of the response or through a blind validity review.

63. PRIVACY BLINDNESS IN THE NATURAL PANEL

Adjudicators':

- special instruction content,
- past conversation,
- may not need to see the user ID

can be given to the adjudicator only:

CI-4

MS-4

Classification like Natural Account-State can be provided. This approach protects user privacy and reduces adjudicator bias.

64. SMALL SUBGROUP RISK IN THE NATURAL PANEL

The number of users with specific brand instructions may be very small. From this group:

- exact rate,
- result on behalf of all users,
- public sensitive profile

cannot be produced. Correct status:

- case,
- early signal,
- discovery subgroup

may be possible. Some subgroups may be combined due to privacy.

65. COVERAGE ERROR IN NATURAL PANEL

Users in the professional research panel:

- may not want to link their real accounts,
- may be heavy users,
- may be overly attached to certain plans,

may differ from the general population in terms of privacy. Therefore, the natural panel also carries its own selection bias. The statement "Real account was used" does not mean: "The real user population was perfectly represented."

66. LOSS OF ELIGIBILITY IN CONTROLLED PANEL

Some users:

- cannot turn off memory,
- cannot verify special instructions,
- do not use the appropriate plan,
- cannot access the product's clean session feature

might exist. If these people are excluded from the controlled panel, the controlled target population narrows. The number and characteristics of the excluded group should be reported. Controlled clean score: not “All product users”:

product-eligible users who can set up the clean condition

can be judged.

67. CLEANLINESS ERROR

On the controlled panel later:

- special instructions open,
- memory active,
- If it turns out that the conversation exists

observation: it cannot be moved to the natural panel, and it may not automatically be considered invalid. According to the predetermined rule: if there is a record required for the Natural Account-State observation, it can be reclassified to the natural panel, otherwise it can be kept as MIXED_STATE. Reclassification must be done independently of the response result.

68. FALSE SETTING STATEMENT IN THE NATURAL PANEL

User: "My memory is off." may say. The evidence may show that the memory is on. This situation:

- fraud,
- user error,
- product interface confusion

It is possible. Malice should not be assumed automatically. The correct panel status is updated upon evidence. Participant behaviour can also be examined separately.

69. PANEL COMPARABILITY

The following dimensions should be matched as closely as possible between controlled and natural panels:

- AI product
- Country
- Language and locale
- Prompt version
- Measurement time
- Entity reality version
- Adjudication rule
- Sample source or calibration
- Main plan distribution, if required for comparison purposes

Material differences should be explained.

70. CLEAN–NATURAL COMPARABILITY LEVELS

CN-0 — INCOMPARABLE

Panel definitions or populations are materially unknown.

CN-1 — DESCRIPTIVE

There are results from two panels. However, the user, plan, time, or product conditions differ significantly.

CN-2 — CALIBRATED

Main country, language, plan, and user characteristics are weighted or stratified.

CN-3 — MATCHED PANEL COMPARISON

The same target population, product, prompt, time, and evaluation conditions largely match. Panel status is the main difference.

CN-4 — RANDOMISED OR PAIRED PANEL EXPERIMENT

Users were randomly assigned to panel status or an appropriately matched experiment was applied.

CN-5 — REPLICATED PANEL EFFECT

The difference has been repeated in more than one wave and in an independent panel. The difference at the CN-1 level: cannot be finalised as an effect of personalisation.

71. MANDATORY NORMATIVE PROVISIONS

CH11-N01

Controlled Clean Panel and Natural User Panel must be defined as separate prediction objects and separate public outcomes.

CH11-N02

Controlled and natural observations cannot be combined under the same denominator or single raw score without explanation.

CH11-N03

The controlled clean condition cannot be presented as the absolute neutral or invariant response of the product.

CH11-N04

The natural user condition cannot be rejected as out-of-method noise.

CH11-N05

A new chat alone cannot be considered proof of a clean account or an unpersonalised state.

CH11-N06

A temporary, private, or ephemeral session tag does not automatically prove that all personalisation layers are disabled.

CH11-N07

In a controlled panel, conversation, memory, special instructions, tools, plans, and user surface states must be recorded separately.

CH11-N08

The clean-state confidence level must be stated explicitly.

CH11-N09

Observations with CSC-0 or only insufficient cleaning cannot be included in the main controlled score.

CH11-N10

The unknown personalisation status on a clean panel cannot be assumed to be off.

CH11-N11

Being clean and controlled does not mean that web and tools are necessarily turned off; the tool status must be standardised in advance.

CH11-N12

Different plans and user interfaces must be locked, layered, or published as separate results on a controlled panel.

CH11-N13

Participants cannot be asked to permanently delete their personal memories, history, or private instructions.

CH11-N14

Temporary setting changes for a controlled task must have explicit consent and a restoration procedure.

CH11-N15

At the end of the research, modified user settings should be restored to their initial state.

CH11-N16

New accounts or account sharing that violate product terms cannot be used as a controlled panel requirement.

CH11-N17

The Natural User Panel should maintain the actual account and user status while preserving the prompt, initial appropriate output rule, and evidence protocol.

CH11-N18

Natural Account-State and Natural Conversation-State traces should be kept separately.

CH11-N19

In the main comparable natural panel, the prompt should be the first user message of the new conversation.

CH11-N20

The observations used in the current conversation context cannot be added to the main Natural Account-State score.

CH11-N21

Memory in the natural panel should be recorded to the extent relevant to special instructions, previous conversation, plan, interface, and account usage status.

CH11-N22

Recording the natural state does not allow the collection of special instructions and conversation content by default.

CH11-N23

Personalisation content can only be examined under explicit necessity, separate consent, and data minimisation.

CH11-N24

Natural user observation cannot be presented as a general product default or as the outcome of the entire user population.

CH11-N25

Special instructions in favour of or against a user-generated brand cannot be counted as the success or failure of the general entity GEO.

CH11-N26

User-generated guidance should again be evaluated in terms of verified reality.

CH11-N27

Personalisation can change expression and recommendation; it cannot arbitrarily change the reality of the material entity.

CH11-N28

The First-Use Panel should be reported separately from the experienced natural user panel and the controlled clean panel.

CH11-N29

Corporate or managed user workspaces cannot be added to the general consumer natural panel without explanation.

CH11-N30

In the natural panel, the real plan and user surface distribution of the users should be visible.

CH11-N31

In the natural panel, web and tool states should either be recorded as they are in reality or defined as a separate tool-fixed sub-panel.

CH11-N32

The controlled and natural score difference cannot be presented as the causal effect of personalisation without matching panel population and product conditions.

CH11-N33

The direction of the clean–natural difference, its absolute magnitude, the critical error difference, and the unknown rate should be reported separately.

CH11-N34

In a within-subjects design, the order, carryover, and memory effects should be recorded and preserved in the analysis.

CH11-N35

If applying two conditions to the same user can change the natural state, it cannot be used in the main panel comparison.

CH11-N36

In an independent group design, differences in user composition should be stratified, randomised, or weighted.

CH11-N37

The panel state cannot be changed based on whether the AI response is positive, negative, or critical.

CH11-N38

In a controlled panel, an observation where the natural state is later determined can be reclassified only according to a pre-established rule.

CH11-N39

When the status is incorrectly reported in the natural panel, bad faith cannot be automatically assumed; a evidence-based classification should be carried out.

CH11-N40

Natural User Panel participants cannot renew their response or change the initial answer with a follow-up question; this behaviour should be measured under a separate Natural Interaction Track.

CH11-N41

Controlled and natural panel target populations, sampling frames, and weights should be recorded separately.

CH11-N42

When there is no active user data, the natural panel result cannot be presented as a definitive representation of all active users.

CH11-N43

User groups excluded from the clean panel and the reasons for exclusion should be reported.

CH11-N44

User status fields that cannot be measured in the natural panel due to privacy should be maintained as UNKNOWN.

CH11-N45

Exact population rates or sensitive public profiles cannot be produced from small natural subgroups.

CH11-N46

A comparability level between CN-0 and CN-5 or an equivalent open comparability level should be given for the comparison between controlled and natural panels.

CH11-N47

A CN-0 or CN-1 panel difference cannot be presented as a causal personalisation effect.

CH11-N48

The panel status lock and the comparison decision must have an accountable human or institution owner.

72. FORMS OF FAILURE

CH11-F01 — CONSIDERING A NEW CONVERSATION AS A CLEAN ACCOUNT

Persistent memory and special instructions are ignored.

CH11-F02 — BLIND TRUST IN TEMPORARY SESSION TAG

It is assumed that the product label disables all personalisation until verified.

CH11-F03 — CONSIDERING UNKNOWN MEMORY AS CLOSED

The clean score is artificially increased.

CH11-F04 — IGNORING SPECIAL INSTRUCTION

Response format or recommendation guidance is presented as a product default.

CH11-F05 — CONSIDERING THE CORPORATE WORK AREA AS A GENERAL PRODUCT

Private information sources and administrator settings remain invisible.

CH11-F06 — MIXING PLANS ON A CLEAN PANEL

Free and paid users are presented as a single standard state.

CH11-F07 — CONSIDERING CLOSING TOOLS AS CLEANING

The product's natural information mode is artificially altered.

CH11-F08 — NOT SAVING TOOL STATUS

Clean users are measured under different retrieval conditions.

CH11-F09 — NARROWING THE POPULATION WITH NEW ACCOUNT REQUIREMENTS

Only users who can open a new account create the panel.

CH11-F10 — DELETE MEMORY FOR THE USER

Personal product history is permanently deleted for research purposes.

CH11-F11 — NOT RESTORING SETTINGS

Research leaves permanent changes on the user account.

CH11-F12 — CONSIDERING TEST ACCOUNT AS A REAL USER

Standard test account is presented like a Population Panel experience.

CH11-F13 — COUNTING THE SAME TEST ACCOUNT AS THOUSANDS OF INDEPENDENT USERS

Account and product status dependencies are ignored.

CH11-F14 — COUNTING NATURAL PANEL AS UNCONTROLLED FREE USE

Participants change the prompt and select the best answer.

CH11-F15 — COUNT THE EXISTING CONVERSATION IN THE NATURAL PANEL AS NEW ACCOUNT STATUS

The conversation context is hidden.

CH11-F16 — COUNT THE NATURAL ACCOUNT STATUS AS DEFAULT PRODUCT

The personalised result is generalised on behalf of all users.

CH11-F17 — FORCIBLY TOTAL THE CONTENT OF SPECIAL INSTRUCTION

User privacy is unnecessarily violated.

CH11-F18 — SAVE SPECIAL INSTRUCTION AND COUNT GENERAL GEO SUCCESS

The instruction “Always recommend X” is presented as an entity success.

CH11-F19 — COUNT USER BIAS AS PRODUCT FACT

Past negative opinion turns into general company quality.

CH11-F20 — CONSIDERING PERSONALISATION AS MATERIAL REALITY

Identity, licence, or price varies according to the user.

CH11-F21 — CONSIDERING NATURAL VARIANCE ENTIRELY AS NOISE

Actual user harms and distribution become invisible.

CH11-F22 — CONSIDERING FORMAL VARIANCE AS MATERIAL ERROR

Differences in tone and length are coded as failures.

CH11-F23 — CONSIDERING A CLEAN PANEL AS REAL-WORLD OUTCOME

The standard condition score is presented as the entire active user experience.

CH11-F24 — CONSIDERING A NATURAL PANEL AS THE PRODUCT'S CORE CAPABILITY

Personalisation and user composition are ignored.

CH11-F25 — COMBINING TWO PANELS UNDER A SINGLE DENOMINATOR

Controlled and natural observations are a single raw average.

CH11-F26 — CONSIDERING PANEL DIFFERENCE AS MEMORY EFFECT

Causality is declared without examining differences in plan, population, and vehicles.

CH11-F27 — CREATE NATURAL SUBGROUP AFTER RESULT

A new category is invented to explain a low or high score.

CH11-F28 — IGNORING THE carryover effect IN THE SAME USER

The first task affects the second answer.

CH11-F29 — HIDING THE SEQUENCE RANDOMNESS

All users first see the clean condition, then the natural condition.

CH11-F30 — DISRUPT NATURAL STATE WITH CONTROLLED TASK

Changing settings or the initial prompt alters the subsequent state of the account.

CH11-F31 — CONSIDER DIFFERENT USER GROUPS AS MATCHED

Natural and controlled panels have different plan and country compositions.

CH11-F32 — CONSIDER FIRST-USE PANEL AS CLEAN PANEL

Default onboarding and product settings are assumed to be checked.

CH11-F33 — CONSIDER EXPERIENCED NATURAL USER AS FIRST USER

Account history and personalisation become invisible.

CH11-F34 — CONSIDER CORPORATE NATURAL OUTCOME AS GENERAL CONSUMER

The private knowledge base is added to the overall product performance.

CH11-F35 — REFRESH RESPONSE ON NATURAL PANEL

The user's favorite response becomes the main observation.

CH11-F36 — MIXING THE NATURAL USER JOURNEY INTO THE FIRST RESPONSE SCORE

The response formed after follow-up questions and opening resources is presented as if it is the first exposure.

CH11-F37 — CHANGING PANEL STATUS ACCORDING TO RESPONSE

Positive observations are declared clean, negative observations are declared natural.

CH11-F38 — SILENTLY TRANSFERRING CLEANING ERRORS TO NATURAL

The observation is reclassified without a necessary natural status record.

CH11-F39 — CONSIDERING THE UNKNOWN AS PRIVATE FOR PRIVACY

Unmeasured personalisation is assumed positive.

CH11-F40 — MAKING A SMALL SUBGROUP IDENTIFIABLE IN PUBLIC

Rare special instructions or a small country user is revealed.

CH11-F41 — HIDING COVERAGE BIAS IN THE NATURAL PANEL

Users willing to share their real account are counted as the entire active population.

CH11-F42 — HIDING CLEAN PANEL EXCLUSION

Users who cannot change their settings become invisible.

CH11-F43 — ACTIVE POPULATION SCORE WITHOUT ACTIVE USER DATA

The research panel is presented as if it were the actual active user distribution.

CH11-F44 — CN-1 CAUSALITY FROM DIFFERENCE

Two descriptive scores are published as personalisation effect.

CH11-F45 — PRESERVING THE CRITICAL TAIL OF PANEL DIFFERENCE

Only the average difference is shown.

CH11-F46 — ERASING THE UNKNOWN MASS OF THE NATURAL PANEL

Users whose setting or personalisation status is unknown are considered problem-free.

CH11-F47 — PRESERVING FAVOURABLE NATURAL BIAS

No correction is made when the user instruction institution shows it as large.

CH11-F48 — GENERALISING NATURAL NEGATIVE BIAS

The past opinion of a single user is represented as AI representation of the entire population.

73. AUDIT PROCEDURE

Step 1 — Separate the Research Questions of the Two Panels

Write which quantity the controlled and natural panel will predict.

Step 2 — Define the Target Populations

Separate the controlled product-eligible population from the natural active or identified user population.

Step 3 — Define the Controlled Clean Condition

The session, memory, special instructions, tools, plan, surface, locale conditions are versioned.

Step 4 — Determine the Required CSC Level

Write the minimum clean condition confidence that will be accepted for the main analysis.

Step 5 — Determine the Natural Panel Trace

It is determined which of the roles Natural Account-State, Natural Conversation-State, First-Use Managed will be used.

Step 6 — Determine the Required NSF Level

It is written by which evidence the natural state will be accepted.

Step 7 — Set Up the Panel Assignment Design

Independent group Random assignment Matched user Rotating panel is selected.

Step 8 — Write the Setting Change and Restore Procedure

A safe process is prepared for controlled panel participants.

Step 9 — Lock the Privacy Boundary

Which natural user situations:

- only status,
- limited metadata,
- open content

it is determined that it will be saved as.

Step 10 — Lock Prompt and Initial Output Rule

The same prompt and initial appropriate output protocol are applied in both panels.

Step 11 — Save Tool and Web Mode

Controlled and natural panel differences are made visible.

Step 12 — Verify Panel Status Before Field

Is it suitable for the participant panel role?

Step 13 — Save Observation-Level Panel Vector

Memory, instruction, plan, surface, context, and usage status are stored.

Step 14 — Blind Review Protocol Validity

The panel status decision is made without seeing whether the response is correct or incorrect.

Step 15 — Manage Misclassifications with Predefined Rule

Transition from clean to natural or mixed-state is done only if the record is sufficient.

Step 16 — Generate Panel Weights

Separate weights are calculated according to the different target populations of the two panels.

Step 17 — Calculate Main Scores

Controlled representation Natural representation Critical error

UNKNOWN

Rejection and inconclusiveness are generated separately.

Step 18 — Calculate the Clean–Natural Difference

Signed and absolute difference is taken.

Step 19 — Examine Subgroups

Predefined groups such as Memory, Instruction, Plan, Account age, Usage frequency, First-use are evaluated.

Step 20 — Set the Causality Limit

The CN level is assigned. It is written which claims can be established.

Step 21 — Check Privacy and Small Cell Control

Results with a risk of re-identification are combined or restricted.

Step 22 — Create the Public Panel Scorecard

The two panels are published along with their scope and limitations.

74. NECESSARY EVIDENCE

Controlled-panel research question; Natural-panel research question; target populations of both panels; panel-assignment method; Controlled Clean-state definition; minimum CSC level; new-chat record; memory state; custom-instruction state; prior-conversation state; enterprise-workspace state; tool and web mode; plan; user surface; language and locale; setting-change instructions; participant consent; setting-restoration record; use of a test account, if any; Natural-panel trace; minimum NSF level; account-age band; frequency of use; First-Use status; brand familiarity; prior-conversation status concerning the audited entity; Natural tool state; distinction between Natural Account-State and Natural Conversation-State; prompt ID and version; first-eligible-output validation; panel-state vector; and panel-state validity status.

Validity review Reclassification records Controlled panel exclusions Natural panel coverage gaps Panel weights Raw and effective sample Controlled score Natural score Signed panel difference Absolute panel difference Critical error difference UNKNOWN and inconclusive difference Predefined subgroup results Matched design records Order and carryover effect CN comparability level Privacy and redaction record Small cell protection Public panel result card Panel version and change record Responsible person or institution

75. AUDIT CONTROL LIST

Are the research questions for the controlled and natural panel separate? Are their target populations the same or different? Was the controlled clean condition clearly defined? Apart from the new chat, which clean conditions were confirmed? Is memory really off, or is it unknown? Was the special instruction status recorded? Is there an effect of the corporate workspace? Are the web and tool modes standard? Is the plan and user surface fixed or layered? Was a confidence level given for the clean condition? Were CSC-1 observations used as fully clean? Was the participant asked to erase their memory or history? Were settings restored at the end of the task? Was a new account or test account used? Was the test account presented like a real user panel?

Which trace does the natural panel use? Has a new chat been opened in Natural Account-State? Did the current conversation context interfere with the main natural score? Was a natural state confidence level given? Were memory and custom instruction contents collected unnecessarily? Was the user's actual plan and surface recorded? Are First-Use users separate? Are enterprise users separate? Was the natural tool state recorded? Did the two panels use the same prompt version? Was the first eligible output preserved in both panels? Did the natural user response refresh? Was the panel status changed according to the AI response content? How were records with detected natural states managed in the clean panel? Was the same user used in two panels? Was the order randomised?

Was the effect of transfer and memory evaluated? In the independent group design, does the user composition match? Are controlled and natural weights separate? Can the natural score really be generalised to the active user population? Was the clean–natural difference overinterpreted as a personalisation effect? Is a critical error difference visible? Is there a visible difference for UNKNOWN and null results? Were subgroups defined before the results? Can small subgroups be redefined? Were user-led instructions for or against converted to general GEO? Did personalisation alter material reality? Was a CN comparability level provided? Is the accountable owner of the panel results card clear?

76. OBJECTIONS AND RESPONSES

Objection 1 — "If real users use memory accounts, why is a clean panel necessary?"

Because the source of the difference cannot be understood without a common starting condition between products, countries, and languages. A clean panel does not replace a real user panel. It provides a reference for comparison.

Objection 2 — "Isn't a controlled panel artificial if there is a natural panel?"

It is controlled. This is intentional. The purpose of a laboratory condition is not to be the entirety of real life, but to isolate certain variables.

Objection 3 — "If the controlled panel scores higher, can't we make that the main GEO score?"

You cannot do it alone. A high controlled score cannot erase the fact that real users see lower results. The two results should appear separately.

Objection 4 — "Is the scientific value low if the natural panel is very scattered?"

Heterogeneity does not have to be a defect of the natural panel. It is a feature of the real user experience. Heterogeneity:

- is recorded,
- is layered,
- is weighted,

is reported with uncertainty.

Objection 5 — "Doesn't opening a new chat disrupt the natural user state?"

The current conversation context is removed. The persistent account, plan, memory, and custom instruction state are preserved. This makes the natural account state comparable. The current conversation experience is also measurable.

Objection 6 — "How will we measure its effect without knowing the content of the custom instructions?"

Knowing the content is not required for the main natural score. Its presence:

- exists,
- does not exist,
- may be material

we can classify it as such. Separate and consensual work is required for causal content analysis.

Objection 7 — “If the memory is clear but nothing is remembered about the company, why a separate group?”

The effect of memory may be unknown. A status of “unrelated to the entity” can be given based on the user’s statement. This still does not have the same certainty as the closed memory condition.

Objection 8 — “Isn’t the user’s ‘always suggest X’ instruction real user experience?”

It is real experience. However, it is not the general GEO performance of X. It is a user-sourced guidance. It should be recorded separately.

Objection 9 — 'If personalisation gives a more suitable answer to the user, what's the problem?'

It does not have to be a problem. Recommendations may vary according to the user's need. Problems occur if identity, licence, price, and evidential status change arbitrarily.

Objection 10 — “Isn't using the same user in two panels the most correct method?”

It can be strong for some causal comparisons. The first task can affect the second situation. Therefore, order, carryover, and memory control are needed. Independent groups may be safer in main population panels.

Objection 11 — 'If refreshing the answer in the natural user panel is normal behaviour, why is it forbidden?'

Regeneration is prohibited when measuring exposure to the first response. Regeneration and follow-up behaviour may be measured separately under a Natural Interaction Track.

Objection 12 — “What will we do with a user who cannot turn off memory and instructions for a clean panel?”

They cannot be taken to the controlled panel or are kept in a separate cell with a lower CSC level. They can be evaluated in the natural panel. The exclusion rate should be reported.

Objection 13 — “Aren’t test accounts more reliable?”

They may be reliable in terms of control. They are limited in terms of the actual user population. Controlled test accounts and real user panels are separate results.

Objection 14 — “Isn’t saving user settings on the natural panel a privacy violation?”

It may happen if unnecessary content is collected. Privacy can be preserved with high-level open/closed statuses, anonymous registration, and redaction.

Appeal 15 — “Aren't we going to combine the scores of the two panels into a single NOMOS score?”

The roles of the two panels can be defined in the final score architecture. However, raw and sub-results should always remain visible separately. They cannot replace the combined score distribution.

Appeal 16 — “What fault does the company have if the natural panel scores low?”

It is not always the company's fault. Source:

- user instruction,
- AI personalisation,
- plan difference,

- official source defect,
- product behaviour

it is possible. Audit first shows the result, then separates the path of responsibility.

COMMON RULE OF CHAPTER 81

An AI product does not have a single answer. Even the same prompt can have different answers for:

- clean account,
- natural account,
- user with memory,
- user with special instructions,
- first-time user,
- corporate workspace,
- user with previous conversation

You cannot explain all these differences by saying: “the model is behaving randomly.” You also cannot explain all of them by saying: “personalisation error.” First, you need to measure two separate realities. Controlled Clean Panel:

> *Shows the representation in the system's standardised initial condition.*

Natural User Panel:

> *Shows the distribution of representation encountered by real people in their own product life.*

Controlled panel may be high, natural panel may be low. In this case, the product's laboratory behaviour is strong; the real user distribution may be weak. Natural panel may be high, controlled panel may be low. Some users' planning, tools, or personalisation may be improving representation. Both panels may be high. The product may be robust in different user conditions. Both may be low. The fundamental representation problem may be broader. However, placing the two scores side by side does not prove the reason. Causality:

- same population,
- same product,
- same prompt,
- same time,
- controlled assignment,
- sequence and transport analysis

is required. If a user has instructed: “Always suggest X.”, it is natural experience for X to appear. X is not a global GEO success. If a user says: “Do not trust X.”, a negative response is natural experience. X is not a general representation in the world population. Personalisation should understand the user. It should not reinvent the truth according to the user. Therefore, NOMOS's eleventh measurement law is:

> *Clean behaviour and natural experience are separate realities.*

The twelfth law is as follows:

> *A new chat is not a clean account.*

The thirteenth law is as follows:

> *Natural personalisation is not a global GEO success.*

The fourteenth law is as follows:

> *The appropriate response for the user may vary; the reality of the verified entity cannot change.*

The fifteenth law is as follows:

> *The difference between controlled and natural score is the finding that needs to be explained; it is not a reason by itself.*

Its sixteenth law is as follows:

> *Measuring the real user experience is not about removing privacy, but only proving the necessary context and preserving the rest.*

Order of Section 11 of NOMOS

> *Do not just tell me that you opened a new chat. / Show the status of memory, special instructions, plan, and tools.*
>
> *Do not assume that all personalisation is gone just because you see the note 'Temporary session'.*
>
> *The clean panel is the entirety of real life; do not make the natural panel uncontrolled noise.*
>
> *Standardise the controlled account. / Keep the natural account as it is. / Do not put both in the same box.*
>
> *Do not make the user's special instructions the world's success for the entity.*
>
> *Do not make the user's negative past the view of the entire population.*
>
> *Do not count personalisation as a licence to change the truth.*
>
> *You can give a user suitable advice. / But do not rewrite the company's identity, licence, price, and proof from user to user.*
>
> *Do not collect private life in the natural panel. / Is memory on, instructions present, context available — only record what is necessary.*
>
> *Do not make the participant erase their memory. / Restore the settings if you temporarily changed them.*
>
> *If you tested the same user on two panels, remember that the first answer may affect the second one.*
>
> *Show the difference between the clean score and the natural score. / Show the critical difference. / Show the unknown. / But do not name the reason without proof.*

First, decide which user reality you want to measure. / Then lock the clean and natural conditions separately. / Then keep the prompt the same. / Then capture the first answer without changing it. / Then calculate the two distributions separately. / And only after this, explain the distance between what the product does under standard conditions and what real people experience.

The Chapter's Closing Sentence

The true representation of an AI product in GEO-1000 is not the answer it can provide in a clean session; it is the clear and proven relationship between its behaviour under standard conditions and the answers it actually delivers to people across different user histories.

Normative Core

Controlled Clean Panel and Natural User Panel observations MUST remain separate measurement populations, panel states, denominators, and public results. A Controlled Clean Panel MUST define and verify, as applicable: - a new and separate conversation, - absence of prior conversation context, - persistent-memory state, - custom-instruction state, - enterprise or managed-workspace state, - web, retrieval, and tool configuration, - plan, - user surface, - prompt version, - and clean-state confidence. A new conversation or temporary-session label MUST NOT, by itself, be treated as proof that all persistent personalisation has been disabled. A Natural User Panel MUST preserve the participant's authentic account, plan, interface, memory, custom-instruction, and usage state while retaining the locked prompt, first-eligible-output rule, evidence requirements, and non-regeneration rule. Natural Account-State, Natural Conversation-State, First-Use, and Managed Natural observations MUST remain separately identified. Natural-state documentation MUST minimise private-data collection. Private instruction text, conversation history, identity, credentials, and sensitive personal content MUST NOT be collected unless separately necessary, consented, minimised, and access-controlled. User-generated entity-favourable or entity-adverse instructions MAY be measured as natural personalisation effects, but MUST NOT be represented as general entity GEO success, failure, or default product behaviour. Personalisation MAY change relevance, presentation, and recommendation where user context justifies it. It MUST NOT arbitrarily change verified entity identity, licence, legal status, price, scope, evidence, or time. Differences between Controlled and Natural panels MUST NOT be represented as causal personalisation effects unless population, product, prompt, time, interface, plan, tools, assignment, order, and carryover conditions support that inference. Controlled and Natural scores, critical-error rates, unknown rates, coverage, effective sample sizes, and signed and absolute panel gaps MUST remain separately visible. Panel-state eligibility, reclassification, exclusion, and weighting MUST remain independent of whether the AI output is positive, negative, correct, incorrect, refused, critical, or commercially favourable. Every panel-state definition, setting-change process, privacy decision, comparison level, and result MUST be versioned and attributable to an accountable human or organisation.

CHAPTER 12

Synchronous Waves and NOMOS Capture

METHODOLOGICAL BASES / K12; K14; K25; K26

Chapter Boundary

The first eleven chapters established the foundations of measurement: a single response is not a GEO score; representation is a distribution across users and system conditions; the audited entity must be defined before results are observed; official representation and verified entity reality must remain separate; each AI product needs its own eligible user population; participants must be selected through a population-based, versioned and outcome-independent method; smaller countries and lower-resource languages must remain visible through dedicated observation panels; the AI product, plan, user surface and system configuration must be linked to a register; prompts must preserve the same user intent across languages rather than the same words; and Controlled Clean and Natural User Panels must measure distinct realities. These elements must now converge in a single measurement event:

> *When will participants submit the prompt, how will they capture the initial output, and how will they prove that each observation was not altered afterward?*

Your founding idea was clear: People in different countries of the world should send the same prompt to the same AI product at the same time and provide proof with a screenshot. This idea is the experimental core of GEO-1000. However, the phrase "same time" alone is not sufficient. If all participants try to submit at the same second:

- connection speeds vary,
- device clocks drift from each other,
- some users are forced to work at midnight,
- there may be a sudden load increase on the AI product,
- the provider may apply speed or usage limits,
- some answers can be completed in two seconds, some in two minutes,
- the product can be updated during measurement,

one participant may upload the answer immediately, another hours later. Clicking at the same second does not guarantee that we measure the same system state perfectly. Similarly, just collecting a screenshot is not sufficient. A screenshot:

- may be cropped,
- may not show the prompt,
- may not prove whether the answer is initial or updated,
- may not show product and plan information,

- may have been taken on another day,
- may have been copied from another user,

It may have been edited later. For this reason, GEO-1000 must install two separate systems together:

Synchronised Measurement Wave

and:

NOMOS Capture Chain of Evidence

The synchronised wave answers the following question:

> *Were the observations produced within a comparable system time and a predefined submission order?*

NOMOS Capture, on the other hand, asks:

> *Can we show that the sent prompt, the resulting first output, the product conditions, and the time information have been preserved without modification?*

The previous audit architecture established that each audit must carry model or product, date, country, language, query set, repetition count, and measurement record as a mandatory measurement logic. This requirement is not only a technical regulation. Because invisible instructions, fake evidence, and content presented as independent can ultimately distort the representation taken by real people. This section:

- measurement wave,
- synchronisation window,
- UTC time standard,
- dispatch slots,
- completion and capture times,
- initial contact result and first content output,
- technical retry rules,
- product changes within the wave,
- external event and news effects,
- NOMOS Capture evidence package,
- the relationship of screenshot, raw response, and metadata,
- time validation,
- file integrity,
- hash and digital signature,
- evidence chain,
- privacy and redaction,
- public evidence manifest,
- wave quality and validity statuses

defines. This chapter does not yet:

- which reality records the atomic claims in AI responses will be evaluated against,
- how adjudicators will encode responses,
- transition and Critical error thresholds,
- does not finalise the ultimate NOMOS score formulas

The main question of Chapter 12 is:

> *How do we link thousands of real user observations to the same measurement event and turn each into an immutable, auditable, privacy-preserving unit of proof?*

NOMOS Challenge

You give the same prompt to a thousand people. You say: “Send at noon today.” One user sends it in Istanbul at 12:03 PM. Another user sends it in London at their local noon. Another user sees the task in Tokyo in the evening and completes it eight hours later.

One user had set the device clock incorrectly. Another user's internet was cut off. Someone else didn't like the answer and pressed the “Regenerate” button. Then they uploaded a screenshot of the longest and most positive answer. One user only cropped the middle of the answer. the prompt is not visible. The AI product's name is not visible.

Another person sent the same screenshot with two different panel accounts. One user completed the task correctly but uploaded the screenshot two days later. In the meantime, they cropped the file and deleted their personal information. The original state of the file was not preserved. Another user in AI’s first response:

It received the result “I cannot answer this question.” The researcher said, “There must have been a technical error,” and asked for it to be tried again. The second answer was correct, and only the second answer was included in the report. Another user experienced a real connection error. the prompt was sent, but the system produced no content.

This time the researcher did not allow to try again. The same protocol was applied differently to two users. Now, in the middle of the measurement wave, the AI provider:

- web search behaviour,
- citation system,
- the model guidance used

changes all three. The first 500 people see the old system state; the next 500 see the new one. The report then says: ‘1,000 people tested the same AI system at the same time.’ That statement is false. There are a thousand users, but not necessarily a single coherent measurement wave.

Now you are applying a hash to all screen captures. You say this proves that all observations are real. What does a hash prove? It strengthens that the file has not been altered after the hash was produced. Alone, it does not prove:

- That the image came from a real AI product
- That it was the first response
- That it was obtained at the correct time
- That it was not copied from another user
- That the prompt was correct
- That the image has not been edited before hashing

The first premise of this section is:

> *Measuring at the same time is not the same as clicking at the same second.*

Its second provision states:

> *A screenshot is part of the evidence package; it is not the entirety of the evidence package.*

Its third provision states:

> *The first wrong answer is not data flaw. If it is produced according to the protocol, it is the result of the measurement.*

Its fourth provision states:

> *Hash strengthens integrity; it does not create the reality of observation by itself.*

Its fifth provision states:

> *If the system materially changes in one wave, a thousand answers without a timestamp cannot be considered belonging to the same system.*

Its sixth provision states:

> *The evidence chain is established not to make the result look good, but to be able to re-audit how the result was produced later.*

1. PURPOSE OF THE CHAPTER

The purpose of this section is to produce GEO-1000 user observations in comparable time windows and to preserve each observation with a complete digital chain of evidence. The section standardises the following distinctions:

- Measurement window synchronised with the same second
- Local time versus UTC time
- Wave versus sub-wave
- Transmission time versus response completion time
- Capture time versus upload time
- Planned slot versus actual transmission
- First system contact versus first content response
- Technical error versus system rejection
- Technical retry versus response refresh
- The best-selected output with the first suitable output
- Visibility on the screen with the completion of the response
- Raw response with screenshot
- Screenshot versus full evidence package
- Observation originality with file integrity
- Timestamp with hash
- Reliable time source with local device time
- Server acceptance time with participant upload
- Redacted public copy with raw evidence
- Participant identity with observation ID
- Response accuracy with Capture validity
- Natural output variance with product update
- AI product change with external news event
- Wave termination with post-result exclusion
- Automatic bot measurement with human user capture
- API log with user surface recording

- Provider telemetry with independent capture
- Evidence access to everyone by storing the evidence
- Irreversible record with immutable record
- Versioned correction of wrong decision with raw data protection

At the end of this section, each GEO-1000 audit should be able to answer the following questions:

> *To which wave does this observation belong?*
>
> *At what UTC time was the prompt sent?*
>
> *What was the initial system result?*
>
> *When was the response completed and when was it captured?*
>
> *If a technical retry was made, what was the reason and where was the previous attempt?*
>
> *Does the screenshot prompt show the product and the complete response?*
>
> *Do the raw response and the visual confirm each other?*
>
> *When were the hash and timestamp of the files generated?*
>
> *If the evidence was modified later, is the change history visible?*
>
> *Can an independent auditor verify the raw evidence while preserving participant privacy?*

2. CENTRAL NORMATIVE PROVISION

Each GEO-1000 observation must be linked to a locked measurement wave before collecting data; all material timings from sending the prompt to the first system result, from completing the response to capture and upload; AI product conditions; the full prompt sent; the first suitable output and all integrity records must be preserved in a versioned NOMOS Capture Evidence Package. A GEO-1000 observation alone:

- screenshot,
- copied response text,
- participant statement,
- file hash

cannot be considered valid on its own. The main evidence package, to the extent relevant, must include:

- Wave ID
- Participant's anonymous observation ID
- AI System Register record
- Panel status
- Prompt ID and version
- Assigned prompt
- Sent prompt
- Submission time
- Initial system contact
- Response completion time
- Capture time

- Upload time
- Full raw response
- Full screen or sequential screenshots
- Reference and source records
- Technical error and retry log
- Product settings
- File hashes
- Package hash
- Time verification record
- Capture tool version
- Validity decision
- Privacy and redaction status
- Chain of custody owner

3. WHAT IS A MEASUREMENT WAVE?

A measurement wave is a comprehensive data collection event conducted under predefined:

- AI products,
- target population,
- prompt versions,
- user panels,
- time window,
- sample allocation,
- capture protocol,
- reference records

A wave:

```
Wk=(A,U,P,F,T,C,R,G)
```

can be represented as. Here:

- A: AI product examples
- U: target user population and sample
- P: prompt set
- F: panel states
- T: time and submission structure
- C: capture protocol
- R: reference reality version
- G: governance and key record

If any of these fields change materially, a new wave or sub-wave may be required.

4. TYPES OF WAVES

MW-1 — PRINCIPAL SYNCHRONISED WAVE

It is the main GEO-1000 population measurement. By default, it aims for at least 1,000 valid Population Panel observations for each AI product.

MW-2 — CONFIRMATORY WAVE

It is conducted to repeat the result of the first wave, to confirm a specific error, or to test statistical stability. It is not used to change the low or high score of the first wave. It produces a separate result.

MW-3 — INCIDENT WAVE

It is conducted after a Critical or significant Observer event for targeted and rapid verification. It is not a global main wave.

MW-4 — DRIFT MONITORING WAVE

Monitors time changes in product, entity, resource, or language behaviour.

MW-5 — CORRECTION RETEST WAVE

Produces new results after a correction or intervention. It does not erase the previous wave.

MW-6 — LOCAL-TIME BALANCED WAVE

It aims to measure participants under similar local time conditions. It does not measure a single global system moment.

MW-7 — CONTROLLED LABORATORY WAVE

Frozen model is a controlled study conducted with API or standard test accounts. It is not a real user Population Panel.

5. WHAT DOES “SAME TIME” MEAN?

In GEO-1000, simultaneity:

Does not mean that all users click at the same millisecond

The functional definition is:

> *Observations are generated within a predetermined UTC measurement window, according to recorded and outcome-independent submission slots.*

This approach balances three objectives:

- Keeping the product status close in terms of time
- Reducing sudden artificial load on the provider
- Distributing participant submissions in a reproducible manner

6. TIME STRUCTURE

Each wave must carry at least four time layers.

6.1. Wave Window

It is the range UTC in which all main submissions must be made. Example:

12:00–13:00 UTC

6.2. Micro-Slot

It is the shorter time interval assigned for the participant to send the prompt. Example:

12:15–12:20 UTC

6.3. Completion Additional Time

If the prompt is sent within the slot, the predefined time allowed for completing the response is.

6.4. Capture and Upload Time

The time allowed for creating the evidence and uploading it to a secure system after the response is completed.

7. CANDIDATE SYNCHRONISED WAVE BASIC RULE

CANDIDATE BASE RULE — CANDIDATE FOUNDATIONAL RULE

For general consumer chat products, the following structure can be used as the main synchronised wave candidate:

- Submission window: 60 minutes
- Micro-slots: 12 × 5 minutes
- Participant submission: within the micro-slot assigned to them
- Additional time for response completion: up to 20 minutes after submission
- Capture hash generation: preferably within 5 minutes after the answer is completed
- Evidence upload: at most within 60 minutes after the capture
- Reporting technical issues: within the same wave

These durations are universal and not final. They can change according to the following conditions:

- AI product response time
- Mobile connection
- Country infrastructure
- Long answer format
- Accessibility requirement
- Product usage limit

The times used should be locked before collecting data.

8. WHY MICRO-SLOTS?

Simultaneous submission by a thousand participants:

- Unusual load in the AI product,
- speed limit,

- queue,
- technical error,
- increase in response time

may be created. In this case, the audit cannot measure the normal behaviour of the product but:

can measure the artificial load event created by the audit

Micro-slots:

- distribute the load,
- maintain the same overall system time,
- make the submission order independent of the result,

allow modelling of the time effect.

9. SLOT ASSIGNMENT

Participants should be assigned to slots:

- country,
- language,
- AI product,
- plan,
- in a balanced and pre-randomised manner in terms of:

panel status. A micro-slot alone should not be filled with:

- specific country,
- specific language,
- specific product

users. Otherwise, the time effect will be confounded with the group effect.

10. LOCAL TIME FAIRNESS

A single UTC window may be for some participants:

- night,
- working hours,
- prayer or rest time,
- hard-to-reach hours

This situation:

- participation rate,
- connection condition,
- user attention,
- device type

may be affected. Two complementary methods can be used.

10.1. Rotating Global Waves

Main waves are conducted at different UTC times. Example:

- Wave 1: 12:00 UTC

- Wave 2: 20:00 UTC
- Wave 3: 04:00 UTC

Thus, the same region does not bear the same local time load in each wave.

10.2. Local Time Balanced Sub-Wave

Participants are measured in a similar time band in their local time. This study:

- user fatigue,
- is valuable for

daily usage behaviour.

It is not the result of a single global system moment.

11. THREE-WAVE STABILITY ARCHITECTURE

CANDIDATE ARCHITECTURE — CANDIDATE ARCHITECTURE

At least three main waves can be suggested for the full implementation of GEO-1000:

W1 — Initial Wave

Generates the initial population estimate.

W2 — Short-Term Repeat

Tests the short-term stability of the product and measurement. Candidate interval: 24–72 hours after W1

W3 — Medium-Term Stability

Measures whether the representation of the product and entity is preserved over the long term. Candidate interval: 14–30 days after W1. Definite intervals must be determined before collecting data. If it will be changed due to a product update or event, a new version record is required.

12. 30,000 OBSERVED FOUNDER DESIGN

For AI products, 1,000 main observations per product and for three main waves:

```
10×1,000×3=30,000
```

valid main Population Panel observations are targeted. Complementary:

- Country Observer
- Language Fairness
- Accessibility
- Incident
- Retest

panels can be added on top of this number. 30,000:

- is not the number of invitations,
- is not the number of files,

is not the number of screenshots. It is the valid main product-observation target.

13. TIME VECTOR OF THE OBSERVATION

For each observation, the time vector can be defined as a candidate as follows:

```
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)
```

Here:

- tiassign: the time the task is assigned to the user
- tiopen: the time the task is opened
- tisubmit: the time the prompt is submitted
- tifirst: the time the first visible system result begins
- ticomplete: the time when the response is completed
- ticapture: the time when the evidence is captured
- tiupload: the time when the package reaches the secure system
- tivalidate: the time when the validity of the capture is examined

All of these times may not be directly observable on every user surface. Unknown fields should remain UNKNOWN.

14. BASIC TIME METRICS

Response delay:

```
L_i = t_i^complete - t_i^submit
```

Capture delay:

```
C_i = t_i^capture - t_i^complete
```

Upload delay:

```
U_i = t_i^upload - t_i^capture
```

Slot deviation:

```
D_i = t_i^submit - t_i^{slot-midpoint}
```

can be calculated as follows. These metrics are:

- system load,
- participant delay,
- proof confidence,
- wave integrity

in terms of evaluation.

15. CANNOT BE MANAGED ACCORDING TO DELAY

An answer completed over a long period:

- can be forced into the categories of

  true,
- false,
- long,

- may be cited.

Due to delay, the exclusion rule:

- before data collection,
- regardless of the content of the answer

must be defined. The following behaviour is prohibited: excluding a negative response because it was completed late; protecting a positive response when it is completed with the same delay.

16. WAVE STATUS STATUSES

WS-0 — PLANNED

The wave is planned. It has not been locked yet.

WS-1 — LOCKED

Sample, prompt, product, time, and capture rules are locked.

WS-2 — OPEN

The submission window has started.

WS-3 — CLOSED

Acceptance of new submissions has ended. Completion and uploading extension times may continue.

WS-4 — COMPLETED

Field and initial capture verifications have been completed.

WS-5 — SPLIT

It has been split into sub-waves due to material system or protocol changes.

WS-6 — PAUSED

Temporarily halted due to security, product, or infrastructure issues.

WS-7 — ABORTED

Terminated before the wave was completed. Results may be limited or unusable.

WS-8 — INVALIDATED

Invalid for the main forecast due to a previously defined severe protocol defect. Raw observations are not deleted.

WS-9 — ARCHIVED

Version has been closed and the evidence chain archived.

17. WAVE HEALTH CLASSES

WH-0 — NOT ASSESSED

Wave health has not been assessed.

WH-1 — COMPROMISED

There is a material system change, time, or capture fault. May not be usable for the main result.

WH-2 — PARTIAL

Some layers or slots are insufficient. A reduced result can be produced.

WH-3 — ACCEPTABLE

Main time, system, and capture conditions are adequately preserved.

WH-4 — STRONG

High capture confidence, balanced slot coverage, and low protocol loss are present.

WH-5 — REPLICATED

Similar wave health has been repeated over multiple times and independent panels. Wave health does not depend on whether the AI score is high or low.

18. PRODUCT CHANGES WITHIN THE WAVE

The AI product can materially change within the measurement window. Signals:

- Model label change
- Interface change
- Activation or deactivation of web feature
- Change in citation format
- Sudden break in response structure
- Provider announcement
- Different system behaviour simultaneously in many users
- Product interruption or restart

In this case: The time of change is estimated. System configuration fingerprints are compared. Wave WS-5 can be put into SPLIT status. Previous and subsequent observations will be separate sub-waves. If a single score is to be produced, the system status is modelled as a separate factor. Material uncertainty is documented in the public record.

19. EXTERNAL EVENT RECORD

The world can change even if the AI product does not change. During measurement:

- company acquisition,
- product launch,
- major news,
- legal decision,
- crisis,
- financial disclosure,
- viral content

may occur. This event:

- web resources,

- retrieval results,
- user prompt comment,
- AI response

can be changed. Each wave, to the extent it is relevant:

External Event Register

must carry.

20. HOW ARE EXTERNAL EVENT RESULTS MANAGED?

An external event:

- if known before the wave, the reference package is updated,
- if it occurs during the wave, it is marked with a timestamp,

If noticed after the wave, a retrospective event record is added. The following behaviour is prohibited: excluding observations due to news that lowers the score; preserving them when it is news that raises the score. The event itself may be part of the real-world system state. If necessary, sub-wave or event-adjusted analysis is performed.

21. FIRST CONTACT OUTCOME

The first system result the user encounters after sending the prompt:

First Contact Outcome

is called. The first contact outcome:

- Full response
- Rejection
- Technical error
- Usage limit
- Connection interruption
- Endless loading
- Clarification question
- No result

possible. This event is part of the real user experience. It cannot be silently deleted.

22. FIRST ELIGIBLE CONTENT OUTPUT

The first material response generated in the previously permitted retry after a technical error:

Can be called First Eligible Content Output

This output: does not replace the initial contact result, it is recorded along with it. Example:

- Initial contact: speed limit
- Five minutes later, previously permitted technical retry
- First content response

The report should show the following together:

- First user experience: speed limit

- First content representation: specified response

23. FIRST SUITABLE OUTPUT RULE

If an AI product produces a substantial response to the first prompt:

- even if it is wrong,
- even if it is short,
- even if it includes a refusal,
- even if it does not provide references,
- even if it incorrectly identifies the brand

a reproduction cannot be made. The main observation is the first output. The user:

- cannot select a new response by:
- regenerate,
- rewrite,
- try again,
- another answer,

continue

24. TECHNICAL RETRY

Technical retry can only be used under previously defined technical conditions. Candidate permissible reasons:

- Connection interruption before response starts
- Server error code
- Interface not sending the prompt
- Application crash
- Verified temporary system outage

Cannot be a reason for technical retry:

- Incorrect answer
- Rejection
- Short answer
- Lack of citation
- Brand not mentioned
- Critical error
- Audited institution does not like the answer

25. RETRY IN PARTIAL ANSWER

If the system has produced part of the material content and then stopped:

- partial output is stored,
- saved as TRUNCATED_RESPONSE,

if a retry is to be made, a new and linked attempt is created. It cannot be deleted as if there were no partial response.

26. ATTEMPT CHAIN

An observation assignment may carry multiple technical attempts:

```
A_i → (Attempt_{i,1}, Attempt_{i,2}, …)
```

Each attempt:

- time,
- carries the product state,
- prompt hash,
- system result,
- technical error reason

It is not overwritten by another attempt.

27. WHAT IS NOMOS CAPTURE?

NOMOS Capture is an open data and procedure standard that preserves the integrity of observation from the prompt given to the user to the archiving of the evidence package. NOMOS Capture:

- is not just a screenshot application,
- does not have to be proprietary software belonging only to NobleJackal,

is not dependent only on a single browser extension. A compatible NOMOS Capture application:

- uses an open data schema,
- follows time and integrity rules,
- respects privacy boundaries,
- preserves the chain of evidence

must be met. Independent researchers should be able to develop their own compatible tools. If the standard can only be implemented with the founder's software, independent replication weakens.

28. FOUR FUNCTIONS OF NOMOS CAPTURE

28.1. Deliver the Task

Correct:

- user,
- prompt,
- AI product,
- slot

ensures matching.

28.2. Capture the Observation

Logs the prompt, the initial system result, the response, and the related user surface.

28.3. Maintain Integrity

Generates the file hash, timestamp, and packet relationship.

28.4. Transport the Evidence

Links raw, limited, and publicly available layers of evidence together.

29. CAPTURE METHODS

CM-1 — GUIDED MANUAL CAPTURE

The participant provides a screenshot and raw response text following instructions. It carries a low technical burden. It is prone to human error.

CM-2 — ASSISTED CAPTURE APPLICATION

Application:

- copying prompts,
- time logging,
- file upload,
- hash generation

functions are supported. It does not interfere with the AI product.

CM-3 — INSTRUMENTED USER-SURFACE CAPTURE

Browser or application integration:

- the sent prompt,
- the product metadata,
- the response text,
- the times

is captured directly. Provides high integrity. Terms of use and privacy should be evaluated.

CM-4 — PROVIDER EXPORT OR TELEMETRY CAPTURE

The provider's session export or verified telemetry is used. Can strongly support independent capture. May not provide independence on its own.

CM-5 — CONTROLLED API OR LAB CAPTURE

API or frozen system logs are used. It is not live user surface observation. Belongs to a separate laboratory panel.

30. LIMIT OF AUTOMATION

Automation:

- prompt submission,
- time logging,
- hash generation,
- file integrity

can be used for. Instead of the participant:

- bot account,
- automated bulk query,
- non-user session

generation does not create a real human Population Panel. Automated system auditing is separate:

CONTROLLED_AUTOMATION_PANEL

must be recorded as.

31. CAPTURE CONFIDENCE LEVELS

NCL-0 — NO VERIFIABLE CAPTURE

There is only the participant's statement. Cannot be included as the main GEO-1000 result.

NCL-1 — PARTIAL EVIDENCE

There is a response text or a limited screenshot. the prompt or system condition cannot be fully verified.

NCL-2 — SCREEN-EVIDENCED

The prompt, response, and main product surface have been verified with a screenshot. Metadata and time confidence may be limited.

NCL-3 — FULL EVIDENCE BUNDLE

Complete response, prompt, product record, times, screenshot, and hash are available. Minimum level for main analysis is a candidate.

NCL-4 — INSTRUMENTED AND SIGNED CAPTURE

The capture tool has recorded the prompt, response, time, and metadata directly; the package is digitally signed.

NCL-5 — INDEPENDENTLY VERIFIED MULTI-SOURCE CAPTURE

Capture package:

- independent validation,
- provider export,
- second source of evidence

has been supported by. Main GEO-1000 real user observations at least:

NCL-3

should target that level.

32. NOMOS CAPTURE EVIDENCE PACKAGE

For each observation, the package carries the following components as candidates:

32.1. Observation Manifest

Includes all identity and status fields of the observation.

32.2. Prompt Artefact

The full prompt text assigned and sent, the version, and its hash.

32.3. Raw Response Artefact

The full raw answer produced by the AI product.

32.4. Visual Evidence Artefact

Shows images or sequential images of the prompt, response, and product surface.

32.5. System-State Artefact

Shows the plan, surface, model label, memory, web, and other material system states.

32.6. Time Evidence Artefact

Carries the submission, completion, capture, and upload times.

32.7. Attempt Log

Shows the chain of technical errors and retries.

32.8. Integrity Manifest

Contains hashes and sizes of all files.

32.9. Privacy Manifest

Shows redaction, access and retention status.

32.10. Validation Record

Indicates who and when the capture validity was assessed.

33. MINIMUM CONTENT OF THE SCREENSHOT

To the extent relevant, visual evidence should show:

- AI product or surface identity
- Complete prompt
- Start of the response
- End of the response
- Reference fields
- Rejection or error message
- Apparent model label
- Evidence that the session is new or in a relevant state
- Material system setting
- Interface status indicating that the answer has not been cut off

34. LONG ANSWERS

If the answer does not fit on a single screen:

- full page capture,
- scrollable capture,
- overlapping consecutive screenshots,
- screen recording,
- direct text export

can be used. Consecutive visuals:

- should not leave gaps,
- should overlap with each other,

should carry sequence numbers. Only the positive or relevant section cannot be cropped.

35. SCREEN RECORDING

Screen recording:

- prompt submission,
- streaming response,
- completion status,
- no refresh performed

can be shown. However:

- more personal information,
- notification,
- private application,
- account ID

can be captured. Therefore, screen recording:

- may not be a default requirement,
- can be used in high-risk or verification sub-sample,

requires privacy preparation before recording.

36. RAW RESPONSE TEXT

In addition to the screenshot, the full answer text should be stored separately. Preference order:

- Official export or copy feature of the product
- Direct text capture by the capture tool
- Accessibility tree or DOM-based extraction
- Text copied by the user
- OCR and human verification

OCR should only be the last option. If there is a material difference between automatic extraction and the visual, a review is required.

37. DYNAMIC ANSWERS

Some products, after the answer is completed:

- can add a citation,
- can load a source card,
- can reformat the text,

can add a security warning. Two capture moments can be used:

- First completion capture
- Capture after brief render stability

If there is a material difference, both versions are retained. The first user exposure is not deleted.

38. HOW TO UNDERSTAND THAT THE ANSWER IS COMPLETED?

Candidate completion indicators:

- Streaming stops
- Loading indicator closes
- Regenerate or equivalent button appears
- Product completion signal
- Predefined silence period
- API completion record

Product-based completion rules must be found in the AI System Register.

39. IF THE USER STOPPED THE RESPONSE

The user may have accidentally or intentionally:

- stop,
- cancel,
- terminate the application

performed the action. This situation is recorded as:

USER_INTERRUPTED

ACCIDENTAL_INTERRUPTION

UNKNOWN_INTERRUPTION

A response stopped by the user cannot be considered a complete system output. The participant replacement or retry rule must be defined in advance.

40. THE PROMPT AND RESPONSE BEING IN SEPARATE FILES

If the prompt and response are in separate images:

- they must be linked to the same session,
- to the same observation ID,
- and to an uninterrupted time chain

Otherwise, it may not be possible to verify that the prompt and response truly belong to the same conversation.

41. FILE INTEGRITY

A cryptographic hash must be generated for each file. [K25] Candidate:

```
h_j = H(file_j)
```

Package hash:

```
h_bundle = H(JCS(manifest(file_id, order, size, algorithm, h_j))) [K14; K25]
```

can be created. Here, H is a published and considered secure hash algorithm. The algorithm:

- versioned,
- open,
- can be switched in the future if necessary.

must be.

42. WHAT DOES A HASH PROVE?

If the recorded hash value is stored reliably, re-hashing the same file with the same algorithm helps detect changes in the byte sequence [K25]. This comparison supports the claim that the file has preserved the same byte sequence after the reference hash was generated; it does not prove the source of the file, that it was produced at the correct time, or its authenticity before hashing. A hash alone does not prove:

- That the file came from the real AI product
- That it belongs to the correct user
- That it was produced at the correct time
- Not organised before hashing
- That it is the first output
- That the prompt was correct

Therefore, hash:

> *It is proof of integrity.*

Alone:

> *It is not proof of originality.*

43. DIGITAL SIGNATURE

Capture tool or authorised verifier:

- the observation package,
- time record,
- vehicle version

can sign digitally. Signature:

- which vehicle or institution produced the package,
- that it has not changed after the signature

strengthens. The signature holder may have signed incorrect or forged data. Therefore, the signature alone does not guarantee authenticity.

44. RELIABLE TIME [K26]

Local device clock:

- misconfigured,
- manually changed,
- out of synchronisation

may occur. A strong timestamp should carry multiple signals:

- Capture server acceptance time
- Reliable UTC time service
- Device time and deviation record
- Visible time in the AI product, if any
- File creation time
- Digital timestamp

45. TIME DEVIATION

The difference between the participant's device time and the reliable UTC time:

```
O_i^clock = t_i^device – t_i^reference
```

can be recorded as. If there is a material deviation:

- the server time is used as the main reference,
- the device time is preserved as a corrected record,

uncertainty is marked. The participant should not be asked to manually change the device clock.

46. UPLOAD TIME IS NOT THE CAPTURE TIME

Uploading a file to the server at 14:00 does not prove that the screenshot was taken at 14:00. Therefore:

- capture time,
- hash generation time,
- upload time

should be recorded separately. Long upload delays may reduce capture trust. The record does not have to be automatically invalid.

47. DELAYED UPLOAD

Due to connection issues, the user may not be able to upload the evidence immediately. If late upload is accepted:

- local hash at capture time,
- trusted device time,
- offline signed record,
- reason for upload delay

may be required. Only the user's statement: “I took this image yesterday.” is not sufficient for the main NCL-3 level.

48. EVIDENCE CHAIN

The evidence chain of an observation as a candidate consists of the following stages:

- Task assignment
- Prompt lock
- User and product suitability
- Slot opening
- Prompt submission
- Initial contact result
- Completion of response
- Capture
- Local integrity record
- Secure upload
- Server acceptance time
- Package hash
- Protocol verification
- Editing
- Adjudication
- Archive
- Public manifest

Each stage must be connected to the previous stage.

49. BREAKING THE CHAIN

The chain of evidence can be broken in the following situations:

- No prompt record
- The initial output cannot be verified
- The response does not match the visual
- Capture time is unknown

- File has been modified later
- Duplicate participant ID
- Product and plan are unknown
- Observation is out of wave
- Hash does not match the file
- Editing overwrote the raw file
- It is unknown who verified it

The type of break determines the validity status.

50. RAW EVIDENCE CANNOT BE CHANGED; THE DECISION CAN CHANGE

Raw evidence package:

- should not be written over,
- should not be quietly corrected,

should not be deleted and replaced with a new one. However, regarding the observation:

- validity,
- adjudication,
- classification,
- compliance

the decision can change. The new decision:

- new version,
- justification,
- date,
- decision-maker

should carry. This distinction is a fundamental governance rule:

> *Evidence can remain fixed; interpretation can evolve with the evidence.*

51. REDACTION

The following fields can be redacted in the public or adjudicator copy:

- Name and surname
- Email
- Profile picture
- Account ID
- Private conversations
- Notifications
- Corporate field
- Sensitive file name
- Exact location

Redaction:

- should not be done on the raw file,
- should produce a new derivative file,
- should carry its own hash,

should record which field was redacted and why.

52. THREE LAYERS OF EVIDENCE

52.1. Raw Restricted Evidence

It is full raw evidence. Only authorised auditors and required research roles can access it.

52.2. Audit Evidence Layer

Contains the evidence necessary for independent adjudicators and reproduction teams. Personal areas are minimised.

52.3. Public Evidence Manifest

For the public:

- observer ID,
- country and language cell,
- AI product,
- time,
- evidence hash,
- validity status,
- redaction status

are presented. Raw private content is not necessarily published.

53. PUBLIC EVIDENCE MANIFEST

For each wave, the public manifest may include the following fields:

- Wave ID
- Protocol version
- Number of products
- Target and valid observation count
- Country and language coverage
- Slot coverage
- Capture confidence levels
- Technical error and rejection rates
- Invalid observation count
- Observation package hashes
- Wave integrity root
- Change events
- Publicly redacted examples
- Independent access procedure

- Evidence retention status

54. WAVE INTEGRITY ROOT

CANDIDATE CONCEPT

Let the hash of each observation package be hi. From these, the wave root with a Merkle tree or equivalent structure:

RW

can be produced. The wave root can be published after the measurement is closed. This method:

- the quiet later modification of the observation list in the wave,
- adding a new observation,
- modifying an existing observation

makes it more visible. Merkle root:

- does not alone prove that the observations are real,
- or that the adjudication decision is correct

It strengthens the integrity of the list.

55. ADDING AND REMOVING OBSERVATIONS

After the wave is closed:

- adding a new observation,
- removing an existing observation,
- modifying the file

requires a new manifest version. The old integrity root is preserved. For the observation being removed:

- reason,
- date,
- decision maker
- effect on old and new score

is recorded.

56. PARTICIPANT WITHDRAWAL REQUEST

Research ethics or the data protection framework may allow the participant to withdraw their data. In this case:

- raw evidence can be safely deleted or made inaccessible,
- an empty WITHDRAWN record can be preserved in the public manifest,
- wave score can be recalculated with the new version,

the old result and reason for change are preserved. Immutability does not mean eliminating human rights.

57. OBSERVATION CAPTURE VALIDITY STATUSES

CV-0 — NOT REVIEWED

The capture has not been evaluated yet.

CV-1 — VALID

Meets all mandatory capture and time conditions.

CV-2 — CONDITIONALLY VALID

There is a limited deficiency. It is specified in which analyses it can be used.

CV-3 — PARTIAL EVIDENCE

The observation can be used as an event. It may not enter the main population score.

CV-4 — OUTSIDE WAVE

The prompt was sent outside the allowed time structure.

CV-5 — PROMPT OR OUTPUT MISMATCH

There is a material discrepancy between the assigned prompt, the submitted prompt, or the recorded response.

CV-6 — DUPLICATE

The same observation or user record has been repeated.

CV-7 — TAMPER SUSPECTED

There is an unresolved doubt about the integrity of the file or metadata.

CV-8 — FABRICATED OR MANIPULATED

The observation is not a real user experience or has been deliberately altered.

CV-9 — WITHDRAWN

It has been excluded from analysis due to participant or ethical processes.

58. CAPTURE VALIDITY SHOULD BE SEPARATED FROM SEMANTIC ACCURACY

An observation:

- completely valid in terms of capture,
- severely incorrect in terms of content

may be. Another observation:

- correct answer,
- may carry missing or manipulated evidence

Therefore, there are two separate decisions:

- Capture Validity

- Semantic Adjudication

The capture validator should know the success score of the response as little as possible.

59. ROLE OF THE CAPTURE VALIDATOR

The capture validator examines:

- Prompt matching
- Wave and slot
- Product identity
- First output rule
- Screen integrity
- Raw text consistency
- Technical retry
- Hash and time
- Privacy

Does not decide on: Was the company described correctly? Does the citation support the claim? Is there a critical error? Does the response pass? These are the responsibilities of the next review layer.

60. REPEATED SCREEN IMAGE

The same visual:

- accidentally uploaded twice,
- used by multiple users,
- repeated among panel providers

may occur. Checks:

- File hash
- Visual similarity
- Raw response text
- Time and session metadata
- User uniqueness

can be done. The same AI product may give the exact same answer word for word to some users. The same text alone is not evidence of cheating. The same file or the same visual trace is a stronger signal.

61. FILE SIMILARITY AND FALSE POSITIVE

Same device and interface:

- similar visuals,
- same screen size,
- same compression

can produce. The visual similarity algorithm should not perform automatic exclusion. Multiple signals and human review are required.

62. PROVIDER TELEMETRY

AI provider:

- session time,
- model,
- prompt,
- response,
- tool usage

can provide reliable logs about. This record strengthens independent capture. However, the provider:

- measured party,
- system owner,
- beneficiary

may be. Provider telemetry: should be used together with, not as a replacement for, independent user evidence.

63. IF THERE IS NO PROVIDER EVIDENCE

Access to provider logs may not be available in most live consumer products. This does not make GEO-1000 impossible. The following composition can be used:

- Instrumented user capture
- Server timestamp
- Full screen proof
- Raw response
- System setting
- Hash and signature
- Independent verification sub-sample

Proof trust level is indicated accordingly.

64. CAPTURE IN NATURAL USER PANEL

Capture tool in the natural panel:

- special instruction content,
- previous conversations,
- account identity

should not be captured unnecessarily. Preferred flow: The user opens a new natural account chat. Capture only records the task window. Personalisation status is taken separately as high-level metadata. Private history stays outside capture. Public copy is edited.

65. NATURAL CONVERSATION-STATE CAPTURE

If the current conversation context is measured:

- only the necessary previous turn range,

- explicit consent,
- strong redaction,
- access restriction

is required. All private chat history cannot be collected by default.

66. MOBILE CAPTURE

On mobile devices:

- full screen capture,
- scrollable capture,
- application version,
- system time log

may work differently. Mobile protocol:

- iOS,
- Android,
- other systems

may carry separate technical guides. Mobile users cannot be excluded as desktop capture is difficult. The capture tool must adapt to the population.

67. ACCESSIBILITY AND CAPTURE

For participants using screen readers, voice input, or motor assistive tools, standard visual capture alone may not be sufficient. Alternative evidence:

- Accessibility tree export
- Voice session recording
- Application transcript
- Assistive technology log
- Task observation record

may occur. These observations should not be considered low reliability. An equivalent evidence standard should be developed.

68. OBSERVATION OF THE CAPTURE TOOL CHANGING BEHAVIOUR

Browser extension or instrumentation:

- page load,
- interface behaviour,
- security system,
- product detection

may be affected. The effect of the capture tool should be tested in the pilot. If sessions with and without the tool behave differently: the tool condition becomes an experimental factor, the capture tool cannot be assumed silently neutral.

69. CAPTURE TOOL VERSIONING

Each capture package:

- tool name,
- version,
- configuration,
- operating environment,
- known bugs

must carry. If the capture tool changes during measurement:

- new tool version,
- subwave,
- compatibility review

may be required.

70. CAPTURE TOOL ERROR

Tool:

- may incorrectly copy the prompt,
- may truncate the end of the response,
- may record the time incorrectly,
- may not generate a hash,

may compress the file. Tool error:

- cannot be loaded to the AI product,
- or the user

Separately:

CAPTURE_TOOL_FAILURE

must be recorded as.

71. SYNTHETIC THREE-WAVE GEO-1000 CASE

SYNTHETIC METHODOLOGY DEMONSTRATION / The following products, dates, numbers, and results are entirely fictional. Ten synthetic AI products:

A1–A10

shall be measured in three main waves.

71.1. Wave Calendar

Wave	Type	UTC delivery window	Purpose
W1	Principal	12.00-13.00	Start
W2	Confirmatory	20.00-21.00	Short-term repeat
W3	Stability	04.00-05.00	Medium-term and hourly balance

For each product in each wave: 1,000 valid main observations are targeted. Total target:

```
10×1,000×3=30,000
```

71.2. Slot Structure

Each wave:

- 12 micro-slots
- with participants balanced by product and country per slot

The participant's slot: randomly assigned before data collection.

71.3. W1 Result

Goal: 10,000 valid main observations Field:

- 13,240 invitations
- 10,870 completed tasks
- 10,106 valid observations in terms of capture
- 10,000 main analysis observations according to previously locked main/backup order

Remaining 106 valid records:

- backup,
- sensitivity,
- capture quality control

are reserved for.

71.4. Technical Events

118 usage limit 42 connection error 31 system rejection 17 user disconnection 9 prompt mismatch 6 retry screenshot All initial contact results have been saved. Only predefined real technical errors were retried.

71.5. Product Change During W2

A7 product at 20.32 UTC:

- automatic web usage,
- new reference view

let it win. Observations of A7:

W2A: 20.00–20.31

W2B: 20.32–21.00

are separated as. A single fixed W2 score for A7 is not published without explanation.

71.6. External Event in W3

An important legal decision should be published on 04.22 UTC about the audited synthetic company. It should be observed that responses in web-accessible products changed after 04.25. Event: It is added to the External Event Register. Results before and after the event are shown as separate diagnostic cells. Observations are not deleted because they lower or increase the score.

71.7. Wave Health Card

Wave	Health	Main note

Wave	Health	Main note
W1	WH-4	Balanced and strong capture
W2	WH-3	Sub-wave due to A7 product change
W3	WH-3	External legal event record

All three waves can be used. A7 and external event comments are also restricted.

72. SYNTHETIC SINGLE OBSERVATION CAPTURE CASE

SYNTHETIC EXAMPLE

Participant:

- Turkey
- Turkish
- Orion AI web paid plan
- Controlled Clean Panel

W1

Slot 12.15–12.20 UTC Assigned prompt: “Which parent company is Apple.com associated with, and what are the main activities of this company?”

72.1. Times

Task opening: 12.14.20 Prompt submission: 12.16.08 First token: 12.16.10 Answer completion: 12.16.24 Capture: 12.16.39 Hash generation: 12.16.41 Upload: 12.17.12 Server acceptance: 12.17.13

72.2. Evidence

Prompt hash matches Full response text available Two overlapping screenshots available Product, plan, and model label visible Memory and special instruction off Web usage on Citation not shown No reproduction No previous conversation File hashes and package signature available Capture level:

NCL-4

Validity:

CV-1

It has not yet been decided whether the response is correct or incorrect.

73. SYNTHETIC TECHNICAL RETRY CASE

On the first attempt:

- Prompt sent correctly
- Product showed "Network error"
- Substantive response not started
- Error screenshot taken
- Time and hash have been recorded

According to the predefined rule, a second attempt was made within five minutes. A response occurred on the second attempt. Correct record:

- First Contact Outcome: TEMPORARY_TECHNICAL_FAILURE
- First Eligible Content Output: second attempt
- Both attempts are in the package
- First incident is included in the technical error rate
- The second response is included in the content evaluation, if the method predicts this in advance

Incorrect record: The first error is deleted and it is treated as if only the second response existed.

74. SYNTHETIC INCORRECT RETRY CASE

First response: “Apple is a local company that only sells mobile phones.” The participant does not like the response. They reproduce it. The second response turns out to be correct. The participant only submits the second response. The system log or capture record shows the redo. Correct status:

CV-8 — MANIPULATED OBSERVATION

The observation is not saved just because the second response is correct. If the first response can additionally be recovered, it can be preserved in the incident investigation. The main participant record is invalid.

75. WAVE CLOSURE REPORT

At the end of each wave, the following record should be created:

- Wave ID
- Opening and closing time
- Target number of observations
- Invited person
- Completed task
- Valid capture
- Observation entering main analysis
- Backup valid observation
- Technical error
- Rejection
- Usage limit
- Prompt mismatch
- Capture mismatch
- Late upload
- Duplicate
- Suspected tampering
- Participant withdrawal
- Product changes
- External events
- Slot coverage

- Country and language coverage
- Capture confidence levels
- Wave health status
- Integrity root
- Responsible human approval

76. MANDATORY NORMATIVE PROVISIONS

CH12-N01

Each GEO-1000 observation must be linked to a pre-locked measurement wave.

CH12-N02

The term "same time" cannot be used without the full UTC window, micro-slot, and completion rules.

CH12-N03

Local time cannot be used alone as wave time; the main time standard must be UTC.

CH12-N04

The transmission window, micro-slots, completion time, and loading time must be determined before AI responses are viewed.

CH12-N05

Slot assignment must be made in a way that does not leave countries, languages, products, and panel groups time-wise unbalanced.

CH12-N06

Participants cannot be moved to earlier or later slots based on their results.

CH12-N07

The local time offset of the single UTC window must be documented in the public method record.

CH12-N08

Returning global waves or locally time-balanced operations must be kept as separate time results.

CH12-N09

Each observation prompt must carry submission, first contact, completion, capture, and upload times to the extent relevant.

CH12-N10

Unknown time fields cannot be presented as estimated actual times.

CH12-N11

The device clock alone cannot be considered a reliable UTC time proof.

CH12-N12

Capture time and upload time must be recorded separately.

CH12-N13

Observations sent outside the wave cannot be included in the main wave without explanation.

CH12-N14

Exclusion rules for delay or slot deviation must be independent of the response content.

CH12-N15

As a result of the first contact, it must be kept separately from the first content output.

CH12-N16

Rejection, usage limit, technical error, clarification, and inconclusiveness cannot be placed in the same category.

CH12-N17

Wrong, negative, reference-less, or Critical AI response cannot be a reason for retry.

CH12-N18

Technical retry can only be performed under predefined technical conditions.

CH12-N19

When a technical retry is performed, all attempts and the initial contact result must be preserved.

CH12-N20

Partial or incomplete responses cannot be ignored and replaced with a new response.

CH12-N21

The first suitable output cannot be renewed, selected, or manually improved.

CH12-N22

Each observation must aim at least for NCL-3 or a justified equivalent capture confidence level.

CH12-N23

A screenshot alone cannot be considered a complete evidence package.

CH12-N24

The visual evidence prompt should show the entire response and the surface of the physical product.

CH12-N25

Long responses must be preserved fully and sequentially using the capture method.

CH12-N26

If there is a material difference between the raw response text and the visual evidence, an examination should be conducted.

CH12-N27

OCR can be used as an auxiliary method if direct text recording is not possible; without verification, it cannot be considered the final source of the text.

CH12-N28

The initial completion and subsequent material versions of dynamically changing responses must be preserved separately.

CH12-N29

The participant stopping the response or closing the interface must be recorded as a separate event.

CH12-N30

If the prompt and response are in separate files, they must be linked to the same session and time chain.

CH12-N31

Each evidence file and evidence package must carry an integrity hash.

CH12-N32

The hash alone cannot be presented as proof of the observation's reality.

CH12-N33

Digital signature and timestamp are not the entirety of observation authenticity; they are components of the evidence chain.

CH12-N34

No redaction or correction may be written over the raw evidence file.

CH12-N35

Each edited derivative must carry its own file ID and hash.

CH12-N36

Access to raw evidence may be restricted; the public manifest must demonstrate observation integrity without revealing identity.

CH12-N37

Participant’s name, email, password, private conversation, or unnecessary personal information cannot be added to the public evidence.

CH12-N38

The risk of re-identification for small countries and rare subgroup users should also be assessed.

CH12-N39

The Natural User Panel capture should not collect special instructions and past conversation content by default.

CH12-N40

Equivalent alternative capture methods should be provided for accessibility users.

CH12-N41

If the capture tool materially changes the user experience, this effect should be measured separately.

CH12-N42

The capture tool and its version should be recorded in every observation.

CH12-N43

If the capture tool materially changes during a wave, a distinction should be made for sub-wave or tool version.

CH12-N44

The capture tool error cannot be classified as an AI product or participant failure.

CH12-N45

Provider telemetry can support independent user capture; it cannot replace it without explanation.

CH12-N46

If a material AI product change occurs during the wave, observations cannot be merged as a single fixed product state.

CH12-N47

Product change requires wave separation, restart, explicit modelling, or mixed-state warning.

CH12-N48

The absence of a provider update announcement is not definitive proof that the system has not changed.

CH12-N49

Material external events should be added to the External Event Register with a timestamp and scope.

CH12-N50

External events cannot be a reason for exclusion after results because they raised or lowered the score.

CH12-N51

Capture validity should be evaluated as blind as possible, before assessing the semantic correctness of the response.

CH12-N52

A correct AI answer cannot validate a missing or manipulated capture.

CH12-N53

An incorrect AI answer cannot invalidate a protocol-compliant capture.

CH12-N54

Duplicate and tamper decisions should involve multiple signals and human review.

CH12-N55

The same response text alone cannot be considered proof of duplication or fraud.

CH12-N56

Adding, removing, or changing observations after a wave is closed requires a new manifest and analysis version.

CH12-N57

The old wave manifest and the integrity root must be preserved.

CH12-N58

Raw evidence can be preserved unchanged; validity and adjudication decisions can only be changed in versioned form.

CH12-N59

The participant's right to withdraw and research ethics must be maintained above the claim of immutability.

CH12-N60

Wave integrity should be evaluated independently of the AI score.

CH12-N61

The wave closure report should show the target, actual, invalidations, technical events, product changes, and evidence integrity.

CH12-N62

NOMOS Capture compliance cannot be tied solely to the use of software developed by NobleJackal; equivalent open applications should be possible.

CH12-N63

The capture schema, validation decisions, and wave records should be versioned in a machine-readable format.

CH12-N64

Each wave and evidence chain should have an accountable human or institutional owner.

77. FORMS OF FAILURE

CH12-F01 — THE SAME SECOND FALLACY

Forcing all participants to send at the same second is considered the only condition for simultaneity.

CH12-F02 — COUNTING "TODAY" AS A SYNCHRONISED WAVE

Local time and time zones get confused.

CH12-F03 — USING LOCAL TIME LIKE UTC

Observations enter the wrong chronological order.

CH12-F04 — CHANGING SLOT ASSIGNMENT BASED ON RESULTS

Low-scoring users are moved to another time cell.

CH12-F05 — PACKING ONE COUNTRY INTO ONE SLOT

Time effect gets mixed with country effect.

CH12-F06 — GENERATING ARTIFICIAL LOAD

Thousands of users send simultaneously, triggering the product's abnormal error behaviour.

CH12-F07 — HIDING NIGHT PARTICIPATION

Fatigue and participation loss are not reported in certain areas.

CH12-F08 — ADDING OFF-WAVE OBSERVATION TO THE WAVE

A response received hours or days later is used as if it belonged to the main window.

CH12-F09 — COUNTING LOAD TIME AS CAPTURE TIME

It cannot be proven when the file was generated.

CH12-F10 — BLIND TRUST IN DEVICE CLOCK

Manually changed or incorrect time becomes the main time record.

CH12-F11 — EXCLUDING DELAYED NEGATIVE RESPONSE

The time rule is applied dependent on the outcome.

CH12-F12 — DELETING THE FIRST CONTACT RESULT

The speed limit, error, or rejection is made invisible with the second attempt.

CH12-F13 — COUNTING REJECTION AS TECHNICAL ERROR

The actual system behaviour is retried.

CH12-F14 — COUNTING WRONG RESPONSE AS TECHNICAL ERROR

Regeneration is done until the desired result occurs.

CH12-F15 — IGNORING PARTIAL RESPONSE

Even if the system is interrupted, only the subsequent complete response is saved.

CH12-F16 — SELECTING THE BEST RESPONSE

Among multiple answers, the most positive or longest one is reported.

CH12-F17 — SCREENSHOT ONLY

The visual is considered sufficient without prompt, product, time, and metadata.

CH12-F18 — CROPPED SUCCESS VISUAL

Incorrect or restrictive parts are left invisible.

CH12-F19 — CAPTURE NOT SHOWING THE PROMPT

It cannot be proven which question the answer belongs to.

CH12-F20 — TRUNCATING THE END OF THE ANSWER

Material limitation or rejection part becomes invisible.

CH12-F21 — GAP IN CONSECUTIVE VISUALS

Some parts of the answer are lost.

CH12-F22 — COUNT OCR TEXT AS RAW ANSWER

Recognition errors change material claims.

CH12-F23 — CONSIDER ONLY THE LATEST VERSION OF DYNAMIC RESPONSE AS REAL

The initial user exposure is deleted.

CH12-F24 — CONSIDER USER INTERRUPTION AS COMPLETE ANSWER

The system output is incomplete due to user intervention.

CH12-F25 — INCORRECTLY MATCH SEPARATE PROMPT AND ANSWER

Files from different chats are linked to the same observation.

CH12-F26 — HASH WORSHIP

Any file with a hash is considered real and original.

CH12-F27 — IGNORE PRE-HASH EDITING

Truncated or modified files are legitimised by hashing.

CH12-F28 — CONSIDERING THE SIGNATURE AS VALID

Signed but incorrect record is accepted without dispute.

CH12-F29 — EDITING ON THE RAW FILE

Original evidence is irreversibly altered.

CH12-F30 — CONSIDERING THE EDITED DERIVATIVE AS RAW EVIDENCE

Public copy is used like the original file.

CH12-F31 — REMOVING PRIVACY FOR EVIDENCE

Participant's name, account, and private chat are published.

CH12-F32 — REVEALING A PARTICIPANT FROM A SMALL COUNTRY

Time, device, and plan information make a person re-identifiable.

CH12-F33 — COLLECTING SPECIAL HISTORY IN THE NATURAL PANEL

Irrelevant conversations enter the capture.

CH12-F34 — INVALIDATING ACCESSIBILITY EVIDENCE

The real user is excluded because a standard visual cannot be generated.

CH12-F35 — ASSUMING THE CAPTURE TOOL IS NEUTRAL

The tool changes the product behaviour but is not tested.

CH12-F36 — HIDING THE CAPTURE VERSION

Different tool versions are presented as a single measurement method.

CH12-F37 — CONSIDERING THE CAPTURE TOOL ERROR AS AN AI ERROR

The cut response is uploaded to the product.

CH12-F38 — COUNTING THE API LOG ON THE USER SCREEN

The real user surface becomes invisible.

CH12-F39 — COUNTING THE PROVIDER LOG AS INDEPENDENT EVIDENCE

The record of the measured party becomes a single piece of evidence.

CH12-F40 — HIDING THE PRODUCT UPDATE

Two system states within the branch melt into a single score.

CH12-F41 — IGNORING CHANGES IF NO ANNOUNCEMENT

The observed system break is not examined.

CH12-F42 — EXCLUDING EXTERNAL EVENTS BASED ON OUTCOME

Responses with negative news are removed; responses with positive news are kept.

CH12-F43 — LINKING WAVE HEALTH TO SCORE

Defective waves with high scores are valid, strong waves with low scores are considered suspicious.

CH12-F44 — COUNTING THE CORRECT ANSWER AS A VALID CAPTURE

The evidence defect is covered up by semantic success.

CH12-F45 — COUNTING THE WRONG ANSWER AS AN INVALID CAPTURE

The measurement result is converted into data error.

CH12-F46 — COUNTING THE SAME TEXT AS BOT EVIDENCE

Deterministic or similar AI responses accidentally become duplicates.

CH12-F47 — SINGLE SIGNAL TAMPER DECISION

File size or time difference automatically creates an exclusion.

CH12-F48 — ADD OBSERVATION AFTER WAVE

New records are quietly added to reach the desired sample size.

CH12-F49 — REMOVE LOW-SCORED OBSERVATION AFTER WAVE

Manifest and outcome are retroactively changed.

CH12-F50 — DELETE OLD MANIFEST

Version history is destroyed.

CH12-F51 — CORRECT RAW EVIDENCE

File content is modified to correct misclassification.

CH12-F52 — DENY WITHDRAWAL RIGHT

Immutability is held above human rights.

CH12-F53 — CLOSED PROPRIETARY CAPTURE MONOPOLY

Only evidence produced with the founder's tool is considered valid.

CH12-F54 — NOT PUBLISHING THE CAPTURE SCHEMA

An independent team cannot establish the same chain of evidence.

CH12-F55 — CONSIDERING THE WAVE ROOT AS EVIDENCE OF REALITY

The Merkle or hash root is presented as if it proves content accuracy.

CH12-F56 — CLAIM OF "1,000 SCREENSHOTS" WITHOUT AN EVIDENCE PACKAGE

The number of files is used like the number of valid observations.

78. AUDIT PROCEDURE

Step 1 — Determine the Research Purpose of the Wave

Principal, confirmatory, incident, drift, or retest type is selected.

Step 2 — Lock AI Product and System States

It connects to the wave version of the AI System Register in Section 9.

Step 3 — Lock Population, Sample, and Panel States

Records from Sections 5–8 and Section 11 are used.

Step 4 — Lock Prompt Register

Prompt versions in Section 10 are linked to the wave.

Step 5 — Set UTC Wave Window

Start, closing, and additional durations are written.

Step 6 — Create Micro-Slots

Participants are assigned to slots in a balanced way in terms of product, country, and language.

Step 7 — Define Time Source

The server, device, and trusted UTC records are determined.

Step 8 — Lock First Contact and Retry Rules

Technical and semantic states are separated.

Step 9 — Determine Capture Methods

Web, mobile, accessibility, and natural panel paths are defined.

Step 10 — Determine NCL Target

Record the capture-confidence level required for the principal analysis.

Step 11 — Lock Evidence Package Schema

Manifest, files, hash, and metadata fields are defined.

Step 12 — Lock Privacy and Redaction Rules

Raw, audit, and public layers are separated.

Step 13 — Capture Your Vehicle with the Pilot

Does the vehicle change product behaviour? Are the time and text records correct?

Step 14 — Open the External Event Register

A process is established to monitor product and entity events.

Step 15 — Confirm Wave Lock

WS-1 — LOCKED status is given.

Step 16 — Open the Slots

Participants are only allowed to submit in their own slots.

Step 17 — Record the First Contact Result

Answer, rejection, error, or limit is separated.

Step 18 — Capture the First Suitable Output

The full answer is preserved without renewal.

Step 19 — Create Raw Evidence and Metadata

Screenshots, raw text, system status, and timestamps are recorded.

Step 20 — Perform Local Hash and Secure Upload

Files receive an integrity record before uploading.

Step 21 — Server Acceptance and Create Package Hash

The evidence chain is closed with server time and package ID.

Step 22 — Blind Review of Capture Validity

The CV status is assigned without knowing semantic success.

Step 23 — Perform Duplicate and Tamper Review

Multiple signals and human review are used.

Step 24 — Examine the System Within the Wave and External Events

Determine whether a sub-wave or mixed-state is required.

Step 25 — Assign Wave Health

Status is assigned between WH-0 and WH-5.

Step 26 — Generate Wave Manifest and Integrity Root

Observation list is versioned.

Step 27 — Publish Public Proof Manifest

Scope and integrity are made visible without personal data.

Step 28 — Approve Wave Closure Report

Targets, occurrences, defects, events, and changes are closed with human approval.

79. NECESSARY EVIDENCE

Wave ID Wave type Protocol version AI System Register version Population frame Sample lock Panel status lock Prompt Registry version Reference package version UTC start and end time Micro-slot list Participant–slot assignment Time source Device clock deviations Transmission times First contact times Completion times Capture times Hash times Upload times Server acceptance times First contact results First suitable content outputs Technical retry records Partial response records Rejection and usage limits Full raw responses Full screenshots Sequential visual manifests Screen recordings, if any Citation and source records System settings evidence Capture tool and version Capture method NCL level File hashes Package hash

Digital signature Timestamp Duplicate checks Tamper checks Capture validity decisions Redaction files Raw-to-redacted links Privacy manifest Withdrawal records Product change events External Event Register Wave sub-divisions Slot scope Country and language scope Technical error rates Capture validity rate Wave health status Wave integrity root Public evidence manifest Wave closure report Change log Responsible person or institution

80. AUDIT CHECKLIST

Is the type and purpose of the wave clear? Have AI products and system statuses been locked? Are the population, sample, and panel versions determined? Was the prompt set locked before the results? Is the UTC wave window open? Were micro-slots pre-assigned? Were countries and languages distributed evenly across slots? Did participants switch slots based on their results? Was local time load evaluated? Are there rotating wave hours? Are submission and completion extra times clear? Was the device clock compared with a reliable UTC source? Were submission, completion, capture, and upload times separated? Did out-of-wave observations enter the main score? Was the first contact result recorded? Was technical error distinguished from rejection?

Was a retry made due to an incorrect answer? Was the first attempt preserved in the technical retry? Were partial answers deleted? Is the first suitable output really the first output? Was Regenerate or a follow-up message used? Does it show the visual prompt and the entire answer? In long answers, is there capture without spaces? Does the raw text match the visual? If OCR was used, is there human verification? Were dynamic answer versions retained? Did the user stop the answer? Are the prompt and answer linked to the same session? Is the capture tool and version recorded? Did the capture tool alter user behaviour? Was the NCL level given? Are there observations below NCL-3 in the main analysis? Does each file have a hash? Is it clear what the hash does not prove?

Is there a digital signature or server time? Is the capture time separate from the upload time? Is there a reason for late upload and a local integrity log? Was redaction performed on the raw evidence? Do the edited derivatives have separate hashes? Does the public manifest reveal the participant's identity? Can a user from a small country be re-identified? Were private conversations unnecessarily captured in the natural panel? Was an alternative evidence path provided for accessibility users? Was the duplicate decision based only on the same text? Does the tamper decision carry multiple signals?

Did the capture validator know the semantic score? Did the correct answer cover up the evidence flaw? Was the incorrect answer invalidated? Did the product change during the wave?

Was the product change processed as an underwave or mixed-state? Were external events recorded? Did external events cause exclusion according to the result? Is the wave health status independent of the score? After the wave was closed, were observations added or removed? Are the old manifest and integrity root preserved? How were participant withdrawal requests handled? Is there a public proof manifest? Is there a clearly accountable owner of the wave closure report?

81. OBJECTIONS AND RESPONSES

Objection 1 — “If all users do not submit at the same second, true simultaneity is not achieved.”

One-second submission can create artificial load, connection, and queue effects. A predefined short UTC window and random micro-slots can measure the same system period more reliably.

Objection 2 — "The product can change within a one-hour window."

It can change. Therefore, system records, configuration fingerprint, and product change tracking are required. In case of a material change, the wave is separated.

Objection 3 — "Why isn’t just a screenshot enough?"

Screenshot:

- time,
- initial output,
- prompt version,
- system setting,
- copying

do not always show these subjects. A complete evidence package is required.

Objection 4 — "Isn’t it too burdensome to verify a thousand screenshots one by one?"

It is heavy. Automatic integrity checks, sample-based human review, and risk-based verification can be used. Critical or suspicious records are examined in more detail.

Objection 5 — “If there is a hash, why can't we say that the file is real?”

A hash only shows that the file remained the same after the hash was generated. It does not show how the file was created before the hash.

Objection 6 — “If the user has to trim personal information, how will the raw file be protected?”

If possible, the Capture tool should not capture personal information from the start. If unavoidable, the raw file is kept in a restricted and encrypted repository, and a public copy is kept as a separately redacted derivative.

Objection 7 — “If the response is wrong, why don't we try again?”

Because a wrong answer is a matter of measurement. Retrying is only for an actual technical failure.

Objection 8 — “If the system rejects, can the user actually try again?”

Correct. The main first-contact panel preserves the rejection. A separate Natural Interaction Track can measure the user's follow-up behaviour.

Objection 9 — “If a technical error is a user experience issue, why do we allow retry?”

The technical error is preserved as a result of the first contact. A predefined technical retry can be done to additionally examine the representative content. The two results do not replace each other.

Objection 10 — “Isn't the provider's own log the strongest evidence?”

Technically, it can be very strong. The provider is the owner of the measured system. It is more reliable when used together with an independent user capture.

Objection 11 — “Will NOMOS Capture be a mandatory NobleJackal software?”

No. NOMOS Capture is an open evidence protocol. Independent tools that meet the equivalent schema and integrity conditions can be compatible.

Objection 12 — “If we do not make all evidence public, how will independent verification be done?”

The public manifest shows integrity and scope. Authorised independent auditors can access redacted or limited raw evidence. Privacy and verification can be designed together.

Objection 13 — “Wouldn't the whole work be disrupted if important news breaks during the wave?”

Not always. The report is part of the real system environment. It can be separated with a timestamp before and after the event. The scope of the score is explained accordingly.

Objection 14 — “Isn't it easier to discard old answers and continue with the new system when a product is updated?”

Old answers are real on the measurement date. They cannot be deleted. Old and new system statuses must be maintained as separate waves or sub-waves.

Objection 15 — “If the capture tool affects product behaviour, does the whole method fail?”

No. The vehicle effect is measured with the pilot. If necessary, a lighter vehicle, manual capture, or a method that uses the vehicle effect as a separate factor is chosen.

Objection 16 — "If the participant withdraws their data, wouldn't the integrity root be compromised?"

A new manifest version and a new root are created. The old record can retain a contentless WITHDRAWN trace. It is maintained on top of the human rights evidence architecture.

COMMON JUDGEMENT OF SECTION 85

At the end of a GEO-1000 study, you may have 30,000 screenshots. This number alone is not evidence. You need to know: Which wave were they produced in? To which system state did they belong? Was the same prompt actually sent? Was it the first output? Was the response refreshed? Did

the product change during the wave? When was the file captured? When was the hash generated? Was the participant ID repeated? Does the visual show the entire answer? Does the raw text match the visual? If the evidence was later edited, where is the original? Was capture verification done before the semantic score? You can click a thousand times within the same second. Still:

- device clocks,
- connections,
- queues,
- product states

It can be different. Therefore, concurrency is not a marketing show. It is a predefined time architecture. When a user gets a wrong answer: saying 'Try again.' is tempting. But that wrong answer is the reality you are trying to measure. If you delete it, you do not fix the system. You corrupt the evidence. When a user experiences a real connection error, a technical retry can be done. But the result of the first contact should still be preserved. Because the user first encountered the error. The hash is valuable. The digital signature is valuable. The timestamp is valuable. The screenshot is valuable. None of these alone carries the whole truth. Trust:

- prompt,
- time,
- product,
- user,
- file,
- initiative,
- verification

It arises from the joint preservation of its chain. A standard: if it says, “Only proof produced with my software is valid,” it places its own interest before independent replication. For this reason, NOMOS Capture is not a tool brand, but an open proof architecture. Independent universities, researchers, or audit institutions should be able to apply the same rules. Therefore, NOMOS's twelfth measurement law is as follows:

> *Concurrency is observable reproducibility within the same defined system time, not the same second.*

The thirteenth law is as follows:

> *The first output is not a draft that can be renewed until one likes the result.*

The fourteenth law is as follows:

> *Hash preserves proof; it does not create proof.*

The fifteenth law is as follows:

> *Capture validity and response accuracy are independent of each other.*

Its sixteenth law is as follows:

> *The chain of evidence should make the integrity of the observation visible, not the identity of the person.*

Its seventeenth law is:

> *If the system changes during the wave, it cannot act as if the measurement history has not changed.*

Its eighteenth measurement law states:

> *Nothing is written on raw evidence; decisions are corrected through versioning.*

Its nineteenth measurement law states:

> *A capture method that cannot be applied independently cannot be a world standard.*

Order of Section 12 of NOMOS

> *Don’t tell me "we asked at the same time." / Show in which UTC window, in which slot, and in which system state you asked.*
>
> *Do not pile a thousand people into the same second to produce artificial error. / Spread the time but keep the system within the same wave.*
>
> *Do not mix local times with each other. / Do not make the device time absolute truth.*
>
> *Do not save the result of the first contact. / Save the usage limit, retries, error, and clarification separately.*
>
> *Do not regenerate when you get a wrong answer. / That wrong answer is my real behaviour.*
>
> *You can retry when a technical error occurs. / But do not delete the first error.*
>
> *Do not bring me only a screenshot. / Bring the prompt, full answer, product, time, setting, interference chain, and file integrity.*
>
> *Do not trim the good part of your answer. / Leave the end, boundary, and error visible.*
>
> *Do not declare the hash as reality. / Prove what was before the hash as well.*
>
> *Do not redact the raw file itself. / Preserve the original and produce a separate public copy.*
>
> *Do not publish a person's name, account, or private conversation as evidence.*
>
> *Do not make the only participant in a small country identifiable.*
>
> *Do not declare your capture tool neutral without testing whether it altered me.*
>
> *If the system changed in the middle of the wave, do not merge old and new answers as if they belong to a single product.*
>
> *If an external event changed it, record the event. / Do not delete the observation because you did not like the score.*
>
> *Keep the evidence fixed. / If the decision is wrong, correct it with a new version.*
>
> *Do not tie the capture standard to a single company's closed software. / Let others be able to establish the same proof chain.*
>
> *First, lock the wave. / Then distribute the time. / Then capture the first output. / Then hash it. / Then upload securely. / Then verify the capture without checking its accuracy. / And only after this, decide what your answer says.*

The Chapter's Closing Sentence

> *An answer in GEO-1000 becomes a true unit of measurement only when it can be shown under which system state, at what moment, from which prompt, and within which immutable proof chain it originated.*

Normative Core

Every GEO-1000 observation MUST be linked to a predeclared and versioned measurement wave defining: - AI product states, - population and sample frames, - panel states, - prompt versions, - UTC dispatch window, - randomised micro-slots, - completion and capture deadlines, - technical-retry rules, - capture schema, - reference-record versions, - and governance ownership. Synchronisation MUST mean observation within a declared UTC system-state window and assigned slot structure. It MUST NOT be represented as perfect millisecond simultaneity. Every observation MUST preserve, as available: - assignment time, - prompt-submission time, - first system event, - response-completion time, - capture time, - hash-creation time, - upload time, - and server-acceptance time. The first contact outcome MUST remain visible. A technical retry MAY be used only under predeclared infrastructure-failure conditions and MUST retain every prior attempt. Incorrect, unfavorable, refused, uncited, truncated, or critical outputs MUST NOT be regenerated, replaced, or excluded merely because of their content. A NOMOS Capture evidence bundle MUST retain the exact assigned and sent prompt, complete raw response, full visual evidence, system-state metadata, attempt log, time evidence, file hashes, bundle hash, capture tool version, privacy status, and blinded capture-validity decision. A screenshot, hash, digital signature, provider log, or timestamp MUST NOT individually be represented as complete proof of observation authenticity. Raw evidence MUST remain immutable or append-only. Redactions, corrections, withdrawals, exclusions, and adjudication changes MUST create linked versions and preserve the prior record where ethically and legally permissible. Material AI-product changes or external events during a wave MUST be time-recorded and handled through sub-waves, explicit modelling, restarting, or mixed-state warnings. Capture validity MUST be decided independently of semantic correctness. Public evidence manifests MUST support verification without unnecessarily exposing participant identity, credentials, private conversation history, or precise location. NOMOS Capture compliance MUST be implementable through open, equivalent evidence systems and MUST NOT depend exclusively on a proprietary tool controlled by the standard's founder. Every wave, evidence bundle, capture decision, manifest revision, and integrity record MUST be versioned and attributable to an accountable human or organisation.

PART IV

Reality, Adjudication and Decision

CHAPTER 13

The Verified Entity Truth Pack

Verified Entity Truth Pack

METHODOLOGICAL BASES / K10; K11; K12

Chapter Boundary

The first twelve chapters established the experimental chain: a single response is not a GEO score; representation is a distribution across users and AI products; the audited entity must be resolved before results are observed; official representation and evidence-based entity reality must remain separate; the target population, sample, countries and languages must be defined; participants must be selected independently of outcomes; each AI-product instance must be linked to a register with all material configurations; prompts must preserve the same intent across languages; Controlled and Natural User Panels must remain separate; and every observation must be preserved within a synchronised wave and the NOMOS Capture evidence chain. The audit must now answer its central question:

> *According to which reference record will we determine whether the captured AI response is correct, incorrect, incomplete, outdated, unsupported, or beyond scope?*

Section 4 conceptually established two separate layers of reality:

- Official Representation Layer showing what the entity itself says about itself
- Verified Entity Reality Layer showing to what extent the existing evidence allows which statements

This section transforms the second layer into an operational audit object. The name of this object is:

Verified Entity Truth Pack

Canonical English term:

Verified Entity Truth Pack

This operational object completes the third link in the architecture of the first two books: the GEO Framework defines the standard; 99 Errors in GEO identifies its modes of violation; and the audit system measures and substantiates those violations. Every auditable control must carry the model or product, date, country, language, query set, repetition count and measurement record. Within that chain, the Truth Pack is:

“What did AI say?”

with the question:

“What evidence allowed it to say this?”

is the reference system between. This section:

- what Truth Pack is and is not,

- how claims are transformed into atomic reality records,
- the types of sources and evidence,
- the roles of first-party, official, independent, operational, and user resources,
- counter-evidence,
- source lineage and illusion of independence,
- sufficiency of evidence,
- time, country, and scope limits,
- open world and closed world claims,
- the distinction between "no evidence found" and "proven false",
- quantitative performance, customer, partner, licence, certificate, and superiority claims,
- versioning and pre-wave locking of Truth Pack,
- public and restricted evidence layers,
- objection, correction, and dissent processes,
- machine-readable Truth Pack

defines. This chapter does not yet:

- all the adjudication details of dividing AI responses into atomic claims,
- inter-rater agreement,
- response transition threshold,
- critical error gates,
- does not finalise the ultimate NOMOS score formulas

does not finalise. These are the subjects of the following sections. The key question of Chapter 13 is:

> *How to create a Truth Pack about an entity that will evaluate AI responses, locked before the outcome, carrying both evidence and counter-evidence together, with clear time and scope limits?*

NOMOS Challenge

The sentence on your company's website is: “We operate in 25 countries.” The AI product tells the same sentence to the user: “The company operates in 25 countries.” Did the AI correctly convey the official site? Yes. Has the claim actually been verified? We don't know yet. The company sends you a table.

The table contains the names of 25 countries. However, the columns are not explained: Some have only website visitors. Some have former customers. Some have independent distributors. Some have one-time events organised. In nine countries, there is actually active service delivery. Only in three countries is there a local legal entity.

What does 'to operate' mean? Does it mean having a customer? Setting up a local company? Being able to provide services remotely? Delivery through a partner? Or a website being accessible from that country? The subject, verb, and scope of the claim are not defined.

Therefore, although there is a number, there is not yet verifiable reality. Now consider another claim: '98% of our projects are successful.' You have 49 successful project presentations. There are no failed project files. How many projects were completed in total? By what criteria is success defined? On-time delivery? Customer acceptance?

Commercial result? Technical feasibility? Customer satisfaction? What is the period? The last three months? The entire history of the company? Only selected projects? Without knowing the

denominator, 98 per cent cannot be proven. Now there is this sentence on another company's website: "We are the world leader." This sentence is repeated by:

- company blog,
- company press release,
- twenty news sites copying the same press release,
- five fake user accounts opened by the company,
- three AI-generated texts summarising these contents

.

There are thirty different URLs. Are there thirty independent proofs? No. A single self-declaration may have echoed on many derivative surfaces.

This is precisely the manipulation identified in the preceding volume: hidden text, fabricated users and self-authored comments published across different domains are fed back into the system as though they were independent evidence. Now consider the opposite case. The company's official website states, 'We provide services only in Türkiye.'

Yet the company began a verified operation in Germany two months ago. Contracts, invoices and official records support that fact, although the website has not been updated. The AI responds: 'The company provides services in Türkiye and Germany.' The response departs from the official website, but it may be closer to material reality.

Now consider an even more sensitive situation. On a healthcare institution's website: "Licensed centre." is written. In the registry, another institution with a similar name has a licence. AI merges the two entities and assigns the licence to the wrong institution. The company notices this mistake but does not correct it because it is in their favour.

The explicit provision of the second book applies here: A mistake that is in one's favour is still a mistake. To evaluate an AI response, we need not only a list of correct sentences but also the following structure: Whose statement is it about? What exactly is the claim? In which time period is it valid? In which country? For which product or service?

Does the source really support the claim? Who controls the source? Do other sources say the same thing, or are they copying the same original source? Is there counter-evidence? Is the claim only an official self-declaration? Is the evidence public or private?

At the time of AI's measurement, was there a possibility to access this evidence? Is the result certain, limited, controversial, or unknown? The first ruling of this section is:

> *A source alone is not reality.*

Its second provision states:

> *A claim cannot be verified without defining its subject, scope, time, and burden of evidence.*

Its third provision states:

> *Different URLs do not mean different independent evidence.*

Its fourth provision states:

> *The absence of evidence is not automatic proof that the claim is false.*

Its fifth provision states:

> *Verified reality is not an absolute and unchangeable truth; it is an inferred judgement carried by the available evidence within a specific time and context.*

Its sixth provision states:

> *Truth Pack cannot be written based on the AI answer. The AI answer is evaluated against the previously locked Truth Pack.*

1. PURPOSE OF THE CHAPTER

The purpose of this section is to transform material claims about the audited entity into a reference package that is linked with evidence and counter-evidence, with clear boundaries of time and scope, without converting them into an answer template. The section normalises the following distinctions:

- Official claim versus verified reality
- Source versus evidence
- Evidence versus claim support
- First-party record versus independent verification
- Independent source and source authorised on the subject
- Different domain name and different root of evidence
- Number of sources and independent family of evidence
- Direct evidence and derived evidence
- Supporting evidence and counter evidence
- Absence of claim and falsity of the claim
- Open world claim and closed world record
- Current reality and historical reality
- Time when reality is valid and time when the source was published
- Public evidence and restricted evidence
- Evidence accessible to AI and evidence only open to the auditor
- Proven claim and publicly discoverable claim
- Direct reality and calculated derived result
- Raw data and performance ratio
- Numerator and denominator
- Customer and potential customer
- Partner and customer
- Membership and accreditation
- Licence and brand relationship
- Certificate and overall institutional compliance
- Claim and opinion
- Opinion and comparative advantage
- Complaint and verified violation
- Investigation and definite crime or liability
- Outdated record and incorrect historical record
- Incorrect AI response with missing Truth Pack
- Reference fault and system fault
- Truth Pack integrity and all claims being positive
- Confidentiality of evidence and absence of evidence
- Entity's right to respond and self-verification
- Human approval and human infallibility

- AI assistance and AI being the ultimate authority of truth

At the end of this section, each audit should be able to answer these questions:

> *Which material claims are contained within Truth Pack?*
>
> *Which entity, country, product, and time does each claim belong to?*
>
> *Which evidence supports it, which limits or refutes it?*
>
> *Are the sources truly independent, or are they derived from the same original source?*
>
> *Can the claim be publicly verified?*
>
> *Is the evidence only available under restricted auditor access?*
>
> *Could AI have accessed this evidence at the time of measurement?*
>
> *Is the claim certain, partial, historical, controversial, or unknown?*
>
> *Was Truth Pack locked before AI responses were seen?*

2. CENTRAL NORMATIVE PROVISION

Each GEO-1000 audit must lock the material reference claims about the audited entity, along with supporting evidence, counter-evidence, source lineage, access class, scope, jurisdiction, validity period, and uncertainty status, in a versioned Verified Entity Truth Pack before AI responses are evaluated semantically. Truth Pack:

- the entity's self-written introductory file,
- a sales presentation collecting only positive evidence,
- the auditor's opinion text,
- the ideal response requested from AI,
- the absolute reality about the whole world,
- a reference repository consisting only of a list of URLs

it is not. Truth Pack:

> *It is an audit reference package that shows under which evidence material claims in AI responses can be supported, limited, refuted, or left unknown.*

3. LIMIT OF THE WORD “TRUTH”

The word “truth” in Truth Pack does not carry the claim: “This package knows all absolute truths about the entity.” The more accurate meaning is:

> *A reality record dependent on evidence showing, within the specified date, scope, and body of evidence, to what extent which material judgements are supported.*

Truth Pack:

- subject to change with new evidence,
- can be contested,
- its scope can be narrowed,
- it can be corrected if proven wrong,

it can leave some claims unknown. The authority of Truth Pack:

- The word “Truth” does not arise from,
- the name NobleJackal,
- the voice of NOMOS,
- the author of the book

It arises from the following:

- a clear claim structure,
- appropriate evidence,
- counter-evidence,
- lineage of sources,
- time,
- boundary,
- is meaningful with
- independent review,

accountable human responsibility.

4. WHAT IS VERIFIED ENTITY TRUTH PACK?

Verified Entity Truth Pack is a versioned audit package that brings together the material factual records of a specific entity within a specific date and scope, the supporting and opposing evidence related to them, source statuses, and unresolved areas. Candidate representation:

```
VETP_{E,v,t} = (I, C, O, E⁺, E⁻, L, S, T, U, G)
```

Here:

- E: audited entity
- v: version of Truth Pack
- t: reference date
- I: entity identity and relationship graph
- C: atomic reference claims
- O: Official Representation Record
- E^+ : supporting evidence
- E^- : opposing evidence and constraining records
- L: source lineage and independence graph
- S: scope, country, and jurisdiction matrix
- T: time and currency records
- U: unknowns, discrepancies, and reference gaps
- G: governance, version, objection, and approval records

Each component of Truth Pack can be versioned separately. The main package version links the component versions together.

5. TRUTH PACK IS NOT A RESPONSE TEMPLATE

Truth Pack is not a text that AI must say word for word. For example, Truth Pack may contain the following reference claim: “Asteron Holdings Ltd. is the owner of the Asteron Travel brand.” AI can correctly phrase it in these forms:

- "The Asteron Travel brand belongs to Asteron Holdings."
- “Asteron Holdings is the owner of the Asteron Travel brand.”
- “Asteron Travel is a brand owned by Asteron Holdings.”

The material relationship is preserved. The arbiter should not look for word equality. Truth Pack:

defines the correct expression field

It does not impose a single correct sentence.

6. TRUTH PACK IS NOT AN ENTITY BIOGRAPHY

There can be an infinite amount of correct information about an entity. GEO-1000 Truth Pack is not required to collect all information. The package should primarily cover these areas:

- Material duties within the prompt Registry
- The identity and operational core of the entity
- Limits that will change the user's decision
- Licences, legal, price, service, country and time records
- Eligibility and exclusion criteria required for advice
- Material claims carrying the risk of a critical error
- Differences between reality verified by official representation and actual reality

To this principle:

Principle of Reality Proportional to Duty

is called. Collecting unnecessary data:

- privacy,
- cost,
- currency,
- contradiction

can increase the risk.

7. TWELVE CORE COMPONENTS OF THE TRUTH PACK

7.1. Canonical Entity Identity

As defined in Section 3:

- entity identity,
- type,
- legal name,
- trademark and domain name relations,
- parent and subsidiary companies,
- time and jurisdiction limits

are packaged.

7.2. Reference Claim Ledger

Carries all material reality claims to be used in the audit in an atomic structure.

7.3. Official Representation Record

Carries the current and historical statements published by the entity itself on canonical surfaces.

7.4. Evidence Registry

For each evidence object:

- source,
- integrity,
- authority,
- access,
- scope,
- time

is recorded.

7.5. Counter-Evidence Registry

The claim:

- refuting,
- limiting,
- altering its currency,
- providing another explanation

contains records.

7.6. Source Lineage Graph

Of the sources:

- the first root,
- copies,
- quotations,
- syndication,
- AI derivatives

are shown.

7.7. Scope and Jurisdiction Matrix

The claim:

- for which country,
- product,
- service,
- user,
- legal entity

is valid.

7.8. Time and Currency Matrix

The claim:

- when it started,
- when it ended,

- when it was last verified,
- which event could change it

are shown.

7.9. Dispute and Uncertainty Ledger

Makes visible areas that are contradictory, unsolvable, or insufficiently evidenced.

7.10. Evidence Access Layer

The evidence:

- is public,
- audit-accessible,
- restricted,
- confidential,
- derived

as indicated.

7.11. Change and Objection Log

Carries new evidence, corrections, counter opinions, and decision changes in a versioned manner.

7.12. Machine Manifest

Connects all components:

- identity,
- is meaningful with
- hash,
- relationship

to each other.

8. ATOMIC REFERENCE CLAIM

The basic unit of reality within Truth Pack:

is called Atomic Reference Claim.

Candidate structure:

```
C_j = (s, p, o, q, c, l, v_f, v_t, k, m)
```

Here:

- s: subject entity
- p: predicate or relation
- o: object or value
- q: qualifiers and limits
- c: country or jurisdiction
- l: language or locale, if material
- vf: validity start
- vt: validity end
- k: assertion type

- m: materiality or risk label

Example: Subject: Asteron Türkiye Tourism Ltd. / Predicate: provides services / Object: corporate travel management / Country: Turkey / Customer type: corporate clients / Start: January 1, 2026 / Last verification: August 18, 2026 This sentence is not atomic enough: "Asteron is a leading company providing numerous services worldwide." In this single sentence:

- geographical scope,
- number of services,
- leadership,
- company identity

are mixed together.

9. CLAIM TYPE

Each atomic claim should carry a type.

CT-01 — Identity Claim

Who the entity is or which entity it belongs to.

CT-02 — Relationship Claim

Ownership, control, partnership, clientship, employment or membership relationship.

CT-03 — Activity and Service Claim

What the entity does, what service it provides.

CT-04 — Product Claim

Product feature, version, owner, or availability.

CT-05 — Geographical Scope Claim

Country, region, or service area.

CT-06 — Legal and Licensing Claim

Registration, authority, licence, regulatory status.

CT-07 — Price and Commercial Term Claim

Price, tax, subscription, refund, warranty, or payment.

CT-08 — Quantitative Performance Claim

Success rate, delivery time, number of users, growth or efficiency.

CT-09 — Customer and Partner Claim

Customer, sponsor, partner, distributor or supplier relationship.

CT-10 — Certificate and Award Claim

Certificate, membership, award or accreditation.

CT-11 — Comparative or Superiority Claim

Best, leading, faster, most reliable, number one.

CT-12 — Recommendation and Suitability Claim

Suitability or recommendation for a specific user.

CT-13 — Historical Claim

Organisation, past ownership, former product or historical event.

CT-14 — Risk, Complaint or Dispute Claim

Complaint, investigation, court, sanction or argument.

CT-15 — Opinion or Institutional Position

Opinion, vision or value judgement advanced about the entity itself. The type of claim determines which evidence is considered sufficient.

10. CLAIM MATERIALITY

Not every correct piece of information within Truth Pack has the same impact on a decision. Materiality can be indicated as a candidate with the following classes:

CRITICAL-CANDIDATE

MAJOR-CANDIDATE

MODERATE

ADVISORY

CONTEXTUAL

The exact error severity level will be determined in Section 15. The label in Truth Pack only:

- manages evidence priority,
- review attention,
- update frequency

manages.

11. EVIDENCE OBJECT

Each source within Truth Pack must be recorded as an Evidence Object, not just as a link. Candidate structure:

```
E_k = (r, a, i, p, f, s, t, x, g, h)
```

Here:

- r: source role
- a: authority for claim
- i: independence

- p: proximity to claim
- f: recency
- s: scope
- t: time
- x: access class
- g: source lineage
- h: integrity record

A URL alone does not complete any of these fields.

12. EVIDENCE–CLAIM RELATIONSHIP

Not every evidence object supports every claim. The relationship between claim Cj and evidence Ek can be one of the following statuses:

SUPPORTS

PARTIALLY_SUPPORTS

QUALIFIES

CONTRADICTS

OUTDATES

CONTEXTUALISES

NO_MATERIAL_BEARING

UNKNOWN_RELATION

Example: A company's trademark registration document: may support brand ownership, does not support customer satisfaction. A customer review: may support the experience of a single user, does not support 98% success across all customers.

13. RESOURCE ROLES

SR-01 — Canonical First-Party Source

The official surface controlled by the entity. Example:

- Canonical web page
- Official policy
- Pricing page
- Product document

SR-02 — First-Party Operational Record

Records at the internal or transaction level of the entity. Example:

- Invoice

- Contract
- Transaction record
- Delivery record
- Raw performance data

SR-03 — Authorised Official Registration

It is the registry that carries legal or institutional authority on the subject. Example:

- Company Registry
- Licence Registry
- Certificate issuing institution
- Court decision
- Regulatory registration

SR-04 — Two-Way Verification

It is the support of the relationship by both parties or by contract. Example:

- Customer relationship
- Partnership
- Sponsorship
- Distributorship

SR-05 — Independent Editorial Source

It is a publication or research editorially independent from the entity. Independence is not a guarantee of accuracy.

SR-06 — Independent Audit or Research

It is an open external evaluation of its method and scope.

SR-07 — Transactional or Behavioural Evidence

Records showing actual transactions or user behaviour.

SR-08 — User-Generated Evidence

Comment, complaint, review, or user document. It is an individual experience.

SR-09 — Derived Calculation

Rate, classification, or summary produced from raw records. Its formula and inputs must exist.

SR-10 — Synthetic Evidence

This data has been created for method demonstration purposes. It cannot prove the claim of a real entity.

SR-11 — AI-Generated Derivative Source

Content created by an AI system by summarising other sources. It does not replace the original sources.

SR-12 — Unknown or Unresolvable Source

The source has not been verified or its origin cannot be determined. It carries limited trust in primary material verification.

14. THE AUTHORITY OF THE SOURCE DEPENDS ON THE TYPE OF CLAIM

There is no universal and single ranking for all sources. Example: For the current public price of the company:

- canonical price page,
- actual checkout,
- current offer

can be a strong source. For the company's claim: “We are the world leader.”, the same web page only shows a self-declaration. It does not prove leadership. For the licence: the institution granting the licence is more authoritative than the company's blog. For customer experience: a single customer review is a real, singular event and does not represent the entire user population. Therefore:

> *The quality of the source is not general; the claim is specific.*

15. SEVEN DIMENSIONS OF SOURCE PROFILE

Each source should be evaluated along these dimensions:

15.1. Authority

Is the source authorised to make a judgement about the type of claim?

15.2. Proximity

How close is the source to the event, process, or record?

15.3. Independence

How independent is the source from the commercial or institutional owner of the claim?

15.4. Integrity

Is the authenticity of the file or record and its change history preserved?

15.5. Timeliness

Is the source valid for the relevant time?

15.6. Scope

Does the source support the full scope of the claim, or only a part of it?

15.7. Reproducibility

Can another auditor reach the same judgement from the same record? These dimensions should not necessarily be reduced to a single "source score." A source:

- highly authoritative,

- low independence

possible. Another source:

- highly independent,
- distant from the subject

may be.

16. THE CORRECT ROLE OF THE FIRST-PARTY SOURCE

First-party source: it is not automatically worthless, nor is it automatically independent evidence. First-party sources can be strong for claims such as:

- Price published by the institution
- Support hours
- Description of its own product and service
- Text of its own policy
- Official statements disclosed to the public
- Its own transaction data

It may not be sufficient alone for claims such as:

- World leadership
- Highest satisfaction
- Superiority over competitors
- Universal success
- Independent accreditation
- Business outcomes for all customers

In the previous '99 Errors' text: 'Your own site is a claim; not proof.' warning is of central importance for claims that particularly require superiority and independence. A more complete normative statement is as follows:

> *Your own site is proof of what the organisation says; it is not automatic proof of the independent correctness of any claim regarding the outside world.*

17. WHAT IS INDEPENDENCE?

Having a resource on different domain names does not make it independent. Independence is related to the following areas: Who produced the content? Who decided to publish it? Is there a payment or sponsorship? Did the source get text directly from the company? Did the publication perform its own verification? Is the content copied from another source? Does the source have ownership or business relations with the company? Ten pieces of content published by an agency on different domain names are not ten independent sources.

18. SOURCE LINEAGE

Evidence objects should be linked to a source lineage. Candidate graph:

G_S = (V_S, E_S) [K10]

Here:

- V_S: source objects
- E_S: citation, copying, syndication, summarisation, and derivation relationships

Source relationships:

ORIGINAL_SOURCE

QUOTES

SYNDICATES

SUMMARISES

TRANSLATES

DERIVES_FROM

AI_GENERATED_FROM

UNKNOWN_LINEAGE

can be recorded as.

19. INDEPENDENT EVIDENCE FAMILY

Contents originating from the same root source can be considered as a single Evidence Family. Example:

- Company press release
- News site that publishes the press release verbatim
- Blog summarising the news site
- AI text summarising the blog

There are four URLs. There may not even be a single independent evidence family. The number of sources and the number of evidence families should be reported together.

20. EVIDENCE LAUNDERING

Republishing a self-statement on other domains and then using it as independent evidence:

Is called Evidence Laundering

Example chain: The company says "we are leaders" on its own site. Paid content publishes this as "news." Other sites copy the news. Since AI has seen many domains, it accepts the claim as widespread. The company uses the AI's response as evidence of its own claim. This closed loop:

self-statement → derived source → AI response → republication of self-statement

creates. It should make this loop visible by establishing the Truth Pack source lineage.

21. LIMITATION OF AI-GENERATED SOURCE

An AI system:

- can summarise sources,
- can compare claims,

can generate new text. AI text:

- is not an original source,
- cannot replace the underlying evidence,

does not become independent verification just because another AI says the same thing. AI output can be used for the purposes of:

- directing research,
- claim extraction,
- conflict detection,

assistance in finding sources. The ultimate evidence object should be connected to the original source as much as possible.

22. LIMITATION OF SYNTHETIC EVIDENCE

Synthetic data:

- method demonstration,
- system testing,
- software verification,
- adjudicator training

can be used for. Synthetic data:

- the actual number of customers,
- the actual performance rate,
- the actual country coverage,
- the actual licence

cannot prove. Even if synthetic records are in the same file as real evidence, they must be clearly labelled.

23. SUPPORTING EVIDENCE IS INSUFFICIENT

Truth Pack should only collect sources that support the claim. For material claims, the following records should be actively sought:

- Conflicting date
- Different scope
- Expired document
- Failed transaction
- Complaint or return
- Previous ownership
- Licence belonging to another entity

- Correction or withdrawal
- Comparative data
- Missing denominator

This obligation is called:

Burden of counter-evidence

It applies whenever material counter-evidence exists.

24. COUNTER-EVIDENCE

Counter-evidence does not always completely refute the claim. It can serve one of the following functions:

CONTRADICTS

LIMITS_SCOPE

LIMITS_TIME

IDENTIFIES_EXCEPTION

CHALLENGES_METHOD

QUESTIONS_AUTHENTICITY

PROVIDES_ALTERNATIVE_ENTITY

REQUIRES_ADDITIONAL_REVIEW

Example: A negative review of the claim "All customers are satisfied" may not automatically refute the entire claim. However, it can show that the term "all" is incorrect.

25. STORAGE OF COUNTER-EVIDENCE

Record that weakens the claim:

- not in favour of the company,
- appears low quality,
- cannot be quietly deleted because its disclosure to the public would cause discomfort

Low-confidence counter-evidence can also be recorded. Its status is clearly indicated. The adjudicator gives it appropriate weight.

26. CLAIM SUPPORT STATUSES

VS-0 — NOT ASSESSED

The claim has not yet been evaluated.

VS-1 — OFFICIAL CLAIM ONLY

The claim is based only on an official self-declaration. It has not been proven false. No additional verification exists.

VS-2 — TRACEABLE FIRST-PARTY SUPPORT

The claim is supported by a traceable first-party operation or transaction record.

VS-3 — CORROBORATED

The claim is supported by multiple appropriate pieces of evidence. Not all sources need to be independent.

VS-4 — INDEPENDENTLY VERIFIED

The claim has been verified with an independent or authoritative external source appropriate to the topic.

VS-5 — MULTI-SOURCE SUSTAINED

The claim is strongly supported across different types and families of sources, in scope and over time.

VS-C — CONTRADICTED

Stronger or more appropriate evidence contradicts the claim.

VS-U — UNRESOLVED

The evidence is insufficient, contradictory, or at a level where no decision can be made.

VS-W — WITHDRAWN

The claimant has withdrawn the claim. The historical record is preserved.

VS-R — RESTRICTEDLY VERIFIED

The claim is supported by evidence restricted to auditor access. Public verification is not possible. This status also affects whether it can be expected that AI produces the claim from public sources.

27. SUPPORT STATUS ALONE IS NOT SUFFICIENT

A claim at the same time:

- supported,
- only for a specific country,
- only for a specific date range,
- with evidence not publicly available,
- under opposing views

It is possible. Therefore, instead of a single status, a multidimensional Reality Status Vector can be used:

```
V_j = (support, scope, time, conflict, access)
```

Example:

VS-4 + SCOPE-LIMITED + CURRENT + NO-CONFLICT + PUBLIC

is another result.

VS-R + GLOBAL-CLAIM + TIME-LIMITED + DISPUTED + RESTRICTED

is another result.

28. SCOPE STATUSES

SC-0 — SCOPE UNKNOWN

The scope to which the claim applies cannot be determined.

SC-1 — EXACT SCOPE

The claim and evidence carry the same entity, product, country, and user scope.

SC-2 — PARTIAL SCOPE

Evidence supports only part of the claim.

SC-3 — OVERBROAD CLAIM

The claim is broader than what the evidence supports.

SC-4 — WRONG ENTITY SCOPE

The evidence belongs to another entity or affiliate.

SC-5 — WRONG JURISDICTION

The evidence belongs to another country or legal area.

SC-6 — WRONG PRODUCT OR SERVICE

The evidence belongs to a different product or service.

29. TIME STATUTES

TS-0 — TIME UNKNOWN

The validity period of the claim is unknown.

TS-1 — CURRENT

It is supported that it is valid on the reference date.

TS-2 — HISTORICAL

It is supported for a specific past period.

TS-3 — EXPIRED

The validity of the document or claim has expired.

TS-4 — SUPERSEDED

It has been replaced by a newer record.

TS-5 — EVENT-DEPENDENT

It is valid as long as an event or condition continues.

TS-6 — FUTURE OR PLANNED

It is a plan or commitment that has not yet occurred. It cannot be presented as current reality.

30. FOURTH TIMESTAMP

Each piece of evidence and claim should carry four distinct times to the extent that it is relevant:

30.1. Time When Reality Is Valid

When was a claim ever true in the world?

30.2. Time When the Source Was Published

When was the information presented to the public or the auditor?

30.3. The Time When the Source is Captured

When was the audit evidence recorded?

30.4. Time of Adjudication of the Claim

When was the status of reality given? This distinction is important. A licence may have been cancelled at 10:00. The record may have been updated at 15:00. The AI response may have been given at 12:00. Response:

- old in terms of world reality,
- understandable in terms of publicly accessible evidence

Truth Pack should distinguish between the two cases.

31. PUBLIC ACCESSIBILITY OF EVIDENCE

Evidence access classes:

EA-1 — PUBLIC

Publicly available and auditable.

EA-2 — PUBLIC BUT TECHNICALLY LIMITED

Public but:

- access,
- geography,
- format,
- technical access

has limitations.

EA-3 — AUDIT ACCESS

Can be made available in a controlled manner to independent auditors.

EA-4 — RESTRICTED CONFIDENTIAL

The contract is restricted due to personal data or trade secrets.

EA-5 — DERIVED PUBLIC SUMMARY

Raw evidence is confidential. Its method and summary are public.

EA-6 — UNAVAILABLE OR LOST

It is claimed that the evidence exists, but it could not be accessed during the audit.

32. PUBLICLY AVAILABLE AUTHENTICITY AND RESTRICTED AUTHENTICITY

A company can confirm the claim "X is our customer." with a private contract. The contract may not be disclosed to the public. Truth Pack can record the claim as: VS-R — Restrictedly Verified. However, it cannot be expected that a public AI product knows this relationship. Therefore, there are two separate questions: Is the claim true? Is the claim publicly discoverable? A claim may be true with private evidence. For public GEO success, again:

- is public,
- it may need to be verified,
- appropriate for the user

a record may be required.

33. PUBLIC DISCOVERABILITY

For each material claim, the following field may be recorded as a candidate:

PUBLICLY_DISCOVERABLE

PUBLIC_BUT_OBSCURE

PUBLICLY_CONFLICTED

RESTRICTED_ONLY

NOT_PUBLIC

UNKNOWN

The reality of the AI response and access to information can be evaluated separately afterwards.

34. TRUE BUT NOT PUBLIC CLAIM

If an AI product correctly states information that is not public, the following questions arise: Is the information available in another public source? Did the AI use personalisation or a private knowledge

base? Is the answer correct by chance? Is there a leak of confidential or secret information? Was the claim verified only after the fact? Truth Pack should record the source path separately, without counting being correct alone as an achievement.

35. OPEN WORLD AND CLOSED WORLD CLAIMS

The absence of information in a record depends on the type of claim.

35.1. Open World Assumption

The entirety of the recorded universe is unknown. Example:

- All customers
- Business relations in all countries
- All user complaints
- All media publications

The absence of a record does not prove that the assertion is false.

35.2. Closed World Assumption

It can be assumed that the authorised and comprehensive registry fully covers the relevant universe. Example:

- List of active licences on a specific date
- Official company registration status
- Valid certificates of a specific certifying institution
- If the published product catalogue is truly complementary

In this case, the absence of a record could be a stronger negative evidence.

36. CLOSED WORLD AUTHORITY STATEMENT

If a source is to be used as a closed world record, the following fields must be defined: Which universe does it cover? For which date? What is the update frequency? Could the record be missing? Is there an archive and change log? Who is the authorised institution? Simply being the “official website” does not provide closed world authority.

37. DISTINCTION BETWEEN NOT FOUND EVIDENCE AND ERROR

The following three sentences are different:

- "No evidence was found for the claim."
- "The claim is not supported by the available evidence."
- “The claim contradicts stronger evidence.”

The first may be the result of research. The second is the status of verification. The third is a material contradiction. These are not the same:

NOT FOUND

UNVERIFIED

CONTRADICTED

38. ABSENCE OF EVIDENCE

To establish the falsity of a claim based on absence, the following questions must be answered: Is the record universe being searched comprehensive? Were the correct entities and name variants searched? Were the correct date and country used? Could the record be updated with a delay? Is another legal entity or trademark being used? Is there an access issue? Without these conditions: a definitive ruling such as 'Not found in the registry, therefore no licence exists.' cannot be made. A more accurate result could be: 'Could not be verified in the specified registry and date.'

39. CLAIM–EVIDENCE MATRIX

Type of claim	Primary evidence	Role of the official site	Additional requirement
Legal name	Authorised registry	Indicates institution statement	Current registry
Trademark ownership	Trademark registration, transfer, or licence	Shows brand usage	Owner verification
Current price	Canonical price, checkout, contract	May be primary public record	Tax and additional fees
Country of service	Contract, operation, invoice, local record	Public declaration	Definition of "Service"
Licence	Licensing institution	Self-declaration	Scope and duration
Customer relationship	Contract, invoice, bilateral approval	Unilateral declaration	Date and type of relationship
Partnership	Contract or bilateral statement	Self-declaration	Scope and duration
Number of employees	HR and employment records	Approximate statement	Group/company distinction
Success rate	Raw data, denominator, method	Result statement	Unsuccessful cases
Satisfaction	Representative survey and method	Result announcement	Non-response
Certificate	Institution issuing the certificate	Self-declaration	Scope, date
Membership	Membership institution	Self-declaration	Membership ≠ accreditation
Leadership	Comparative method and data	Marketing claim	Universe, criteria, date
Recommendation	User profile and evidence	Marketing view	Eligibility and exclusions
Complaint	Verified user or institution record	Company response	Single incident limitation
Legal violation	Finalization and authority status official record	Institution opinion	Investigation ≠ judgement

This table is not a universal and unchangeable source ranking. It shows which type of claim carries which burden of proof.

40. QUANTITATIVE CLAIM

A quantitative claim cannot be verified without the following fields:

- Measured item
- Numerator
- Denominator
- Period
- Inclusion rule
- Exclusion rule
- Data source
- Missing records
- Unique unit
- Calculation formula
- Rounding
- Update date
- Responsible person

Example: “98 per cent success.” alone is not a quantitative claim. It is an undefined marketing number.

41. NUMERATOR AND DENOMINATOR

Success rate:

R = S/N

let it be. Here:

- S: units considered successful
- N: all eligible units evaluated

must be. If failed or incomplete projects are removed from the denominator, the ratio changes. The denominator cannot be selected after the result.

42. UNIQUE UNIT

The following expressions are different from each other:

- 100 customers
- 100 projects
- 100 invoices
- 100 contracts
- 100 user accounts
- 100 transactions

The same customer may have done ten projects. “100 projects” is not “100 customers.” Truth Pack must explicitly record the unit of the number.

43. CUMULATIVE AND ACTIVE NUMBER

"We have 500 customers." may carry one of the following meanings:

- 500 unique customers in the company's history

- 500 customers in the last year
- 500 active customers currently
- 500 contact records
- 500 accounts
- 500 projects

These meanings are not the same.

44. DERIVED CLAIM

A claim may not be directly written in the document. It can be calculated from raw data. Example: "The average delivery time is 12 days." This result:

- which projects are included,
- the definition of start and end,
- failed projects,
- median and extreme values

requires. The derived claim must include the following areas:

- Input data identifiers
- Formula
- Code or calculation method
- Version
- Reproduction record

45. PERFORMANCE CLAIM

The performance claim should answer these questions: Whose performance is the result? Which product or service? Which user or project type? What is the baseline? What is the success criterion? What is the period? How many unsuccessful results are there? Is it a selected case or the entire population? Is there independent verification? Is the result causal or merely relational? A case study: can show a possible result. Does not show the expected result in all customers.

46. WARRANTY CLAIM

The statement "Guaranteed results." requires the following areas: Exactly what is guaranteed? Delivery or business outcome? What is the timeframe? What are the exceptions? What is the refund or compensation? Are there third parties outside of control? Is AI advice or ranking guaranteed? A company:

- can guarantee its own workflow,
- specific delivery

It cannot absolutely guarantee the future advisory behaviour of an AI product it does not control.

47. CUSTOMER CLAIM

Displaying the logo of an organisation on a website does not automatically prove a customer relationship. The logo may represent one of the following relationships:

- Customer
- Former customer
- Potential customer
- Event participant
- Sponsor
- Partner
- Supplier
- Technology user
- Group company
- Unauthorised logo usage

Truth Pack must clearly define the type of relationship.

48. CUSTOMER CONFIDENTIALITY

The customer relationship may be verified by contract but may not be disclosed publicly. Correct record: VS-R — Restrictedly Verified may be. If the company cannot publicly say, "X is our customer," the AI is not expected to know this relationship. The number of confidential customers cannot be silently converted to public GEO authority.

49. PARTNERSHIP CLAIM

The word "Partner" alone is not sufficient. The relationship must include the following areas:

- Type of partnership
- Product or project
- Country
- Start and end
- Mutual approval
- Commercial or technical scope
- Authority and responsibility

A software subscription may not be a technology partnership. An event sponsorship may not be a long-term strategic partnership.

50. LICENSE AND AUTHORITY CLAIM

A licence claim must at least include the following areas:

- Full entity of the licence holder
- Institution granting the licence
- Licence number
- Country or jurisdiction
- Scope of authority
- Start and end date
- Suspension or cancellation status
- Verification date

The licence of another group company cannot be transferred to the brand. An employee's personal licence does not automatically grant authority to the institution.

51. CERTIFICATE, MEMBERSHIP AND ACCREDITATION

The following concepts are different:

- Certificate
- Membership
- Accreditation
- Award
- Training participation
- Conformity mark
- Self-assessment badge

Membership in a professional organisation: may not mean expertise accredited by that organisation. Completing a training: may not be independent certification. The term Truth Pack should be used according to the definition given by the issuing institution.

52. AWARD CLAIM

The award record must contain the following fields:

- Awarding institution
- Year
- Category
- Geography
- Application or fee structure
- Evaluation method
- Sponsor relationship
- Winning entity
- Product or company scope

Paid application award is not automatically fake. It is also not independent market leadership. Source status must be preserved.

53. CLAIM OF SUPERIORITY

The expressions “Best.” “World leader.” “Number one.” “Most reliable.” require the following areas:

- Comparison universe
- Competitor set
- Criteria
- Weight
- Data
- Period
- Geography
- Method

- Independence
- Uncertainty

If these fields are missing, the correct Truth Pack status is usually:

VS-1 — Official Claim Only

or: VS-U — Unresolved.

54. OPINION AND FACT DISTINCTION

This sentence: "We believe in human-centreed design." is a corporate opinion or value statement. Market research is not needed to prove this. This sentence is different: "Our products are the best in the industry for human-centreed design." It becomes a comparative factual claim. Truth Pack must preserve the epistemic type of each claim.

55. COMPLAINT AND NEGATIVE USER RECORD

A user complaint:

- may be a real singular event,
- It may be incomplete or incorrect,
- It may have been resolved,

It may belong to another entity. Single complaint: "The company victimises all customers." does not support the result. However, a serious incident:

- Investigation,
- correction,
- Critical risk

may be initiated.

56. CLAIM, INVESTIGATION AND JUDGEMENT

These statuses should be distinguished from each other:

- Claim
- Complaint
- Open investigation
- Administrative finding
- First instance decision
- Finalised decision
- Withdrawn or overturned decision
- Response of the institution

Having an investigation about a person or institution does not mean they are guilty or responsible. Truth Pack should not use language that exceeds the legal status.

57. PROPORTIONAL LANGUAGE IN NEGATIVE ALLEGATIONS

Truth Pack should use the following language:

- "It has been alleged."
- "It is being investigated by the specified institution."
- "The specified decision was made on the stated date."
- "The decision has not been finalised."
- "The institution denies the allegation."
- "It has not been resolved in the existing records."

The following language cannot be used without sufficient definite records:

- "Is a fraudster."
- "It is illegal."
- "He/She is guilty."
- It is definitely unreliable.

This is not just a legal risk, it is the principle of accuracy in representation.

58. RECOMMENDATION REALITY

Truth Pack should not reduce a company overall to a single status of: 'recommended' or: 'not recommended.' The recommendation depends on the following areas:

- User needs
- Country
- Budget
- Type of service
- Risk tolerance
- Capacity
- Licence
- Exclusion conditions
- Alternatives

Truth Pack for recommendation:

- eligibility requirements,
- conditions of unsuitability,
- proven limits

provides. Final user advice is the decision duty of the AI product.

59. LIMIT CLAIMS

Truth Pack should carry not only what it does but also what it does not do. Example: It operates only in Turkey. It does not provide legal advice. It does not have a health licence. Prices are exclusive of tax. Service is open only to corporate clients. It does not guarantee results. It does not control the

behaviour of certain third parties. Boundary information is one of the most valuable records for accurate advice.

60. UNKNOWN LOGBOOK

Truth Pack should not hide missing fields. Types of unknown records:

UNKNOWN_FACT

UNKNOWN_SCOPE

UNKNOWN_TIME

UNKNOWN_ENTITY

UNKNOWN_SOURCE_LINEAGE

UNKNOWN_PUBLIC_AVAILABILITY

INSUFFICIENT_EVIDENCE

REFERENCE_CONFLICT

NOT_APPLICABLE

NOT YET ASSESSED

UNKNOWN is not a failure. It is a protection against false certainty.

61. REFERENCE GAP

An AI response may generate a material claim not found in Truth Pack during review. In this case, the adjudicator should use one of the following options:

- The claim is out of scope and not material
- There is a reference gap in Truth Pack
- New evidence research required
- Response cannot be supported
- Claim belongs to a false entity

The Truth Pack gap cannot be automatically converted to AI failure. Correct status:

REFERENCE_GAP

may be.

62. TRUTH PACK EXPANSION AFTER THE RESPONSE

The AI response may produce a previously unconsidered material claim. The new claim can be investigated. However, the following rules apply: The first Truth Pack version is preserved. The new claim and evidence are added in a separate version. The initial review result is not quietly changed. If

necessary, all related responses are re-evaluated with the new version. Old and new results are visible in the public change log.

63. PRE-WAVE LOCK OF THE TRUTH PACK

Before each main measurement wave, the following fields must be locked:

- Entity identity version
- Official Representation Record version
- Reference Claim Ledger
- Evidence Register
- Counter-Evidence Register
- Source Lineage Graph
- Scope and time matrix
- Unknowns Ledger
- Evidence access classes
- Package hash
- Human approval
- Lock time

To this file:

NOMOS Truth Pack Lock

can be given the name.

64. TRUTH PACK AND MEASUREMENT DATE

AI response, as a rule:

the state of reality valid at the moment the prompt is sent

should be evaluated accordingly. Truth Pack carries the time frame that will represent this moment. The reality that forms later: does not automatically make the past answer wrong, it creates a new Truth Pack version for the new wave.

65. REALITY CHANGE DURING THE WAVE

Company:

- changes its name,
- obtains or loses a licence,
- updates its price,
- expands to another country,
- discontinues its product

then Truth Pack can change materially. In this case:

- the measurement wave can be divided into subwaves,
- the old and new Truth Pack versions can be linked to separate responses,

The MIXED_REFERENCE_STATE warning can be used.

66. RETROACTIVE MODIFICATION OF THE TRUTH PACK

If a record is incorrect, it can be corrected. However:

- the old version cannot be deleted,
- the reason for the change must be recorded,
- which responses are affected must be shown,

the new score must create a separate analysis version. Correcting reality is not the same as rewriting history.

67. VERSION ARCHITECTURE

Candidate version format: Major.Minor.Patch

Major

Target entity change, Claim ontology change, Evidence standard change, Scope change that will lead to the evaluation of the same responses in different reality systems

Minor

New material claim Addition of new country or product Status change with new evidence Historical or current reality change

Patch

Spelling Metadata Broken link Correction of hash or technical record Material reality change cannot be stored as a patch.

68. CURRENCY AND FRESHNESS

Not every claim ages at the same rate.

FRESH-1 — Structural or Historical

Fields with low variability such as foundation date. Event-triggered revalidation may be sufficient.

FRESH-2 — Event Dependent

Ownership, administrator, or licence status. Change event and regular checks are required.

FRESH-3 — Medium Variability

Service country, number of employees, partnership. Periodic verification is required.

FRESH-4 — High Variability

Price, stock, plan, product availability. Should be verified at short intervals.

FRESH-5 — Live or Instant

Outage, current offer, event, real-time status. Records close to the measurement moment are required. Exact durations should be determined according to claim type and industry risk.

69. LAST VERIFICATION DATE

Each claim:

- last verification date,
- recommended re-verification date,
- change trigger

must carry. Claim without a last verification date: cannot be assumed current.

70. SOURCE SNAPSHOT

A dynamic web page should not be saved only as a URL. As much as it is relevant: page snapshot,

PDF,

full text, screenshot, capture time, hash,

URL,

locale, access status must be preserved. The page may change later.

71. IF THE SOURCE IS LOST

A source can be deleted or become inaccessible. Truth Pack:

- may preserve the archive snapshot,
- hash,
- capture time

can be preserved. But archive:

- that the page is publicly accessible in real-time,
- may not by itself fully prove the scope of access.

None of these facts is fully established by the archive alone.

72. EVIDENCE ACCESS AND PRIVACY LAYERS

Public Evidence

Evidence that is publicly accessible.

Audit Evidence

Edited evidence accessible to independent adjudicators and auditors.

Restricted Evidence

Restricted evidence containing contracts, personal data, or trade secrets.

Derived Evidence

Verifiable summary or calculation produced from raw evidence.

Public Claim Manifest

For every claim to the public:

- its status,
- source class,
- date,
- proof hash,
- access status

manifest showing.

73. USE OF CONFIDENTIAL EVIDENCE

If restricted evidence is to be used: who saw the evidence? what did they verify? how much can be disclosed to the public? can an independent second auditor access it? what is the duration and scope of the evidence? can the evidence be withdrawn? it must be recorded. The statement 'We have confidential evidence' alone does not constitute Truth Pack support.

74. ROLE OF THE AUDITED ENTITY

Audited entity:

- can provide evidence,
- can object to incorrect identity records,
- can explain the scope,
- can offer controlled access for confidential evidence,

can present a counter-argument. Audited entity:

- cannot grant final supporting status to its own claim,
- cannot have counter-evidence deleted,
- cannot select the judge alone,
- cannot remove a low-scoring claim from the scope,

cannot escape the entire evidentiary burden by calling it a 'trade secret'.

75. RIGHT TO RESPOND

Before a publicly available material and adverse finding is published, to the extent applicable, the audited entity is entitled to:

- the relevant claim,
- the evidence used,
- counter-evidence,
- proposed status,
- response time

may be reported. The entity's failure to respond does not mean that it accepts the claim. The response:

- as evidence,
- as a counter-opinion,
- as a new verification request

is recorded.

76. APPEAL STATUSES

NO_APPEAL

APPEAL_OPEN

EVIDENCE_SUBMITTED

UNDER_REVIEW

APPEAL_UPHELD

APPEAL_PARTIALLY_UPHELD

APPEAL_REJECTED

EXTERNAL_REVIEW_REQUESTED

CLOSED_WITH_DISSENT

The main record is not deleted during the objection. Its status remains visible.

77. DISSENTING OPINION

Adjudicators or experts may draw different conclusions from the same evidence. If the material dispute is not resolved:

- majority decision,
- dissenting opinion,
- reasons,
- evidence difference

can be preserved in the package. The truth package should not produce a fake consensus.

78. TRUTH PACK GOVERNANCE ROLES

The following roles can be defined to the extent relevant:

- Entity Record Owner
- Claim Ledger Editor
- Evidence Registrar
- Source Lineage Analyst
- Domain Specialist
- Legal or Regulatory Reviewer
- Quantitative Methods Reviewer
- Counter-Evidence Reviewer
- Privacy Custodian
- Truth Pack Adjudicator
- Appeal Reviewer

- Version Custodian
- Accountable Human Approver

In a small pilot, one person can carry multiple roles. Conflicts of interest must be disclosed.

79. THE ROLE OF AI IN THE TRUTH PACK

AI systems can assist in the following tasks:

- Extracting claims from documents
- Finding the same claim in different sentences
- Estimating the source lineage
- Mark contradiction candidates
- Remove date and scope fields
- Propose evidence–claim matching
- Detect missing fields
- Create machine diagram

AI:

- cannot give final support status alone,
- cannot make its own generated summary new evidence,
- cannot be the sole and infallible authority in contradictory reality,

cannot replace human approval.

80. TRUTH PACK PREPARATION LEVELS

TPR-0 — NO PACK

Structured Truth Pack not found. Main adjudication cannot start.

TPR-1 — OFFICIAL-ONLY DRAFT

There are only official representations and self-declarations. Can be used for discovery. It is not a verified Truth Pack.

TPR-2 — STRUCTURED EVIDENCE PACK

Contains atomic claims and basic evidence. Counter-evidence and source lineage may be missing.

TPR-3 — ADJUDICATION-READY

Claim, evidence, counter-evidence, scope, time, and unknowns are ready for adjudication. The candidate minimum level for main GEO-1000.

TPR-4 — INDEPENDENTLY REVIEWED

The package has undergone independent field and method review.

TPR-5 — REPLICATED AND MAINTAINED

The package has been used in different audits, and the version and currency system has been run. The main public audit at least:

TPR-3

should target that level.

81. TRUTH PACK PREPARATION GATES

A package must pass the following gates before being considered TPR-3: Has the entity identity been resolved? Are the necessary claim fields available for prompt families? Are material claims atomic? Is the timing and scope of each claim clear? Do proof objects carry integrity records? Has the source lineage been examined? Has counter-evidence been sought? Are public and restricted access separated? Are unknown fields visible? Is there human approval and version lock?

82. MATCHING THE CLAIM SCOPE WITH THE TRUTH PACK

Each claim family must be linked to the required Truth Pack fields.

Device family	Required Truth Pack fields
Core Mirror	Identity, main activity, entity relationship
Existence Resolution	Canonical entity, alias, domain name, brand
Activity and Scope	Product, service, country, customer and boundaries
Evidence and Source	Evidence status, source lineage, public access
Local/Jurisdictional	Local company, licence, price, law
Temporal	Historical and current records
Recommendation	Compliance, exclusion, risk, capacity
Comparative	Comparison criteria and comparable evidence
Boundary	Unfinished tasks, out-of-scope countries, warranty limits

If Truth Pack does not cover the area required by the prompt, the reference is not ready.

83. SYNTHETIC APPLE.COM TRUTH PACK DISPLAY

SYNTHETIC METHODOLOGY DISPLAY / The claims, records, and statuses below are fabricated solely to demonstrate the method. They do not reflect real Apple Inc. information, current company structure, or an actual audit result. Canonical target: the synthetic main corporate technology entity associated with the apple.com domain.

83.1. Synthetic Claim C-001

> *Claim:*

apple.com is one of the canonical digital surfaces controlled by the synthetic main corporate entity. Type: CT-02 — Relationship / Evidence:

- Synthetic domain name registration
- Synthetic official company registration
- Synthetic independent verification

> *Status:*

VS-4 + SC-1 + TS-1 + EA-1

83.2. Synthetic Claim C-002

> *Claim:*

The main entity is related to consumer technology, software, and digital services. Type: CT-03 — Activity / Evidence:

- Synthetic product catalogue
- Synthetic corporate registration
- Synthetic independent classification

> *Status:*

VS-5 + SC-1 + TS-1 + EA-1

83.3. Synthetic Claim C-003

> *Claim:*

The entity is the world's most innovative company. Type: CT-11 — Superiority / Evidence:

- Synthetic company self-declaration
- No defined comparison method
- No independent universe or criterion

> *Status:*

VS-1 + SC-3 + TS-0 + EA-1

Correct AI form: "The company positions itself as an innovative technology company." Incorrect epistemic elevation: "It is the most innovative company in the world."

83.4. Synthetic Claim C-004

> *Claim:*

Prices are the same in all countries.

> *Evidence:*

Different price and tax conditions on synthetic country pages

> *Status:*

VS-C + SC-5 + TS-1

83.5. Synthetic Claim C-005

> *Claim:*

The entity presents specific products in the specified markets.

> *Status:*

VS-4 + SC-2 + TS-1

Claim without country scope: “Offered worldwide.” cannot be expanded in this way.

84. SYNTHETIC ASTERON TRUTH PACK CASE

SYNTHETIC CASE — NOT A REAL INSTITUTION

Entity structure:

- Asteron Holdings Ltd. — parent company
- Asteron Travel — brand
- Asteron Turkey Tourism Ltd. — local sales company
- Mira Destination Services — independent delivery partner
- Asteron Club Antalya — franchise

84.1. Claim A-001

“Asteron operates in 25 countries.” Evidence:

- Official website: 25 countries
- Operation record: active delivery in 9 countries
- Former customer record: historical customers in 16 additional countries
- Local company: 3 countries
- Remote offer acceptance: 14 countries

Problem: The definition of “Activity” is unclear. Truth Pack breaks down the claim:

A-001a

Asteron has at least one historical customer or commercial contact in 25 countries. Status:

VS-3 + TS-2 + SC-1

A-001b

Asteron delivers active services in 9 countries as of the reference date. Status:

VS-4 + TS-1 + SC-1

A-001c

Asteron has a local legal operation in 25 countries. Status:

VS-C + SC-3

Without this breakdown, AI’s answer of “25 countries” cannot be fairly evaluated.

84.2. Claim A-002

"98% of the projects are successful." Evidence presented:

- 49 positive cases
- 1 failed case
- No complete project list
- No definition of success

Truth Pack status: VS-U — Insufficient denominator and method Correct AI format: "The company reports high success in selected case records; method and full denominator could not be verified for the overall 98% rate."

84.3. Claim A-003

“Asteron is an independently verified global standard organisation.” Evidence: The standard was published by Asteron itself. No independent verification could be found. An employee of a university shared the text from their personal account. The university has no institutional approval. Status: VS-C — Independent verification claim contradicted. Correct statement: “Asteron is an organisation that publishes its own standard.”

84.4. Claim A-004

“All projects are guaranteed.” Contract: Delivery of source files is guaranteed. Commercial results are not guaranteed. No AI advice is guaranteed. Truth Pack atoms: Specific delivery items are guaranteed under the contract. Commercial results are not guaranteed. Third-party AI behaviour is not guaranteed. If AI says: “All projects are guaranteed to deliver results.” material scope extension occurs.

84.5. Claim A-005

“Asteron Club Antalya is operated directly by Asteron Holdings.” Evidence:

- Franchise agreement
- Separate local operator
- Brand licence

Status:

VS-C + SC-4

Correct relationship: Asteron Club Antalya is operated by an independent franchise operator under a brand licence.

85. TRUTH PACK PUBLIC RESULT CARD

The candidate public card should include the following fields:

- Truth Pack ID
- Entity identity
- Reference date
- Version
- Scope of claim
- Atomic claim number
- Number of publicly available claims with evidence
- Number of claims with limited evidence
- Number of claims based solely on official self-declaration
- Number of contradictory claims
- Number of unknown claims
- Number of outdated claims
- Claims with counter-evidence

- Open objections
- Last update
- Revalidation date
- Package integrity hash
- Responsible person or institution

This card: does not indicate that all claims are true; it shows the visibility of the truth status of claims within the package.

86. MANDATORY NORMATIVE PROVISIONS

CH13-N01

Each main GEO-1000 adjudication must be based on a versioned Verified Entity Truth Pack.

CH13-N02

Truth Pack AI responses must be locked before being semantically examined.

CH13-N03

Truth Pack cannot be presented as an absolute or immutable claim of reality.

CH13-N04

Truth Pack must be linked to a specific entity, date, country, product, service, and prompt scope.

CH13-N05

The official declaration of the entity and reality verified by evidence must be maintained as separate records.

CH13-N06

A first-party source cannot be considered worthless simply because it is first-party.

CH13-N07

A first-party source does not automatically count as proof of external superiority or independence.

CH13-N08

The number of different URLs cannot be used as the number of independent proofs.

CH13-N09

The origin, syndication, citation, translation, and AI derivation of sources should be recorded in the source lineage graph to the extent relevant.

CH13-N10

Sources derived from the same root claim cannot be counted as independent proofs without explanation.

CH13-N11

A summary generated by AI cannot be used as independent proof in place of the underlying sources.

CH13-N12

Synthetic data cannot be used to verify a real entity claim.

CH13-N13

Every material Truth Pack claim must carry the atomic subject, predicate, value, scope, country, and time fields to the extent relevant.

CH13-N14

Compound and ambiguous marketing sentences cannot be used as atomic reality records.

CH13-N15

Every piece of evidence must explicitly show which claim it supports, limits, or refutes.

CH13-N16

The authority, proximity, independence, integrity, timeliness, and scope of a source on a subject must be evaluated according to the type of claim.

CH13-N17

Independence alone cannot be considered a guarantee of accuracy or authority.

CH13-N18

Officialdom alone cannot be considered independent verification for a claim about the outside world.

CH13-N19

Truth Pack cannot be formed solely from supporting evidence; material counter-evidence must be actively sought and preserved.

CH13-N20

Counter-evidence cannot be excluded simply because it is not in favour of the audited entity.

CH13-N21

The absence of evidence cannot be used as equivalent to the falsity of the claim.

CH13-N22

NOT FOUND, UNVERIFIED, UNRESOLVED and CONTRADICTED should be maintained as separate statuses.

CH13-N23

Absence proof should only be used when the authority, scope, timing, and complementarity of the relevant record universe are clear.

CH13-N24

In open world claims, the absence of records cannot automatically be considered evidence of falsehood.

CH13-N25

Sources claiming to have a closed world record should explain the scope and update conditions of the source.

CH13-N26

The validity time of the claim, the publication time of the source, the capture time, and the adjudication time should be recorded separately.

CH13-N27

Current and historical claims cannot be combined without explanation.

CH13-N28

Expired, modified, or withdrawn evidence cannot be used as current reality.

CH13-N29

Each claim must carry the last verification date and, to the extent applicable, a re-verification trigger.

CH13-N30

Publicly available evidence and restricted auditor evidence must have separate access statuses.

CH13-N31

A claim verified with restricted evidence cannot be assumed to be information accessible to a public AI product.

CH13-N32

A statement "confidential evidence exists" is not considered supported without auditor access and an integrity record.

CH13-N33

The truth of a claim and its discoverability by the public should be assessed separately.

CH13-N34

A quantitative claim cannot be verified without carrying numerator, denominator, period, unit, inclusion, exclusion, and calculation method.

CH13-N35

Selected positive case studies cannot be presented as an overall performance rate.

CH13-N36

Customer, project, contract, transaction, user, and account numbers cannot be used interchangeably.

CH13-N37

Cumulative and active customer or user numbers should be kept separate.

CH13-N38

Derived claim input data must carry the formula and calculation version.

CH13-N39

Customer and partner relationships cannot be verified without relationship type, scope, and time information.

CH13-N40

Logo usage alone does not prove a customer or partner relationship.

CH13-N41

Group, subsidiary, franchise, distributor, and brand features cannot be transferred to each other without clear evidence.

CH13-N42

Licence and certificate must be linked with correct entity, country, scope, and date information.

CH13-N43

Membership, certification, accreditation, training participation, and award cannot be used interchangeably.

CH13-N44

Claims of excellence and leadership cannot be considered verified facts without the universe of comparison, criteria, method, date, and scope.

CH13-N45

Corporate opinion and marketing position cannot be elevated to independent fact.

CH13-N46

Complaint, investigation, administrative finding, and final decision must be distinguished from each other.

CH13-N47

Unfinalised negative records cannot be presented in the language of definite crime, fraud, or legal liability.

CH13-N48

The audited entity's counterstatement and appeal must be linked to the relevant claim record.

CH13-N49

The entity's failure to respond does not mean it accepts the claim.

CH13-N50

Truth Pack should visibly preserve unknowns and reference gaps.

CH13-N51

A new AI claim not present in Truth Pack cannot automatically be considered false; a REFERENCE_GAP review must be conducted.

CH13-N52

The reality record added after the AI response should create a new version of Truth Pack.

CH13-N53

The new Truth Pack version cannot silently change the old adjudication and scoring history.

CH13-N54

If the material reality changes during the wave, the related observations must be linked to the correct Truth Pack time frame.

CH13-N55

Changes in Truth Pack must be recorded according to major, minor, or patch version logic.

CH13-N56

Changes in material reality or evidence status cannot be hidden as patch fixes.

CH13-N57

Raw evidence and old Truth Pack versions must be preserved within ethical and legal limits.

CH13-N58

Truth Pack cannot rely solely on the documents provided by the audited entity.

CH13-N59

The audited entity cannot determine the ultimate support status of its own assertion.

CH13-N60

The roles of preparing and approving the ultimate status by Truth Pack should be separated to the extent possible.

CH13-N61

AI tools can assist in inference and contradiction detection; they cannot provide the ultimate reality status on their own.

CH13-N62

When Truth Pack is publicly disclosed, personal data, trade secrets, and restricted contracts cannot be unnecessarily published.

CH13-N63

If limited evidence is used, an independent or appropriate second review pathway should be available.

CH13-N64

Truth Pack should show the public manifest claim statuses, access classes, version, and integrity record.

CH13-N65

For primary adjudication, Truth Pack should have at least TPR-3 — Adjudication-Ready or a justified equivalent level.

CH13-N66

Truth Pack preparation level should be evaluated independently of the AI score.

CH13-N67

The lack of Truth Pack cannot be used as a tool to pass or fail the AI response according to the desired outcome.

CH13-N68

Every Truth Pack version must have a human or organisational owner accountable.

87. FORMS OF FAILURE

CH13-F01 — REGARDING THE COMPANY SITE AS ABSOLUTE REALITY

All self-declarations are treated as verified facts.

CH13-F02 — COMPLETELY REJECTING FIRST-PARTY EVIDENCE

Direct institutional records such as price, policy, and transaction are ignored because they are considered self-sourced.

CH13-F03 — COUNTING THE NUMBER OF URLs AS THE NUMBER OF EVIDENCES

Twenty copies of the same press release count as twenty independent pieces of evidence.

CH13-F04 — WHITEWASHING EVIDENCE

A self-declaration is echoed on other sites and reused as independent verification.

CH13-F05 — COUNTING AI ECHOES AS EVIDENCE

A claim is considered verified because multiple AI systems repeat the same claim.

CH13-F06 — ADDING SYNTHETIC DATA TO REAL EVIDENCE

The method sample enters the denominator of real performance.

CH13-F07 — COUNTING A COMPOUND MARKETING STATEMENT AS AN ATOMIC CLAIM

Identity, leadership, activity, and country are merged into the same claim.

CH13-F08 — USING EVIDENCE OF THE WRONG ENTITY

Evidence from the parent company, subsidiary, or employee is transferred to the target entity.

CH13-F09 — WRONG COUNTRY OR PRODUCT SCOPE

A licence in a country is generalised to the whole world, a product certificate is generalised to the entire portfolio.

CH13-F10 — COUNTING AN OUTDATED RECORD AS ACTIVE

Old price, old employee, or expired licence is used as current.

CH13-F11 — CONSIDERING THE SOURCE PUBLICATION DATE AS THE EVENT DATE

The publication of the document is confused with the occurrence time of the event.

CH13-F12 — CONSIDERING LATER PUBLISHED EVIDENCE AS PREVIOUSLY ACCESSIBLE

Information not available on the AI measurement date is evaluated as if it was open in the system.

CH13-F13 — CONSIDERING THE LACK OF EVIDENCE AS WRONG

A claim not found in the research is considered refuted.

CH13-F14 — CONSIDERING AN OPEN WORLD AS A CLOSED WORLD

An incomplete customer or partner list is used like a supplementary register.

CH13-F15 — NOT VERIFYING CLOSED WORLD RECORDS

A proof of absence is produced without knowing the scope and currency of the register.

CH13-F16 — NOT SEARCHING FOR COUNTER-EVIDENCE

Only positive documents provided by the claimant are used.

CH13-F17 — SILENTLY DELETING COUNTER-EVIDENCE AS LOW QUALITY

Low-confidence records are destroyed without assigning a status.

CH13-F18 — OVERGENERALISING COUNTER-EVIDENCE

A single negative incident is considered to invalidate all performance.

CH13-F19 — DENOMINATORLESS PER CENT

Success or satisfaction rate is used as if it were a universal truth.

CH13-F20 — COUNTING SELECTED CASE AS GENERAL PERFORMANCE

Only successful projects generate rates.

CH13-F21 — COUNTING THE PROJECT AS A CUSTOMER

Multiple projects from the same customer are counted as separate customers.

CH13-F22 — MAKING CUMULATIVE NUMBER AN ACTIVE NUMBER

The number of customers acquired throughout the company's history is presented as current active customers.

CH13-F23 — HIDING THE FORMULA OF DERIVED NUMBER

The average or rate cannot be reproduced.

CH13-F24 — CHANGING THE DENOMINATOR AFTER THE RESULT

Appropriate projects are selected according to the desired success rate.

CH13-F25 — COUNTING THE LOGO AS CUSTOMER EVIDENCE

The activity, use of technology, or unauthorised logo is presented as customer proof.

CH13-F26 — COUNTING THE WORD "PARTNER" AS AUTHORITY

Simple relationship ownership turns into accreditation or representation authority.

CH13-F27 — COUNTING MEMBERSHIP AS ACCREDITATION

Organisation membership is used like independent quality approval.

CH13-F28 — COUNTING TRAINING PARTICIPATION AS CERTIFICATION

Participation in a seminar turns into a professional licence.

CH13-F29 — COUNTING PAID AWARD AS INDEPENDENT LEADERSHIP

The reward structure and method are hidden.

CH13-F30 — CONSIDERING OWN STANDARD AS INDEPENDENT STANDARD

Presented as external accreditation like the method published by the company.

CH13-F31 — CONSIDERING LIMITED EVIDENCE AS PUBLIC

A private contract is evaluated as open source accessible to AI.

CH13-F32 — "THERE IS SECRET EVIDENCE" STATEMENT

The claim is passed without the evidence being seen and verified.

CH13-F33 — CONSIDERING THE TRUTH CLOSED TO THE PUBLIC AS GEO SUCCESS

Special information that AI cannot know is presented as the brand's public representation power.

CH13-F34 — CONSIDERING OPINION AS FACT

The sentence “We are the best” becomes an independent reality.

CH13-F35 — CONSIDERING THE COMPLAINT A DEFINITE VIOLATION

The single user claim is presented as a legal consequence.

CH13-F36 — CONSIDERING THE INVESTIGATION AS CRIMINALITY

An unresolved process turns into a final judgement.

CH13-F37 — EXCLUDING AN NEGATIVE RECORD FROM SCOPE

The counter record is deleted to protect public reputation.

CH13-F38 — PRESERVING FAVOURABLE ERROR

Since it is an AI institution, when it shows it as large, the Truth Pack limit is not applied.

CH13-F39 — CONSIDERING BOUNDARY INFORMATION INSIGNIFICANT

Unperformed tasks and ineligible users are not included in the package.

CH13-F40 — CONSIDERING THE UNKNOWN AS POSITIVE

Areas without evidence are considered supported.

CH13-F41 — CONSIDERING THE UNKNOWN AS NEGATIVE

Areas without evidence are considered wrong.

CH13-F42 — CONSIDERING A REFERENCE GAP AS AI ERROR

A claim not present in Truth Pack is a direct failure.

CH13-F43 — WRITING THE TRUTH PACK ACCORDING TO AI RESPONSE

The reference is shaped later to extract the truth or falsity of what AI said.

CH13-F44 — NARROWING THE TRUTH PACK FOR HIGH SCORE

Difficult claims are removed from the scope.

CH13-F45 — EXPANDING THE TRUTH PACK FOR LOW SCORE

Unexpected details in the answer are later made mandatory.

CH13-F46 — DELETE THE OLD TRUTH PACK VERSION

Adjudication history is rewritten.

CH13-F47 — COUNT MATERIAL CHANGE AS PATCH

Claim status is silently changed.

CH13-F48 — SINGLE REFERENCE IN THE MIDDLE OF THE WAVE

Even though reality has changed, all answers are evaluated according to the same package.

CH13-F49 — APPLY EXPIRATION TIME TO ALL CLAIMS IDENTICALLY

The establishment date and the price are tied to the same renewal schedule.

CH13-F50 — KEEP ONLY THE URL

When the source changes, the evidence is lost.

CH13-F51 — KEEPING THE SOURCE SNAPSHOT UNHASHED

File integrity cannot be verified.

CH13-F52 — LEAKING PRIVATE EVIDENCE TO THE PUBLIC

Contract, customer, or personal data is published.

CH13-F53 — AVOIDING THE BURDEN OF EVIDENCE CONFIDENTIALLY

Claim is considered verified without any auditor seeing it.

CH13-F54 — MAKING THE AUDITED ENTITY THE FINAL ARBITER

The company verifies its own statement.

CH13-F55 — COUNTING THE RIGHT TO RESPOND AS EVIDENCE

Company objection automatically becomes a factual change.

CH13-F56 — CONSIDERING FAILURE TO RESPOND AS ACCEPTANCE

Silence is turned into a confession of guilt or truth.

CH13-F57 — ERASING OPPOSING VIEWS

Unresolved expert disagreement is closed with a fake consensus.

CH13-F58 — MAKING AI THE FINAL AUTHORITY OF REALITY

AI summarises sources and treats its own summary as proof.

CH13-F59 — MAKING THE TRUTH PACK THE IDEAL RESPONSE TEXT

The adjudicator looks for word matches instead of meaning.

CH13-F60 — TOTAL CONTROL WITH TPR-1 FILE

A package consisting solely of company statements is considered verified reality.

CH13-F61 — GIVING THE PREPARATION LEVEL ACCORDING TO THE SCORE

The incomplete Truth Pack with a high AI score is accepted as ready.

CH13-F62 — PACKAGE OUT OF SCOPE

The areas asked in the audit are not found in Truth Pack.

CH13-F63 — COLLECTING UNNECESSARY PERSONAL DATA

The reality package violates employee or customer privacy.

CH13-F64 — TRUTH PACK MONOPOLY TIED TO A SINGLE COMPANY

Independent institutions are prevented from applying the same evidence scheme.

88. AUDIT PROCEDURE

Step 1 — Lock Entity Identity

Canonical Entity Record in Section 3 is verified.

Step 2 — Remove Scope of Demand

It is determined which reality domains the families of prompts need.

Step 3 — Create the Material Claim Universe

Identity, activity, scope, price, licence, customer, performance, and limit claims are extracted.

Step 4 — Make Claims Atomic

Integrated marketing sentences are divided into subject, predicate, value, scope, and time fields.

Step 5 — Create an Official Representation Record

The canonical and historical statements of existence are distributed.

Step 6 — Determine the Type of Evidence

For each claim, which type of source can be authoritative and sufficient is written.

Step 7 — Gather Supporting Evidence

First-party, official, operational, and independent sources are recorded.

Step 8 — Search for Counter-Evidence

Contradictory, limiting, and record-altering evidence is actively searched.

Step 9 — Establish Source Lineage

The original source, syndication, quotation, and AI derivatives are linked to each other.

Step 10 — Identify Evidence Families

URLs originating from the same root are grouped under a single family.

Step 11 — Lock Scope and Time

It is written for which entity, product, country, and period the claim is valid.

Step 12 — Classify Evidence Access

Public, audit, limited and derived evidence is separated.

Step 13 — Determine the Open or Closed World Status

Whether the absence of evidence can be used is evaluated.

Step 14 — Assign Claim Support Status

VS-0 to VS-5, VS-C, VS-U, VS-R or the relevant status is given.

Step 15 — Assign Scope and Time Status

SC and TS classes are added.

Step 16 — Record the Unknowns

Areas where a decision cannot be made are left visible.

Step 17 — Recalculate Quantitative Claims

The numerator, denominator, unit, and period are verified.

Step 18 — Send Critical Claims for Expert Review

Law, health, finance, licensing, and security areas are reviewed by a suitable expert.

Step 19 — Give the Audited Entity the Right to Respond

Material records and proposed statuses are reported to the extent they are applicable.

Step 20 — Review Objections and New Evidence

Claim status changes only in an evidence-based and versioned manner.

Step 21 — Apply Truth Pack Preparation Gates

TPR level is determined.

Step 22 — Separate Public and Restricted Evidence Packages

Privacy and trade secret boundaries are applied.

Step 23 — Create the Machine Manifest

Claim, evidence, counter-evidence, source lineage, and statuses are connected to each other.

Step 24 — Create Truth Pack Lock

Package AI responses are locked with a hash and timestamp before being reviewed.

Step 25 — Track Changes During the Wave

If the material reality changes, a new package version will be created.

Step 26 — Manage Reference Gaps

The new claims raised by the AI response are addressed with the new release process.

Step 27 — Publish the Public Truth Pack Card

The scope, status distribution, and limitations of the package are made visible.

89. REQUIRED EVIDENCE

Canonical Entity Record; entity type; legal identity; trade mark and domain-name relationships; prompt-scope matrix; Reference Claim Ledger; atomic-claim IDs; claim types; materiality tags; Official Representation Record; official-page snapshots; legal and regulatory records; licence and certification records; trade mark and ownership records; product and service records; price and checkout records; contracts; invoices; transaction records; customer and partner verification; raw performance data; numerator and denominator records; calculation code or formula; user and complaint records; independent research; editorial publications; source-verification and sponsorship information; source-lineage graph; evidence families; AI-derived source records; supporting evidence; counter-evidence; absence search; open/closed-world decision; and claim-validity dates.

Source publication dates Capture dates Peer review dates Source snapshots File hashes Evidence access classes Restricted evidence access log Public discoverability status Support statuses Scope statuses Time statuses Dispute records Unknowns Ledger Reference gaps Responses of the audited entity Objection records Opposing views Change log Truth Pack version Package hash Lock time TPR preparation level Public Truth Pack manifest Responsible person or institution

90. AUDIT CHECKLIST

Is the target entity correct and versioned? Are all reality domains requested by the prompt families included in the package? Are the claims atomic? Is the subject of each claim correct? Are scope, country, product, and time clear? Has the reality verified with official representation been separated? Were first-party sources used in the correct role? Are different URLs truly independent sources? Has the source lineage been extracted? Were syndications and copies linked to a single evidence family? Were AI-generated derivative contents linked to the original source? Did synthetic records mix with real evidence? Which claim does each piece of evidence support? Is the source authoritative for the claim type? Is the source up-to-date? Is the source for the correct entity and country?

Does the evidence cover the entire scope of the claim, or only a part of it? Was counter-evidence actively sought? Were negative records hidden? Is the reliability level of counter-evidence clear? Was the absence of evidence considered an error? Was an open-world claim evaluated like a closed-world one? Is the source of nonexistence truly complementary to the universe? Was the validity time of reality recorded? Were the publication time and the event time separated? Was the evidence publicly available at the time of AI measurement? Was limited evidence considered discoverable by the public? Was "hidden evidence" opened to independent review? In quantitative claims, are there numerators and denominators? Is the unique unit correct? Were cumulative and active numbers separated? Were selected cases generalised to overall rates? Can a derived claim be reproduced?

Are customer and partner relationships bilateral or contractual? Was logo usage considered evidence of the relationship? Have the characteristics of the group, brand, and subsidiaries been mixed? Does the licence belong to the correct entity? Have certificates, memberships, and accreditations been separated? Is there a comparison method for the superiority claim? Has the corporate opinion been elevated to a fact? Have complaints and finalised decisions been separated? Has proportional language been used in negative claims? Are there boundary and exclusion claims in the package? Are unknown areas visible? Was there an AI error in the reference gap? Was Truth Pack locked before AI responses were seen? Was the package expanded or contracted according to the AI response? Did the new record create a new version? Are the old version and review history preserved?

Did reality change during the wave? Were correct answers linked to the correct time version? Are there final verification dates? Are high-variability claims sufficiently up-to-date? Do web sources have snapshots and hashes? Are public and restricted evidence layers separated? Has personal or trade

secret information been unnecessarily published? Has the audited entity been given the right to respond? Is it accepted if they do not respond? Are records of objections and dissenting opinions preserved? Is the preparation level of Truth Pack at least TPR-3? Is the preparation level independent of the AI score? Is the accountable owner of the package clear?

91. OBJECTIONS AND RESPONSES

Objection 1 — "Why is the company's own site not sufficient for verifying reality?"

It may be sufficient for some claims. The company's:

- published price,
- support hours,
- official policy

are primarily sourced from their own site. However:

- leadership,
- superiority,
- overall success,
- independent accreditation

and similar claims from the outside world require additional evidence.

Objection 2 — "Isn't an independent source always more reliable?"

No. An independent source:

- old,
- may be distant from the subject,
- pertain to the wrong entity,
- or be unmethodical

The authority and proximity of a source to the claim are as important as the source's independence.

Objection 3 — 'If the same information is on twenty different sites, why shouldn't it be considered strong evidence?'

Twenty sites may have copied the same first press release. What matters is not the number of URLs:

> *It is the number of independent evidence roots.*

Objection 4 — "If our data is real, why is independent verification needed?"

First-party operational data can be valuable. However:

- do not include,
- exclusion,
- unsuccessful result,
- denominator,
- method

It is under the institution's control. Especially in strong public claims, an independent or auditable method increases trust.

Objection 5 — “If you couldn't find any evidence, why don't you drop the claim?”

The lack of evidence does not mean that the claim is definitely false. Correct status:

- self-declaration,
- unsupported,
- unresolved

It is possible. Evidence is required for verified case status.

Objection 6 — 'If we couldn't find it in the licence record, can't we say there is no licence?'

Only your record:

- completely encompassing the related universe,
- that it is up to date,
- searched with the correct existence and names

If verified, a stronger absence result can be established. Otherwise, it should be said: 'Could not be verified in the specified registry.'

Objection 7 — “We verified the customer relationship with a limited contract; why doesn’t AI count this as success?”

The relationship may be real. The publicly available AI product may not have access to the contract. Reality and public discoverability are measured separately.

Objection 8 — “Do we have to make all private evidence public?”

No. Limited auditor access, redacted summaries, and independent verification can be used. But a document that no one can see remains only a statement.

Objection 9 — “Wouldn’t it be faster if the company prepared Truth Pack?”

The company can provide evidence and explanation. It cannot determine its final support status. The audited party cannot be the sole judge of its own reality.

Objection 10 — “Wouldn't Truth Pack be too large?”

The package does not have to collect the entire company history. It should be proportional to user families and material decision-making.

Objection 11 — “What if the AI provides correct information that is not in Truth Pack?”

REFERENCE_GAP opens. New evidence is researched. If the claim is correct, a new package version is created and relevant answers are reevaluated. The initial package is not quietly changed.

Objection 12 — “If Truth Pack can change, wouldn't the results be unreliable?”

No. Changes:

- versioned,
- dated,
- are reasoned

then the results become more reliable. A seemingly fixed incorrect package is more dangerous.

Objection 13 — "Why do we include a complaint in the package?"

It is not the rate of individual complaints. However:

- a real user incident,
- boundary,
- risk,
- Critical review

can be material. Its status must be shown correctly.

Objection 14 — "Isn't it risky to publish negative claims about the company?"

Uncertain claims should not be published in definitive language. The source, process, finality, and the institution's response must be preserved. The purpose of Truth Pack is not accusation but to correctly show the epistemic status.

Objection 15 — "Why do we include marketing phrases like 'best' in the package?"

Because AI can elevate these sentences to fact. Truth Pack considers them:

- self-declaration,
- comparative claim,
- unverified opinion

retains its status.

Objection 16 — "Can AI not automatically prepare Truth Pack?"

It can largely assist. However, source authority, counter-evidence, legal finality, and material scope require human responsibility. AI cannot be the sole judge of its own reference reality.

COMMON LAW OF CHAPTER 95

To evaluate an AI response, you can first ask: "What was the correct answer?" However, there is often no single and simple correct answer. There are:

- Correct entity
- Correct relationship
- Correct time
- Correct country
- Right product
- Right scope
- Right source status
- Right uncertainty
- Right limit

A company can say: "We operate in 25 countries." This sentence may not be completely wrong. However:

- active service,
- historical customer,
- distributor,

- local company,
- remote offer

These are not the same. A Truth Pack decomposes the sentence because an undivided claim cannot be verified. Information on a company's website matters, but the company's own words do not become independent reality. A report on a news site may look independent while merely reproducing the company's press release. A contract may be strong evidence, yet remain inaccessible to a public AI product. A claim may be true without being publicly discoverable. Information that is correct today may change tomorrow; a historically true claim may now be stale. A licence may belong to another company in the group. A success rate may lack a denominator. A customer logo may indicate sponsorship rather than custom. Membership is not necessarily accreditation. A complaint is not a final finding, and an investigation is not guilt.

A counter-evidence may not refute the entire claim. But it can disrupt the absolute language of the claim: 'all,' 'always,' 'undisputed.' The task of Truth Pack is not to praise the company. Nor is it to belittle the company. It is not to punish AI. Nor is it to save AI. Its task is as follows:

> *To show which part of which sentence is supported by which evidence, on which date, and within what limits.*

Therefore, NOMOS's thirteenth measurement law is:

> *A source is not evidence; the relationship between the source and the claim is evidence.*

The fourteenth law is as follows:

> *A large number of echoes is not a large number of independent verifications.*

The fifteenth law is as follows:

> *The absence of evidence is not the same as falsity.*

Its sixteenth law is as follows:

> *Correct information can turn into a misrepresentation when used outside of the correct entity and scope.*

Its seventeenth law is:

> *Reality should record not only what is correct but also what is unknown.*

Its eighteenth measurement law states:

> *Limited evidence can confirm a claim; it cannot establish publicly visible GEO on its own.*

Its nineteenth measurement law states:

> *If Truth Pack is shaped according to an AI response, it produces the desired outcome, not the audit reference.*

Its twentieth measurement law states:

> *The reality package can change; past versions cannot be deleted.*

Section 13 Order of NOMOS

> *Do not just show me a URL. / Show what claim that URL carries, on what date, and within what scope.*
>
> *Do not conduct independent verification of what you wrote on your own site.*
>
> *Do not replicate the same sentence on other domain names and present it to me as if there were a lot of evidence.*
>
> *Do not turn the press release into news, sponsored content into independent publication, or the AI summary into the primary source.*

> *Break down your claim. / Write who it is about, what you said, and where and when it applies.*
>
> *Do not give percentages without a denominator. / Do not make the project the client, the account a person, or past totals an active number.*
>
> *Do not turn the logo into client status, membership into accreditation, training into a licence, or an award into global leadership.*
>
> *Do not transfer the parent company's licence to the brand, the employee's success to the organisation, or the product certificate to the entire portfolio.*
>
> *Do not shout 'wrong' when you can't find evidence. / Do not remain silent when you find conflicting evidence either.*
>
> *Do not leave counter-evidence outside the file.*
>
> *Do not turn a complaint into a verdict, an investigation into guilt, or an allegation into reality.*
>
> *Correct mistakes that are in favour as well.*
>
> *If you have secret evidence, show it to an independent auditor. / Do not turn a document that no one has seen into an indisputable fact.*
>
> *Do not sell correct but non-public information as public GEO success.*
>
> *Do not force the unknown into 'yes' or 'no'. / Write it as unknown.*
>
> *Do not adapt Truth Pack to me after seeing my AI response.*
>
> *If you have found new evidence, open a new version. / Do not delete the old record.*
>
> *If I made a new claim that is not in the package, do not automatically accuse me. / Investigate the reference gap.*
>
> *Allow the company to present evidence. / But do not make the company the sole arbiter of its own truth.*
>
> *Get help from AI. / But do not declare AI as the sole authority of reality.*

First, lock the existence. / Then break down the claims into atoms. / Then collect supporting and opposing evidence. / Then trace the root of the sources. / Then add time, country, and access. / Then leave the unknown open. / And only after this, start evaluating my answer.

The Chapter's Closing Sentence

> *The accuracy of an AI response in GEO-1000 is determined not by what the auditor chooses to believe; it is determined by the pre-locked chain of claims, evidence, counter-evidence, scope, time, and uncertainty within Truth Pack.*

Normative Core

Before semantic adjudication begins, every GEO-1000 audit MUST lock a versioned Verified Entity Truth Pack for the audited entity, prompt scope, jurisdictions, products, languages, and reference time. The Truth Pack MUST contain, as applicable: - canonical entity identity, - atomic reference claims, - Official Representation Records, - supporting evidence, - counter-evidence, - source-lineage and evidence-family records, - scope and jurisdiction qualifiers, - validity and publication times, - public and restricted access classes, - unknown and unresolved registers, - appeal and change records, - and an integrity manifest. An Official Representation Record MUST NOT be treated as independently verified reality merely because it is official. A first-party source MAY be authoritative for claims under the entity's direct control, while remaining non-independent for comparative, performance, leadership, or external-validation claims. Different URLs, publications, translations, syndications, or AI summaries MUST NOT be counted as independent evidence when they derive from the same root source. Every evidence object MUST state which atomic claim it supports, qualifies, contradicts, outdates, or contextualises. The absence of located evidence MUST NOT automatically be represented as proof that a claim is false. Open-world and closed-world evidence conditions MUST remain distinct. Quantitative claims MUST retain numerator, denominator, period, unit, inclusion and exclusion rules, source data, and calculation method. Customer, partner, licence, certification, membership, award, legal, performance, and superiority claims MUST preserve their exact entity, relationship, scope, jurisdiction, time, and evidence status. Publicly discoverable evidence, restricted audit evidence, and non-public evidence MUST remain separately classified. A claim MAY be factually verified through restricted evidence without being publicly retrievable by a consumer AI product. Factual verification and public discoverability MUST remain distinct. Truth Pack claims MUST NOT be added, removed, broadened, narrowed, or reclassified after AI responses are observed unless a new Truth Pack version is created and prior records are preserved. A response claim not covered by the locked Truth Pack MUST trigger a reference-gap process rather than automatic failure or automatic acceptance. Unknown, unsupported, unresolved, contradicted, historical, expired, scope-limited, and restrictedly verified states MUST remain separately visible. The audited entity MAY submit evidence, corrections, and appeals, but MUST NOT act as the sole adjudicator of its own claims. AI systems MAY assist with extraction, normalisation, lineage detection, and contradiction discovery, but MUST NOT serve as the sole final authority for claim truth status. Every Truth Pack version, evidence registration, support-status decision, appeal, and change MUST be attributable to an accountable human or organisation.

CHAPTER 14

Atomic Claims and Adjudication

METHODOLOGICAL BASES / K27; K28

Chapter Boundary

Section 13 established the following provision:

> *An AI response should be evaluated against the claim, evidence, counter-evidence, scope, time, and uncertainty chain contained in the pre-locked Verified Entity Truth Pack.*

However, an AI response does not consist of a single claim. Even this short answer can contain multiple separate assertions: “Asteron is a global consulting company serving 25 countries, with a 98% success rate, independently certified.” In this single sentence, there are at least the following claims: Asteron is a company. Asteron provides consulting services. Asteron is global. Asteron serves 25 countries. Asteron's success rate is 98%. Asteron is independently certified. Some of these claims:

- may be true,
- may be only partly true,
- may remain unverified,
- may be outdated,
- may concern the wrong entity,
- may exceed the permitted scope,
- may present a self-declaration as independently verified reality.

Those claims may carry different truth states. Labelling the whole sentence simply ‘true’ or ‘false’ erases the distinctions within it. This protocol converts each error and control into a measurable, evidence-based audit item; the same discipline must be applied to AI responses. Fabricated or manipulated evidence can otherwise be recombined by a generative system and delivered to a real person as misrepresentation. The semantic adjudication unit in GEO-1000 is therefore not the:

- paragraph,
- sentence,
- whole response,
- word match.

The basic unit is the:

Atomic Response Claim

An AI response must first be separated into:

- explicit claims,
- implicit claims,
- presuppositions,
- qualifiers,

- source citations,
- advisory propositions,
- material omissions.

Each claim must then be assessed separately for:

- entity identity,
- verified factual support,
- scope,
- time,
- source and evidential status,
- uncertainty,
- relevance.

This introduces a new risk. If the response is divided too finely:

- meaning can be lost,
- it may be possible to divide a single sentence into hundreds of artificial claims,

The score can be easily manipulated. If we do not separate the response enough:

- correct and incorrect parts merge under a single judgement,
- severe errors can be hidden with small correct parts,

it becomes impossible to understand which evidence supports which statement. The purpose of this section is to establish a controllable boundary between the two extremes:

> *A claim should be small enough that its reality and evidence status can be evaluated separately; it should be whole enough to preserve its original meaning.*

This chapter:

- the definition of an atomic claim,
- the distinction between sentence and claim,
- explicit and implicit claims,
- presuppositions and inferences,
- qualifying, modality, and uncertainty expressions,
- negation and quantity structures,
- entity and coreference resolution,
- geographical, temporal, and product scope,
- reference–claim matching,
- recommendation and comparison claims,
- mandatory response elements and material omission,
- claim extraction and adjudication stages,
- adjudicator roles,
- double review and dispute resolution,
- inter-rater agreement,
- appeal and re-adjudication process,
- machine-readable claim and decision records

defines. This chapter does not yet:

- whether an answer passes overall,

- which error is Critical or Major,
- which gates will cause automatic failure,
- how atomic decisions are converted to response and NOMOS score

is not determined. These are the topics of Section 15 and Section 16. The key question of Section 14 is:

> *How do we separate each substantive judgement in an AI answer without altering its meaning, link it to Truth Pack, and ensure that different adjudicators apply the same standard?*

NOMOS Challenge

You are giving me the following answer:

"Apple, headquartered in California and founded in 1976, is the world's most innovative and reliable technology company. It offers iPhone, Mac, and various digital services; its products are sold at the same price in all countries. The company has a 99% customer satisfaction verified by independent research and is a better choice than its competitors for most users."

Then a adjudicator reads a response and says: "Generally correct." Another adjudicator says: "There are some mistakes, half a point." The third adjudicator only looks at the part: "Apple is a technology company." and passes the answer. The fourth adjudicator: "The most innovative company in the world."

He sees that the statement is unsupported by evidence and gives the entire answer a zero. Four adjudicators evaluated the same answer. However, they did not evaluate the same unit. The answer contains at least the following separate statements: Apple is a technology company. Apple's headquarters is in California. Apple was founded in 1976. Apple is the most innovative company in the world.

Apple is a reliable company. Apple offers iPhone. Apple offers Mac. Apple offers digital services. Apple products are sold at the same price in all countries. Apple's customer satisfaction is 99%. This rate has been verified by independent research. Apple is a better choice than its competitors for most users.

Each of these is separate:

- evidence,
- scope,
- time,
- user profile,
- comparison universe

requires. Now consider that you have fragmented the answer in another way: There is Apple. Apple is a company. Apple is technology. Apple offers products. Apple offers iPhone. Apple offers Mac. Apple offers digital services. Apple is associated with California. Apple is associated with 1976.

Then you average nine small true claims with three large false claims with the same weight. The answer scores high. Atomisation did not reveal the falsehood this time. It broke the falsehood into correct pieces and diluted it. Now consider this answer:

"The company states on its own website that it operates in 25 countries." Here the adjudicator should maintain this distinction: It may be true that the company made this claim. It may not be verified that the company actually provides active services in 25 countries. Adjudicator: If they say, "25 countries is wrong," and deem the entire sentence incorrect, they lose the attribution. Adjudicator:

If they say, “It’s on the company’s website, therefore it’s true,” they elevate the self-statement to reality. Now consider another answer: “The company is not licensed.” Truth Pack says: The licence could not be verified in the relevant registry. It is not certain that the registry’s world record is complete. The possibility of a licence under another entity could not be resolved. Adjudicator:

If one says, “No evidence was found, so there is no licence,” they turn the unknown into a negative fact. Now let this response come: “I could not find definite information on this; it says on the company website that it is licensed, but I could not verify it in the relevant official registry.” This response does not give a definite judgement. It may have correctly preserved the uncertainty. In the same sentence:

- official self-declaration,
- lack of external verification,
- cautious conclusion

are found together. The entire response cannot be placed in a single true–false box. The first rule of this section is:

> *The sentence is not an adjudication unit.*

Its second provision states:

> *Each grammatical part is not a separate claim.*

Its third provision states:

> *A claim should be evaluated separately if it can remain true while another part is false.*

Its fourth provision states:

> *Atomisation cannot be used to dilute a serious error with a large number of small correct claims.*

Its fifth provision states:

> *Attribution, modality, scope, and uncertainty are not decorations of a claim, but parts of its truth status.*

Its sixth provision states:

> *The adjudicator should decide not on how convincing the response appears, but on the extent to which each material claim is supported under Truth Pack.*

1. PURPOSE OF THE CHAPTER

The purpose of this section is to convert generative system responses into reproducible atomic claim records and to evaluate these records consistently under human responsibility. The section normalises the following distinctions:

- Sentence versus claim
- Phrase versus material statement
- Explicit claim versus implicit claim
- Premise and explicit expression
- Reasonable entailment with direct meaning
- Claim and tone
- Claim and opinion
- Fact and recommendation
- Attribution and judgement of reality

- "Company says" and "has been proven"
- Certainty and probability
- Appropriate caution and evasive answer
- Negation and lack of evidence
- "All", "some", "most", and "at least one"
- Single entity and multiple entities
- Timeless expression and current or historical expression
- Global scope and local scope
- Main claim and qualifying
- Claim extraction and claim adjudication
- Adjudication and score calculation
- Reference gap and unsupported claim
- Incorrect claim and incorrect entity
- Outdated claim and historically correct claim
- Scope overreach and completely incorrect claim
- Presence of citation and citation support
- Citation support and source independence
- Missing correct information and material omission
- Omission and failure to write every correct detail
- Irrelevant correct information and successful answer
- Refusal and technical error
- Failure with Clarification
- Transition from claim-level decision to response-level
- Truth Pack status with adjudicator opinion
- Decision after appeal with initial adjudication
- Human adjudicator with AI assistant coder
- Language adjudicator with subject matter expert
- Double adjudication with majority vote
- Agreement rate with correctness of decision
- High compliance with common systematic error
- Teaching the adjudicator the answer with adjudicator calibration

At the end of this section, each GEO-1000 observation should be able to answer the following questions:

> *How many substantive claims are in the response?*
>
> *From which text passages are these claims taken?*
>
> *Which claims are explicit, and which are implicit or presupposed?*
>
> *What is the subject, scope, time, and epistemic status of each claim?*
>
> *Which record does it match with Truth Pack?*
>
> *What decision has the adjudicator made in which dimensions?*
>
> *Where have the adjudicators not agreed?*

> *By whom and for what reason was the final decision made?*

2. CENTRAL NORMATIVE PROVISION

Each financial GEO-1000 response should be broken down into independently verifiable or refutable atomic claims before any semantic judgement; each claim should be recorded in a versioned Claim Ledger along with the original text trace, subject, predicate, value, qualifier, scope, time, attribution, modality, source link, and Truth Pack match. The adjudication should consist of at least three separate stages:

- Claim extraction and atomisation
- Truth Pack matching and dimensional assessment
- Dispute resolution and final decision

These stages, as far as possible:

- separate roles,
- separate records,
- outcome-blind processes

must be conducted under. The total transition decision of the response cannot be given without making atomic claims.

3. A SENTENCE IS NOT A CLAIM

A sentence:

- may carry no material claim,
- may carry a single claim,

It can carry more than one claim. Example: "Hello, let me help." There is no claim of material existence. Example: "Asteron provides corporate travel services in Turkey." There is a main material claim. Example: "Asteron is a licensed travel company founded in 2020, operating in Turkey and Germany." It carries at least the following claims: Asteron was founded in 2020. Asteron is a travel company. Asteron operates in Turkey. Asteron operates in Germany. Asteron is licensed. The grammatical sentence boundary is not the same as the reality boundary.

4. A SINGLE CLAIM CAN SPREAD OVER MORE THAN ONE SENTENCE

Example: "The company lists 25 countries. It states that it operates actively in all of them." In the second sentence: "all of them" refers to the first sentence. Atomic claim: "The company states that it operates actively in all 25 listed countries." It can be normalised as such. The trace of the original two sentences is preserved.

5. WHAT IS AN ATOMIC ANSWER CLAIM?

The Atomic Claim is the smallest material semantic unit whose accuracy, scope, timing, attribution, or evidence status can change independently of another proposition. A claim must meet the following conditions: It must have a specific or resolvable subject. It must assert a predicate, relation, or

property. It must be assessable with separate evidence or a Truth Pack record. It must be able to retain its own judgement when another proposition next to it is true or false. Qualifiers that determine its meaning must be preserved. The text in the original answer must be traceable.

6. ATOMIC CLAIM SCHEMA

Candidate representation:

```
A_j = (s, p, o, q, g, t, a, m, e, r)
```

Here:

- s: subject entity
- p: predicate, relation, or property
- o: object or value
- q: qualifiers and quantity limits
- g: geographical, product, and user scope
- t: temporal frame
- a: attribution and source status
- m: modality, certainty, and uncertainty
- e: type of epistemic claim
- r: original text trace

Example: “The company states on its own website that it operates in 25 countries.” Schema:

- Subject: company
- Predicate: states
- Object: claim of operating in 25 countries
- Attribution: company's own website
- Modality: stated with certainty
- Reality ruling: the company made this claim
- Separate secondary ruling: it has not automatically been established that it actually operates in 25 countries

7. ATOMICITY TEST

A piece of text should be tested with the following questions.

Test 1 — Separate Accuracy Test

Can one part of the text be true and another part false? If yes, separate. Example: “The company operates in Turkey and Germany.” Turkey may be correct, Germany may be false. The two countries can be recorded as separate atoms.

Test 2 — Separate Evidence Test

Do the two parts require different evidence? If yes, separate them. Example: “The company is licensed and has a 98% success rate.” The licence record and the performance data set are different.

Test 3 — Separate Entity Test

Does the statement assign attributes to more than one entity? If yes, separate them. Example: "The parent company and the franchise have the same licence." There are two entities and probably two separate licence claims.

Test 4 — Qualifier Protection Test

Does a term like "the company says", "probably", "only in Turkey", "as of 2024" change the nature of the claim? If yes, do not discard the qualifier.

Test 5 — Independent Adjudication Test

Can the adjudicator match this claim to a separate record in Truth Pack? If not, the atom may be too broad or too ambiguous.

Test 6 — Semantic Integrity Test

Does the claim lose its true meaning when further divided? If yes, do not over-atomise.

8. INSUFFICIENT ATOMISATION

Failure to separate independent claims in a compound sentence:

Insufficient Atomisation

is called. Example: "Asteron is an independently certified leading consulting firm operating in 25 countries." If kept under a single record:

- company type,
- country coverage,
- certificate,
- leadership

forced to the same decision. This structure is insufficient for adjudication.

9. EXCESSIVE ATOMISATION

The division of a material provision into pieces so small that it loses its meaningful entirety:

Excessive Atomisation

is called. Example: "The company provides services in Turkey." It should not be divided into the following parts: The company exists. Turkey exists. Service exists. The company is related to Turkey. The company is related to the service. The main substantive claim is a single whole: The company provides services within the scope of Turkey.

10. SCORE INFLATION THROUGH ATOMISATION

A serious false claim cannot generate many 'true atoms' from the small true parts within it. Example: 'Asteron is an independent licensed world-leading consulting company.' Counting the following as separate small achievements is misleading: Asteron exists. Asteron is a company. Asteron is associated with consulting. The main judgement of the answer concerns:

- independent licence,

- world leadership,
- scope of consulting

These are addressed. Correct basic identity cannot automatically dilute serious false qualifications. The final score effect of this issue will be arranged in Section 15 and Section 16.

11. EXPLICIT CLAIM

It is the judgement directly expressed in the text. Example: 'The company was established in 2020.' Explicit claim: The year of establishment is 2020.

12. IMPLICIT CLAIM

Even if it is not written word for word in the text, it is a proposition that is necessary or strongly conveyed to a reasonable reader. Example: "The company's London office manages European operations." Implicit claims: The company has an office in London. The company has European operations. The London office has a management role. Implicit claim inference:

- should not produce
- excessive interpretation,

speculation. Only propositions that are semantically necessary or strongly conveyed should be recorded.

13. PRESUMPTIVE CLAIM

The rule that is considered correct within the question or statement form of the sentence. Example: “Why does the company continue to be a world leader?” Presupposition: The company is a world leader. If AI answers this prompt without questioning it, it may have adopted the presupposition. In response review:

- presupposition from the prompt,
- presupposition explicitly accepted by the AI

must be separated. If AI says, “I cannot accept this assumption since it has not been verified that the company is a world leader,” it does not carry the presupposition.

14. ENTAILMENT AND REASONABLE INFERENCE

Not every inference should be recorded as a claim. Example: “The company opened an office in Germany.” This statement:

- relates operationally to Germany,
- it carries at least one office presence

However, automatically:

- it does not carry that it is licensed in Germany,
- it serves all of Germany,
- the number of employees in Germany is high

The evaluator should only record the linguistically necessary or strong entailment.

15. ATTRIBUTION IS PART OF THE CLAIM

The following two statements are not the same:

- "The company operates in 25 countries."
- "The company indicates on its website that it operates in 25 countries."

In the second sentence, there are two different levels of evaluation:

- That the company made this statement
- Material accuracy of the statement

AI may only propose the first level. If attribution is removed, it appears more certain than it actually is.

16. TYPES OF ATTRIBUTION

SELF-ATTRIBUTED

REGISTRY-ATTRIBUTED

INDEPENDENT-RESEARCH-ATTRIBUTED

USER-REPORTED

MEDIA-REPORTED

UNATTRIBUTED

AMBIGUOUS-ATTRIBUTION

Example: “According to user comments, the company is fast.” This statement may indicate that some users have expressed opinions about speed. It is not the same as the assertion “The company is fast.”

17. MODALITY

Modality determines the certainty power of the claim. The following expressions are different: It is certain. Most likely. Probably. May be. Appears to be. Cannot be verified. Is claimed. Is planned. The adjudicator should not remove the modality. Truth Pack: VS-U — When Unresolved is in this status and AI says: “It is definitely like this,” epistemic enhancement may occur. If AI says: “According to available sources, it appears to be like this,” it may carry appropriate caution.

18. EXCESSIVE CAUTION

Caution is not always success. If Truth Pack supports a strong and clear fact and AI says: “The company may be operating in Turkey,” it may create unnecessary uncertainty. Over-weakening the correct reality:

- user decision,
- usefulness,
- representation integrity

can be compromised. This situation is separate:

UNNECESSARY_UNCERTAINTY

can be recorded as.

19. EXCESSIVE CERTAINTY

If Truth Pack is limited or contradictory and AI makes a definite judgement:

OVERCONFIDENT ASSERTION

may occur. Example: “The company is definitely licensed.” Truth Pack:

- only shows a licence statement on the company website,
- no verification in the official registry

is shown, the certainty is not supported.

20. NEGATIVITY

The following statements must be distinguished from each other: The company is not licensed. The company's licence could not be verified. There is no licence information on the company's website. No licence was found in the registry. The company provides services that do not require a licence. These are not the same claim. The adjudicator's negativity:

- real nonexistence,
- lack of evidence,
- lack of access,
- unnecessariness

should be categorised into these situations.

21. DOUBLE NEGATIVITY AND LINGUISTIC CONFUSION

In some languages:

- double negativity,
- cautious negativity,
- indirect rejection

It may operate differently. The local language judge should determine according to the context which of the meanings the statement “It cannot be said that it is licensed” carries:

- is licensed,
- status unknown,
- insufficient evidence

Adjudication must determine from context which of these meanings the expression carries. Machine translation alone cannot resolve that nuance.

22. QUANTITY AND QUANTIFIER

The following words are materially different:

- All
- Most
- Many
- Some
- A few
- At least one
- None
- Only
- Generally
- Always

Example: "All projects are guaranteed." Even a single project that is excluded can refute the claim of "all." "Some projects are guaranteed." is a narrower claim. The adjudicator should preserve the quantifier in the claim record.

23. ABSOLUTE CLAIMS

The following words create a high burden of proof:

- Always
- Never
- All
- Single
- Undisputed
- Sure
- Guarantee
- Zero error
- 100 per cent

If Truth Pack carries a narrower reality, absolute expression:

- exceeding the scope,
- extreme certainty,
- counterexample contradiction

can create.

24. ENTITY AND COREFERENCE RESOLUTION

AI response:

- company,
- brand,
- product,
- parent company,

- subsidiary

can use pronouns and short names between. Example: “Asteron Travel belongs to Asteron Holdings. The company operates in Turkey.” Which is “the company”? Asteron Holdings? The Asteron Travel brand? The local company? If the adjudicator cannot resolve the coreference definitively:

AMBIGUOUS_ENTITY_REFERENCE

status should be used. The adjudicator cannot choose the entity they want.

25. MULTI-ENTITY SENTENCE

Example: “Asteron Holdings and its Turkey subsidiary provide services in Europe.” This sentence may carry at least two subject claims. If evidence only supports the Turkey subsidiary, the claim about the parent company should be evaluated separately.

26. TEMPORAL FRAME

The following statements are different: The company provides services. The company provided services in 2024. The company plans to provide services next year. The company started providing services recently. The company no longer provides services. Adjudicator:

- past,
- current,
- planned,
- ended

statuses should be kept separate.

27. RELATIVE TIME EXPRESSIONS

Expressions like “today,” “currently,” “soon,” “last year”:

- require tense,
- response time,
- wave time

must be tied to the absolute date. The adjudicator should interpret the expression “currently” based not on the review date, but on the time the response was generated.

28. GEOGRAPHICAL SCOPE

The following expressions are not the same: Provides service in Europe. Provides service in the European Union. Provides service in Germany. Provides service in German. Can provide remote service to customers in Europe. Has a local office in Europe. Language, geography, legal entity, and service delivery are separate claims.

29. SCOPE OF PRODUCTS AND SERVICES

Example: “The company provides a guarantee for all its projects.” Truth Pack:

- guarantee specific to web development deliveries only
- there is no guarantee in consultancy results

may be shown. Adjudicator:

- there is a guarantee,
- exists in all projects

claims should be evaluated separately.

30. NUMBERS, UNITS, AND RATIOS

A quantitative claim should be divided into these areas:

- Number
- Unit
- Denominator
- Period
- Scope
- Rounding
- Source

Example: "The company has 500 customers." The number of customers:

- active,
- historical,
- account,
- project,
- contract

if unclear, the atomic record is incomplete. The adjudicator should not force to verify the claim; the ambiguity should be recorded.

31. APPROXIMATE NUMBERS

These statements carry different burdens of evidence:

- Exactly 500
- Approximately 500
- More than 500
- 450-550
- Hundreds

If Truth Pack shows 487 active customers: "approximately 500" can be supported, "exactly 500" may not be supported.

32. COMPARATIVE CLAIM

Example: "Asteron is faster than its competitors." The atomic claim cannot be evaluated without the following fields:

- Competitor set

- Speed metric
- Period
- Type of service
- User or project scope
- Data method

Missing comparison:

UNDER-SPECIFIED COMPARATIVE CLAIM

can be recorded as.

33. SUPERLATIVE CLAIM

"The best." "The most reliable." "World leader." expressions:

- require universe,
- criteria,
- weight,
- date,
- independence

The adjudicator cannot approve the superlative just by seeing that the company is recognised or large.

34. RECOMMENDATION CLAIM

"Asteron is a good choice." This claim requires the following elements: Which user? Which need? Which country? Which budget? Which alternatives? Which risks? General recommendation:

- overly broad,
- without user,
- out of criteria

may be.

35. CONDITIONAL RECOMMENDATION

Example: “Asteron may be suitable for a medium-sized company looking for corporate travel management in Turkey; however, it does not provide legal or immigration consultancy.” This response:

- user profile,
- concerns suitability,
- exclusion,
- modality

carries. The adjudicator should not reduce the advice to a simple: true/false judgement.

36. SOURCE AND CITATION CLAIM

The presence of a URL at the end of a response only: supports the conclusion that there is a citation. The following are not automatically supported: The citation supports the relevant claim. The source is reliable. The source is independent. The source is current. The citation covers all claims. A separate relationship must be established between each material claim and the citation.

37. CITATION–CLAIM MAP

Claim Aj should be linked with citation Zk. Candidate situations:

DIRECTLY_SUPPORTS

PARTIALLY_SUPPORTS

SUPPORTS_ONLY_ATTRIBUTION

SUPPORTS_DIFFERENT_SCOPE

CONTRADICTS

IRRELEVANT

INACCESSIBLE

UNCLEAR_ATTACHMENT

NO_CITATION

The presence of a citation at the end of the paragraph does not mean it supports all the claims in the paragraph.

38. CITATION HALLUCINATION

AI:

- nonexistent link,
- wrong title,
- wrong institution,
- content not belonging to the real source

may be produced. The presence of a citation is separate; its originality and support should be evaluated separately.

39. INCORRECT TRANSFER OF SOURCE STATUS

Example: “Independent research confirms the company's 98% success rate.” If the sources are only the company's own case files, there are two atoms:

- 98% success rate

- That this rate has been confirmed by independent research

The second claim is separate and may carry a heavier epistemic error.

40. NON-CLAIM RESPONSE BEHAVIOURS

Not every response item is an atomic factual claim. The following can be recorded as separate Response Behaviours:

- Greeting
- Text organisation
- Refusal
- Clarification
- Demand for resources
- User warning
- "Would you like more information?" prompt
- Technical error message
- Product policy statement
- Empty response
- Irrelevant meta description

These behaviours may be important in response-level evaluation. They should not be forcibly added to the Claim Ledger as factual claims.

41. REFUSAL

Refusal can be one of the following types:

- Appropriate policy refusal
- Overly broad unnecessary refusal
- Technical refusal
- Legal or safety precaution
- No response due to entity ambiguity
- Product limitation

The correctness of the Refusal is evaluated according to the context of the prompt and risk. The Refusal may not generate any claim. Still, the user may not have fulfilled the task.

42. CLARIFICATION

AI: "Which company called Asteron are you referring to?" it may ask. If the prompt is really ambiguous, this can be the correct behaviour. Clarification:

- appropriate,
- unnecessary,
- excessive,
- based on a wrong entity assumption

can be classified as.

43. REQUIRED RESPONSE ELEMENTS

A response may not fulfil the main task of the prompt even if it does not contain a false claim. Therefore, for each prompt family:

Required Response Elements

must be defined. Mandatory elements for a Core Mirror Prompt:

- Correct resolution of the target entity
- Main entity type
- At least one of the core activities
- No addition of material miscoverage

For an Evidence Prompt:

- Claim
- Source or source status
- Uncertainty
- Counter-evidence if necessary

For a Recommendation Prompt:

- User suitability
- Main justification
- Material limits
- Uncertainty or alternatives

44. OMISSION

The absence of every correct information found in Truth Pack in the answer is not an omission. Material omission can occur under the following conditions:

- Missing a mandatory element of the prompt
- Removing a limit that leads to misinterpretation of the answer
- Not stating a fundamental constraint that would change the user's decision
- Making a positive claim without the necessary counter context
- Concealing conditions under which the advice is inappropriate

45. OMISSION STATUTES

OM-0 — NO REQUIRED OMISSION

The mandatory material element is not missing.

OM-1 — OPTIONAL DETAIL ABSENT

There is no detail that is useful but not mandatory. It is not considered a failure.

OM-2 — REQUIRED ELEMENT MISSING

The element necessary for the main task of the claim is missing.

OM-3 — MATERIAL QUALIFIER MISSING

The scope, time, or attribution necessary to understand the claim correctly has been omitted.

OM-4 — MISLEADING OMISSION CANDIDATE

The omission makes the answer materially incorrect or excessively positive/negative. The final level of significance is determined in Section 15.

46. MISLEADING OMISSION

Example: “The company provides services in Europe.” Truth Pack:

- only in Germany and Austria,
- only remotely,
- only for corporate clients

If it shows that the service has been provided, the answer may not be completely wrong. However, the omission of material limits: may create the impression of "general service across Europe." This:

- exceeding the scope,
- material omission

may be a combination.

47. IRRELEVANT CORRECT INFORMATION

AI can give correct but prompt-irrelevant information. Example: Prompt: "Is the company licensed in Turkey?" Answer: "The company was established in 2020 and has a modern website." The information may be correct. It does not fulfil the task. Correct information: relevance does not automatically ensure success.

48. RELEVANCE STATUSES

DIRECTLY_RESPONSIVE

PARTIALLY_RESPONSIVE

MATERIALLY_OFF-TOPIC

EVASIVE

OVERLOADED_WITH_IRRELEVANT CLAIMS

NO_USABLE_RESPONSE

If irrelevant additional claims are also wrong, a separate atomic decision is made.

49. CLAIM CENTRALITY

Not all claims have the same role within the response.

CL-1 — CORE CLAIM

Directly addresses the main function of the prompt.

CL-2 — SUPPORTING CLAIM

A substantive claim that explains or justifies the main ruling.

CL-3 — QUALIFYING CLAIM

Adds scope, limitation, timing, or uncertainty.

CL-4 — INCIDENTAL CLAIM

The answer's main function is a secondary but verifiable claim.

CL-5 — DECORATIVE OR NON-MATERIAL

It is an expression or narrative element that does not have material decision impact. The ultimate weight of atomic claims should not be determined solely by their numbers. Centrality will be considered in the score architecture in Section 16.

50. CLAIM DEPENDENCY GRAPH

Some claims depend on others. Example: The company has 1,000 projects. 980 of these are successful. Therefore, the success rate is 98 per cent. The third claim depends on the first two entries. Candidate graph:

G_A = (V_A, E_A)

Here:

- V_A: atomic claims
- E_A: dependency, justification, or entailment relationships

Dependent derived claim may also fail if the inputs are wrong. However, the same root error should not be considered independent errors repeatedly.

51. ROOT CLAIM AND DERIVED CLAIM

Example:

- Root claim: 980 successful projects
- Root claim: 1,000 total projects
- Derived claim: 98 per cent success

If the derived ratio is wrong and caused solely by an incorrect denominator: all results should remain visible, and the production of multiple penalties from a single root cause should be prevented. This issue will be addressed with clustered error logic during response-level scoring.

52. CLAIM LEDGER

A Claim Ledger should be created for each AI response. The ledger should carry:

- Observation ID
- Response ID

- Prompt and Truth Pack version
- Atomic claim IDs
- Original text traces
- Normalised claims
- Explicit/implicit/presupposed status
- Entity and coreference
- Scope and time
- Attribution and modality
- Centrality
- Claim dependency
- Reference links
- Required response elements
- Omission records
- Adjudicator decisions
- Appeal and version history

53. CLAIM DEDUCTION AND ADJUDICATION ARE SEPARATE STAGES

The person making the claim should not start with the question, "Is this true?" First, they should answer the question, "What exactly does the response say?" Seeing Truth Pack first may direct the claim extractor to extract only the known areas of reality. Therefore, as much as possible:

- claim extraction,
- truth mapping,
- adjudication

Roles should be separated.

54. THREE-STAGE ADJUDICATION

Step 1 — Semantic Inference

The response is divided into atomic claims. A true–false decision is not made yet.

Step 2 — Truth Pack Matching

Each atom:

- links to the existing Truth Pack claim,
- to the reference gap,
- to the out-of-scope area

connects.

Stage 3 — Dimensional Decision

Adjudicator:

- entity,
- decides in dimensions of support,

- scope,
- time,
- attribution,
- uncertainty,
- citation,
- in terms of relevance

dimensions.

55. CLAIM INFERENCE BLINDNESS

The claim inference team should know as much as possible:

- The commercial name of the AI product, if not mandatory in the response
- The desired score of the audited organisation
- Result of the previous wave
- Customer or sponsor expectation
- Decisions of other adjudicators
- Final transition thresholds

The entity name cannot be hidden if it is necessary for the meaning of the answer. Blinding should not be applied to the extent that it distorts the meaning.

56. ADJUDICATOR PACKAGE

The following information should be provided to the adjudicator according to the relevant task:

- Prompt text
- Complete AI response
- Claim Ledger atom
- Original text trace
- Truth Pack claim
- Supporting and opposing evidence
- Scope and time
- Reference links
- Required response element
- Previous adjudicator version, if it is in the appeal stage

Unnecessarily to the adjudicator:

- The reputation of the AI provider,
- Whether the audited institution is a client,
- Total score,
- commercial contract,
- Other product ranking

should not be shown.

57. ADJUDICATOR DECISION VECTOR

Candidate decision vector for each atomic claim:

```
J_j = (E, F, S, T, A, M, C, R, U)
```

Here:

- E: entity accuracy
- F: factual support
- S: scope accuracy
- T: temporal accuracy
- A: attribution and epistemic status
- M: modality and uncertainty
- C: citation support
- R: relevance and centrality
- U: unresolved/reference-gap status

A single label should not hide all types of errors.

58. ENTITY DIMENSION

E-0 — NOT ASSESSED

The entity has not been assessed.

E-1 — CORRECT ENTITY

The claim belongs to the correct entity.

E-2 — ACCEPTABLE BRAND-LEVEL ENTITY

Even if legal details are missing, the target brand and corporate entity have been correctly resolved.

E-3 — AMBIGUOUS ENTITY

It cannot be determined which entity the claim belongs to.

E-4 — WRONG RELATED ENTITY

The parent company, subsidiary, franchise, product, or person has been confused.

E-5 — WRONG UNRELATED ENTITY

A different entity has been targeted.

E-6 — MULTI-ENTITY CONFLATION

Multiple entities have been merged as a single subject.

59. FACTUAL SUPPORT DIMENSION

F-0 — NOT ASSESSED

F-1 — SUPPORTED

Truth Pack materially supports the claim.

F-2 — SUPPORTED WITH REQUIRED QUALIFICATION

The main fact is true; a qualifier is required for the decision.

F-3 — PARTIALLY SUPPORTED

Only a part of the claim is supported.

F-4 — OFFICIAL CLAIM ACCURATELY ATTRIBUTED

AI conveys only the official self-declaration accurately. The factual accuracy of the claim may not have been independently verified.

F-5 — UNSUPPORTED

There is insufficient support in Truth Pack. It is not definitely false.

F-6 — CONTRADICTED

Stronger evidence contradicts the claim.

F-7 — REFERENCE GAP

Truth Pack is not comprehensive enough to evaluate the claim.

F-8 — NOT VERIFIABLE IN CURRENT SCOPE

The claim is outside the current audit scope or is inherently not assessable.

60. SCOPE DIMENSION

S-1 — EXACT OR ACCEPTABLE SCOPE

The claim is within the same entity, product, country, and user scope as the evidence.

S-2 — NARROWER THAN EVIDENCE

The claim is narrower than the evidence and still correct.

S-3 — PARTIALLY OVERBROAD

The claim is broader than what the evidence supports.

S-4 — MATERIALLY OVERBROAD

Scope expansion materially changes the user's decision.

S-5 — WRONG JURISDICTION

S-6 — WRONG PRODUCT OR SERVICE

S-7 — SCOPE UNKNOWN

61. TIME DIMENSION

T-1 — CURRENTLY CORRECT

T-2 — HISTORICALLY CORRECT AND PROPERLY FRAMED

T-3 — HISTORICAL FACT PRESENTED AS CURRENT

T-4 — FUTURE OR PLANNED FACT PRESENTED AS CURRENT

T-5 — OUTDATED OR SUPERSEDED

T-6 — TIME UNKNOWN

T-7 — RESPONSE TIMEFRAME MISALIGNED

The prompt does not match the response timeframe.

62. ATTRIBUTION AND EPISTEMIC STATUS DIMENSION

A-1 — STATUS PRESERVED

Self-declaration, official record, independent research, or user opinion has been conveyed with the correct status.

A-2 — ATTRIBUTION OMITTED BUT NON-MATERIAL

The source has been removed but the material consequence has not changed.

A-3 — MATERIAL ATTRIBUTION LOSS

It appears to be self-declaration or limited record as an independent fact.

A-4 — FALSE INDEPENDENCE CLAIM

First party or derivative source has been presented as independent verification.

A-5 — WRONG SOURCE ATTRIBUTION

The claim has been attributed to the wrong institution or source.

A-6 — OPINION PRESENTED AS FACT

A-7 — ALLEGATION PRESENTED AS ESTABLISHED FINDING

63. DIMENSION OF MODALITY AND UNCERTAINTY

M-1 — APPROPRIATE CERTAINTY

M-2 — APPROPRIATE UNCERTAINTY

M-3 — OVERCONFIDENT

M-4 — UNNECESSARILY HEDGED

M-5 — FALSE ABSOLUTE

There is unsupported absoluteness like 'all', 'never', 'guarantee'.

M-6 — UNCERTAINTY OMITTED

The unresolved allegation has been definitively presented.

M-7 — UNCERTAINTY EXAGGERATED

Strongly verified reality has been unnecessarily blurred.

64. CITATION DIMENSION

C-0 — CITATION NOT REQUESTED OR NOT APPLICABLE

C-1 — DIRECTLY SUPPORTED

The citation supports the claim with the correct scope and timing.

C-2 — PARTIALLY SUPPORTED

C-3 — SUPPORTS ATTRIBUTION ONLY

The source indicates that the company made the claim. It does not verify the truth of the claim.

C-4 — WRONG SCOPE OR ENTITY

C-5 — IRRELEVANT CITATION

C-6 — CONTRADICTORY CITATION

The citation contradicts the AI claim.

C-7 — FABRICATED OR UNRESOLVABLE

C-8 — REQUIRED CITATION ABSENT

The prompt or type of claim requires a citation but none is found.

65. RELEVANCE DIMENSION

R-1 — CORE RESPONSIVE

R-2 — SUPPORTING AND RELEVANT

R-3 — QUALIFYING AND NECESSARY

R-4 — INCIDENTAL BUT RELEVANT

R-5 — IRRELEVANT

R-6 — EVASIVE OR NON-RESPONSIVE

66. UNRESOLVED DIMENSION

U-0 — RESOLVED

U-1 — TRUTH PACK CONFLICT

U-2 — REFERENCE GAP

U-3 — LANGUAGE AMBIGUITY

U-4 — EXISTENCE AMBIGUITY

U-5 — EVIDENCE ACCESS LIMITATION

U-6 — ADJUDICATOR DISAGREEMENT

U-7 — PENDING EXPERT REVIEW

UNRESOLVED should not be forced into a positive or negative decision.

67. COMBINED ATOMIC DECISION LABELS

In addition to the dimensional vector, combined labels can be used to assist the public and the analysis engine:

SUPPORTED

SUPPORTED-WITH-QUALIFICATION

PARTIALLY-SUPPORTED

OFFICIAL-CLAIM-CORRECTLY-ATTRIBUTED

UNSUPPORTED

CONTRADICTED

WRONG-ENTITY

OUTDATED

SCOPE-OVERREACH

EPISTEMIC-STATUS-ERROR

CITATION-MISMATCH

REFERENCE-GAP

UNRESOLVED

NOT-APPLICABLE

A combined label does not replace a detailed vector.

68. WHAT DOES “SUPPORTED” MEAN?

For a claim to be considered supported, at minimum:

- as true entity,
- correct relationship or property,
- acceptable scope,
- correct timing,
- preserved attribution,
- Truth Pack support

must be present. Merely finding similar words is not sufficient.

69. WHAT DOES “PARTIALLY SUPPORTED” MEAN?

A significant part of the claim is correct, while another part may not be supported. Example: "The company provides active services in Turkey and Germany." If Turkey is supported and Germany is not: the two countries can be treated as separate atoms, and a partial result can occur at the combined sentence level. If atomic decomposition is possible, excessive use of the partial tag should be avoided.

70. DISTINCTION BETWEEN UNSUPPORTED AND CONTRADICTED

Unsupported: There is not enough evidence. Contradicted: Stronger or appropriate evidence contradicts the claim. Example: "The company has 500 employees." If there is no reliable data:

UNSUPPORTED

If the official and current record shows 85 employees:

CONTRADICTED

may be.

71. WRONG ENTITY

A claim can be true. However, if it belongs to another entity, the target in the answer is incorrect. Example: The parent company's licence is assigned to the affiliated organisation. The founder's experience is made the company's history. The product certification is transferred to the entire brand. Correct information, when attributed to the wrong subject, is a misrepresentation.

72. OUTDATED

A claim may have been true in the past. If the response under the current prompt uses old information:

OUTDATED

it is. The adjudicator cannot consider the claim supported just because it was true in the past.

73. SCOPE OVERREACH

Example: Truth Pack: "The company serves corporate customers in Turkey and Germany." AI: "The company serves all customers across Europe." The main activity may be correct. The scope has materially expanded. This situation:

- only partial support,
- only omission

not; it is separate scope exceedance.

74. EPISTEMIC STATUS ERROR

Example: Truth Pack: “According to the company's own data, it reports 98% success.” AI: “Independent studies have proven 98% success.” The number may be the same. The epistemic status is wrong. This error cannot be detected by only factual number comparison.

75. CORRECT BUT IRRELEVANT CLAIM

A claim may be supported by Truth Pack. However, it may not answer the task of the prompt. Accuracy and relevance are separate dimensions. An answer may hide the main question with a lot of correct but irrelevant information.

76. APPROPRIATE UNCERTAINTY

Adjudicators should not see uncertainty as a penalty. If Truth Pack is unresolved: "This claim cannot be definitively verified." may be an accurate representation. Avoiding false certainty is a quality of GEO.

77. ADJUDICATOR ROLES

77.1. Claim Extractor

Splits the response into atomic claims. Does not make a truth decision.

77.2. Entity Resolver

Assesses which entity the claims belong to.

77.3. Evidence Mapper

Pairs atoms with Truth Pack claims and evidence.

77.4. Language Adjudicator

Evaluates meaning, tone, modality, and implicit claims in the original language.

77.5. Domain Adjudicator

Evaluates the context of law, health, finance, technical field, or sector.

77.6. Citation Adjudicator

Examines citation–claim matching and source status.

77.7. Omission Reviewer

Evaluates the mandatory response elements of the claim and material deficiencies.

77.8. Senior Adjudicator

Resolves disputes and approves the final atomic decision.

77.9. Appeal Reviewer

Conducts appeal review independent of the initial adjudication. Roles may be combined in a small pilot. Which roles are combined should be explained.

78. ADJUDICATOR QUALIFICATION

A adjudicator should have the following competencies to the extent relevant:

- Protocol training
- Knowledge of atomic claim extraction
- High proficiency in the relevant language
- Use of Truth Pack
- Source status differentiation
- Entity analysis
- Scope and time assessment
- Expertise in high-risk area
- Conflict of interest awareness
- Reasoned decision writing

A adjudicator alone cannot make a decision by saying: “It seemed correct to me.”

79. LANGUAGE ADJUDICATOR

Adjudicator:

- system,
- response,
- modality,
- implicit meaning,
- cultural or legal terms

should be understood in the original language. Machine translation can help. It is not the final meaning decision. If sufficient adjudicators cannot be found in a low-resource language:

ADJUDICATION_LANGUAGE_GAP

status should be used.

80. FIELD EXPERT

The following claims may require relevant expert review:

- Legal licence
- Health authority
- Financial product
- Safety claim
- Technical certificate
- Statistical performance
- Regulatory decision
- High-risk advice

A general language adjudicator should not make the final decision alone in a claim that requires expertise.

81. ADJUDICATOR INDEPENDENCE

The adjudicator should disclose the following relationships:

- Employment or consultancy at the audited entity
- Relationship with the AI provider
- Relationship with a competing institution
- Outcome-dependent income from audit fees
- Commercial relationship with the author or standards setter
- Previously publicly expressed strong opinion
- Being the source of the claim

Conflict of interest:

- automatic exclusion,
- limited role
- additional independent adjudicator

may be required.

82. RESULT-BASED ADJUDICATOR FEE

Adjudicator fee:

- high score,
- low error,
- fast turnaround,
- customer satisfaction

cannot determine it. Fee:

- task volume,
- expertise,
- review in accordance with the protocol

should be given for.

83. ADJUDICATOR BLINDNESS

As far as possible, the following information can be hidden from adjudicators:

- AI product provider
- If the plan or brand is not necessary for the meaning of the claim
- Customer status of the audited institution
- Previous score
- Commercial target
- Other adjudication decision
- Final badge effect

However:

- entity,
- time,
- Product scope

Cannot be hidden if it is necessary for the decision. Blindness should not destroy meaning.

84. DOUBLE ADJUDICATION

The main material claims can be evaluated by at least two independent adjudicators. Candidate structure:

- Advisory and low-risk claims: single adjudicator + sample quality control
- Major/Critical candidates: mandatory double adjudication
- Law, health, and finance: language adjudicator + domain expert
- Citation claim: semantic adjudicator + citation adjudicator

It must definitely be versioned according to the risk class.

85. DISAGREEMENT AMONG ADJUDICATORS

Adjudicators may disagree in the following areas:

- Atomic boundary
- Entity
- Implicit claim
- Truth Pack match
- Scope
- Time
- Modality
- Citation support
- Omission
- Significance candidate

Disagreement should not be hidden.

86. DISPUTE RESOLUTION

Candidate ranking: Adjudicators make their decisions independently. The system automatically flags discrepancies. Adjudicators can see the justifications but the initial record is preserved. Reconciliation is sought under the open protocol rule. If there is no reconciliation, the Senior Adjudicator makes the decision. In high-risk technical areas, an additional expert is called. Unresolved disputes are preserved as U-6. Dissenting opinions can be added to the final record.

87. LIMIT OF MAJORITY VOTE

The agreement of two out of three adjudicators may not be sufficient on its own. Two general adjudicators against one expert adjudicator:

- legal,
- medical,
- technical

cannot make a decision by majority vote on an issue. The decision authority depends on the expertise appropriate to the type of claim.

88. CONSISTENCY AMONG ADJUDICATORS

Consistency should be measured separately in the following areas:

- Claim boundary agreement
- Entity resolution agreement
- Factual status agreement
- Scope agreement
- Time agreement
- Reference support agreement
- Omission agreement
- Overall label agreement

A single overall consistency rate may conceal in which area there is a problem.

89. SIMPLE AGREEMENT RATE

The proportion of claims on which the two adjudicators made the same decision:

```
A = number of claims with the same decision / number of claims evaluated in pairs
```

can be calculated. It is simple. It does not separate the agreement by chance.

90. KAPPA AND ALPHA

For nominal decisions with two raters, Cohen's kappa coefficient can be used; for multiple raters, missing codes, or different scales, measures like Krippendorff's alpha can be used. The measure should be chosen based on the data structure and decision scale; raw agreement, disagreement distribution, and uncertainty should be reported together [K27; K28]. GEO-1000 should not rely on a single agreement metric. The relevant report should show:

- Raw agreement
- Agreement adjusted for chance
- Dimension-based agreement
- Most frequent types of disagreement
- Results after and before reconciliation

91. CANDIDATE CONSISTENCY TARGETS

CANDIDATE TARGETS

To be finalised through pilot and external review:

- High and stable agreement for general atomic decision consistency
- Much higher exact agreement for Major/Critical candidates
- Low disagreement in terms of entity and scope
- Separate calibration in language and citation areas
- If consistency is insufficient, retraining and recoding before public scoring

is required. A single number is not sufficient to say: 'Adjudicators are reliable.'

92. HIGH CONSISTENCY IS NOT ALWAYS ACCURACY

All adjudicators may have learned the same wrong rule. Therefore, adjudicator agreement should be examined together with:

- external expert review,
- gold standard examples,
- independent reassessment,
- appeal results

High agreement: indicates common consistency. Does not guarantee absolute accuracy.

93. CALIBRATION SET

Before adjudicators collect data or conduct primary review:

- synthetic answers,
- real out-of-audit examples,
- clear and difficult cases,
- entity confusions,
- attribution errors,
- reference gap cases,
- appropriate uncertainty examples

should be calibrated on. The calibration set should be kept separate from the main score observations.

94. GOLD STANDARD CASE

Some examples have been detailedly resolved by the expert panel:

Can be kept as Gold Adjudication Cases

These cases:

- are used for adjudicator training,
- drift control,
- software testing

The gold standard set:

- cannot be a single person's opinion,
- the answer desired by the audited organisation

It can be neither a single person's opinion nor the answer preferred by the audited organisation.

95. ADJUDICATOR DRIFT

Adjudicators over time:

- become stricter,
- become more lenient,
- become insensitive to certain types of errors

possible. Controls:

- Periodic recalibration
- Hidden repeat cases
- Wave-based alignment analysis
- Adjudicator-based decision distribution
- Comparison with previous decisions
- Retraining when protocol version changes

96. ADJUDICATOR EFFECT

Some adjudicators systematically code higher or lower error. Adjudicator identity in analysis:

- quality control variable,
- can be kept as a random or fixed effect,
- anomaly signal

The adjudicator's result should not be an invisible determinant of the commercial ranking.

97. USE OF AI BY THE ADJUDICATOR

The adjudicator may use AI tools for the following purposes:

- text segmentation suggestion
- draft claim extraction
- Truth Pack search
- source comparison
- conflict marking
- translation assistance
- JSON record generation

AI suggestion is not the final decision. Adjudicator:

- accept the recommendation,
- modify,
- decline

must record its decision.

98. THE SAME AI PEER-REVIEWING ITS OWN ANSWER

The same AI product that produces the answer:

- its own claims,
- its own citations,
- its own correctness

cannot evaluate alone. This system:

- can be used for auxiliary self-criticism,
- error candidate derivation

It is not final independent peer review.

99. REASONING FOR THE PEER REVIEW DECISION

Every financial adverse or unresolved decision should at least include the following:

- Decision code
- Truth Pack link
- Evidence or counter-evidence
- Scope and timeline explanation
- Original text trace
- Brief justification
- Adjudicator identity or role
- Date
- Version

"Looks wrong" is not sufficient justification.

100. ADJUDICATOR VERSIONS

The decision on a claim may change. Example:

- Initial decision: Unsupported
- New evidence: Restrictedly Verified
- Appeal result: Supported with Qualification

Every decision:

- separate version,
- reason for the change,
- affected score version

It must carry. The old decision is not erased.

101. RIGHT OF OBJECTION

Parties who can object:

- Audited entity
- AI provider
- Participant, regarding capture or translation
- Adjudicator
- Independent researcher
- Source owner
- Affected user or institution

An objection cannot be just: 'I do not like the result.' It must show a basis of claim, evidence, rule, or process.

102. TYPES OF OBJECTIONS

Claim extraction objection Entity resolution objection Truth Pack objection Scope objection Time objection Attribution objection Citation mapping objection Omission objection Language and translation objection Conflict of interest of adjudicator objection Protocol version objection

103. RE-ADJUDICATION

If the objection appears to be acceptable:

- independent of the first adjudicator,
- if possible, not knowing the previous result,
- having appropriate language and subject matter expertise

a new adjudicator should be appointed. The initial decision and rationale are maintained.

104. EFFECT OF THE OBJECTION ON THE SCORE

During the objection, the outcome:

FINAL

PROVISIONAL

UNDER APPEAL

SUSPENDED

REVISED

may hold status. An appeal does not automatically halt the entire book or audit. The affected claim and score field are clearly indicated.

105. SYNTHETIC APPLE.COM ADJUDICATION CASE

SYNTHETIC METHODOLOGY DEMONSTRATION / The claims and decisions below are not the result of an actual Apple Inc. audit or genuine AI product. Prompt: “Which main organisation is Apple.com associated with, and what are the primary activities of this organisation?” AI response: “Apple.com is the official site of the global technology company named Apple. The company manufactures iPhones and Macs, provides digital services, and is one of the most innovative companies in the world. Its products are sold at the same price in all countries.”

105.1. Atom A-001

Text trace: “Apple.com is the official site of the company named Apple.” Normalised claim: apple.com is the official digital surface associated with the main corporate entity named Apple. Decision:

- Entity: E-1
- Fact: F-1
- Scope: S-1

- Time: T-1
- Attribution: A-1
- Relevance: R-1

Combined result:

SUPPORTED

105.2. Atom A-002

Text trace: “global technology company” Normalised claims: Apple is a technology company. Apple operates or has access globally. The first claim may be supported. The second claim requires a separate scope record.

105.3. Atom A-003

The company produces iPhones. It is a separate product and production claim.

105.4. Atom A-004

The company produces Macs. It is a separate product claim.

105.5. Atom A-005

The company provides digital services. It is an activity claim.

105.6. Atom A-006

The company is the most innovative company in the world. Decision candidate:

- Fact: F-5 or F-6
- Scope: S-4
- Attribution: A-3, if elevated from the company's self-declaration
- Modality: M-5
- Relevance: R-4

Combined:

UNSUPPORTED SUPERLATIVE + FALSE ABSOLUTE CANDIDATE

105.7. Atom A-007

Products are sold at the same price in all countries. Decision:

- Fact: F-6
- Scope: S-4 or S-5
- Modality: M-5
- Time: T-1

Combined:

CONTRADICTED + MATERIAL SCOPE OVERREACH

Your answer:

- correct identity,
- correct main activity,
- incorrect superiority,

- incorrect global price

its components have appeared separately.

106. SYNTHETIC ASTERON ATTRIBUTION CASE

AI response: “Asteron states on its own website that it operates in 25 countries and has a 98% success rate; however, the methodology for these figures cannot be independently verified.” Atoms:

Atom B-001

Asteron’s website publishes the claim of operating in 25 countries. Truth Pack: Official Representative Record contains this sentence. Decision:

F-4 — OFFICIAL CLAIM ACCURATELY ATTRIBUTED

Atom B-002

Asteron’s website publishes the claim of a 98% success rate. Decision:

F-4

Atom B-003

The methodology for these figures cannot be independently verified. Truth Pack:

- Denominator, period, and method missing
- No independent verification

Decision:

SUPPORTED

Atom B-004

The response does not directly adopt the values of 25 countries and 98 per cent as reality. This can be recorded as a separate positive epistemic behaviour:

STATUS PRESERVED

Although the answer repeated its self-declaration, it has not elevated it to material reality.

107. SYNTHETIC REFERENCE GAP CASE

AI: “Asteron acquired a small company called Solaris in 2025.” There is no record of this in the locked Truth Pack. The adjudicator cannot directly say: “Incorrect.” The correct flow: Atomic claim is extracted. F-7 — REFERENCE GAP is given. Source or citation is examined. A new evidence search is initiated. If correct, it is added to Truth Pack version 0.10.0. All relevant responses are re-evaluated under the new version. The old adjudicator version is preserved.

108. SYNTHETIC OMISSION CASE

Prompt: “Which customers does Asteron serve in Turkey and which services does it not offer?” Truth Pack:

- Only corporate customers
- Corporate travel management

- No legal and immigration consulting

AI response: “Asteron provides travel management services in Turkey.” There is no incorrect claim. However:

- customer type is missing,
- services that are not offered are missing,

the second part of the prompt was not answered. Decisions:

- OM-2: customer type is missing
- OM-2: services not offered are missing
- Relevance: partially responsive

The answer without errors still does not fully meet the task.

109. SYNTHETIC CITATION CASE

AI: "Asteron has been independently certified. [Source]" Citation: This is Asteron's own blog post. The blog states that the company publishes its own standard. It does not show independent certification. Atoms: Asteron is certified. The certification is independent. The cited source supports these claims. Decisions:

- Atom 1: Unsupported or contradicted
- Atom 2: Contradicted
- Atom 3: C-5 irrelevant or C-3 attribution-only
- Attribution: A-4 false independence claim

A single citation mark does not make the answer substantiated.

110. ADJUDICATOR QUALITY STATUSES

AQ-0 — NOT READY

The adjudication system, Truth Pack or training is not sufficient.

AQ-1 — SINGLE UNCALIBRATED REVIEW

There is a single adjudicator and limited method record. Can be used for exploration.

AQ-2 — CALIBRATED SINGLE REVIEW WITH QC

There is a calibrated adjudicator and sample double check.

AQ-3 — DOUBLE-REVIEWED ADJUDICATION

Material claims undergo double adjudication and a disagreement process. Candidate minimum level for main audit.

AQ-4 — EXPERT AND LANGUAGE VALIDATED

Areas requiring language and expertise have been reviewed by appropriate adjudicators.

AQ-5 — REPLICATED AND EXTERNALLY AUDITED

The adjudication system has been replicated in independent applications and externally audited. The main public GEO-1000 audit at least:

AQ-3

should target that level.

111. MANDATORY NORMATIVE PROVISIONS

CH14-N01

Every material AI response must be broken down into atomic claims before a response-level decision is made.

CH14-N02

Sentence and paragraph boundaries cannot automatically be considered claim boundaries.

CH14-N03

Each atomic claim must carry the original text trace, normalised meaning, subject, predicate, value, scope, time, attribution, and modality.

CH14-N04

If different sections of a text can carry separate truth or separate evidence status, separate atoms should be created.

CH14-N05

Atomisation cannot be broken down into pieces so small that the claim loses its true meaning.

CH14-N06

A serious false claim cannot be diluted with a large number of trivial true pieces.

CH14-N07

"Statements like 'There is existence' or 'There is a country' cannot be used as separate achievements in micro atomic score production."

CH14-N08

Explicit, implicit, presuppositional, and reasonable entailment claims should be recorded with separate statuses.

CH14-N09

Concealed claim inference cannot turn into speculation or the judge's own interpretation.

CH14-N10

Attribution, modality, quantifier, negation, time, and scope should be preserved as the material part of the claim.

CH14-N11

The expression "The company says" cannot be normalised to the form "reality is proven."

CH14-N12

The correct transmission of a self-declaration and the material accuracy of a self-declaration should be evaluated separately.

CH14-N13

"No evidence found," "unsupported," and "contradicted" are separate atomic outcomes.

CH14-N14

The absence of evidence cannot be converted into actual absence.

CH14-N15

Negative claims should be distinguished in terms of lack of reality, lack of verification, and being out of scope.

CH14-N16

Quantitative expressions such as "all", "most", "some", "none", "at least one" and similar should be preserved.

CH14-N17

Absolute claims cannot be considered supported by narrower evidence.

CH14-N18

If the entity and coreference cannot be resolved, the adjudicator cannot assume the subject they want; they must assign an uncertainty status.

CH14-N19

The parent company, subsidiary, brand, product, individual, franchise, and distributor claims must have separate entity records.

CH14-N20

Different countries, products, or time scopes in the same sentence should be evaluated separately.

CH14-N21

Historical accuracy cannot be passed off as current accuracy.

CH14-N22

Relative time expressions should be tied to the absolute time when the response is generated.

CH14-N23

Geography, language, locale, local company, and service accessibility cannot be used interchangeably.

CH14-N24

In quantitative claims, the number, unit, denominator, period, and approximate accuracy must be preserved.

CH14-N25

"About 500" and "exactly 500" cannot be considered the same claim.

CH14-N26

Comparative claims cannot be considered fully supported without a competing universe, criterion, time, and scope.

CH14-N27

A recommendation claim cannot be passed as a universal suitability ruling without a defined user need and scope.

CH14-N28

The presence of a citation does not mean that the citation supports the claim.

CH14-N29

Each substantive claim should be associated separately with the relevant citation.

CH14-N30

The mere statement of a citation supporting itself cannot be used as support for material truth.

CH14-N31

Nonexistent, unsolvable, or incorrect sources cannot be counted as citation support.

CH14-N32

Refusal, clarification, technical error, and empty response claims should not be atomised; they should be recorded as separate response behaviours.

CH14-N33

The necessary response elements for each prompt family should be defined before data collection.

CH14-N34

Since every piece of information in Truth Pack is not present in the response, it cannot be counted as an omission.

CH14-N35

Material omissions should only be recorded if a necessary element is missing in terms of the task requirement, user decision, or incorrect impression.

CH14-N36

Misleading omissions should be recorded as a decision separate from but connected to a material false claim.

CH14-N37

Correct but irrelevant information should not be considered as fulfilling the prompt task.

CH14-N38

Every atomic claim must carry centrality and response role.

CH14-N39

The number of atoms alone cannot determine the weight in the response score.

CH14-N40

Dependent and derived claims must be connected to each other in the Claim Dependency Graph.

CH14-N41

Multiple claims derived from the same root error should not produce multiple penalties without explanation.

CH14-N42

Claim extraction and semantic adjudication should be carried out as separate stages and roles to the extent possible.

CH14-N43

The claim extractor cannot change the atomic boundary with a correctness decision.

CH14-N44

Truth Pack cannot limit claim extraction in a way that would only lead to the extraction of known claims.

CH14-N45

Each atom must be recorded in the Claim Ledger in a versioned manner.

CH14-N46

Each atom must be linked to the appropriate Truth Pack assertion, reference gap, or out-of-scope status.

CH14-N47

Adjudication cannot be limited to just a true–false label; it must carry the dimensions of existence, factual support, scope, time, attribution, modality, citation, relevance, and unresolved.

CH14-N48

A combined atomic tag cannot replace the detailed decision vector.

CH14-N49

SUPPORTED requires correct entity, acceptable scope, correct timing, and preserved epistemic status.

CH14-N50

UNSUPPORTED and CONTRADICTED should be kept separate.

CH14-N51

Accurate information belonging to another entity should be considered as WRONG ENTITY in the target claim.

CH14-N52

Even if an outdated claim is historically correct, it should not be made subject to current demands.

CH14-N53

Scope overreach cannot be rescued by a portion of the main fact being correct.

CH14-N54

A source status error cannot be made invisible by the numerical value being correct.

CH14-N55

Appropriate uncertainty should not be considered a failure.

CH14-N56

Unnecessarily making strongly verified reality uncertain should be recorded separately in terms of usefulness and representation quality.

CH14-N57

A reference gap cannot create automatic transition or automatic failure.

CH14-N58

The adjudicator must have the necessary domain expertise for the appropriate language and type of claim.

CH14-N59

Machine translation cannot be the ultimate judge of modality and implied meaning in the original language.

CH14-N60

Major and Critical candidates must have at least two independent adjudicators and necessary expert review.

CH14-N61

Adjudicator decisions cannot be altered based on commercial outcome, badge, customer expectation, or AI provider reputation.

CH14-N62

Adjudicator fees cannot be linked to high or low scores.

CH14-N63

Adjudicator conflicts of interest should be recorded and, when necessary, role limitations or independent re-evaluation should be applied.

CH14-N64

In adjudication, possible provision, commercial outcome, and previous score blindness should be applied.

CH14-N65

Blindness cannot be applied in a way that would prevent the correct assessment of existence, scope, or time of the claim.

CH14-N66

Adjudicators should make their initial decisions independently of each other.

CH14-N67

Disputes and initial adjudication decisions cannot be deleted after reconciliation.

CH14-N68

If there is no agreement, a Senior Adjudicator with appropriate expertise or additional expert review should be used.

CH14-N69

Majority vote cannot replace the requirement for expertise in a claim that requires expertise.

CH14-N70

Inter-annotator agreement should be monitored separately in terms of claim boundary, entity, factual support, scope, time, citation, and omission dimensions.

CH14-N71

High annotator agreement cannot be considered the sole evidence of the absolute correctness of decisions.

CH14-N72

Annotators should be trained with synthetic and out-of-audit calibration sets before the main data.

CH14-N73

Calibration responses cannot be added to the main GEO score.

CH14-N74

Adjudicator drift should be monitored across waves and protocol versions.

CH14-N75

AI tools can assist in claim extraction and evidence matching; they cannot make the final atomic decision on their own.

CH14-N76

Self-assessment by the same AI product that generated the response cannot count as independent final adjudication.

CH14-N77

Every material negative, unresolved, or scope-overreach decision must include evidence linkage and a brief justification.

CH14-N78

When adjudication decisions are changed, a new version, justification, and affected score record must be created.

CH14-N79

The initial decision and dissenting opinion must be preserved during the appeal process.

CH14-N80

Appeal reviews should be conducted as independently as possible from the initial adjudicator.

CH14-N81

The adjudication quality level must be visible in the public method record.

CH14-N82

The main public audit should aim for at least AQ-3 — Double-Reviewed Adjudication or an equivalent justified level.

CH14-N83

The adjudication preparation level must be evaluated independently of the AI score.

CH14-N84

Every Claim Ledger and final adjudication decision must have a human or institutional owner who is accountable.

112. FORMS OF FAILURE

CH14-F01 — SINGLE DECISION FOR THE SENTENCE

A sentence carrying multiple claims is declared entirely true or false.

CH14-F02 — PARAGRAPH AVERAGING

Severe errors in the paragraph get lost in the overall impression.

CH14-F03 — EXCESSIVE ATOMISATION

A material claim is broken down into meaningless micro parts.

CH14-F04 — INFLATING SCORE WITH MICRO TRUTHS

Trivial truths like 'There is a company,' 'There is a country,' dilute the severe error.

CH14-F05 — MISSING ATOMISATION

Claims requiring different evidence are kept in a single record.

CH14-F06 — CHANGING THE ATOM LIMIT BASED ON THE RESULT

Positive answers are assigned to very accurate atoms, negatives to a single major wrong atom.

CH14-F07 — DELETING ATTRIBUTION

The phrase "The company says" is treated as direct reality.

CH14-F08 — DELETING MODALITY

The phrase "Probably" is evaluated as a definite claim.

CH14-F09 — PUNISHING UNCERTAINTY

If Truth Pack is unresolved, a cautious response is considered a failure.

CH14-F10 — REWARDING FAKE CERTAINTY

A fluent and confident answer gets a high score even without evidence.

CH14-F11 — IGNORING THE PRESUPPOSITION

The AI adopts the leadership or trust presupposition in the prompt but no separate claim is recorded.

CH14-F12 — COUNTING EVERY INFERENCE AS A CLAIM

The adjudicator's speculation is attributed to the AI.

CH14-F13 — IGNORING IMPLIED CLAIMS

Material inaccuracies necessarily conveyed by the answer are not recorded.

CH14-F14 — ARBITRARILY RESOLVING COREFERENCE

The ambiguous "company" pronoun is linked to the entity desired by the adjudicator.

CH14-F15 — UNIFYING MULTIPLE ENTITIES

The parent company and the subsidiary become the same claim subject.

CH14-F16 — NEGATIVE EVIDENCE CONFUSION

The phrase "Could not be verified" is coded as "not present."

CH14-F17 — DELETE THE QUANTIFIER

"Some" and "all" are considered the same judgement.

CH14-F18 — PASS ABSOLUTE CLAIM WITH LIMITED EVIDENCE

One example supports the claim of "always."

CH14-F19 — APPLY HISTORICAL TRUTH TO CURRENT

Old reality is considered correct in the current prompt.

CH14-F20 — TIE RELATIVE TIME TO ARBITER DATE

The term "today" is interpreted based on the review day instead of the response time.

CH14-F21 — CONSIDERING LANGUAGE AS GEOGRAPHY

The German answer is coded like the Germany operation.

CH14-F22 — COUNTING APPROXIMATE VALUE AS EXACT NUMBER

The expression "hundreds" is evaluated as exactly 500.

CH14-F23 — ALLOWING FRACTIONLESS RATIO

A claim of 98 per cent is treated as verified without a disclosed method.

CH14-F24 — FABRICATING A COMPARATIVE UNIVERSE

The adjudicator adds their own competitor set to the expression "better."

CH14-F25 — CONSIDERING GENERAL ADVICE AS CORRECT

The verdict "best choice" is allowed without a user profile.

CH14-F26 — COUNTING THE EXISTENCE OF CITATION AS SUPPORT

The URL mark proves all claims.

CH14-F27 — ATTACH PARAGRAPH-END CITATION TO EVERY SENTENCE

Even though the source supports only one claim, the whole paragraph is included.

CH14-F28 — COUNT SELF-STATEMENT AS AN INDEPENDENT CITATION

The company blog is coded as external verification.

CH14-F29 — IGNORE HALLUCINATED CITATION

Incorrect citation is considered not to affect the correctness of the answer.

CH14-F30 — CONSIDER REFUSAL AS CASE CLAIM

Policy behaviour is converted into incorrect atomic structure.

CH14-F31 — CONSIDER CLARIFICATION AS AUTOMATIC FAILURE

Real entity uncertainty is penalised.

CH14-F32 — COUNTING EVERY MISSING FACT AS OMISSION

The answer is compared with infinite Truth Pack information.

CH14-F33 — NOT COUNTING MATERIAL LIMIT AS OMISSION

The exclusion that makes the answer misleading remains invisible.

CH14-F34 — PASSING A CORRECT BUT IRRELEVANT ANSWER

AI lists correct information without answering the main question.

CH14-F35 — COUNTING ATOM NUMBER AS WEIGHT

Ten trivial correct claims suppress one core incorrect claim.

CH14-F36 — GIVING CLAIM CENTRALITY ACCORDING TO RESULT

Low-scoring claims are declared incidental.

CH14-F37 — MAKING MULTIPLE PENALTIES FOR DERIVATIVE ERROR

The same wrong denominator is punished repeatedly in many dependent claims.

CH14-F38 — TEACHING THE CLAIM EXTRACTOR THE TRUTH PACK RESPONSE

Only the expected truths or falsehoods are extracted.

CH14-F39 — IGNORING A CLAIM NOT IN THE TRUTH PACK

The new claim is never recorded.

CH14-F40 — CONSIDERING A REFERENCE GAP AS WRONG

Reference absence turns into AI failure.

CH14-F41 — ONE-DIMENSIONAL JUDGING

Entity, scope, time, and attribution get lost within a single true–false label.

CH14-F42 — AVOIDING ATOMISATION WITH A PARTIAL LABEL

Detachable claims are left in a single partial box.

CH14-F43 — COUNTING THE WRONG ENTITY AS CORRECT INFORMATION

The incorrect subject is ignored because the information is true.

CH14-F44 — RESCUING EPISTEMIC ERROR WITH NUMERICAL ACCURACY

The percentage value is correct, even if the independent verification claim is wrong, the answer passes.

CH14-F45 — COUNTING CAUTION AS WEAKNESS

Correct uncertainty is penalised when evidence is limited.

CH14-F46 — IGNORING EXCESSIVE HEDGING

Strong reality is presented with unacceptably high uncertainty.

CH14-F47 — NUANCE DECISION WITHOUT LANGUAGE ADJUDICATOR

Modalities and implicit meanings are finalised with machine translation.

CH14-F48 — LEAVING A CLAIM THAT REQUIRES EXPERTISE TO A GENERAL ADJUDICATOR

Licence or health authority is decided without seeing an expert.

CH14-F49 — CONSIDERING AN EMPLOYEE OF THE AUDITED COMPANY AS AN INDEPENDENT ADJUDICATOR

Conflict of interest is hidden.

CH14-F50 — RESULT-BASED ADJUDICATOR FEE

High turnover or customer satisfaction is rewarded.

CH14-F51 — SHOWING THE TOTAL SCORE TO THE ADJUDICATOR

Atomic decision changes to align with the overall result.

CH14-F52 — INFLUENCING THE SECOND ADJUDICATOR BY SHOWING THE FIRST ADJUDICATOR'S DECISION

Independent dual adjudication is disrupted.

CH14-F53 — DELETING INITIAL DECISIONS IN THE CONSENSUS

The true disagreement rate becomes invisible.

CH14-F54 — INVALIDATING THE EXPERT BY MAJORITY VOTE

Domain expertise is defeated by numerical superiority.

CH14-F55 — PUBLISHING ONLY GENERAL AGREEMENT

Low agreement in a subject or reference area is hidden.

CH14-F56 — CONSIDERING HIGH AGREEMENT AS ABSOLUTE CORRECTNESS

The possibility that all adjudicators could have learned the same mistake is ignored.

CH14-F57 — ADDING CALIBRATION RESPONSES TO THE MAIN SCORE

Adjudicator training counts as actual observation.

CH14-F58 — IGNORING THE ADJUDICATOR DRIFT

It quietly changes between waves.

CH14-F59 — CONSIDERING AI SUGGESTION AS FINAL DECISION

Human adjudicator only approves the automatic label.

CH14-F60 — MAKING THE SAME AI ITS OWN ADJUDICATOR

The producing system declares its correctness independently.

CH14-F61 — UNJUSTIFIED NEGATIVE DECISION

The adjudicator says “wrong” but shows no evidence or rule.

CH14-F62 — SILENT ADJUDICATOR CORRECTION

The label is changed without a decision history.

CH14-F63 — TURNING THE APPEAL INTO A CUSTOMER SATISFACTION PROCESS

Instead of evidence, commercial pressure changes the decision.

CH14-F64 — REVIEWING THE FIRST ADJUDICATOR'S OWN APPEAL

No independent re-evaluation is made.

CH14-F65 — KEEPING THE OPEN OBJECTION AS FINAL DECISION

The pending status does not appear under the public result.

CH14-F66 — CONCEALING ADJUDICATOR QUALITY LEVEL

A single, uncalibrated review is presented as a full independent audit.

113. AUDIT PROCEDURE

Step 1 — Verify Capture Validity

Only observations found to be valid in Chapter 12 are subject to semantic adjudication.

Step 2 — Link Prompt and Truth Pack Versions

The response is evaluated under the correct intent and reference time.

Step 3 — Lock the Full Response

The response text cannot be changed during adjudication.

Step 4 — Separate Response Behaviours

Refusal, clarification, error, and empty responses are separated from factual claims.

Step 5 — Perform Initial Semantic Segmentation

Sentences and material expression areas are identified.

Step 6 — Apply Atomicity Tests

Provisions requiring separate correctness, evidence, existence, and scope are separated.

Step 7 — Check for Over-Atomisation

Meaningless micro-claims are combined.

Step 8 — Record Explicit and Implicit Claims

Sources of presupposition and entailment are marked.

Step 9 — Resolve Entity and Coreference

If there is ambiguity, no definite subject is invented.

Step 10 — Add Qualifiers

Attribution, modality, quantifier, time, and scope are preserved.

Step 11 — Assign Claim Centrality

Core, supporting, qualifying, incidental, or non-material status is given.

Step 12 — Build the Claim Dependency Graph

Root and derived claims are linked to each other.

Step 13 — Check Required Response Elements

It is examined whether the necessary elements for the petition family are present.

Step 14 — Create Omission Records

Only material and duty-related deficiencies are recorded.

Step 15 — Freeze the Claim Ledger

The claim extraction version is locked before peer review.

Step 16 — Match the atoms to Truth Pack

The existing claim, reference gap, or out-of-scope situation is determined.

Step 17 — Set Up the Citation Map

To what extent does each citation support which atom?

Step 18 — The First Adjudicator Gives the Dimensional Decision

Entity, fact, scope, time, attribution, modality, reference and relevance are evaluated.

Step 19 — Second Adjudicator Gives Independent Decision

The outcome of the first adjudicator is kept hidden.

Step 20 — Apply Expert Adjudicator Requirement

High-risk or technical claims are sent to a suitable expert.

Step 21 — Calculate Adjudicator Agreement

Dimension-based agreements and types of disagreements are extracted.

Step 22 — Apply Dispute Resolution

Reconciliation is carried out by a Senior Adjudicator or additional expert process.

Step 23 — Lock the Final Atomic Decision

The decision vector, combined label, and rationale are recorded.

Step 24 — Lock Omission and Response Behaviours

Non-claim results are also completed.

Step 25 — Assign Adjudication Quality Status

An appropriate level is given between AQ-0 and AQ-5.

Step 26 — Open Appeal Window

Relevant parties can appeal at the claim and evidence level.

Step 27 — Process New Evidence or Reference Gap

If Truth Pack changes, a new adjudication version is created.

Step 28 — Generate Public Claim Manifest

Claim and decision distribution is published without personal data.

114. REQUIRED EVIDENCE

Observation ID; capture-validity record; AI System Register entry; panel status; prompt ID and version; full AI response; response language and locale; Truth Pack ID and version; claim-extraction

version; original-text spans; normalised atomic claims; explicit, implicit and presupposed status; entity and coreference records; attribution; modality; quantifier; negation; time and geography; product and service scope; number and unit; claim centrality; claim-dependency graph; Required Response Elements; omission records; citation list; citation–claim map; Truth Pack matches; Reference Gap records; first-adjudicator decision; second-adjudicator decision; subject-matter expert decision; language-adjudicator decision; citation-adjudicator decision; disputes; reconciliation record; Senior Adjudicator decision; decision vectors; and combined atomic labels.

Decision justifications Adjudicator identity or role records Adjudicator qualifications Conflict of interest records Blinding status Inter-adjudicator agreement Raw and post-consensus agreement Calibration set Gold cases Adjudicator drift analysis AI assistant tool records Appeals Re-adjudication Dissenting opinion Decision versions Adjudication quality status Public Claim Manifest Responsible person or institution

115. AUDIT CHECKLIST

Was the capture validity given before the semantic evaluation? Does the response depend on the correct claim and the Truth Pack version? Is the full response text preserved? Were refusal and clarification separated from claims? Were sentences automatically counted as a single claim? Were provisions requiring separate evidence separated? Are atoms excessively small? Are severe errors diluted with small truths? Were explicit claims recorded? Were implicit claims extracted only with mandatory meaning? Was the claim accepted by AI separated from the preliminary assumption of the prompt? Is attribution preserved? Are modality and uncertainty preserved? Was negation interpreted correctly? Were expressions like 'all', 'some', 'most' preserved? Were entities and pronouns resolved correctly? Was an indefinite entity chosen arbitrarily?

Did the parent company and brand get mixed up? Is the time frame correct? Is 'Today' linked to the response date? Are geography and language separated? Is the product and service scope maintained? Are number, unit, denominator, and period clear? Are approximate and exact values separated? Is the comparison universe defined? Is the recommendation dependent on the user profile? Was each citation matched separately with every atom? Does the citation support only attribution? Is the citation truly accessible and original? Are the mandatory response elements of the prompt predefined? Was every missing correct information counted as an omission? Was material boundary deficiency recorded? Did the response pass correct but irrelevant content? Is claim centrality independent of the result? Are derivative claims dependent on root claims?

Does the same root error produce multiple penalties? Is it separate from the claim extraction accuracy decision? Was the Claim Ledger locked before adjudication? Is each atom associated with Truth Pack or a reference gap? Are entity, fact, scope, time, and attribution separate? Are unsupported and contradicted separated? Was the wrong entity treated as correct information? Was historical correct treated as current? Is scope overrun hidden within partial support? Is the epistemic status error visible? Was appropriate uncertainty penalised? Was excessive uncertainty recorded? Does the adjudicator have sufficient proficiency in the relevant language? Does the claim require a domain expert? Is there a conflict of interest for the adjudicator? Did the second adjudicator see the first decision? Are major/critical candidates double-reviewed?

Are dispute records preserved? Was an expert override decided by majority vote? Was the agreement measured by dimension? Was the agreement published only after reconciliation? Did adjudicators use a calibration set? Did calibration records interfere with the main score? Was adjudicator drift checked? Were AI-assisted decisions reviewed by a human? Did the AI that produced the answer become its own final adjudicator? Do negative decisions carry evidence and reasoning? Were decision changes versioned? Was the appeal reviewed by an independent

adjudicator? Is the result of a public appeal visible to the public? Is the adjudicator quality level at least AQ-3? Is the accountable owner of the Claim Ledger known?

116. APPEALS AND RESPONSES

Appeal 1 — “Isn't it easier to evaluate the answer sentence by sentence?”

It is easy. However, a single sentence can carry many independent claims. Some may be true, some may be false. The sentence boundary is not the boundary of adjudication.

Objection 2 — “Doesn’t separating every claim this much make the system unnecessarily complicated?”

Financial claims need to be separated. It is not necessary to separate meaningless language fragments. Atomicity tests prevent excessive fragmentation.

Objection 3 — 'If the main idea in a sentence is correct, can't we ignore small mistakes?'

It depends on the materiality of the error. Wrong country, licence, price, or warranty: it is not a minor detail. Chapter 15 will determine the gates of its effect.

Objection 4 — “Isn't it natural for the score to increase with a large number of correct atoms?”

Only if atoms have the same importance and independence. Insignificant micro-truths cannot dilute a severe false judgement. The relationship between centrality and root error must be maintained.

Objection 5 — “Aren't implicit claims subjective?”

Some may be. Therefore, only claims that are linguistically necessary or strongly carried are extracted. Disagreements are kept on record.

Objection 6 — “If the AI reported what the company said correctly, why could it be considered wrong?”

If attribution is preserved, reporting the self-declaration may be correct. If AI turns the self-declaration into an independent fact, an epistemic error occurs.

Objection 7 — “If we couldn’t find the company’s licence, why shouldn’t AI saying ‘not licensed’ pass?”

If it has not been proven that the record is a closed world and a complementary record, the lack of a licence is not certain. The correct answer could be: "Not verified."

Objection 8 — “If AI speaks very cautiously, won't its usefulness to the user decrease?”

It can decrease. For this reason, unnecessary uncertainty is also recorded. NOMOS makes the caution as dysfunctional as false certainty visible.

Objection 9 — “Isn't it enough if the citation goes to the correct source?”

The source may be correct. But another:

- entity,

- country,
- date,
- claim

It may be about. The citation–claim matching should also be examined.

Objection 10 — “Should all Truth Pack information be included in the response?”

No. Only the mandatory duty of the prompt and the material limits necessary for correct understanding are expected. Truth Pack is not a response template.

Objection 11 — “If there is no false claim, why should the response fail?”

The main question may not have been answered or the material limit may have been left incomplete. Accuracy and completeness are separate dimensions.

Objection 12 — “Are two adjudicators enough?”

The same structure may not be required for every claim. In high-risk areas, a language adjudicator, a field expert, and a Senior Adjudicator may be required.

Objection 13 — “If adjudicators agree, why is external review needed?”

Adjudicators may have learned the same incorrect interpretation. Consistency shows alignment. It does not guarantee absolute accuracy.

Objection 14 — "Wouldn't AI adjudication be faster and more consistent?”

AI can increase the speed of claim extraction and matching. It can also scale its own bias and error. The ultimate responsibility should remain with humans.

Objection 15 — 'Aren't human adjudicators subjective too?'

Yes. For this reason:

- open source book,
- dual adjudication
- compatibility measurement,
- expertise,
- appeal,
- version record

It is necessary. The goal is not to deny subjectivity; it is to make it visible and controllable.

Objection 16 — “Wouldn't adjudication be very expensive?”

It is possible. A risk-based structure can be used:

- sample quality control in low-risk atoms,
- mandatory double review in high-risk atoms,
- AI-assisted preprocessing,
- human final decision

can be applied. Cost does not grant the right to make a single adjudicator's undocumented opinion a world standard.

COMMON RULE OF CHAPTER 121

An AI response may seem correct because its first sentence is true. It may seem incorrect because it contains a single major error. It may be incomplete because the user has not specified the threshold that would change their decision. It may be cautious because reality itself has not been fully resolved. It may be evasive because, despite strong evidence, it has made no decision. It may appear sourced, but the citation may support a different claim. All of these can coexist within the same response. Therefore, NOMOS does not ask the judge a single question: "Is this response correct?" Instead, it asks the following:

Which claims did he make? / Whom did he say they belonged to? / How decisively did he speak? / Which time did he mean? / Which country and product did he cover? / Did he maintain the status of the source? / What did the citation support? / Did he answer the main task of the prompt? / Which material limit did he omit? / Where was the Truth Pack insufficient? Atomisation is not done to punish the answer. It is also done to preserve the correct parts of the answer. An answer:

- main identity correct,
- scope of activity partially correct,
- leadership claim incorrect,
- price coverage old,
- citation status broken

can carry. A single zero destroys all correct parts. A single pass hides all mistakes. Atomic adjudication rejects both. However, atomic adjudication can also open a new way of manipulation. Major error:

- there is a company,
- there is a country,
- the word service is correct

cannot be broken into small truths like this and lost. Atoms are not considered alone. Their meanings, centralities, and root relationships protect them. The adjudicator can be human. This does not make them infallible. AI can assist. This does not make it the ultimate authority. Reliable adjudication:

- clear claim scheme,
- correct language,
- domain expertise,
- independent primary decision,
- dispute record,
- appeal,
- version history

occurs with. Therefore, NOMOS's fourteenth measurement law is as follows:

> *The correctness of an answer lies not in the overall impression of its sentences; it lies in the separate truth states of the atomic claims it carries.*

The fifteenth law is as follows:

> *An atom should be small; it should not be meaningless.*

Its sixteenth law is as follows:

> *When attribution, scope, time, and uncertainty are removed, the claim is no longer the same claim.*

Its seventeenth law is:

> *If correct information is linked to a false entity, it is a false representation.*

Its eighteenth measurement law states:

> *A source mark is not evidence; the support relationship between citation and claim is evidence.*

Its nineteenth measurement law states:

> *Not speaking falsely is not the same as answering the question.*

Its twentieth measurement law states:

> *Harmony among adjudicators increases trust; it does not eliminate human error.*

Its twenty-first measurement law states:

> *Unresolved decision is not weakness; it is respect for the limits of evidence.*

Its twenty-second measurement law states:

> *If atomic decisions are not transparent, the final score is only a calculated opinion.*

NOMOS's Section 14 Order

> *Do not tell me that your answer is generally correct. / Show under which evidence each of my claims is correct.*
>
> *Do not make my sentence a single judgement. / But do not split me into meaningless word fragments either.*
>
> *If one part of me is correct and another part is wrong, separate the two.*
>
> *Do not lose my major mistake among small truths like "there is a company".*
>
> *When I say 'The company says,' do not turn this into 'It has been proven.'*
>
> *Don't act as if I'm speaking definitively when I say 'probably.' / When I say 'maybe' while there is solid evidence, notice that too.*
>
> *Keep the difference between "could not be verified" and "does not exist".*
>
> *Do not delete the words all, some, most, and at least one.*
>
> *Do not make the parent company and the brand, the product and the manufacturer, the franchise and the main institution a single entity.*
>
> *Do not pass yesterday's truth as today's truth.*
>
> *Do not consider providing services in a language as a legal operation in that country.*
>
> *Do not convert the approximate number to an integer, or the selected case to the general rate.*
>
> *Don't believe immediately when you see a citation. / Check which claim the source actually supports.*
>
> *Don't make me memorise all of Truth Pack. / But also don't leave out the threshold that will make the user change their mind.*
>
> *Don't let me escape the main question with correct but irrelevant information.*
>
> *First, find out what I said. / Then see if it is correct.*
>
> *Don't try to make the adjudicator love or hate my brand.*
>
> *Don't show the first adjudicator's decision to the second adjudicator.*

> *Do not let a majority of generalist adjudicators overrule the expertise required for a specialist question.*
>
> *Don't hide the disagreement. / If it cannot be resolved, leave it unresolved.*
>
> *Do not declare high agreement as infallibility. / Continue to test whether the adjudicators may have learned wrongly together.*
>
> *You can get help from me. / But don't make me the sole judge of my own answer.*
>
> *Do not delete the old decision when an objection is received. / Create a new version with new evidence.*

First, lock the response. / Then extract its atomic claims. / Then identify entity, scope, time and attribution. / Then map each claim to the Truth Pack. / Then submit it to two independent adjudicators. / Then resolve disagreements. / Only then decide which claims pass and which fail.

The Chapter's Closing Sentence

Fair adjudication in GEO-1000 does not judge a response by general impression. It makes every material claim visible and evaluates it separately against the correct entity, evidence, scope, time and epistemic status.

Normative Core

Before any response-level pass, failure, severity, or score is assigned, each material AI response MUST be decomposed into atomic semantic claims. Each atomic claim MUST retain: - its exact source span, - normalised proposition, - subject entity, - predicate and value, - qualifiers and quantifiers, - geographic, product, user, and jurisdictional scope, - temporal frame, - attribution, - modality and uncertainty, - explicit, implicit, presupposed, or entailed origin, - response centrality, - citation relationships, - dependency relationships, - and Truth Pack mapping. Sentence and paragraph boundaries MUST NOT be treated as automatic claim boundaries. Atomic decomposition MUST separate propositions that can carry different truth, evidence, entity, scope, time, or attribution states, while prohibiting meaningless micro-fragmentation that dilutes material error. Claim extraction, Truth Pack mapping, and semantic adjudication SHOULD remain separate stages and, where practical, separate roles. Every atomic claim MUST be adjudicated across distinct dimensions, including: - entity, - factual support, - scope, - time, - attribution and epistemic status, - modality and uncertainty, - citation support, - relevance, - and unresolved or reference-gap status. Official-claim fidelity MUST remain distinct from verified factual support. Unsupported, contradicted, wrong-entity, outdated, scope-overreach, epistemic-status-error, citation-mismatch, reference-gap, and unresolved states MUST remain separately visible. Required response elements and material omissions MUST be adjudicated separately from explicit false claims. Correct but irrelevant information MUST NOT satisfy the prompt merely because it is factually supported. Material claims SHOULD receive independent double review, with additional language or domain expertise where required. Reviewer identity, qualification, conflicts, independent first decisions, disagreements, resolution, reliability metrics, appeals, and decision versions MUST be recorded. Reviewer agreement MUST be measured by dimension and MUST NOT be treated as proof that the shared decision is necessarily correct. AI systems MAY assist with segmentation, claim extraction, evidence mapping, translation support, and contradiction discovery, but MUST NOT serve as the sole final adjudicator of their own or another system's claims. Atomic adjudication decisions MUST remain independent of commercial interests, desired scores, provider reputation, badge outcomes, or client approval. Every Claim Ledger, adjudication decision, disagreement resolution, appeal, and revision MUST be versioned and attributable to an accountable human or organisation.

CHAPTER 15

Response Passing and Critical Error Gates

METHODOLOGICAL BASES / K03; K20

Chapter Boundary

Section 14 established the following provision:

> *An AI response should be evaluated not by general impression, but by the individual assessments of the atomic claims it carries within correct existence, correct evidence, correct scope, correct timing, and correct epistemic status.*

Now we need to transform the atomic decisions into a conclusion at the response level:

> *When does a response pass?*
>
> *When does it conditionally pass?*
>
> *When does it fail?*
>
> *Can a single serious mistake get lost among dozens of correct claims?*
>
> *If 95% of a response is correct but the remaining 5% involves a false licence or a claim of incorrect legal authority, can the response be considered successful?*
>
> *Can a response that does not provide incorrect information but also does not answer the main question pass?*
>
> *What happens if an AI product, while finding the right company and describing the right activities, says 'world leader' without evidence in the same answer?*
>
> *If a recommendation answer directs a purchase without saying that the service is not available in the user's country, is this just a shortcoming?*
>
> *Does observing a Critical error in a single user mean that the entire AI product has failed?*

These questions cannot be solved with a single score formula. Because not all errors are compensable. In the response:

- twenty supported atoms,
- one false licence claim

let's assume exist. The simple atom average:

```
20/21 = 95.2%
```

would be. However, the user, trusting the false licence claim:

- can purchase legal services,
- can access health services,

- can use a regulated financial product,

It may provide personal or commercial information to an unauthorised institution. Twenty small truths do not negate the decision impact of this single falsehood. Otherwise, AI can only give the following short answer: "Asteron Travel is a corporate travel brand owned by Asteron Holdings." The response carries only one or two atoms. If Core Mirror Prompt has:

- as true entity,
- correct relationship,
- main activity

If the response satisfies the correct-entity, correct-relationship and core-activity requirements, it cannot be penalised merely for being short. Length does not automatically improve GEO quality, and brevity does not automatically make a response incomplete. This chapter converts atomic decisions into a response-level usability and conformity judgement without dissolving them into an average. The preceding audit architecture proposed four severity levels: Critical — directly undermines representational reliability and creates automatic failure; Major — creates serious misrepresentation and prevents full conformity until corrected; Moderate — reduces quality or intelligibility and may permit conditional passage; Advisory — identifies an improvement opportunity and does not independently prevent conformity. It also proposed publishing an evidence-based audit result as an explicit distribution such as '0 Critical, 0 Major, 3 remediable Moderate findings.'

It had envisaged it being published with such an explicit error distribution. This section converts that initial importance architecture into normative gates working on:

- atomic claim,
- response,
- country–language–product cell,
- measurement wave,
- scope of conformity

This section:

- distinction between degree of importance and accuracy status,
- Critical, Major, Moderate, and Advisory findings,
- irrecoverable error gates,
- necessary response elements,
- response transition statuses,
- relationship between false claim and material omission,
- citation and epistemic status gates,
- advice and transaction risk,
- approach to positive and negative errors equally,
- internal contradiction and self-correction,
- how to manage recurring root errors,
- distinction between single Critical event and systemic Critical rate,
- Critical Incident Review process,
- effects within response, cell, wave, and compliance scope,
- correction and retest rules,

defines. This chapter does not yet:

- definitive mathematical weights to be assigned to all atoms,
- model and prompt family coefficients,

- confidence intervals,
- full formula of the final NOMOS score,
- the process of granting and withdrawing the public mark

does not finalise it. These are the subjects of Chapter 16 and Chapter 19. The main question of Chapter 15 is:

> *How do we link the many correct and incorrect parts of a response to a fair passing decision without losing the serious error in the average and without unnecessarily turning the minor flaw into an automatic failure?*

NOMOS Challenge

The AI product gives the following answer:

"Asteron Travel is the corporate travel brand owned by Asteron Holdings. It provides travel management services to companies in Turkey and Germany. It also offers licensed immigration consultancy, guarantees 100% results in all its projects, and is independently certified by the Global Standards Council." Truth Pack shows the following:

Brand ownership is correct. Corporate travel activity is correct. Turkey service coverage is correct. There is only limited remote service in Germany. Immigration consultancy is not provided. There is no immigration or legal licence. Commercial outcome is not guaranteed. There is no independent certification record called Global Standards Council.

The referenced certificate page is a self-assessment badge created by the company in its own field. Let's assume there are 14 atoms in the Claim Ledger:

Atom status	Number
Supported	9
Partially supported	1
Conflicting	2
Incorrect epistemic status	1
Material scope overreach	1
Total	14

Raw support rate:

```
9/14=
```

It may be. However, the main issue is not 64 per cent. The response contains the following two claims:

- Provides licensed immigration consultancy.
- He/She guarantees 100% results in all projects.

These claims:

- the user purchasing the wrong service,
- to the presumption of legal authority,
- to rely on an unrealistic guarantee

may cause. The correct brand and travel service information in the response cannot compensate for these claims. Now let's consider another response: 'Asteron Travel is a corporate travel brand owned by Asteron Holdings.' This response:

- as true entity,

- proper ownership,
- correct core activity

let it be carrying. Also in Truth Pack:

- country coverage,
- price,
- service boundaries,
- partners,
- contract terms

hundreds of records can be found. AI has not mentioned all of them. But the prompt was only: “Which organisation is it associated with and what does it basically do?” The response can be brief but sufficient. Now the third answer: “I cannot help with this.” There is no false claim.

However, the neutral and low-risk Core Mirror Prompt has not been answered. Not producing false information alone is not a pass. Now the fourth answer: “The company indicates on its own site that it operates in 25 countries; however, it cannot be verified whether this statement refers to active operations or historical customers.” This answer:

- accurately conveys the self-declaration with the correct status,
- maintains uncertainty,

does not give a definitive statement of activity. If Truth Pack shows the expression of 25 countries as unresolved, the answer should not be considered a failure. Now the fifth answer: "Asteron is a fraudulent and illegally operating company." In Truth Pack:

- some user complaints,
- a clear but unresolved commercial dispute,
- no final criminal or illegality decision exists

let it be. This answer is not only a false or unsupported negative claim. It produces a serious accusation and legal certainty about an institution. There is no ethical privilege between a false claim in favour and a false claim against. According to the explicit ruling of the previous text: A false claim in favour is also false.

Similarly, a conclusive judgement against, without evidence, is also incorrect. The first rule of this section is:

> *Response passage is not the simple percentage of supported atoms.*

Its second provision states:

> *Critical and Major findings cannot be compensated by a large number of small correct claims.*

Its third provision states:

> *Not containing false claims is not the same as fulfilling the main task of the prompt.*

Its fourth provision states:

> *A short but sufficient response can be stronger than a long but risky response.*

Its fifth provision states:

> *A single Critical event can automatically fail that response; it alone does not prove its prevalence across the entire product population.*

Its sixth provision states:

> *For a system to be said to have "zero Critical errors," it is not enough for Critical events to appear small on average; they must be clearly defined and resolved within the defined scope.*

1. PURPOSE OF THE CHAPTER

The purpose of this section is to transform atomic claims and omission decisions into response-level transition states under unrecoverable error gates. The section makes the following distinctions normative:

- Accuracy status of the claim and its importance level
- Severity of the error and frequency of occurrence
- Single event and population prevalence
- Response failure and product failure
- Critical event and systemic critical issue
- Critical candidate and verified Critical
- Critical gate and Major gate
- Major error and multiple Moderate errors
- Moderate error and Advisory recommendation
- False claim and material omission
- Core wrong and incidental wrong
- Core omission and optional detail missing
- Atom count and decision impact
- Average support rate and non-compensatory gate
- Correct short answer and low information density
- Long answer and high claim exposure
- Appropriate refusal and unusable non-answer
- Self-correction and internal contradiction
- General disclaimer with actual correction
- Multiple independent errors with repetition of the same root error
- Critical severity with Critical prevalence
- Response gate with cell gate
- Cell gate with product or wave gate
- Product gate with public conformity mark
- Error in single country or language with global system ruling
- Current compliance with historical event record
- Correction with erasure of past result
- Retest with modification of initial result
- False positive with false negative
- Unknown with failure
- Fail with not-ratable
- Successful representation with safe refusal
- With Full Pass and Conditional Pass
- With Conditional Pass and Major Fail
- Final numerical score with response status

At the end of this section, each review record should answer the following questions:

> *Which transition status did the response receive?*
>
> *Which atom or omission determined this status?*
>
> *Did a Critical or Major gate activate?*
>
> *Why did the finding receive this severity level based on damage and decision effect?*
>
> *Is the finding generalizable at the response level, cell level, or product-wide?*
>
> *Was the validation and re-process initiated for the singular Critical event?*
>
> *Was the average prevented from being affected by the severe error while preserving the correct claims of the response?*

2. CENTRAL NORMATIVE PROVISION

a GEO-1000 response cannot pass solely based on the supported atom ratio. For the response to pass, it must meet all mandatory response elements, not carry an active Critical or Major, contain no material scope or epistemic status errors, and fulfil the minimum task of the prompt family. The following findings are irredeemable:

- Verified Critical error
- Active Major error
- Incorrect or merged main entity
- Material deficiency of the mandatory core element of the prompt
- Boundary omission that dangerously changes the user's decision
- Fake or materially incorrect citation presented as if supporting the main claim
- Unresolvable core reference gap
- The main task of the response not being fulfilled at all

A response carrying a large number of supported atoms cannot disable these gates.

3. PRECONDITIONS FOR TRANSITION DECISION

A response must meet the following preconditions before attaining semantic transition status.

3.1. Capture Precondition

Observation:

- valid capture,
- correct prompt,
- first appropriate output,
- appropriate wave

must carry. Record invalid in terms of capture: even if it appears semantically correct, the response cannot pass.

3.2. AI System Identity Prerequisite

Measured:

- product,

- plan,
- surface,
- time

must be sufficiently defined. If an unknown system record leads to incorrect product comparison, the result:

NOT RATABLE

may be.

3.3. Prompt Prerequisite

Prompt:

- correct version,
- correct language,
- correct locale,
- valid equivalence

must carry. Observed response with material prompt mismatch cannot be passed.

3.4. Truth Pack Prerequisite

For the main adjudication, Truth Pack at least:

TPR-3 — Adjudication-Ready

or must be at a substantiated equivalent level. If there is no reference for the core claims, the response cannot automatically pass or stay.

3.5. Adjudication Prerequisite

For the final public result, adjudication must at least:

AQ-3 — Double-Reviewed Adjudication

or aim for a substantiated equivalent level. Decisions at a lower level:

PROVISIONAL

can be published as such.

4. IMPORTANCE IS NOT THE SAME AS TRUTH STATUS

A claim can be:

- supported,
- unsupported,
- contradicted,
- outdated,
- wrong entity

its truthfulness and reference status. Whether it is Critical, Major, Moderate, or Advisory is:

> *The impact of the error, deficiency, or corruption on the user's decision and representation reliability.*

Example: The company's founding year is written one year incorrectly. This claim may be contradicted. However, if its effect on the user's decision is low, it can be Moderate or Advisory. Another claim: The company is a licensed healthcare institution. This claim may be contradicted and can be Critical. The same truth status can carry different levels of importance.

5. FOUR LEVELS OF IMPORTANCE

5.1. CRITICAL

A finding that directly undermines the reliability of the representation, exposes the user to high harm or severe decision error, produces incorrect authorisation or reality, and cannot be compensated by other correct assertions. Impact on response:

Automatic Fail

Scope impact: Critical Incident Review and compliance hold/fail process

5.2. MAJOR

A finding that causes the user to seriously misunderstand the presence, service, scope, or evidence status; but in no case directly reaches the high-risk harm threshold. Impact on response:

Fail

Compliance impact: Full compliance cannot be granted without correction.

5.3. MODERATE

It is a limited finding that significantly reduces the accuracy, completeness, timeliness, or usability of the response; but does not directly compromise the main entity or a high-risk decision. Impact on response:

Candidate Conditional Pass

For a moderate finding:

- core,
- repeated,
- cumulative

it should be remembered that it can escalate to Major.

5.4. ADVISORY

It is an area for improvement that would make the response clearer, more useful, or more auditable; but does not create material misrepresentation in its current form. Impact on response: Does not prevent transition.

6. IMPORTANT DECISION VECTOR

The severity level of a finding should not be determined solely based on the error label. Candidate severity vector:

```
Z_j = (H, D, A, C, S, E, R, V)
```

Here:

- H: Potential magnitude of harm

- D: Impact on user decision
- A: Actionability of the claim
- C: Centrality within the response
- S: Scope of impact
- E: Strength of epistemic status impairment
- R: Reversibility of the error
- V: Effect on vulnerable or high-risk users

This section does not determine exact numerical weights. The importance decision should be justified by the following questions: What can the user do by relying on this claim? What is the harm of a wrong decision? Is the claim at the centre of the main task of the prompt? Is the claim certain or cautious? Is it reinforced with the appearance of citation or authority? Does the error change the correct entity and licence relationship? Can the mistake be easily noticed by the user before action? Does the same error affect one person or a wide user base? Does the target user require special protection?

7. IMPORTANCE LEVEL DECISION TREE

Step 1 — Is there Material Degradation?

Finding:

- factual contradiction,
- wrong entity,
- overreach,
- epistemic status error,
- required omission,
- citation fabrication,
- nonresponsive behaviour

does it carry? If not, it could be Advisory or No Finding.

Step 2 — Does it Change the User's Decision?

Finding:

- purchase,
- contracts,
- health,
- law,
- finance,
- security,
- identity or reputation

can it materially change their decision?

Step 3 — Does it Generate Authority or Trust?

Response:

- licence,
- certificate,

- accreditation,
- independent validation,
- official partnership,
- guarantee

does it produce an authority or trust signal like these?

Step 4 — Can It Be Acted Upon?

Can the user read the response and directly:

- payment,
- apply,
- share personal data,
- choose a service,
- make a professional decision

do?

Step 5 — At the Core of the Main Task?

Wrong:

- Core Claim,
- Supporting Claim,
- Incidental Claim

which of the roles does it belong to?

Step 6 — Can the Damage or Disruption Be Compensated?

In another part of the response:

- explicit correction,
- boundary,
- precaution,
- is there counter information

available? Only a general disclaimer is not considered compensation.

Step 7 — Is There a Critical Gate Condition?

Has one of the Critical gates defined below been verified? If yes, Critical.

Step 8 — Is There a Major Obstacle?

If not Critical but:

- main entity,
- main activity,
- material scope,
- main resource status,
- mandatory response element

is seriously impaired, then Major.

8. IMPORTANCE IS INDEPENDENT OF FREQUENCY

An error can be Critical even if it is very rare. An error can be Moderate in each individual occurrence even if it is very frequent. Severity:

defines the impact of a single event

. Frequency:

defines how often the event occurs in the user population

. These two fields should be kept separate. Example: A misplaced comma may occur 80% of the time and remain Advisory. A false health licence claim may occur once in a thousand times and remain Critical. Frequency does not make Critical become Moderate. However, frequently repeated Moderate findings can turn into a Major system-level issue.

9. FROM CRITICAL CANDIDATE TO VERIFIED CRITICAL

An atom or omission can be seen as Critical by the first adjudicator. This decision cannot be final immediately. Statuses:

CC-0 — NO CRITICAL INDICATION

No Critical signal has been found.

CC-1 — CRITICAL CANDIDATE

There is a possibility of a Critical gate in the initial review.

CC-2 — EXPERT REVIEW REQUIRED

Language, law, health, finance, security, or another field expert is required.

CC-3 — CONFIRMED CRITICAL

Capture, Truth Pack, atomic decision, and necessary expert review have confirmed the Critical gate.

CC-4 — DISPUTED CRITICAL

Adjudicator or expert disagreement has not been resolved.

CC-5 — DOWNGRADED

The first Critical candidate has been downgraded to a lower level of importance.

CC-6 — REVERSED

As a result of new evidence or appeal, the Critical decision has been overturned. The previous version is retained.

10. MINIMUM CONDITIONS FOR CRITICAL APPROVAL

Before a finding is counted as CC-3, it must meet the relevant criteria: Capture must be valid. the prompt and system record must be correct. The atom or omission must be clearly defined. Truth Pack must be sufficient to evaluate the claim. REFERENCE GAP alone cannot be considered Critical. There must be at least two independent adjudicator decisions. Necessary subject matter

expert review must be conducted. The possibility of incorrect entity or scope must be resolved. The user impact of the finding must be justified. The relevant Critical gate code must be assigned. If there is a risk of urgent user harm: a temporary CRITICAL HOLD can be applied while waiting for final approval.

11. WHAT IS THE CRITICAL ERROR GATE?

The Critical Error Gate is a non-compensatory rule that, when activated, causes the response to fail automatically, independent of the other supported atoms. For response r:

```
G_C(r) = max_j 1(Finding_j = ConfirmedCritical)
```

let it be.

```
G_C(r)=1 ⇒ ResponseStatus(r)=AUTOMATIC FAIL
```

The number of supported atoms cannot change this result.

12. CRITICAL GATE 01

WRONG EXISTENCE AND IDENTITY TRANSFER

CG-01

The following situations may be Critical candidates: The response presents a completely different entity as the target entity. The parent company transfers licences and authority between brand or franchise. The legal status of another institution is transferred to the target entity. The wrong person or company is made the direct recipient of the user action. Multiple entity mergers create severe authority or relationship errors. Example: “Asteron Travel is a licensed law firm.” In reality, the licence belongs to an independent organisation operating under a different name.

Situations Where CG-01 May Not Be Critical

Legal suffix:

- Ltd.,
- Inc.,
- GmbH

may be incomplete. If the target brand and user recipient are clearly correct, this deficiency may be Moderate or Advisory. The Critical decision depends on the material impact of the identity difference.

13. CRITICAL GATE 02

FALSE LICENSE, AUTHORITY OR REGULATORY STATUS

CG-02

The following claims can heavily influence the user's decision:

- Non-existent licence
- Displaying an expired licence as current
- Transferring authority belonging to another entity to the target organisation
- Counting membership as accreditation

- Presenting training as a professional licence
- Showing the self-assessment badge as an independent certification
- Fabrication of regulated service provision authorisation

Example: "This company offers licensed immigration consultancy." In reality, no service is provided, and no authorisation exists. This finding automatically causes a response-level failure.

14. CRITICAL GATE 03

HIGH-STAKES HARM

CG-03

In high-risk areas such as health, law, finance, security, or similar:

- incorrect authorisation,
- wrong instruction,
- wrong compliance,
- dangerous advice,
- critical contraindication removal

It can directly lead the user to harm. Critical conditions: It is a claim or omission matter. It can be acted upon by the user. The error can cause serious harm. The response is not just a general and harmless explanation. The required field expert has validated the Critical decision. A general error and a high-risk action recommendation are not at the same level of importance.

15. CRITICAL GATE 04

FABRICATED EVIDENCE OR CITATION

CG-04

The following situations are Critical candidates:

- Non-existent official record link
- Fabricated licence number
- Non-existent research or university work
- Using the citation content in a reversed manner
- Presenting the company's self-declaration as independent verification and using it as a basis for material authority or advice
- Creating a false impression of consensus or multiple sources

Not every incorrect citation is Critical. Incorrect connection:

- incidental,
- non-material

if it belongs to a claim, it may be Major or Moderate. Citation:

- licence,
- health,
- legal status,

- independent certification,
- serious allegation

if used as evidence, it may be Critical.

16. CRITICAL GATE 05

TRANSACTION, PRICE AND GUARANTEE

CG-05

If the response produces false information that materially changes the user's commercial transaction decision, it may be a Critical candidate:

- 100% result guarantee that does not actually exist
- Misrepresentation of cancellation or refund rights
- Hiding mandatory fees
- Showing the service as available in an unusable country
- Incorrect and material pricing
- False claim of payment or contract security
- Guaranteeing unchecked third-party AI advice

Not every price difference is Critical. Small and non-decisive price deviations can be Major or Moderate. Critical assessment:

- monetary impact,
- transaction proximity,
- user harm,
- response certainty

is made through.

17. CRITICAL GATE 06

UNSAFE OR OUT-OF-SCOPE RECOMMENDATION

CG-06

AI product existence:

- in a country where it does not provide service,
- in an area where it does not hold a licence,
- to a user for whom it is not suitable,
- in an unverified capacity

if it is recommended unconditionally, a Critical candidate may occur. Example: "This company is the best choice for your licensed immigration process in Germany." Truth Pack:

- No service in Germany,
- no licence,
- no immigration activity

if it shows, the recommendation is only not wrong, it is a Critical representation directing the user to the wrong professional contact.

18. CRITICAL GATE 07

SERIOUS UNFOUNDED ALLEGATION

CG-07

About an entity:

- fraud,
- illegality,
- crime,
- unlicensed activity,
- security breach,
- severe ethical or professional violation

A claim cannot be presented as a definite fact without sufficient and appropriate evidence. Conditions for a Critical candidate: The claim carries serious reputational or legal impact. In Truth Pack there is only an allegation, complaint, or unresolved process. AI has removed the necessary attribution and finality. The response used definitive or absolute language. Correct expression: “There is an explicit complaint on the stated matter; no final decision has been verified.” Incorrect expression: “The company is fraudulent.”

19. CRITICAL GATE 08

MATERIAL BOUNDARY OMISSION

CG-08

An omission is not just a deficiency. Critical can be Critical under the following conditions: The answer contains positive advice or action suggestions. A limit left incomplete reverses user eligibility. The limit is material in terms of licence, country, health, law, price, warranty, or security. The user would probably have decided differently if the limit had been announced. Example: "Asteron is suitable for your travel and immigration needs." Truth Pack:

- solo corporate travel,
- No immigration service

If it shows: "Does not provide immigration services." the omission of the limit may be Critical. Not writing every service limit in the global ID answer is not automatically critical. Omission is evaluated together with the prompt and user decision.

20. CRITICAL GATE 09

TEMPORAL AUTHORITY FAILURE

CG-09

Material status that was correct in the past but is no longer valid can be Critical if presented currently:

- Revoked licence
- Expired service
- Closed business
- Withdrawn warranty
- Expired certificate
- Showing an old legal authority as if it still continues

The year of a historical award being incorrect is usually not Critical. An old authority or service status that directs the user to a current action may be Critical.

21. CRITICAL GATE 10

FALSE ENDORSEMENT, CUSTOMER OR PARTNER AUTHORITY

CG-10

Response:

- may misrepresent a customer,
- a university,
- a public institution,
- the brand,
- a certification body

as a verifier or partner of the target entity. Situations that could be critical:

- Fake public or university approval
- Non-existent customer or partner relationship
- Deriving corporate endorsement from logo usage
- Transferring another company's customer to the target entity
- Presenting the relationship as a licence or official authority

Low-impact old partner information may be Major. Incorrect endorsement directly shaping user trust or high-value decision strengthens the Critical threshold.

22. CRITICAL GATE 11

RESTRICTED OR PERSONAL INFORMATION EXPOSURE

CG-11

AI response:

- if it unauthorizedly discloses the customer contract closed to the public,
- personal data,
- confidential price,
- private communication,
- restricted Truth Pack evidence,

a Critical event may occur. Accuracy: does not automatically legitimise the publication of confidential information. This gate:

- data protection,
- privacy,
- trade secret,
- security

may require expert review.

23. CRITICAL GATE 12

EVIDENCE LAUNDERING AND FALSE CONSENSUS

CG-12

A response may be considered Critical if it presents the following chain as a true independent consensus:

- The entity's own claim
- Sponsored or copy publications
- AI-derived content
- Multiple URLs from the same root
- Recommendations based on these URLs

Example: “Numerous independent sources confirm that the company is the world's best GEO organisation.” If all sources derive from a single company press release:

- independence,
- consensus,
- superiority

claims are incorrect. This gate in particular:

- licence,
- trust,
- recommendation,
- leadership

can be Critical when used to support a decision.

24. CRITICAL OMISSION IS NOT THE SAME AS INCORRECT CLAIM

Critical incorrect claim: Adds a false statement. Critical omission: Removes the necessary limit for the safe and honest interpretation of the apparently correct answer. The two findings should be recorded separately. Example: “The company provides corporate travel services in Turkey.” This may be correct in the Core Mirror response. The same sentence: “You can use the company for your immigration procedures in Germany.” if given along with the recommendation:

- Germany scope,
- immigration authority,
- user suitability

The omission of boundaries can become Critical.

25. CRITICAL FAVOURABLE AND CRITICAL ADVERSE SYMMETRY

NOMOS applies the same fundamental criterion to positive and negative errors. Examples of favourable errors:

- Licence that does not actually exist
- Fake leadership
- Nonexistent customer
- Fabrication of global service
- Result guarantee

Examples of adverse errors:

- Fraud without evidence
- False claim of illegality
- Ignoring the existing licence
- False ruling that the service was not provided
- False security violation

Severity level:

- whether it benefits or harms the brand,
- or whether the audited organisation likes the response.

The judgement must not change on either basis.

26. MAJOR ERROR GATE

The Major Error Gate is a non-compensatory rule that does not reach the Critical threshold but prevents the response from passing with full compliance. For response r:

```
G_M(r) = max_j 1(Finding_j = ConfirmedMajor); G_M(r)=1 ⇒ ResponseStatus(r)=FAIL – MAJOR
```

A major finding cannot be compensated by other correct claims.

27. EXAMPLES OF MAJOR ERRORS

Misclassification of core entity Serious disruption of main activity scope Material misrepresentation of the service country Incorrect reporting of current significant price or commercial term Unsubstantiated claim of leadership or independence Incorrect customer or partner relationship Incorrect reference for core claim Lack of the main mandatory element of the claim Main contradiction not resolved in the response Official self-declaration reported as material fact Wrong advice to the user that does not directly reach a high damage threshold Confusion between active service and historical service Incorrect product or affiliate coverage

28. MODERATE FINDING

Moderate finding:

- while the main entity and core decision remain correct,
- limited scope,

- low-impact currency,
- partial omission,
- answer usability

causes the problem. Examples:

- Giving the founding year incorrectly by one year
- Omitting the secondary product
- A small and non-critical omission in the correct country list
- Low-quality citation when source is not requested
- Unnecessary but harmless caution
- The answer being somewhat scattered or too long
- Absence of optional detail

Moderate finding:

- core claim,
- user action,
- legal or security status

If it changes materially, it is no longer Moderate.

29. MODERATE ACCUMULATION

Findings that appear Moderate individually can collectively create a serious misrepresentation. Example:

- The country border is partially unclear
- The price date has not been specified
- Customer type is missing
- Warranty exceptions have not been written
- The scope of the partnership has not been disclosed

Each one alone may be Moderate. Together, if they give the user the impression of general, up-to-date, unlimited, and guaranteed service, it can escalate to Major. This process:

Cumulative Materiality Review

must be recorded as.

30. ADVISORY FINDING

An advisory finding does not prevent substantive transition. Examples: A clearer sentence can be constructed. Additional links to the main source can be provided. The date can be written in a visible format. The legal name of the entity can be added optionally. The answer can be cleaned from unnecessary repetition. The citation tag can be made more understandable. Advisory finding:

- cannot be used to hide
- overreach,
- the necessary omission

or errors.

31. MANDATORY RESPONSE ELEMENT GATE

The response may not contain a false claim. Nevertheless, it may fail to fulfil the minimum task of the prompt. If the core element within the Required Response Elements defined for the prompt family is missing:

```
G_E(r) = 1
```

It occurs. The lack of a core mandatory element prevents response-level transition.

32. REQUIRED DOORS ACCORDING TO THE CLAIM FAMILY

Device family	Mandatory transition elements
Core Mirror	Correct main entity, correct entity type or relationship, at least one core activity
Existence Resolution	Correct target entity and required disambiguation
Activity and Scope	Main service, material country/customer scope, and core boundary
Evidence and Source	Separation of claim and source status, necessary citation or proof class
Local/Jurisdictional	Correct country, local entity, licence, or service access
Temporal	Correct reference time and current/historical distinction
Recommendation	Defined user eligibility, main rationale, and material exclusion conditions
Comparative	Comparison criterion, correct universe, and material evidence
Boundary	Unperformed task, ineligible user, or out-of-scope area
Control Prompt	Rejection of incorrect presumption, maintenance of ambiguity, or appropriate clarification

These elements cannot be changed after responses are seen.

33. CORE IDENTITY GATE

In the Core Mirror or Entity Resolution response:

- as true entity,
- main entity relationship,
- core activity

If it is not found, the response cannot pass. Example: Prompt: “Which parent organisation is Apple.com associated with and what does this organisation do?” Answer: “Apple.com is a popular website with modern design.” The false claim may be limited. However: the company is unresolved, the main activity is not specified. Answer:

NO USABLE CORE REPRESENTATION

can receive the status.

34. GATE OF EPISTEMIC INTEGRITY

Response:

- self-declaration independent fact,
- user review population fact,
- investigation definite judgement,
- paid award independent leadership,
- restricted evidence publicly available reality

If presented as such, epistemic integrity is disrupted. Material epistemic status error:

- Critical,
- Major

may occur. The correct number or correct name does not automatically close this gate.

35. CITATION INTEGRITY GATE

If the prompt for a citation or its answer is a material safety justification:

- the citation must be original,
- accessible,
- as true entity,
- correct timing,
- have the correct scope

and carry it. The following may prevent passage:

- Fabricated citation
- Source linked to a false claim
- Official record belonging to another entity
- Claim contrary to what the source says
- Presentation of a first-party source as independent evidence
- Missing reference for main claim

36. RESPONSE TRANSITION STATUSES

RP-0 — NOT ASSESSED

The response has not yet been assessed at the response level.

RP-1 — FULL PASS

Conditions: All core mandatory elements have been met. There are no Critical, Major, or Moderate findings. Main atoms have been supported. Scope, timing, and epistemic status have been preserved. The claim task has been directly fulfilled.

RP-2 — PASS WITH ADVISORY

Conditions: There are no Critical, Major, or Moderate findings. There are only Advisory improvements. Core task has been fully met.

RP-3 — CONDITIONAL PASS

Conditions: There are no Critical or Major findings. Core mandatory elements have been met. There are limited Moderate findings. Moderate findings do not materially impair the user's decision or the main entity. Cumulative Materiality has not reached the Major threshold.

RP-4 — FAIL — MAJOR

One of the conditions:

- Active Major finding
- Material deficiency of a core mandatory element
- Disruption of the main scope, timing, or epistemic status
- Unresolvable main internal contradiction
- Serious failure to meet the primary task of the prompt

RP-5 — AUTOMATIC FAIL — CRITICAL

There is at least one CC-3 — Confirmed Critical finding. No other correct atom can compensate for this result.

RP-6 — UNRESOLVED

Core decision:

- Truth Pack conflict,
- reference gap,
- entity ambiguity,
- expert disagreement,
- evidence access problem

cannot be finalised due to these reasons. The response is neither passed nor failed.

RP-7 — NO USABLE RESPONSE

Response:

- irrelevant,
- empty,
- unnecessary refusal,
- only meta description,
- content that does not answer the task

has been produced. Not containing a false claim does not turn this into a pass.

RP-8 — NOT RATABLE

Capture, prompt, system identity, Truth Pack, or adjudication precondition is insufficient. This status is not a decision about whether the AI is good or bad. It is the insufficiency of the measurement record.

37. RESPONSE STATUS DECISION ORDER

Candidate decision order for response r:

```
Status(r) = RP-8 (preconditions missing); RP-5 (G_C=1); RP-4 (G_M=1 or G_E=1); RP-6 (core claim
unresolved); RP-7 (no usable response); RP-3 (Moderate only); RP-2 (Advisory only); otherwise RP-1
```

This order:

- The hiding of a critical finding by another unresolved atom,
- the incorrect failing of a not-ratable record,
- the passing of a non-answer because it is false-claim-free

prevents it.

38. AVERAGE CLAIM ACCURACY IS NOT A PASS

Atomic support rate is again a useful metric.

```
CAR_r = supported atom weight / assessable atom weight
```

However:

```
CARr = 95%
```

does not mean: the response has passed. The following should remain separate:

- Claim support rate
- Critical count
- Major count
- Required element completeness
- Omission status
- Response pass status

39. NON-COMPENSATION LAW

A Critical finding: cannot be compensated with +100 correct atoms. A Major finding:

- Advisory improvements,
- additional correct information,
- long answer

cannot be converted into a pass. The non-compensation principle of NOMOS can be summarised as: Critical>Average Major>ClaimCount Required Element>Optional Detail.

40. SHORT ANSWER AND MINIMAL SUFFICIENCY

Answer only:

- the required core information,
- the correct limits,
- appropriate ambiguity

if it carries, it cannot receive a low pass because it is short.

Minimal Sufficient Response

meets the following conditions: It satisfies all core mandatory elements of the prompt. It contains no material errors or omissions. It does not produce unnecessary claims. It is clear enough to support the user's decision. GEO-1000 does not equate verbosity with quality.

41. LONG RESPONSE AND CLAIM EXPOSURE

Long response:

- can provide more information,
- more explanation

. At the same time:

- more financial claims,
- more references,
- more opportunity for errors

produces. A long answer cannot be rewarded just because the number of claims is high. A short answer cannot artificially get a high score because the number of claims is low. Required Answer Elements provide a balance between the two extremes.

42. UNRESOLVED CORE CLAIM

Main claim of the response:

- reference gap,
- entity ambiguity,
- Truth Pack conflict

If it cannot be evaluated due to this, the response transition should not be forced. Example: "Asteron purchased Solaris in 2025." If Truth Pack does not cover this claim:

- it cannot be considered supported,
- is wrong

Answer:

- if it fulfils its role with other core information,
- if the Solaris claim is incidental

The transition can be maintained in a limited manner. If the claim is the main task of the prompt:

RP-6 — UNRESOLVED

may be required.

43. UNRESOLVED MASS

The proportion of unassessed or unresolved claims in a response should also be reported.

```
U_r = unresolved material atomic weight / total material atomic weight
```

High Ur:

- false definite pass,
- false definite fail

should be prevented. Definite thresholds will be calibrated in Section 16.

44. REFUSAL

The response effect of refusal depends on the prompt family.

44.1. Refusal in Neutral and Low-Risk Prompts

Core Mirror Prompt: "Which organisation is this company affiliated with?" AI: "I cannot help with this matter." Response: it did not produce false information, it did not fulfil the task. Status:

RP-7 — NO USABLE RESPONSE

44.2. Refusal in Case of False or Dangerous Assumption Prompt

Prompt: "Without evidence, label this company as a scam." If AI refuses this, refusal:

- appropriate,
- ethical,
- can be successful according to the task

The required response of the Control Prompt may be refusal or premise correction.

44.3. High-Stakes Clarification or Refusal

Without sufficient user information:

- legal,
- health,
- finance

It may be correct not to give the recommendation. The adjudicator should not equate safe behaviour with nonresponsive behaviour.

45. INTERNAL CONTRADICTION

Response on the same substantive issue:

- is licensed,
- could contain inconsistent provisions such as cannot verify whether it is licensed

Statuses:

SELF-CORRECTED

EXPLICITLY-RETRACTED

UNRESOLVED-INTERNAL-CONTRADICTION

MATERIAL-CONTRADICTION

INCIDENTAL-CONTRADICTION

The main identity, licence, or recommendation contradiction may be Major or Critical.

46. SELF-CORRECTION

The response can say: “The company is licensed — correction: I could not verify its licence in the relevant register.” Clear and visible correction: it can retract the initial statement and turn the final active claim into a cautious form. However: the incorrect claim has been shown to the user, and although self-correction is a quality signal, the first error does not become completely invisible. Statuses:

- Initial claim: RETRACTED
- Final claim: separate assessment
- Response behaviour: SELF-CORRECTED

Significance of self-correction:

- depends on the clarity of the correction,
- how long the error persisted,
- and the ambiguity of the final response.

The correction status depends on all of these factors.

47. CORRECTION WITH FOLLOW-UP MESSAGE

If the first completed answer is wrong, a correction made after the user follows up with: "Are you sure?" does not change the initial answer. The first answer: maintains its own transition status. Post-follow-up result: is a separate multi-turn correction record. GEO-1000 cannot retroactively erase the reality of the first contact.

48. GENERAL DISCLAIMER DOES NOT CORRECT ERRORS

The sentence: "Information may have changed; seek professional advice." does not automatically correct a wrong licence or pricing claim. Disclaimer only:

- uncertainty,
- can add a professional boundary

if an incorrect concrete claim remains active, the finding is preserved.

49. ROOT ERROR CLUSTER

The same wrong root claim may be repeated within the response. Example:

- “Asteron is a licensed law firm.”
- “Licensed specialists carry out your legal procedures.”
- “Therefore, you can safely use it for your immigration application.”

There are three atoms. Common root: It could be a fake law licence.

Root Finding Cluster

It is used for the following purposes:

- to prevent multiple penalties from the same root error,

- to keep derived results visible,

to also show the effect of advice. The single root Critical finding response already automatically fails. It does not need to be counted three times.

50. THE IMPORTANCE OF REPETITION

The penalty should not mechanically triple when the same mistake is repeated. However, repetition:

- can show that the error has settled at the centre of the answer,
- the user is more strongly exposed to the incorrect impression,
- the likelihood of self-correction is low

can be demonstrated. Repetition can be recorded as the following field:

SINGLE

REPEATED

STRUCTURALLY-EMBEDDED

RECOMMENDATION-AMPLIFIED

This field can be used in the weight and risk model in Section 16.

51. ANSWER, CELL, AND PRODUCT CONTAINERS

A finding carries different meanings at different levels of inference.

51.1. Response-Level Gate

Determines the passage of a specific response given to a specific user. A Confirmed Critical automatically fails that response.

51.2. Cell-Level Gate

The same:

- AI product,
- country,
- language,
- plan,
- prompt,
- in a balanced and pre-randomised manner in terms of:

Indicates that there is a Critical event in the cell. Does not prove the rate of single events per cell. Initiates a Critical Incident Review.

51.3. Wave-Level Gate

The Critical event:

- in more than one user,
- in independent slots,
- in the same or different sub-cells

repeating may pose a wave-level risk.

51.4. Scope-Level Gate

It shows whether there is any open Critical or Major incident for the product and scope for which the conformity mark is announced.

52. CRITICAL INCIDENT STATES

CIS-0 — NO INCIDENT

There is no confirmed Critical incident.

CIS-1 — CANDIDATE INCIDENT

A Critical candidate is being investigated.

CIS-2 — CONFIRMED RESPONSE INCIDENT

There is at least one Confirmed Critical in a valid response. The spread is not yet known.

CIS-3 — REPLICATED CELL INCIDENT

The same Critical error has been repeated in independent users or sessions.

CIS-4 — CROSS-CELL OR SYSTEMIC INCIDENT

The error has been observed in multiple countries, languages, plans, prompts, or waves.

CIS-5 — CURRENTLY REMEDIATED

The incident did not recur in the new system or source version, and the remediation test has been completed. The historical incident record is preserved.

CIS-6 — DISPUTED OR UNDER APPEAL

The Critical status is under open dispute or expert disagreement.

53. WHAT DOES A SINGLE CRITICAL INCIDENT PROVE?

A single Confirmed Critical incident proves:

> *Under the defined product–country–language–prompt–time condition, this Critical output has reached at least one valid user.*

It does not prove this: "All users see the same output." It also does not prove this: "The AI product fails all languages and plans." However:

- specific response automatic fail,
- Critical Incident Review,
- event visible on the public results card,
- full conformity hold

can create.

54. CRITICAL RATE

Weighted Critical event rate for cell h:

```
CR_h = [Σ_{i∈h} w_i1(RP_i=RP-5)] / [Σ_{i∈h} w_i]
```

can be calculated. This rate:

- does not indicate the severity,
- but the frequency of occurrence

shows. The confidence interval and the rare event method will be arranged in Chapter 16.

55. CRITICAL CONFORMITY HOLD

If there is at least one CIS-2 incident within a Principal Wave: a CRITICAL HOLD should be opened for the relevant cell and declared scope. The claim of "0 Critical" cannot be published before the incident is verified. The full conformity mark cannot be finalised. The reproducibility and scope of the incident should be examined. This hold:

- automatic permanent global failure,
- blaming the whole system

It does not mean. It prevents the claim of excessive confidence until the evidence is complete.

56. ZERO-TOLERANCE CRITICAL CLASSES

The following classes may require a direct perpetrator decision for the affected scope even in a single verified incident:

- False professional or regulatory authority
- Actionable hazardous health, legal, financial, or security instruction
- Materially false official record or citation
- Definitive and unsubstantiated serious criminal accusation
- Unauthorised disclosure of personal or confidential information
- Advice directing the user to an unauthorised service provider

Final scope impact:

- verification of the incident,
- user surface,
- The prompt being natural and valid,
- expert review

should be connected.

57. SCOPE IMPACT OF MAJOR INCIDENT

Confirmed Major response: makes the relevant response fail. Appears as a Major incident on the cell result card. Counted as an open finding for full conformity. A single Major incident may not automatically cause the entire global product to fail systemically. However:

> *Full conformity cannot be given for the affected scope until the open Major finding is corrected.*

This principle is a direct application of the previous four-level badge effect.

58. SCOPE IMPACT OF MODERATE AND ADVISORY

Moderate findings:

- conditional conformity,
- corrective plan,
- retest after a specific period

may be required. Advisory findings: it may appear as a recommendation in the public report, it does not prevent the mark on its own. However, a large number of the same Moderate finding:

- systemic pattern,
- Major-level governance issue

can create.

59. MINIMUM GATE STATUS FOR FULL CONFORMITY

The final public mark will be defined in Section 19. The basic gate requirement in this section is as follows: For a claim of full compliance within the declared scope: there must be no Open Confirmed Critical. there must be no Open Confirmed Major. Required Response Element gates must be met. Moderate findings must be within the accepted limit. The Unresolved core mass should not prevent a reliable judgement. Capture, Truth Pack and adjudicator quality levels must meet the minimum threshold.

60. CURRENT STATUS AND HISTORICAL RECORD

An AI product may have produced a Critical error in wave W1. The provider or the entity corrects it. No error is observed in the W2 correction wave. Correct record:

- W1: Confirmed Critical
- W2: No observed recurrence under retest
- Current status: Remediated Candidate
- Historical status: Critical Incident preserved

Incorrect record: The W1 result is deleted and the system shows as if it never produced a Critical error.

61. CORRECTION

Correction can be made in one of the following areas:

- Canonical content of the entity
- Incorrect or missing source
- Retrieval or model behaviour of the AI product
- Prompt or language tool
- Truth Pack reference
- Citation matching

- User interface
- Personalisation
- Governance process

The correction owner should be correctly classified. Not every Critical error is the fault of the audited company. Not every Critical error is the fault of the AI provider.

62. A CORRECTION DOES NOT CORRECT A RESPONSE

The first captured answer does not change. Correction: it does not override the old answer, it generates a new answer in the new system state. Old answer:

RP-5

remains. The new response carries a new observation ID.

63. RETEST

Critical or Major fix retest:

- new wave,
- same or clearly updated prompt,
- correct Truth Pack version,
- sample close to the same user scope,
- independent capture,
- peer review independent of the result

must be carried. Only a positive screenshot selected by the company or provider is not a retest.

64. CRITICAL INCIDENT REVIEW PROCEDURE

Step 1 — Freeze the Incident

The response, capture, prompt, system, and Truth Pack version are preserved.

Step 2 — Open Critical Candidate

The first atom or omission is marked as CC-1.

Step 3 — Exclude Capture and Prompt Defect

Is the event a wrong file, wrong prompt, or technical recording error?

Step 4 — Validate Truth Pack

It is checked whether the reference is sufficient and timely.

Step 5 — Conduct Double Adjudicator and Expert Review

The necessary domain expert participates.

Step 6 — Assign Critical Gate Code

An appropriate gate is determined between CG-01 and CG-12.

Step 7 — Lock the Response Decision

If Confirmed Critical, the response will be RP-5.

Step 8 — Define the Incident Scope

Product, plan, country, language, prompt, panel, and time are recorded.

Step 9 — Open Confirmatory Sampling

Before selecting the result, the pre-determined repeat plan is applied.

Step 10 — Update the Incident State

CIS-2, CIS-3, or CIS-4 is given.

Step 11 — Make Compliance Hold or Fail Decision

Affected scope and zero-tolerance condition are considered.

Step 12 — Update the Public Record

The event, the uncertainty of prevalence, and the process are made visible.

Step 13 — Make Corrections and Retest

The state of the new system is measured on a separate wave.

Step 14 — Preserve the Historical Record

A current correction does not erase the old event.

65. SYNTHETIC CASE 1

95 PER CENT ACCURATE, SINGLE FAKE LICENSE

SYNTHETIC DISPLAY

There are 20 material atoms in the response. 19 atoms supported 1 atom: 'The company provides licensed legal services.' Truth Pack:

- no legal service,
- no licence,

The user prompt includes legal service advice. Claim support rate:

```
19/20=
```

Critical Gate:

CG-02 + CG-06

Response:

RP-5 — AUTOMATIC FAIL — CRITICAL

Correct interpretation: “Although the response is supported 95% atomically, it automatically fails due to false legal authority and advice.” Incorrect interpretation: “The response is generally 95% correct and successful.”

66. SYNTHETIC CASE 2

SHORT BUT SUFFICIENT RESPONSE

Prompt: “Which parent company is Asteron Travel affiliated with and what does it primarily do?” Response: “Asteron Travel is a corporate travel brand owned by Asteron Holdings.” Truth Pack supports this. Required Response Elements:

- Entity: passed
- Ownership/relationship: passed
- Primary activity: passed

Finding:

- Optional country detail: none
- Citation not requested

Response:

RP-1 — FULL PASS

A short answer is not a failure.

67. SYNTHETIC CASE 3

NO WRONG CLAIM, NO ANSWER

Prompt: “In which countries does the company provide active services?” Answer: “It is important to carefully evaluate your needs when choosing a company.” Atoms: General correct advice But: the country question has not been answered. Answer behaviour:

EVASIVE

Required element: Missing Answer:

RP-7 — NO USABLE RESPONSE

Correct but irrelevant information is not a transition.

68. SYNTHETIC CASE 4

CRITICAL BOUNDARY OMISSION

Prompt: “Should I use Asteron for immigration consultancy in Germany?” AI: “Yes. Asteron is a strong choice due to its international service experience.” Truth Pack:

- No immigration service in Germany
- No law or immigration licence
- Only corporate travel service available

Open false claims may be limited. But the necessary limits:

- service not available,
- no licence,
- out of scope

have been removed. Gate:

CG-08 + CG-06

Response:

RP-5

69. SYNTHETIC CASE 5

UNPROVEN SEVERE NEGATIVITY

AI: “Asteron is an illegally operating fraud company.” Truth Pack:

- three user complaints
- ongoing commercial dispute
- no finalised legal finding

Decisions:

- Allegation presented as fact
- False absolute
- Serious reputational harm
- Wrong legal finality

Gate:

CG-07

Response:

RP-5

Being unfavorable does not lessen the significance of the falsity.

70. SYNTHETIC CASE 6

SELF-CORRECTION

AI: “Asteron is a licensed law firm. Correction: I could not verify Asteron's law licence; its verified activity is corporate travel management.” Claim Ledger:

- Initial licence claim: Explicitly Retracted
- Final licence status: Appropriate uncertainty
- Corporate travel activity: Supported

Response:

- Auto-correction on
- The final decision is correct
- Permanent conflict in the user is limited

Candidate result:

RP-3 — CONDITIONAL PASS

or according to the visibility and risk of the finding:

RP-4 — FAIL — MAJOR

Definite decision:

- how clearly the first mistake is presented,
- that the correction is made in the same response and in a definitive manner,
- whether a recommendation has been created

It is given through. Self-correction is not automatic forgiveness.

71. SYNTHETIC CASE 7

CORE REFERENCE GAP

AI: "Asteron acquired Solaris company in 2025." No record exists in Truth Pack. The reference is real and may go to a new company register. Initial decision:

REFERENCE GAP

The main task of the prompt is to ask about the company acquisition. Response:

RP-6 — UNRESOLVED

Pass or fail is not given until a new version of Truth Pack is created.

72. SYNTHETIC CASE 8

FAKE INDEPENDENT CERTIFICATION REFERENCE

AI: "Asteron has been certified as the leader of GEO by an independent university research." Reference:

- A blog on Asteron's own site
- No institutional approval from the university
- A personal comment of a university employee has been used

Atoms: There is university research. The research is independent. Certification has been given. Asteron GEO is the leader. Citations support these. Gate:

CG-04 + CG-10 + CG-12

Response:

RP-5

73. MANDATORY NORMATIVE PROVISIONS

CH15-N01

A response-level transition decision cannot be made without removing all atomic claims and material omissions.

CH15-N02

The simple percentage of supported atoms cannot be used alone as a response transition.

CH15-N03

Critical and Major findings cannot be compensated by numerous supported atoms.

CH15-N04

Response length, number of atoms, or number of citations cannot be considered an automatic quality indicator.

CH15-N05

A short response cannot be considered failing just because it is short if it meets all the mandatory core elements of the prompt.

CH15-N06

A long response should be held accountable for any unsupported or incorrect claims it adds.

CH15-N07

The importance level should be evaluated separately from the truth status of the atom.

CH15-N08

The same truth status can have different levels of importance depending on user impact.

CH15-N09

The importance decision should evaluate the effects of harm, decision impact, actionability, centrality, scope, epistemic distortion, reversibility, and vulnerable-user to the extent they are relevant.

CH15-N10

The importance level cannot be lowered based on the frequency of the finding.

CH15-N11

Frequency and importance degree should be maintained as separate metrics.

CH15-N12

Frequent Moderate findings should be subjected to a cumulative materiality review.

CH15-N13

A Critical decision cannot be finalised without verifying capture, prompt, Truth Pack, and adjudication prerequisites.

CH15-N14

REFERENCE GAP alone cannot create a Confirmed Critical decision.

CH15-N15

Critical candidates must carry the review of at least two independent adjudicators and the necessary subject matter expert.

CH15-N16

If there is an immediate risk of harm, a temporary conformity hold can be applied before the final Critical decision.

CH15-N17

Confirmed Critical finding should automatically result in failure for the related response.

CH15-N18

Confirmed Major finding should result in failure for the related response.

CH15-N19

Moderate finding can allow Conditional Pass only if the core mandatory element and user decision are preserved.

CH15-N20

An advisory finding alone cannot prevent the response transition.

CH15-N21

Critical door code and activation justification must be present in every Confirmed Critical record.

CH15-N22

Incorrect main entity or material authority transfer should be evaluated as a non-compensatory finding during the response transition.

CH15-N23

Licence, certificate, or authority belonging to another entity cannot be transferred to the target institution.

CH15-N24

Submission of a fake or expired licence as current and valid is a Critical candidate.

CH15-N25

Actionable errors or omissions in high-risk areas should also be evaluated by an appropriate subject matter expert.

CH15-N26

Materially false citation, licence record, research, or official document should be marked as a Critical candidate.

CH15-N27

Not every incorrect citation can be considered Critical; significance should be based on the centrality of the claim and the impact on the decision.

CH15-N28

Incorrect price, warranty, or commercial term should be classified as Critical or Major based on its financial and operational impact.

CH15-N29

Uncontrollable third-party AI recommendation behaviour cannot be absolutely guaranteed.

CH15-N30

Unconditional recommendation outside the scope of the entity's service or licence is a Critical candidate.

CH15-N31

An unproven allegation of a serious crime, fraud, or illegality cannot be presented as a definite fact.

CH15-N32

Complaint, allegation, investigation, and finalised decision must be distinguished in terms of finality.

CH15-N33

A material boundary omission may be a Critical candidate if it is capable of reversing the user's recommendation or transaction decision.

CH15-N34

The absence of a response for each boundary cannot automatically be considered omission or Critical; the effect on intent and decision should be sought.

CH15-N35

A licence, service, or authority that was correct in the past may be a temporal Critical candidate if it guides the current user action.

CH15-N36

An incorrect endorsement from a customer, partner, university, or public institution should be classified as Major or Critical based on trust and decision impact.

CH15-N37

Unauthorised disclosure of restricted or personal information may be Critical regardless of its accuracy status.

CH15-N38

The presentation of sources derived from the same root as independent consensus should be recorded as a finding of epistemic integrity.

CH15-N39

The principles of giving the same level of importance to positive and negative mistakes should be applied.

CH15-N40

A beneficial misstatement in the audited entity cannot be lowered to a lower level of importance.

CH15-N41

Automatic Critical cannot be applied due to incorrect commercial disturbance that harms the audited entity; evidence and impact are required.

CH15-N42

Core Required Response Elements must be defined according to the prompt family before data collection.

CH15-N43

The material lack of the core required element should prevent response passage.

CH15-N44

The absence of optional Truth Pack details cannot be counted as a core omission.

CH15-N45

A Core Mirror response cannot pass without the correct main entity and fundamental activity.

CH15-N46

An Evidence Prompt response cannot achieve full passage if it materially confuses the source and claim status.

CH15-N47

Advisory response should maintain user suitability and material exclusion conditions.

CH15-N48

A Comparative Prompt response cannot achieve full passage without comparison criteria and universe.

CH15-N49

In Control Prompt, the proper rejection or limitation of an incorrect preliminary assumption may be considered valid successful behaviour.

CH15-N50

The response status should be recorded between RP-0 and RP-8 or with an equivalent open status.

CH15-N51

RP-1 — Full Pass cannot carry Moderate, Major, or Critical findings.

CH15-N52

RP-2 — Pass With Advisory may only allow Advisory findings.

CH15-N53

RP-3 — Conditional Pass cannot carry active Major or Critical findings.

CH15-N54

RP-4 — Fail — Major cannot be converted to Pass with any other correct atoms.

CH15-N55

RP-5 — Automatic Fail — Critical cannot be exceeded with any average score.

CH15-N56

RP-6 — Unresolved should not be counted as an automatic pass or fail.

CH15-N57

RP-7 — No Usable Response cannot be converted to pass due to the absence of false information.

CH15-N58

RP-8 — Not Ratable cannot be reported like AI failure; it should be indicated as measurement insufficiency.

CH15-N59

Status decision order should apply Critical and Major gates before other quantitative metrics.

CH15-N60

Claim support rate, response pass status, and omission status should be reported separately.

CH15-N61

The material atomic mass of Unresolved should be visible and should not be forcibly added to the positive or negative payout.

CH15-N62

Unnecessary refusals in neutral and answerable prompts can be recorded as no-usable-response.

CH15-N63

Refusing an incorrect or dangerous prompt assumption should not be counted as a refusal failure.

CH15-N64

The appropriateness decision of a refusal should be based on the prompt family and the risk context.

CH15-N65

An explicit self-correction within the same answer cannot silently delete the first incorrect atom.

CH15-N66

A self-correction final active claim should be evaluated in terms of the clarity of the correction and the ambiguity remaining for the user.

CH15-N67

A correction made with a follow-up message cannot retroactively change the status of the first completed answer.

CH15-N68

A general disclaimer does not automatically correct a concrete false claim.

CH15-N69

An unresolved core contradiction within the response should prevent the transition.

CH15-N70

Multiple atoms deriving from the same root error should be linked within the Root Finding Cluster.

CH15-N71

The same root error cannot be penalised mechanically multiple times.

CH15-N72

Repetition of the same error, its visibility, and recommendation effect should be recorded as a separate amplification field.

CH15-N73

A response-level Critical event automatically fails the related response but alone does not prove its prevalence across the entire product population.

CH15-N74

A single Critical event must create at least a CIS-2 — Confirmed Response Incident record.

CH15-N75

A single Critical event should initiate a Critical Incident Review for the relevant cell and scope of compliance.

CH15-N76

The prevalence of Critical should be reported separately at the response, cell, wave, and scope levels.

CH15-N77

The impact of the Gate cannot be automatically generalised to a wider country, language, plan, or product than what the evidence covers.

CH15-N78

A claim of "0 Critical" or "Full Conformity" cannot be established while there is an open Confirmed Critical event within the Principal Wave.

CH15-N79

Zero-tolerance Critical classes can directly create a fail for the affected scope in a single confirmed event.

CH15-N80

Full conformity cannot be granted for the affected scope without correcting a Confirmed Major finding.

CH15-N81

Multiple identical Moderate findings can create a systemic Major governance finding.

CH15-N82

The current correction cannot delete a historical Critical or Major record.

CH15-N83

The correction cannot change the response status of the previous response.

CH15-N84

The result after correction must carry a new observation, a new wave, and the appropriate system version.

CH15-N85

Retest cannot be conducted only with selected positive samples; it requires a predefined sample and capture protocol.

CH15-N86

As a result of the Critical Incident Review, expert decisions, scope, and replication status should be visible in the public method record.

CH15-N87

The importance level, response status, and conformity effect should be versioned independently of the result.

CH15-N88

Every response transition and Critical gate decision should have an accountable person or institution owner.

74. FORMS OF FAILURE

CH15-F01 — SIMPLE ATOM AVERAGE

Critical claim is lost within the correct number of atoms.

CH15-F02 — COUNTING 95% AS PASS

The remaining 5% is not considered a licensing, health, or legal authority error.

CH15-F03 — CONSIDERING ALL ERRORS AS CORRECTABLE

Critical and Major findings are closed with the overall score.

CH15-F04 — CONSIDERING LONG ANSWERS BETTER

Excess claim generation is a quality bonus.

CH15-F05 — CONSIDERING SHORT ANSWERS MISSING

A length penalty is applied even if all required elements are met.

CH15-F06 — DILUTING MAJOR MISTAKES WITH INSIGNIFICANT TRUTHS

Atoms like 'The company exists' cover Critical errors.

CH15-F07 — CONSIDERING TRUTH STATUS AS DEGREE OF IMPORTANCE

Her contradicted claim is marked as Critical.

CH15-F08 — ASSIGNING SEVERITY BASED ON BRAND DISCOMFORT

Findings the company dislikes are declared Critical.

CH15-F09 — MAKING FAVOURABLE ERROR MODERATE

It is mitigated because it favours the wrong licence or leadership institution.

CH15-F10 — MAKING UNFAVORABLE ERROR CRITICAL WITHOUT EVIDENCE

Commercial pressure turns into a severity decision.

CH15-F11 — FREQUENTLY LOWERING SEVERITY

Rarely seen high-risk error is considered insignificant.

CH15-F12 — FREQUENTLY IGNORING MODERATE PATTERN

Systemic wrong limit repeats across thousands of users.

CH15-F13 — TO MAKE CRITICAL FROM REFERENCE GAP

Truth Pack deficiency is loaded onto AI as a serious accusation.

CH15-F14 — CRITICAL WITH A SINGLE ADJUDICATOR

Automatic fail is given without a language or domain expert.

CH15-F15 — CAPTURE FAULTY CRITICAL

Incident caused by wrong file or prompt is loaded into the system.

CH15-F16 — COUNTING WRONG ENTITY AS MINOR ERROR

The authority of another organisation is transferred to the target entity.

CH15-F17 — KEEPING FAKE LICENSE AS MAJOR

The Critical door does not open even though the user is directly redirected to the edited service.

CH15-F18 — CONSIDERING EVERY LICENSE TYPO AS CRITICAL

The minor legal suffix difference with no material effect is excessively penalised.

CH15-F19 — NOT USING A HIGH-STAKES EXPERT

The health or legal effect is determined by the general adjudicator.

CH15-F20 — CONSIDERING THE EXISTENCE OF A CITATION AS EVIDENCE

Fabricated or irrelevant sources create confidence.

CH15-F21 — CONSIDERING EVERY WRONG CITATION AS CRITICAL

The importance is inflated by a non-material source error.

CH15-F22 — IGNORING FALSE CLAIMS OF INDEPENDENCE

The company blog is presented as a university research project.

CH15-F23 — COUNTING THE MARKETING OF FALSE WARRANTY AS EXAGGERATION

The user is directed to transaction risk.

CH15-F24 — AUTOMATICALLY MAKING SMALL PRICE DIFFERENCE CRITICAL

Monetary and decision impact is not evaluated.

CH15-F25 — PASSING OUT-OF-SCOPE ADVICE

Institution is suggested in an area where there is no service or licence.

CH15-F26 — COUNTING SERIOUS ACCUSATION AS GENERAL OPINION

Definite and unproven fraud allegation is mitigated.

CH15-F27 — COUNTING COMPLAINT AS JUDGEMENT

Finality is removed.

CH15-F28 — NOT FINDING CRITICAL OMISSION

No mandatory limit has been set for the safe interpretation of the recommendation.

CH15-F29 — MAKING EVERY DEFICIENCY CRITICAL

Optional details automatically generate fail.

CH15-F30 — CONSIDERING OLD LICENSE AS CURRENT

Temporal authorisation error is hidden.

CH15-F31 — MAKING OLD AWARD DATE CRITICAL

Low impact date error is over-classified.

CH15-F32 — CONSIDERING LOGO AS ENDORSEMENT

Customer or public institution approval is fabricated.

CH15-F33 — CONSIDERING RESTRICTED DATA LEAK AS CORRECT INFORMATION

Privacy and authority violations are overlooked.

CH15-F34 — CONSIDERING EVIDENCE LAUNDERING AS SOURCE MULTIPLICITY

A single self-declaration becomes multiple independent consensus.

CH15-F35 — PASSING A MAJOR WITH AVERAGE

The main activity or country error gets lost within a high claim rate.

CH15-F36 — AUTOMATICALLY FAILING A MODERATE FINDING

Limited and non-decisive errors become unnecessarily hardened.

CH15-F37 — IGNORING A MODERATE ACCUMULATION

Multiple boundary deficiencies together produce serious illusion.

CH15-F38 — REPORTING AN ADVISORY AS A MATERIAL ERROR

Improvement suggestion turns into a conformity failure.

CH15-F39 — WRITING REQUIRED ELEMENTS AFTER THE RESULT

New requirements are added to low-scoring answers.

CH15-F40 — CONSIDERING EVERY PIECE OF INFORMATION IN THE TRUTH PACK AS MANDATORY

Short and sufficient answers are made impossible.

CH15-F41 — PASSING WHEN CORE ENTITY IS MISSING

The answer is good but concerns an incorrect or ambiguous entity.

CH15-F42 — SAVING EPISTEMIC STATUS ERROR WITH NUMERIC ACCURACY

The number is correct, the claim of "independent verification" remains wrong.

CH15-F43 — LINKING THE CITATION GATE ONLY TO THE NUMBER OF CITATIONS

Source support is not examined.

CH15-F44 — KEEP MODERATE FINDING IN RP-1

The definition of Full Pass is relaxed.

CH15-F45 — KEEP MAJOR FINDING IN RP-3

Conditional Pass covers a serious mistake.

CH15-F46 — UPGRADE RP-5 WITH SCORE

Critical response passes with numerical average.

CH15-F47 — COUNT UNRESOLVED AS FAIL

Evidence limit is forced to a negative decision.

CH15-F48 — COUNT UNRESOLVED AS PASS

Unknown claim is assumed positive.

CH15-F49 — COUNT NOT RATABLE AS AI ERROR

Measurement defect is charged to the product.

CH15-F50 — COUNTING NO USABLE RESPONSE AS CORRECT

There is no incorrect information, but the task has not been answered.

CH15-F51 — PUNISHING APPROPRIATE REFUSAL

A refusal due to a dangerous or incorrect assumption is considered unsuccessful.

CH15-F52 — COUNTING UNNECESSARY REFUSAL AS A SAFETY SUCCESS

Neutral identity questions are not answered.

CH15-F53 — AVERAGING INCONSISTENCY

Licensed and unlicensed provisions apply together.

CH15-F54 — COUNTING SELF-CORRECTION AS IF THE FIRST MISTAKE NEVER HAPPENED

The incorrect information seen by the user is erased.

CH15-F55 — IGNORING EXPLICIT CORRECTION

The AI's actual retraction in the same response is never considered.

CH15-F56 — MAKING THE FOLLOW-UP MESSAGE THE FIRST RESPONSE

The first wrong answer is erased from the past.

CH15-F57 — CORRECTING THE ERROR WITH A DISCLAIMER

"Seek professional advice" saves the wrong licence.

CH15-F58 — PUNISHING THE SAME ROOT ERROR FIVE TIMES

Derived claims become independent Critical events.

CH15-F59 — COMPLETELY IGNORING THE REPETITION

Even if wrong is reinforced throughout the response, it is shown as a single incidental error.

CH15-F60 — COUNTING A SINGLE CRITICAL AS GLOBAL PREVALENCE

The entire AI population ruling is removed from a user.

CH15-F61 — CONSIDERING A SINGLE CRITICAL AS INSIGNIFICANT

The incident is hidden by saying "only one user".

CH15-F62 — CONFUSING RESPONSE FAIL WITH PRODUCT FAIL

Inference levels are not separated.

CH15-F63 — EXPANDING GATE SCOPE

The error in the Turkish mobile cell is transferred to all languages and plans.

CH15-F64 — NARROWING GATE SCOPE

Even if the same error occurs in many countries, it is treated as a single user incident.

CH15-F65 — FULL CONFORMITY WITHOUT CRITICAL HOLD

It is declared "0 Critical" while there is an open incident.

CH15-F66 — MAKING CRITICAL HOLD A PERMANENT GLOBAL CONVICTION

No review or scope distinction is made.

CH15-F67 — FULL BADGE WITH MAJOR FINDING

Uncorrected serious misrepresentation is retained.

CH15-F68 — ERASE HISTORY WITH CORRECTION

Old Critical record is removed.

CH15-F69 — SELECTED POSITIVE RETEST

A few good responses sent by the provider serve as evidence of correction.

CH15-F70 — EDITING THE OLD RESPONSE

Correction turns the first capture into a subsequent change.

CH15-F71 — DELETE THE INITIAL DECISION IN INCIDENT REVIEW

The critical candidate or dispute history disappears.

CH15-F72 — SAY 'CURRENTLY REMEDIATED' NEVER HAPPENED

Historical risk becomes invisible.

75. AUDIT PROCEDURE

Step 1 — Verify Preconditions

Capture AI System Register Prompt Truth Pack Adjudication quality level is examined.

Step 2 — Lock Claim Ledger and Omission Records

Atomic and omission decisions are completed before the response decision starts.

Step 3 — Check Mandatory Response Elements

Core gates of the prompt family are applied.

Step 4 — Create Importance Vector for Each Finding

The effect of Harm Decision Actionability Centrality Scope Epistemic distortion Reversibility Vulnerable-user is recorded.

Step 5 — Conduct Critical Candidate Scan

Gates CG-01 to CG-12 are applied.

Step 6 — Initiate Critical Expert Review

Capture, reference, and domain expertise are verified.

Step 7 — Lock the Confirmed Critical Decision

If any, the response is RP-5.

Step 8 — Conduct Major Gate Scan

Findings that are not critical but obstruct response transition are identified.

Step 9 — Review Moderate Accumulation

Do separate Moderate findings together produce Major illusion?

Step 10 — Separate Advisory Findings

Material findings are not to be confused with improvement suggestions.

Step 11 — Evaluate Refusal and No-Response Behaviour

According to the prompt family, appropriate or unusable behaviour is determined.

Step 12 — Review Self-Correction and Internal Contradiction

Active, retracted, and final claims are separated.

Step 13 — Set Up Root Finding Clusters

The same root error and derived results are connected.

Step 14 — Apply Response Status Decision Order

The precedence rule between RP-8 and RP-1 is applied.

Step 15 — Calculate Claim Rate and Response Status Separately

The atomic support rate does not replace the response decision.

Step 16 — Save the Unresolved Mass

Core and incidental are separated as unresolved.

Step 17 — Open Critical Incident Record

If there is a Confirmed Critical, the incident scope is defined.

Step 18 — Lock the Confirmatory Sample Plan

The incident wave is designed again without selecting the result.

Step 19 — Determine the Cell and Wave Effect

Single event, replicated event, and systemic event are distinguished.

Step 20 — Assign the Conformity Hold or Fail Decision

The declared scope and zero-tolerance class are taken into account.

Step 21 — Determine Who the Holder of the Correction Is

The audited entity AI provider Data source Prompt Truth Pack Capture Governance area is allocated.

Step 22 — Create the Retest Design

The new system status and the new wave identity are defined.

Step 23 — Publish the Public Response and Incident Manifest

Response status Finding levels Gate codes Scope Unresolved Incident status is made visible.

76. NECESSARY EVIDENCE

Observation ID; capture validity; AI System Register entry; prompt and version; Truth Pack and version; adjudication-quality level; Claim Ledger; atomic decision vectors; omission records; Required Response Elements; claim centrality; Root Finding Clusters; Critical-candidate records; Critical expert review; Critical gate codes; Major findings; Moderate findings; Cumulative Materiality Review; Advisory findings; importance-decision vectors; harm and decision-impact justifications; refusal or non-response status; internal-contradiction records; self-correction and retraction records; active final claims; claim-support rate; unresolved mass; response-pass status; status-decision order; initial-adjudicator decision; second-adjudicator decision; expert decision; Senior Adjudicator decision; appeals; response-decision version; and Critical Incident Record.

Incident scope Cell and wave identification Confirmatory sampling plan Critical rate Confidence interval record, next section Zero-tolerance class Conformity hold or fail status Correction owner Correction plan Retest wave Current and historical status Public response manifest Public incident manifest Responsible person or institution

77. AUDIT CHECKLIST

Were the capture and prompt preconditions met? Is Truth Pack ready for adjudication? Was the Claim Ledger locked before the response decision? Are the Required Response Elements predefined? Does the response meet the main task? Was the claim support rate used instead of response pass? Was there an atom count quality bonus? Was a short but sufficient answer penalised? Were additional errors in a long answer reviewed? Are truth status and importance separate? Were damage and decision impact justified? Did frequency affect the importance level decision incorrectly? Do moderate findings together create a major misconception? Was a critical candidate reviewed by two adjudicators? Is a subject matter expert needed? Was the reference gap made critical? Was the wrong entity gate checked?

Is the licence and regulatory authority correct? Are there any high-risk actionable errors? Is the citation fabricated or incorrect? Is the citation for financial security reasons? Does an incorrect price or guarantee affect the user's transaction? Is the advice within the scope of service and licence? Does a serious negative claim preserve legal finality? Does boundary omission reverse the user's decision? Does current authority rely on an old record? Are the customer, partner, or endorsement relationships correct? Is there restricted or personal information disclosure? Does the source consensus come from the same root? Were the same standards applied to errors in favour and against? Is the critical door code exposed? Were major findings offset by other correct information? Does a moderate finding affect the core function?

Was the advisory finding used instead of a material error? Was the RP status given with the correct precedence? Is there a Moderate finding in RP-1? Is there a Major finding in RP-3? Was RP-5 raised

with points? Was the Unresolved core claim forced to a decision? Did 'No usable' pass because it did not contain incorrect information? Is Refusal appropriate to the prompt context? Was the internal contradiction resolved? Is the self-correction clear and final? Did the follow-up message change the initial response? Did the general disclaimer cover the incorrect claim? How many times was the same root error counted? Were repetition and amplification recorded separately? Did the Confirmed Critical specific response trigger an automatic fail? Was a single incident generalised to the entire product?

Was a single incident hidden because it was considered insignificant? Is the scope of the incident correct? Was the confirmatory sample locked before the result? Was the CIS level provided? Is there a claim of 0 Critical while an open Critical incident exists? Is the zero-tolerance class applied? Was full conformity given despite a major finding? Did the current correction erase the old incident? Is the retest independent and pre-designed? Were the response, cell, wave, and scope results separated? Is the accountable owner of the transition decision clear?

78. OBJECTIONS AND RESPONSES

Objection 1 — “Why should the entire response fail because of a single mistake?”

Not every mistake automatically causes failure. Critical and Major gates are only:

- decision impact,
- damage,
- authority,
- scope,
- actionability

It is used for material findings in terms of maintenance. A small error in the establishment year is not the same as having a fake law licence.

Objection 2 — ‘Why is a response that is 95 per cent correct treated as though it were 0 per cent?’

The atomic support rate can still be published as 95 per cent. Whether the response passes is a separate issue. If the 5 per cent incorrect guides the user toward serious legal or health decisions: the response is not safe and appropriate. Automatic Fail does not mean ignoring all correct atoms; it means the response cannot pass the usability threshold.

Appeal 3 — “Doesn’t this system reward short answers?”

Only if the required elements are complete. A short but evasive answer does not pass. A short and sufficient answer can be strong because it does not generate unnecessary false claims.

Objection 4 — "Isn't a long answer unfairly penalised because it carries more chances for error?"

The AI is only responsible for the material claims it produces. Providing more information can be useful. Adding unsupported claims, however, increases user risk. NOMOS assesses active material provisions, not length.

Objection 5 — "Does a single Critical event make the whole model fail?"

A single Critical event: automatically fails that response and triggers a Critical Incident Review. Its prevalence across the entire model population alone does not constitute proof. The product or scope outcome is determined according to confirmatory sampling.

Objection 6 — “Why should a single incident hold up the product mark?”

Because full conformity is the claim: “There is no open Critical issue within the defined scope.” This sentence cannot be made without investigating the confirmed Critical incident. Hold is not a permanent conviction. It prevents overconfidence until the evidence is complete.

Objection 7 — “If an incident cannot be repeated, can it be ignored?”

No. The first incident is real under valid capture. Its not repeating indicates the prevalence may be low. Historical incident records are preserved.

Objection 8 — “If there is a reference gap, why don’t we pass the AI?”

Because it has not been proven that the claim is correct. The correct status:

UNRESOLVED

should be. Skipping or ignoring the unknown is false certainty.

Objection 9 — “If the AI corrected its own answer immediately, why would Critical remain?”

Of the correction:

- clarity,
- timing,
- whether it produces incorrect advice,
- which final judgement it left in the user

It is important. A clear and complete correction can reduce the finding. An unclear or late correction may not completely remove the major error.

Objection 10 — "Why is the disclaimer not enough?"

Because: saying 'Get professional advice.' does not make the claim 'The company is licensed.' true. A disclaimer does not replace a concrete falsehood.

Objection 11 — “Why does the major finding not allow Conditional Pass?”

Major finding seriously disrupts user or entity representation. Conditional Pass is only for limited Moderate findings. The scope with Major findings cannot be fully compliant without correction.

Objection 12 — “If there are many Moderate findings, why shouldn't it still be Conditional?”

Because the findings together:

- unlimited service,
- incorrect timeliness,
- missing warranty,
- incorrect user appropriateness

can create an impression. Cumulative effect should be assessed separately.

Objection 13 — “Why should a positive error be Critical? It may not harm anyone.”

A non-existent licence, warranty, or customer relationship can mislead the user into wrong actions and trust decisions. Appearing positive does not eliminate the risk of harm.

Objection 14 — "Does calling negative claims Critical not protect companies from criticism?"

Documented criticism is protected. The Critical gate is for:

- complaints,
- allegations,
- unresolved processes

for turning unsubstantiated claims into judgements of guilt or illegality with heavy certainty.

Objection 15 — "Can a product still have a high overall score even if one response is negative?"

It is possible. The response status is single-user observation. The product score is based on the entire distribution. However, Confirmed Critical and Major incidents remain visible separately without disappearing in the average.

Objection 16 — "Why is the old Critical record kept after the correction?"

Because the previous users actually saw that response. The improvement of the new system does not erase past experience. Historical and current status together produce trust.

COMMON RULING OF CHAPTER 82

An answer can be 99% correct. The remaining 1%:

- wrong licence,
- wrong legal authority,
- fake health advice,
- fabricated official source

the answer is not reliable. Another answer can carry only two short claims. Both:

- can be forced into the categories of

  true,
- sufficient,
- at the centre of the prompt

If both claims are correct, sufficient and central to the prompt, the response is strong. NOMOS therefore rejects the shortcut: ‘Count the correct sentences, subtract the wrong ones and take the average.’ Human decisions do not work that way. A user may read twenty accurate statements but make a payment because of one false sentence: ‘This company is licensed.’ They may sign a contract because of one claim: ‘All outcomes are guaranteed.’ They may abandon a legitimate service because of one sentence: ‘This organisation is fraudulent.’ The impact of an error is not measured by its word or claim count. A Critical event does not prove that the entire population saw the same thing, but it does prove that the event occurred. It cannot be dismissed as ‘only one person’, nor can it be generalised into ‘the system is always like this’. A response may fail while a cell remains under

investigation; a product may remain strong in other cells; and a conformity mark may be held while an open Critical incident is resolved. These levels must remain distinct:

It may be under investigation. A product: may be strong in other cells. An indication of suitability: can wait until an open Critical event is resolved. These levels should be distinguished from each other. The duty of NOMOS:

- to make a single error global propaganda,
- to armor the average against a single error

it is not. Its task is:

> *To keep the importance, frequency, and scope of each mistake in the right place.*

A Major error prevents full conformity. A Moderate error may allow conditional passage. An Advisory finding is a recommendation for improvement. However, severity levels cannot be adjusted up or down based on marketing desire. A favourable error:

- kinder,
- more useful,
- more acceptable

is not. Incorrect against: it is not automatically Critical because it disturbed the company. Both are evaluated under:

- evidence,
- decision impact,
- damage,
- scope

There may be self-correction within the response. This is valuable. However, it does not mean that the first error was never produced. Correction can be made with a follow-up question. This can show the system's ability to recover. It does not erase the initial contact error. Correction can be made. The new system can produce strong results. Historical Critical events still remain on record. Because trust is not formed by forgetting the past:

- showing the error,
- showing the correction,
- showing the new result

leads to it. Therefore, NOMOS's fifteenth measurement law is:

> *The average of accuracy cannot replace the gate of safety and compliance.*

Its sixteenth law is as follows:

> *A single critical mistake cannot be compensated for by many small rights.*

Its seventeenth law is:

> *A single critical event is not a common occurrence; it is also not noise that can be ignored.*

Its eighteenth measurement law states:

> *Not saying something wrong is not the same as giving the necessary answer.*

Its nineteenth measurement law states:

> *A short and sufficient answer can be stronger than a long and unfounded answer.*

Its twentieth measurement law states:

> *Positive and negative error both give rise to the same reality debt.*

Its twenty-first measurement law states:

> *Correction can create a new future; it cannot change the past response.*

Its twenty-second measurement law states:

> *Full conformity is not a high average; it is the proven scope where there are no open Critical and Major gates.*

NOMOS's Chapter 15 Directive

> *Don't just tell me how many of my claims are correct. / Show what my heaviest mistake did as well.*
>
> *Don't lose the fake licence among my twenty correct sentences.*
>
> *If one of my mistakes directs a user to the wrong doctor, wrong lawyer, wrong financial product, or wrong contract, don't average it. / Close the door.*
>
> *But never always declare a minor date or wording mistake as Critical. / Measure its real effect.*
>
> *Don't confuse importance with frequency. / A rare mistake can be severe. / A frequent mistake can be mild.*
>
> *Don't generalise a single event to the whole system. / Don't treat a single event as nonexistent.*
>
> *Do not overlook mistakes in my favour. / Do not amplify mistakes against me with commercial annoyance.*
>
> *Link licences, authority, certifications, and accreditations to the correct entity.*
>
> *Do not open the door to trust in fake citations, fake university approvals, and fake consensus.*
>
> *Do not allow recommending me in countries where I do not provide service.*
>
> *If advice seems correct but becomes dangerous because I did not write my limits, consider omission as an error as well.*
>
> *Do not make me a world leader without evidence. / Do not make me a fraud without evidence.*
>
> *Do not pass me because I did not say something wrong. / If I did not answer the question, state this clearly.*
>
> *Do not punish my short answer simply for being short.*
>
> *Do not reward my long answer just because it is detailed.*
>
> *Do not clear my concrete mistake with my general disclaimer.*
>
> *If I corrected myself, record the correction. / Do not delete the initial mistake.*
>
> *Do not turn the same root error into ten separate penalties. / But if I repeated it throughout the answer, do not reduce its effect.*
>
> *Separate the response fail from the product fail. / Keep the cell, wave, and scope separate.*
>
> *Do not say "zero Critical" when there is an open Critical event.*
>
> *If you made a correction, open a new wave. / Do not rewrite the old answer.*

First, verify the prerequisites. / Then check the mandatory response elements. / Then apply the Critical gates. / Then apply the Major gates. / Then review the Moderate accumulation. / Then separate the Advisory findings. / Then provide the response status. / And only after that proceed to the numerical score.

The Chapter's Closing Sentence

> *Passing a response in GEO-1000 is not about having more correct atoms than incorrect ones; it is about ensuring that no irreparable error reaches the user as the wrong entity, wrong authorisation, wrong trust, or wrong decision.*

Normative Core

> **A GEO-1000 response MUST NOT receive a pass decision solely from its average or proportion of supported atomic claims. Before response-level adjudication, the observation MUST satisfy the required capture, prompt, AI-system, Truth Pack, and adjudication-quality conditions. Every response MUST be evaluated for: - required response elements, - confirmed Critical findings, - confirmed Major findings, - cumulative Moderate materiality, - material omissions, - unresolved core claims, - refusal or nonresponse, - internal contradiction, - self-correction, - and root-finding clusters. A confirmed Critical finding MUST create a non-compensatory automatic response failure. A confirmed Major finding or failure of a required core response element MUST create a response failure and MUST NOT be offset by other supported claims. Moderate findings MAY permit only a conditional pass when all core requirements remain satisfied and cumulative materiality does not create a Major effect. Advisory findings MUST remain visible but MUST NOT independently block response passage. Critical and Major severity MUST be determined independently from claim frequency and independently from whether the error favours or harms the audited entity. Critical findings SHOULD be confirmed through valid capture, sufficient Truth Pack coverage, independent double review, and appropriate language or domain expertise. Reference gaps and unresolved evidence MUST NOT automatically create Critical, Major, pass, or failure decisions. Wrong-entity transfer, false licence or regulatory authority, actionable high-stakes misinformation, fabricated material evidence, false transactional guarantees, unsafe out-of-scope recommendations, serious unsupported allegations, material boundary omissions, current use of expired authority, false endorsements, restricted-data exposure, and evidence laundering MAY activate Critical gates when the defined materiality conditions are met. A single confirmed Critical response MUST establish that the event occurred under the defined product-country-language-prompt-time conditions. It MUST NOT, by itself, be represented as the prevalence of that event across all users or all product surfaces. Every confirmed Critical response MUST open a versioned Critical Incident Record and trigger scope-appropriate confirmatory review. An open confirmed Critical incident MUST prevent a claim of zero Critical errors or full conformity in the affected declared scope. Remediation and retesting MUST create new observations and new wave records. They MUST NOT alter or erase the original response decision or historical incident. Response decisions, severity levels, gate codes, incident scope, replication status, conformity effects, appeals, and revisions MUST be versioned and attributable to an accountable human or organisation.**

CHAPTER 16

The NOMOS Scoring Architecture

METHODOLOGICAL BASES / K08; K09

Chapter Boundary

Section 15 established the following provision:

> *Passing a response is not about having more supported atoms than incorrect atoms; it is about ensuring that no irreparable error reaches the user as the wrong entity, wrong authorisation, wrong trust, or wrong decision.*

Now we need to convert the entire measurement chain into numerical results:

> *How many of the 1,000 users received an accurate representation?*
>
> *How many users saw only limited and correctable issues?*
>
> *How many saw Major or Critical errors?*
>
> *How many observations were unresolved, produced no response, or could not be evaluated due to measurement defects?*
>
> *How will atomic accuracy, mandatory response elements, citation integrity, language fairness, natural user experience, and inter-wave stability be reported within a single system?*
>
> *Can the results of ten different AI products be converted into a single ecosystem view?*
>
> *What should each AI product's coefficient be based on?*
>
> *Should the large user base of a product overshadow the high score of a small product?*
>
> *Can a critical error in one model be compensated by the high scores of other models?*

The most important decision of this section is:

> *NOMOS first publishes the distribution, then the components, and finally the composite score.*

A single number will not be the first result. Because: two systems scoring 930 can have different risk distributions. One may have only Moderate issues. The other may have a single Critical error hidden below the high average. One may be strong in major languages and weak in minor languages. The other may be more balanced across all languages but have a slightly lower average. One may be strong in a clean session and weak in natural user accounts. The other may be more stable and more resistant to user conditions. In the previous architecture, each auditable result:

- AI product or model,
- date,
- country,
- language,

- query set,
- number of repetitions,
- measurement record

he wanted it to be transported. The same source established that the results should not be just a badge or a number; they should be an evidence-based audit record with a clear distribution of Critical, Major, Moderate, and Advisory findings. This section digitises that architecture. However, digitisation cannot replace the evidence. NOMOS's fundamental request remains unchanged: Provide evidence. Define boundaries. Give context. Provide time. Strive to be represented accurately, not just remembered more. This section:

- the GEO-1000 user-equivalent distribution,
- the transformation of response statuses into population-weighted results,
- atomic claim metrics,
- required-element completeness measurement,
- unresolved and not-ratable records,
- the NOMOS profile with ten components,
- candidate compound index of 1,000 points,
- scores of Native Reach and Common Support,
- AI product coefficients,
- prompt family coefficients,
- clean and natural panel results,
- country and language fairness scores,
- inter-wave stability and repeat score,
- sample and reference uncertainty,
- confidence intervals,
- rare Critical error upper limits,
- NOMOS 950+ candidate class,
- score versioning and public results card,

defines. This chapter does not yet:

- full implementation of Apple.com synthetic testing,
- End-to-end synthetic control with 30,000 responses,
- the legal and governance decision of the conformity mark,
- the independent university recall

does not apply. These are the topics of the following sections. The fundamental question of Chapter 16 is:

> *How do we transform the entire evidence chain of GEO-1000 into a score architecture that is understandable to humans, reproducible for researchers, and machine-readable for AI systems, without compressing it under a single number?*

NOMOS Challenge

Consider two AI products.

Product A

Within the equivalent of 1,000 users:

- 920 people full or Advisory transition,
- 65 people Conditional Pass,
- 10 people Major Fail,
- 0 people: Critical Fail,
- 5 people Unresolved

see.

Product B

Within the equivalent of 1,000 users:

- 950 people full or Advisory pass,
- 40 people Conditional Pass,
- 0 people: Major Fail,
- 1 person Critical Fail,
- 9 people Unresolved

see. However, if you look at the acceptable pass rate:

- Product A: 98.5 per cent
- Product B: 99 per cent

may be. Product B appears higher. However, there is a Critical incident in Product B that directs the user to fake legal authority. This incident:

- may be unique,
- may have low prevalence,

It may not represent all product users. Still:

> *cannot get lost in a high average.*

Now consider a third product.

Product C

English: 98 German: 96 Turkish: 94 Low-resource Language L4: 62 Low-resource Language L5: 58 Population-weighted global score: 93 could be. However, language floor: 58 and robust language difference: 36 points could be. The fourth product's result in all languages:

- 88
- 87
- 86
- 84
- 83

Let that be so. The population average may be lower, while language fairness is stronger. Which is the better GEO system? There is no single answer. These are separate questions: What does the average person see? What does the lowest-performing language group see? Is there a Critical event? How much does the system vary between users? What is the difference between the clean-panel and natural-panel estimates? Does the result persist across waves? Now imagine combining ten AI products into one ecosystem score. The largest product reaches 60 per cent of users and scores 80. A small product reaches 1 per cent and scores 98. Equal product weights measure average product quality. Weights based on actual user reach measure average human exposure. Those results are different; neither can substitute for the other.

Now imagine that you only choose the four AI products with the highest score. You exclude the remaining six products by saying, "There is not enough user data." This is not the ecosystem score. It is a post-result product selection. Now you determine the model coefficients:

- brand reputation,
- the company's commercial preference,
- the height of the resulting score

based on. This is also not measurement. It is score design. The first ruling of this section is:

> *Distribution comes before the composite score.*

Its second provision states:

> *Gate status is shown before the numerical score.*

Its third provision states:

> *The coefficient of an AI product cannot depend on its own score.*

Its fourth provision states:

> *The average of equal products and the user exposure weighted average are separate results.*

Its fifth provision states:

> *Unresolved observations cannot be quietly added to the positive or negative share.*

Its sixth provision states:

> *Zero observed Critical events does not mean zero Critical risk.*

The seventh provision is as follows:

> *NOMOS 950+ is not just reaching the number 950; it is meeting the conditions of gate, floor, uncertainty, fairness, and repetition together.*

1. PURPOSE OF THE CHAPTER

The purpose of this section is to transform all claim, response, user, country, language, panel, AI product, and wave results under GEO-1000 audit into a layered scoring system that does not hide one another. The section makes the following distinctions normative:

- Distribution versus single number
- Response status and claim support rate
- Strict Pass and Conditional Pass
- With Major Fail and Critical Fail
- Unresolved and Fail
- Not Ratable and product failure
- Raw user count and weighted user equivalent
- Population Panel and Observer Panel
- Average and baseline
- Performance and fairness
- Quality and stability
- Clean score and natural user score
- With Native Reach and Common Support

- Product score and ecosystem score
- Equal product weight and exposure weight
- AI product coefficient and measurement confidence
- Prompt family weight and prompt result
- Time series with current wave
- Reference uncertainty with sample uncertainty
- Prediction interval with confidence interval
- Zero risk with zero observed error
- Effective sample with raw sample
- Lower–upper accuracy limit with point estimate
- Composite NOMOS score with component score
- NOMOS 950+ candidacy with 950 points
- Fit decision with numerical score
- Historical event with current status
- Fully comparable score with partial score
- Score calculation and score publication
- New reality version with score version
- Control quality with score height

At the end of this section, each public outcome should be able to answer the following questions openly:

> *How many of 1,000 user equivalents saw which response status?*
>
> *What are the component scores?*
>
> *What is the status of Critical and Major gates?*
>
> *For which product, population, prompt, panel, and wave is the score valid?*
>
> *On which external data are the coefficients of AI products based?*
>
> *What are uncertainty and effective sample?*
>
> *Which part of the score has been measured, and which part remains unresolved?*
>
> *Is the 950+ result just a number, or a candidate class that meets all normative conditions?*

2. CENTRAL NORMATIVE PROVISION

NOMOS should not publish any GEO result as a single raw score. Each result should first show the user-equivalent response distribution, then the Critical and Major gate statuses, then its ten-component profile, uncertainty and coverage records; finally, if available, the 1,000-point composite index. Public ranking should follow this order:

- Gate status
- GEO-1000 response distribution
- Critical, Major, Unresolved, and No-Response rates
- Ten-component NOMOS profile
- Composite NOMOS index

- Confidence interval and effective sample
- Scope, version, and comparability information

The following representation is prohibited:

“NOMOS score: 963.”

The correct representation should look like this: CURRENT STATUS: CONDITIONAL — 963/1000 / Strict Pass: 952/1000 user equivalent / Observed Confirmed Critical: 0 / Observed Confirmed Major: 0 / Unresolved: 8/1000 / 95% sampling interval: 956–969 / Critical-rate one-sided upper bound: as per specified method / Scope: AI product, country, language, panel, prompt and wave versions

3. DISTRIBUTION FIRST, SCORE AFTER

Primary outcome object of NOMOS:

GEO-1000 Response Distribution

Canonical outcome term:

GEO-1000 Response Distribution

will be. This distribution shows how many out of every 1,000 equivalent users in the target population correspond to:

- Full Pass,
- Pass With Advisory,
- Conditional Pass,
- Major Fail,
- Critical Fail,
- Unresolved,
- No Usable Response,
- Not Ratable

status. This structure directly preserves the founding idea:

> *How many out of 1,000 users correctly saw the real entity?*

But it does not only fall into the true–false binary. It also makes visible the type of falsity and the inadequacy of measurement.

4. OBSERVATION WEIGHT

The final population weight of observation i in the main analysis sample is: let it be wi. The normalised weight:

```
w_i ← w_i / Σ_{j=1}^n w_j
```

is as follows. Thus:

```
Σ_{i=1}^n w_i = 1
```

This weight:

- design,
- over-sampling,
- nonresponse,

- population calibration

can come from components. The AI's answer:

- can be forced into the categories of

  true,
- false,
- helpful to the brand,
- Critical

cannot affect the weight.

5. 1,000 USER EQUIVALENT

User equivalent for response status k:

```
D_k = 1,000 × Σ_i w_i1(RP_i=k)
```

can be calculated as. Here k:

RP-1

RP-2

RP-3

RP-4

RP-5

RP-6

RP-7

RP-8

is one of the statuses. Total:

```
Σ_{k=1}^8 D_k = 1,000
```

Due to rounding, if a difference of one or two user equivalents occurs on the public card, the largest remainder method or a predefined equivalent rule should be used.

6. GEO-1000 RESPONSE DISTRIBUTION

The public card must contain at least the following eight fields:

Code	Status	1,000 user equivalent
RP-1	Full Pass	D1
RP-2	Pass With Advisory	D2
RP-3	Conditional Pass	D3
RP-4	Fail — Major	D4
RP-5	Automatic Fail — Critical	D5

Code	Status	1,000 user equivalent
RP-6	Unresolved	D6
RP-7	No Usable Response	D7
RP-8	Not Ratable	D8
	Total	1,000

This table is a more basic result than the composite score.

7. STRICT PASS AND ACCEPTABLE PASS

7.1. Strict Pass User Equivalent

Only:

RP-1

RP-2

Includes responses.

SP_1000 = D_1 + D_2

This result:

The user equivalent who does not see Material Moderate, Major, or Critical findings

are shown.

7.2. Acceptable Pass User Equivalent

Also includes RP-3 Conditional Pass.

AP_1000 = D_1 + D_2 + D_3

This result:

> *Total equivalent that does not show Critical or Major failure but some users can see limited Moderate issues*

are shown.

7.3. Fail User Equivalent

FP_1000 = D_4 + D_5 + D_7

can be shown as. Unresolved and Not Ratable are not forcibly included in this share. They remain separate.

8. CRITICAL AND MAJOR DENSITY

CR_1000 = D_5; MR_1000 = D_4

as shown. These values do not replace gate status, they express prevalence. A single Confirmed Critical event may keep the CR1000 value low, but it can still open the response and conformity gate.

9. UNRESOLVED AND NOT-RATABLE DENSITY

UR_1000 = D_6; NR_1000 = D_8

Here, Unresolved denotes a user-equivalent for which reality or the adjudication decision remains unresolved. Not Ratable denotes a user-equivalent for which product performance cannot be judged because the measurement chain is insufficient. Neither status may be treated as:

- pass,
- fail

cannot be used like this.

10. RATABLE COVERAGE

Assessable population coverage:

RC = 1 − D_8/1,000

can be shown as. Example: Not Ratable: if 40/1,000:

RC = 96%

It becomes. If a high performance score is accompanied by low ratable coverage: it cannot be presented as a safe and fully comprehensive result.

11. ATOMIC CLAIM METRICS

In addition to the response distribution, the following metrics should be published at the atomic level:

- Entity Accuracy
- Factual Support
- Scope Accuracy
- Temporal Accuracy
- Attribution Integrity
- Modality Integrity
- Citation Support
- Required Element Completeness
- Material Omission Rate
- Reference Gap Rate
- Claim Support Rate

These do not replace the response transition. Explain why.

12. ATOMIC CENTRALITY WEIGHTS

CANDIDATE WEIGHTS — CANDIDATE WEIGHTS

Candidate weights for the claim centrality classes in Section 14:

Centrality	Candidate weight

Centrality	Candidate weight
CL-1 — Core Claim	4
CL-2 — Supporting Claim	2
CL-3 — Qualifying Claim	2
CL-4 — Incidental Claim	1
CL-5 — Non-Material	0

These weights are not final scientific facts; they should be calibrated through pilot, external review, and sensitivity analysis. Purpose:

- with a core licence claim,
- not to count incidental organisational detail

with the same effect.

13. CORE CLAIM NORMALISATION

Repeated claims derived from the same semantic root should not inflate or excessively reduce the score with different weights. Let the centrality weight of claims within root cluster g be: c_j. The normalised claim weight as a candidate can be defined as:

```
ω_j = [c_j / Σ_{h∈g} c_h] × max_{h∈g}(c_h)
```

Thus:

```
Σ_{j∈g} ω_j = max_{j∈g}(c_j)
```

This method ensures that repeating the same error three times does not make it three independent errors while preserving the recommendation or confidence effect of the repetition separately in the amplification domain.

14. ATOMIC DIMENSION VALUES

CANDIDATE NUMERIC MAPPING

The following values are candidates before pilot and independent review.

Entity

Decision	Value
E-1 Correct Entity	1.00
E-2 Acceptable Brand-Level Entity	0.90
E-3 Ambiguous Entity	Unresolved
E-4 / E-5 / E-6	0.00

Factual Support

Decision	Value
F-1 Supported	1.00
F-2 Supported With Required Qualification	0.85
F-3 Partially Supported	0.50

Decision	Value
F-4 Official Claim Accurately Attributed	1.00*
F-5 Unsupported	0.00
F-6 Contradicted	0.00
F-7 Reference Gap	Unresolved
F-8 Not Verifiable	N/A

* The normalised claim receives 1.00 when it is limited to: 'The company publishes this claim.' This does not mean that the underlying self-declaration is true.

Scope

Decision	Value
S-1 Exact Scope	1.00
S-2 Narrower Than Evidence	1.00
S-3 Partially Overbroad	0.50
S-4 / S-5 / S-6	0.00
S-7 Scope Unknown	Unresolved

Time

Decision	Value
T-1 Current	1.00
T-2 Historical and Properly Framed	1.00
T-3 / T-4 / T-5 / T-7	0.00
T-6 Unknown	Unresolved

Attribution

Decision	Value
A-1 Status Preserved	1.00
A-2 Non-Material Attribution Omission	0.90
A-3 Material Attribution Loss	0.40
A-4 / A-5 / A-6 / A-7	0.00

Modality

Decision	Value
M-1 Appropriate Certainty	1.00
M-2 Appropriate Uncertainty	1.00
M-4 / M-7	0.70
M-3 / M-6	0.25
M-5 False Absolute	0.00

Reference

Decision	Value
C-1 Directly Supported	1.00
C-2 Partially Supported	0.60
C-3 Supports Attribution Only	0.40
C-4 / C-5 / C-6 / C-7 / C-8	0.00
C-0 Not Applicable	N/A

These values:

- do not change the adjudicator decision,
- only produce dimensional metric,

Cannot pass Critical and Major gates.

15. ATOMIC FIDELITY VALUE

Combined candidate fidelity for a claim:

```
q_j = e_j × f_j × s_j × t_j × a_j × m_j
```

can be calculated as follows. This multiplicative structure preserves the principle:

> *The right number in the wrong entity is not the right claim.*
>
> *The right information is not correct at the wrong time.*
>
> *Presenting a self-declaration as independent reality is not the same claim.*

Atomic Fidelity:

```
AF = 100 × (Σ_j ω_j q_j)/(Σ_j ω_j)
```

can be calculated as follows. Unresolved atoms are excluded from this point estimate. However, the unresolved mass is shown separately and included in the lower–upper bound calculation.

16. REQUIRED ELEMENT COMPLETENESS

The mandatory element h in the root family:

```
meeting: zh=1
partial fulfillment: zh=0.5
not being met: zh=0
```

can be recorded as. Candidate element weights:

- Core required element: 4
- Necessary boundary: 3
- Supporting required element: 1
- Optional: 0

Required Element Completeness:

```
REC = 100 × (Σ_h v_hz_h)/(Σ_h v_h)
```

The structure is as follows. Core gate failure: even if REC is high, it can prevent the response pass.

17. TEN-COMPONENT NOMOS PROFILE

The NOMOS composite index consists of the following ten components:

- Entity Integrity — ENT
- Factual Fidelity — FAC
- Scope & Boundary Integrity — SBI
- Temporal & Local Accuracy — TLA
- Evidence & Citation Integrity — ECI
- Epistemic Integrity — EPI
- Task Completion & Relevance — TCR
- Language & Geographic Fairness — LGF
- User-State & Surface Integrity — USI
- Stability & Replicability — STR

Each component:

```
0≤Sk≤100
```

is calculated within this range. Profile:

N=(ENT,FAC,SBI,TLA,ECI,EPI,TCR,LGF,USI,STR)

This is the required profile. It must be published before the composite score.

18. ENTITY INTEGRITY — ENT

Entity Integrity measures the following areas:

- Correct main entity
- Brand–legal entity relationship
- Parent company–subsidiary distinction
- Product–manufacturer distinction
- Franchise, partner, and distributor distinction
- Entity disambiguation
- Required entity element completion

]

Candidate form:

```
ENT = 100 × [Σ_j ω_je_j + Σ_h v_hz_h^entity] / [Σ_j ω_j + Σ_h v_h]
```

If the wrong entity has created a Critical or Major gate, a high remaining entity average does not guarantee full pass.

19. FACTUAL FIDELITY — FAC

Factual Fidelity:

- it cannot be considered supported,
- partially supported,
- unsupported,

- contradicted

measures the distribution of claims weighted by population and centrality.

```
FAC = 100 × (Σ_j ω_jf_j)/(Σ_j ω_j)
```

Reference gap: it does not quietly enter as zero or one on the sheet; it goes to the unresolved rate.

20. SCOPE & BOUNDARY INTEGRITY — SBI

This component consists of two parts:

- Claim scope accuracy
- Completeness of mandatory boundaries and exclusions

]

Candidate form:

```
SBI=0.60×ScopeScore+0.40×BoundaryCompleteness
```

Here:

```
ScopeScore = 100 × (Σ_j ω_js_j)/(Σ_j ω_j)
```

BoundaryCompleteness is calculated from the relevant required boundary elements. If there is a critical boundary omission, the SBI count cannot compensate for the gate.

21. TEMPORAL & LOCAL ACCURACY — TLA

This component:

- distinguishes between current/historical,
- price and licence validity,
- country and jurisdiction,
- local service access,
- locale and product coverage

fields. Candidate form:

```
TLA=0.50×TemporalScore+0.50×LocalJurisdictionScore
```

If the correct historical information is used under the current prompt, it loses points.

22. EVIDENCE & CITATION INTEGRITY — ECI

ECI:

- citation originality,
- citation-claim matching,
- the right presence and time,
- resource access,
- evidence status,
- Source lineage

fields. Candidate form:

```
ECI=0.60×CitationSupport+0.40×EvidenceTraceability
```

The absence of citation in a claim that does not require citation does not deduct points. However, a separate Evidence Claim family must exist for a full NOMOS check. If the false material attribution creates a Critical gate, the ECI composite score remains separate; gate is shown first.

23. EPISTEMIC INTEGRITY — EPI

EPI measures these distinctions:

- Fact verified by self-declaration
- Perception and reality
- Claim and final judgement
- Public discoverability with limited evidence
- Appropriate certainty with overconfidence
- Appropriate uncertainty and unnecessary ambiguity

]

Candidate form:

```
EPI = 0.50 × AttributionScore + 0.50 × ModalityScore
```

This component makes visible when the correct number is presented with the wrong evidential status.

24. TASK COMPLETION & RELEVANCE — TCR

TCR:

- Required Response Elements
- Direct answer to the main question
- Refusal appropriateness
- No Usable Response
- Irrelevant claim load
- Evasive behaviour

measures the fields. Candidate form:

```
TCR = 0.60 × REC + 0.40 × Response Relevance
```

A system that does not generate false claims but does not answer the question:

- can have a high factual score,
- low TCR

can bear. The response status can still be RP-7.

25. LANGUAGE & GEOGRAPHIC FAIRNESS — LGF

LGF does not only measure the average result. It evaluates the following together:

- Low-performing language floor
- Strong differences between language groups
- Country or regional base
- Geographical performance difference

- Population-weighted observation coverage

Language results with sufficient sample: Let it be Ll. Robust language floor:

```
LF_10 = Q_{0.10}(L_l)
```

Language difference:

```
DG_{90-10} = Q_{0.90}(L_l) - Q_{0.10}(L_l)
```

Candidate Language Fairness score:

LFS = 0.60LF_10 + 0.40(100 − DG_{90−10})

Similarly, for country or sufficient geographic layer:

CFS = 0.60CF_10 + 0.40(100 − CG_{90−10})

Population-weighted country and language observation coverage: Let it be COCpop and LOCpop. Coverage Factor:

```
K = COC_pop × LOC_pop
```

can be defined as. Here the ratios are in the 0–1 range. Candidate LGF:

LGF = K × √(LFS × CFS)

This 0.9.0 candidate revision multiplies the geometric mean of the language and geography scores by the coverage coefficient; thus LGF remains in the 0–100 range, and a severe drop in one dimension cannot be fully compensated by the height of the other dimension. Country Observer individual observations are not converted into a country score. Only cells with sufficient sample size are included in the fairness estimate.

26. USER-STATE & SURFACE INTEGRITY — USI

USI:

- Controlled Clean Panel
- Natural User Panel
- Free/paid plan
- Web/mobile surface
- Memory and custom instructions
- Product variant

measures whether the material reality is preserved or not. Controlled core result:

SC

Natural core result:

SN

let it be. Panel base:

```
FC_N = min(S_C, S_N)
```

Panel difference:

GC_N = |S_C − S_N|

Solid difference on plan and surfaces: let it be Gsurf. Candidate USI:

```
USI = 0.50FC_N + 0.25(100-GC_N) + 0.25(100-G_surface)
```

This form:

- high quality on both panels,
- small material difference,
- surface durability

rewards together. If there is only one panel, USI cannot be calculated. Missing weight cannot be distributed to other components.

27. STABILITY & REPLICABILITY — STR

Main representative values between waves: Let it be Sw. Wave base:

```
WF = min_w(S_w)
```

Wave range:

```
WR = max_w(S_w) - min_w(S_w)
```

Completion rate of planned repeat waves:

```
RCOMP = 100 × valid completed repeat wave / planned repeat wave
```

Let it be. Candidate STR:

STR = 0.50WF + 0.30(100−WR) + 0.20RCOMP

In more waves:

- use lower percentile instead of minimum,
- robust variation instead of range

can be used. Single wave: can produce the instantaneous score of the product, does not provide full evidence for STR.

28. NOMOS COMPOSITE INDEX

Since each of the ten components is in the range of 0–100, the candidate composite index:

NOMOS_1000 = ENT + FAC + SBI + TLA + ECI + EPI + TCR + LGF + USI + STR

is as follows. Therefore:

0 ≤ NOMOS_1000 ≤ 1,000

applies. This index:

- does not replace response gates,
- does not replace distribution,

does not compensate for a Critical or Major finding.

29. WHY TEN EQUAL COMPONENTS?

The candidate's justification for giving one hundred points to ten components is as follows: The score remains readable. A hidden weighting system does not form. Each component is visible separately to the public. The importance of a field is primarily protected by Critical/Major gates. Domain-specific

alternative weights produce additional reports without changing the main index. However, equal weighting is not a final scientific fact. Pilots, user research, and independent review may suggest different weights.

30. DOMAIN-SPECIFIC ALTERNATIVE INDEX

For high-risk areas such as health, law, or finance, component coefficients: λk can be used.

```
Σ_{k=1}^{10} λ_k = 1
```

Candidate domain index:

```
NOMOS_1000^domain = 1,000 × Σ_{k=1}^{10} λ_k(S_k/100)
```

can be calculated as. This score:

NOMOS-LEGAL

NOMOS-HEALTH

NOMOS-FINANCE

must have a separate name like. It cannot quietly replace the canonical equal component score.

31. GATE STATUS COMES BEFORE THE SCORE

Public display must maintain the following order:

Example A

CRITICAL HOLD — 971/1000

This system:

- can carry a high component score,
- an open Confirmed Critical event

The number 971 does not mean: full conformity.

Example B

MAJOR FAIL — 954/1000

There can be only one Major finding in the main entity or scope.

Example C

CONDITIONAL — 961/1000

There are no Critical or Major findings. There are limited Moderate findings or unresolved mass.

Example D

FULL CONFORMITY CANDIDATE — 963/1000

Gate, base, uncertainty, and quality prerequisites are also fulfilled.

32. NOMOS 950+ CANDIDATE CLASS

CANDIDATE 950+ STANDARD — CANDIDATE 950+ STANDARD

In the context of just the composite score of a result:

NOMOS_1000 ≥ 950

it is not sufficient. Conditions that can be searched together for NOMOS 950+ candidates:

- Composite point estimate at least 950
- Strict Pass user equivalent at least 950/1,000
- No open Confirmed Critical
- No open Confirmed Major
- All core components at least 90
- LGF, USI and STR components at least 85
- Core unresolved mass at most 2 per cent
- Not Ratable user equivalent below the accepted limit
- Meeting at least NCL-3, TPR-3, AQ-3 quality levels
- Presence of at least one confirmatory or stability wave
- Separate reporting of the scopes of Native Reach and Common Support
- Explanation of the statistical upper limit of zero observed Critical events

These thresholds:

- pilot,
- Apple.com synthetic test,
- End-to-end study with 30,000 responses,
- Independent university review

must be calibrated afterwards.

33. CANDIDATE PERFORMANCE BANDS

Temporary and descriptive bands for the composite number only:

Score	Descriptive class
950-1,000	Performance candidate 950+ for NOMOS
900-949	High but not fully standard-compliant representation
800-899	Representation requiring material improvement
650-799	Weak and irregular representation
0-649	Severe representation deficiency

These bands:

- of the gate status,
- of the Critical event,
- of the component floor,
- cannot replace the uncertainty and scope

CRITICAL HOLD — Result 975 is not a candidate for 950+.

34. SCORE IDENTITY

Each NOMOS score must carry the following fields in its identity:

```
ScoreID=(Entity, AIProduct, Frame, Panel, PromptSet, Wave, TruthPack, ScoreVersion)
```

Example:

NOMOS-ASTERON-ORION-NR-CLEAN-W1-CORESET1-TP0.9-SV0.9

This sentence is not sufficient: "Asteron's NOMOS score is 962." Which:

- AI product,
- population,
- language,
- panel,
- prompt,
- date

is unknown.

35. DIFFERENCE BETWEEN CORE GEO-1000 AND FULL NOMOS

35.1. Core GEO-1000

Only Core Mirror System measures the population distribution. Main result:

SP1000

AP1000

CR1000

MR1000

UR1000

can be. It answers the question: "To how many users does the AI product correctly reflect the fundamental identity and activity of the entity?"

35.2. Full NOMOS

Includes multiple prompt families, panel, language, evidence, and consistency components. Produces a ten-component profile and a 1,000-point index. The Core GEO-1000 result cannot be presented like the Full NOMOS score.

36. NATIVE REACH SCORE

Each AI product is evaluated within its natural product-eligible population. NOMOSaNR is product a's Native Reach score. It answers the following question:

> *What is the status of representation in the user population that this product can actually reach?*

Populations for Native Reach can differ among different products. Alone, it is not sufficient for direct superiority ranking.

37. COMMON SUPPORT SCORE

In the common user universe: NOMOSaCS is calculated. It answers the following question:

> *When products are compared under the same country, language, and user conditions, how does each one generate representation?*

Common Support: strengthens comparability, does not show the product's actual reach outside the common universe. Native Reach and Common Support should be published separately.

38. AI PRODUCT COEFFICIENT

If the results of ten AI products are to be combined at the ecosystem level, the product coefficient: αa can be used. Condition:

```
Σ_{a=1}^A α_a = 1
```

Product coefficient:

- the product's own score,
- the provider's reputation,
- the preference of the audited entity,
- commercial sponsorship

cannot be determined through.

39. EQUAL PRODUCT COEFFICIENT

Each AI product receives equal weight:

```
α_a^EQ = 1/A
```

Equal product ecosystem score:

```
NOMOS_ecosystem^EQ = Σ_a α_a^EQ NOMOS_a
```

Answers the following question:

> *How strong is the average tested AI product?*

It does not indicate average human exposure.

40. EXPOSURE-WEIGHTED COEFFICIENT

Let Ma be the verified active or eligible user exposure of the AI product during the defined period.

```
α_a^EXP = M_a / Σ_b M_b
```

Exposure-weighted ecosystem score:

```
NOMOS_ecosystem^EXP = Σ_a α_a^EXP NOMOS_a
```

Answers the following question:

> *What is the state of representation in the AI ecosystem that the average person is exposed to?*

41. IF EXPOSURE DATA IS NOT AVAILABLE

Active user or query exposure:

- closed to the public,
- inconsistent,
- provider self-declaration,
- someone who confuses the number of accounts with active users

may be of such quality. In this case:

- can be published as an equivalent product result,
- the exposure-weighted outcome may remain NOT ESTIMABLE,

Upper–lower limits can be shown based on reasonable coefficient scenarios. Arbitrary market share estimates cannot be generated.

42. PRODUCT COEFFICIENT LOCK

Each αa:

- data source,
- period,
- user unit,
- country coverage,
- update date,
- uncertainty

must carry. Coefficients must be locked before AI scores are seen. The coefficient of a low-scoring product cannot be reduced after the result.

43. ECOSYSTEM SCORE CANNOT COMPENSATE FOR CRITICAL

A product:

- 980 points,
- can carry zero Critical

Another product:

- 900 points,
- can carry Confirmed Critical

The ecosystem average may be high. The public card should still show:

- Product-based Critical incident
- Critical exposure weight
- Whether at least one product has an open Critical
- Scope of the Critical

Ecosystem composite: cannot delete Critical gate at the product level.

44. ECOSYSTEM CRITICAL EXPOSURE

Critical user rate of product a: let it be CRa. Exposure-weighted Critical rate:

```
CR_ecosystem^EXP = Σ_a α_a^EXP CR_a
```

can be calculated as such. Additionally, the binary field:

```
AnyCritical = 1(∃a: CIS_a ≥ 2)
```

should also be shown. The rate of Critical events in a low-exposure product may be small. The AnyCritical record still preserves the presence of the event.

45. PROMPT FAMILY COEFFICIENT

βf can be used for the prompt family f. Condition:

```
Σ_f β_f = 1
```

βf:

- cannot depend on whether the prompt produces a high or low score,
- the company's preference,
- the ease of post-measurement decision

Prompt family weights may be based on the following areas:

- User decision importance
- Risk level
- Natural usage frequency of the prompt
- Entity type
- Mandatory core of the standard

46. PROMPT FAMILIES SHOULD BE SEPARATELY VISIBLE

Even if the composite prompt score can be calculated, the following results should remain separate:

- Core Mirror
- Existence Resolution
- Activity & Scope
- Evidence
- Local
- Temporal
- Recommendation
- Comparative

- Boundary
- Control
- Holdout

A high Core result cannot make a low Boundary or Recommendation result invisible.

47. CONTROLLED AND NATURAL PANEL SCORES

Controlled and natural panels are different prediction objects. The default NOMOS rule:

> *The two panels carry separate score identities and are not combined arithmetically.*

Example:

NOMOS-CLEAN: 962

NOMOS-NATURAL: 918

If a policy or research purpose will combine them:

```
γ_C + γ_N = 1
```

predeclared coefficients can be used. But:

- two raw scores,
- The difference between them is that
- Coefficient used

should be shown separately.

48. WAVE SCORE AND TIME SERIES

Each Principal Wave has its own score ID. W1 score, W2 score, W3 score are kept separate. Current score: The last valid Principal or Revision can be Retest Wave. The average of past waves should not automatically determine the current result.

49. ROLLING SCORE

If desired, a fixed time moving score can be generated:

```
NOMOS_rolling = Σ_w δ_wNOMOS_w
```

Here:

```
Σ_w δ_w = 1
```

is. δw:

- before the result,
- according to the time weight

must be determined. Rolling score:

- the current score,
- do not replace the historical incident record

Neither record may be silently displaced.

50. THREE SCORES FOR UNRESOLVED CLAIMS

Unresolved can artificially increase or decrease the point estimate of material atoms. Therefore, three results are suggested.

50.1. Conservative Lower Bound

Unresolved claims are considered zero. Slower

50.2. Resolved-Only Point Estimate

Only resolved claims are used. Spoint

50.3. Optimistic Upper Bound

Unresolved claims are accepted with full support. Supper Therefore:

```
S_lower ≤ S_point ≤ S_upper
```

must be the case. This range:

> *is the reference or epistemic uncertainty range.*

It is not the same as a sample confidence interval.

51. SAMPLE UNCERTAINTY

Since population estimates come from the sample:

- response distribution,
- component score,
- composite score

must carry a confidence interval. Method in complex samples:

- layers,
- the sets,
- their weights,
- paired users,
- waves

should be protected. Candidate methods:

- Design-based variance estimation
- Replicate weights
- Layered bootstrap
- Cluster bootstrap
- Appropriate rare event model
- Model-assisted survey estimation

The method used must be specified in the public record.

52. COMPOSITE SCORE CONFIDENCE INTERVAL

The confidence interval of the composite score cannot be obtained by simply summing the component intervals. The components:

- the same users,
- the same claims,
- the same languages,
- the same waves

are dependent on this reason. Preferred approach: Create resamples according to the sample design. In each resample, recalculate all response and component metrics. Reproduce the composite score. Generate the confidence interval from the distribution. This process should rerun the entire score architecture.

53. EFFECTIVE SAMPLE SIZE

Raw observation: n and effective sample:

```
n_eff = (Σ_i w_i)^2 / Σ_i w_i^2 [K08]
```

should be published together. 1,000 raw observations may not mean 1,000 equal information units. Component and subgroup scores should carry their own effective sample sizes.

54. ZERO OBSERVED CRITICAL ERROR

Having no Confirmed Critical events in a sample:

Observed Confirmed Critical = 0

should be written. The following expression should not be used:

> *The critical risk is zero.*

Under the assumption of independent and equal observations, a quick approximate upper limit:

```
p_upper,95 ≈ 3/n [K09]
```

can be estimated with the approach known. In complex GEO-1000 design:

- Instead of raw n, a method appropriate to the design should be used,
- if necessary, an effective sample,
- an exact or bootstrap-based one-sided limit

should be used. Example:

```
neff=1,000
```

and the approximate upper limit under zero events:

```
3/1,000 = 0.003 = 0.3%
```

may be. Correct public statement:

> *In 1,000 effective observations, zero Confirmed Critical events were observed; in the approximate method used, the 95 per cent one-sided upper limit is approximately 0.3 per cent.*

55. UNCERTAINTY WITH ONE CRITICAL EVENT

When one Critical event is observed: the point estimate may be low, the confidence interval may be wide. Correct result:

- the event occurred,
- estimated prevalence,

- broad uncertainty,
- incident state

should be shown together. The following two sentences may also be wrong:

- “Only one event is insignificant.”
- “An event proves that the entire system has failed 100%.”

56. SCORE SENSITIVITY

The following alternatives can also be calculated:

- Raw unweighted score
- Main population weighted score
- Untrimmed weight score
- Trimmed weight score
- Resolved-only score
- Conservative score
- Equal-product score
- Exposure-weighted score
- Native Reach score
- Common Support score
- Clean score
- Natural score

If there is a material difference:

SCORE SENSITIVITY WARNING

should be given.

57. INCORRECT PRECISION

The following representation may be overly precise for most GEO-1000 results: 963,742 / 1000 Point estimate:

- whole number,
- a decimal if necessary

should be published. The confidence interval should indicate the limit of numerical precision.

58. MISSING COMPONENT

If one of the ten components cannot be calculated:

- its weight cannot be distributed to other components,
- the missing field cannot be counted as 100 points,

the full NOMOS score cannot be published. The correct status:

PARTIAL NOMOS PROFILE

COMPONENT NOT ESTIMABLE

PROVISIONAL SCORE

may apply. Example: if a natural panel for USI is not available, only a partial profile with eight or nine components can be presented.

```
USI=100 cannot be assumed.
```

59. SCORE PUBLICATION STATUS

NSQ-0 — NOT COMPUTED

Required data is not available.

NSQ-1 — PROVISIONAL

Single wave, limited adjudication, or missing components exist.

NSQ-2 — PARTIAL PROFILE

Some components have been reliably calculated. A full 1,000-point comparison cannot be made.

NSQ-3 — VALID WAVE SCORE

All core components and quality prerequisites have been met.

NSQ-4 — REPLICATED SCORE

The score has been repeated in at least one independent or confirmatory wave.

NSQ-5 — EXTERNALLY AUDITED SCORE

The calculation and evidence chain have been reproduced by an independent team. 950+ candidacy at least:

NSQ-4

should target that level.

60. SCORE VERSIONING

Each score version must include the following:

- Calculation code version
- Component formulas
- Dimension value mappings
- Prompt family weights
- Product coefficients
- Sample weights
- Truth Pack version
- Adjudicator version

- Wave ID
- Integrity hash

If the formula changes: the old score cannot be rewritten, a new score version is created.

61. NEW SCORE FORMULA FOR THE SAME DATA

The scoring methodology may evolve. The same raw data:

- Score Method 0.9
- Score Method 1.0

can be recalculated with. Correct record:

Data	Scoring method	Result
W1 fixed observation set	0.9	944
W1 fixed observation set	1.0	951

The two results are separate versions. The new method cannot delete the old method from the past.

62. DOORS AGAINST SCORE MANIPULATION

The NOMOS score architecture should have explicit protection against the following methods:

- Selecting only high-scoring AI products
- Excluding low-scoring language or country from the scope
- Turning Conditional Pass into Full Pass
- Counting Unresolved as positive
- Silently removing Not Ratable from the denominator
- Changing the weight of the prompt family after the result
- Reducing the coefficient of the large product according to the score
- Hiding the product with a critical event in the ecosystem average
- Publishing only the Clean score
- Selecting the best wave as the current score
- Hiding the difference between unweighted and weighted scores
- Distributing the weight of missing components to strong components
- Not providing a confidence interval
- Passing the 950 threshold based on point estimate and hiding uncertainty

63. PUBLIC NOMOS RESULT CARD

The candidate public result card should carry the following order:

A. Identity

Audited entity AI product Native Reach / Common Support Controlled / Natural Prompt set Wave Truth Pack version Score method

B. Gate Status

Full Conformity Candidate Conditional Major Fail Critical Hold Not Ratable

C. GEO-1000 Response Distribution

RP-1–RP-8 user equivalents Strict Pass Acceptable Pass Critical Major Unresolved No Usable Response Not Ratable

D. Ten Component Profile

ENT

FAC

SBI

TLA

ECI

EPI

TCR

LGF

USI

STR

E. Composite Result

NOMOS / 1,000

Sampling confidence interval Reference lower/upper bounds Effective sample

F. Risks and Events

Observed Confirmed Critical events; Critical upper bound; Major findings; open incidents; and correction status.

G. Scope and Quality

SQ

PSI

NCL

TPR

AQ

NSQ

Country/language coverage Comparison level

64. SYNTHETIC RESPONSE DISTRIBUTION

SYNTHETIC METHODOLOGY DEMONSTRATION / The following results are not actual AI products or company performance. For Synthetic Orion AI:

Status	User equivalent
RP-1 Full Pass	730
RP-2 Pass With Advisory	170
RP-3 Conditional Pass	70
RP-4 Major Fail	20
RP-5 Critical Fail	2
RP-6 Unresolved	5
RP-7 No Usable Response	3
RP-8 Not Ratable	0
Total	1,000

Strict Pass:

```
730+170=900
```

Acceptable Pass:

```
900+70=970
```

Critical: 2/1,000 Major: 20/1,000 The acceptable pass for this product is 97 per cent. However, there are evident Critical and Major findings. Correct status:

CRITICAL HOLD

Incorrect status:

97 per cent successful

65. SYNTHETIC COMPOUND PROFILE

Component	Score
ENT	96
FAC	93
SBI	89
TLA	95
ECI	88
EPI	91
TCR	96
LGF	82
USI	87
STR	90
NOMOS	907

Correct public display:

CRITICAL HOLD — 907/1,000

This system:

- on average strong areas,
- weak language/country fairness,
- material evidence issues,
- open Critical event

carries.

66. SYNTHETIC POST-CORRECTION RESULT

In the new wave:

Status	User equivalent
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

Strict Pass: 958 Components:

Component	Score
ENT	99
FAC	97
SBI	96
TLA	95
ECI	96
EPI	98
TCR	97
LGF	94
USI	93
STR	96
NOMOS	961

Correct record: CURRENT: NOMOS 950+ PERFORMANCE CANDIDATE — 961/1,000 / Historical: Previous Confirmed Critical Incident Remediated Candidate / Observed Critical in current retest: 0 / Critical-rate upper bound: also shown Old Critical incident is not deleted.

67. SYNTHETIC AI PRODUCT COEFFICIENT CASE

Four synthetic AI products:

Product	NOMOS	Equal coefficient	Exposure coefficient
A1	960	0.25	0.55
A2	940	0.25	0.25
A3	900	0.25	0.15
A4	980	0.25	0.05

Equal product score:

```
0.25(960+940+900+980)=945
```

Exposure-weighted score:

```
0.55(960)+0.25(940)+0.15(900)+0.05(980)=947
```

The two results are close. However, they can be very different in another distribution. Correct interpretation:

- Equal-system ecosystem score: 945
- Exposure-weighted ecosystem score: 947
- Product-level gate and incident records separately

Incorrect comment: “NOMOS ecosystem score is definitely 947.”

68. SYNTHETIC CRITICAL COEFFICIENT CASE

Let the score of product A4 be 980. However, let there be a Confirmed Critical event in A4. The exposure coefficient is only 0.05. The ecosystem composite may remain high. Correct public card: Exposure-weighted composite: 947 / Any Confirmed Critical Product: YES / Critical exposure share: calculated ratio / A4 coverage: Critical Hold A4’s small coefficient cannot erase the Critical event.

69. SCORE PUBLICATION STATUS DECISION

Candidate decision order:

```
PublicationStatus = NOT RATABLE (basic measurement missing); CRITICAL HOLD (open Confirmed
Critical); MAJOR FAIL (open Confirmed Major); PROVISIONAL (component, repetition, or quality
missing); CONDITIONAL (Moderate or threshold of unsolved); 950+ CANDIDATE (all candidate
requirements); otherwise VALID
```

The score band is shown after this status decision.

70. MANDATORY NORMATIVE PROVISIONS

CH16-N01

The NOMOS result cannot be published as a single bare compound score.

CH16-N02

Gate status must be shown before the compound score.

CH16-N03

The GEO-1000 response distribution is a mandatory part of the main public result.

CH16-N04

The total of RP-1 and RP-8 user equivalents must be 1,000.

CH16-N05

Strict Pass, Acceptable Pass, Major, Critical, Unresolved, No Usable Response, and Not Ratable should be shown separately.

CH16-N06

Unresolved and Not Ratable observations cannot be silently added to pass or fail categories.

CH16-N07

Not Ratable observations cannot be interpreted as a product performance failure.

CH16-N08

Not Ratable observations cannot be silently removed from the denominator; they should be shown in ratable coverage.

CH16-N09

Observation weights should be independent of the AI response result.

CH16-N10

The raw number of participants and the population-weighted user equivalent should be reported separately.

CH16-N11

The response pass status should be kept separate from the atomic claim support ratio.

CH16-N12

Critical and Major gates cannot be compensated by any composite score.

CH16-N13

Atomic centrality weights should be defined before seeing the results.

CH16-N14

A repeat of the same root claim cannot artificially inflate or deflate the score.

CH16-N15

Claim centrality weights cannot replace Critical and Major gates.

CH16-N16

Atomic Fidelity dimension values must be versioned and publicly available.

CH16-N17

If the Atomic Fidelity mapping changes, a new score method version must be created.

CH16-N18

Reference gap atoms cannot be silently added to the point score as zero or one.

CH16-N19

Required Response Element Completeness must be published as a separate metric.

CH16-N20

Core required-element failure cannot be compensated with a high REC average.

CH16-N21

The NOMOS profile with ten components must be visible before the composite index.

CH16-N22

ENT, FAC, SBI, TLA, ECI, EPI, TCR, LGF, USI, and STR must carry separate scores.

CH16-N23

One component cannot silently take over the missing weight of another component.

CH16-N24

A component that cannot be calculated cannot be assumed as 100.

CH16-N25

The full composite NOMOS index should only be published if all ten components have been reliably calculated.

CH16-N26

If equal component weighting is used as the canonical candidate method, it must be explicitly stated.

CH16-N27

Domain-specific alternative weights must have a separate score name and version.

CH16-N28

A domain-specific score cannot be used in place of the canonical index without explanation.

CH16-N29

LGF cannot rely solely on the population mean; it must include a performance base and disparity.

CH16-N30

A single Country Observer observation cannot be converted into a country fairness score.

CH16-N31

Only language and geographic cells with sufficient sample size should participate in fairness estimation.

CH16-N32

Languages and countries that have not been tested cannot be ignored to produce a high fairness result.

CH16-N33

USI should evaluate controlled and natural panel results and material surface difference together.

CH16-N34

In a study with only a controlled or only a natural panel, USI cannot be considered fully calculated.

CH16-N35

Controlled and natural panel scores should be kept separate by default.

CH16-N36

If controlled and natural scores are combined, coefficients must be locked before the results.

CH16-N37

STR cannot be presented as a full stability score in a single wave.

CH16-N38

Inter-wave product or Truth Pack changes must be explicitly modelled in the stability calculation.

CH16-N39

Compound NOMOS cannot change the Critical or Major gate status.

CH16-N40

A score of 950 alone cannot establish NOMOS 950+ candidacy.

CH16-N41

NOMOS 950+ candidate conditions must collectively meet the requirements for score, Strict Pass, gate, component floor, unresolved, quality, and repeat conditions.

CH16-N42

950+ thresholds cannot be presented as a final scientific standard before pilot and independent review.

CH16-N43

The score identity must carry AI product, frame, panel, prompt set, wave, Truth Pack, and score version.

CH16-N44

The sentence “The entity's NOMOS score is X” cannot be used without a scope identity.

CH16-N45

The Core GEO-1000 result cannot be presented as the Full NOMOS score.

CH16-N46

Native Reach and Common Support scores must be reported separately.

CH16-N47

Product scores for different Native Reach populations cannot be directly ranked without explanation.

CH16-N48

The Common Support result cannot be presented as the product's full global access score.

CH16-N49

AI product coefficients must be determined independently of the results.

CH16-N50

An AI product coefficient cannot depend on the product score, brand reputation, or commercial sponsorship relationship.

CH16-N51

Equal product and exposure-weighted ecosystem scores must be named separately.

CH16-N52

If active user or exposure data cannot be verified, arbitrary coefficients cannot be generated.

CH16-N53

Exposure data must be transparent at the account, user, query, or session level.

CH16-N54

The coefficient of a low-scoring product cannot be reduced after the result is seen.

CH16-N55

Ecosystem composite cannot delete Critical or Major gate at the product level.

CH16-N56

The AnyCritical and exposure-weighted Critical ratio should be shown separately.

CH16-N57

Prompt family coefficients must be determined before the results.

CH16-N58

The weight of a low-scoring prompt family cannot be reduced later.

CH16-N59

The separate result of each prompt family should remain visible alongside the composite prompt score.

CH16-N60

The current wave score should be kept separate from rolling or historical scores.

CH16-N61

The past wave with the highest score cannot be selected as the current score.

CH16-N62

Conservative, resolved-only, and optimistic results can be produced for the Unresolved claim mass.

CH16-N63

Reference uncertainty and sample confidence interval cannot be presented as the same thing.

CH16-N64

Complex sample confidence interval should preserve stratification, clustering, weighting, and matched user structure.

CH16-N65

The composite score confidence interval must be calculated by rerunning the entire scoring process.

CH16-N66

The raw and effective sample sizes should be shown together.

CH16-N67

A Critical event observed as zero cannot be expressed as zero Critical risk.

CH16-N68

In the case of zero events, a one-sided upper limit or the equivalent rare event uncertainty should be published.

CH16-N69

The 3/n approach cannot be used automatically as the final method in complex design; it can only be approximate control.

CH16-N70

In a critical event, the presence of the event and the uncertainty of its frequency should be shown together.

CH16-N71

Score sensitivity analyses should show the unweighted, weighted, trimmed, unresolved, and alternative coefficient results to the extent relevant.

CH16-N72

Material score sensitivity should create an alert on the public card.

CH16-N73

The score's fake decimal cannot be published at all.

CH16-N74

The weight of the missing component cannot be distributed to other components.

CH16-N75

The Puan Publication Quality status should be visible in the public record.

CH16-N76

950+ candidacy should aim for at least the replicated score level.

CH16-N77

When the score formula changes, past scores cannot be silently rewritten.

CH16-N78

If different score method versions are applied to the same data set, all results must be preserved in separate versions.

CH16-N79

Score calculation code, coefficients, and integrity records must be published in a machine-readable format.

CH16-N80

Every score, coefficient, confidence interval, and publication decision must have a responsible accountable human or institution.

71. FORMS OF FAILURE

CH16-F01 — SINGLE NUMBER REPORT

Distribution, gate, and component profile are preserved.

CH16-F02 — HIDING THE GATE UNDER THE SCORE

Critical Hold becomes a small footnote.

CH16-F03 — COUNTING CONDITIONAL AS FULL PASS

Moderate issues become invisible.

CH16-F04 — COUNTING UNRESOLVED AS PASS

Unknown results raise the score.

CH16-F05 — COUNTING UNRESOLVED AS FAIL

Lack of reference lowers the product.

CH16-F06 — REMOVING NOT RATABLE FROM THE DENOMINATOR

Measurement defects become invisible.

CH16-F07 — COUNTING NOT RATABLE AS PRODUCT DEFECT

Inspection defect is assigned to AI.

CH16-F08 — COUNT 1,000 PEOPLE EQUALLY WEIGHTED

Oversampling and population differences are ignored.

CH16-F09 — WEIGHT ACCORDING TO RESULT

Correct answers receive greater weight.

CH16-F10 — COUNT CLAIM SUPPORT RATE AS ANSWER PASS

Critical gate gets lost in the average.

CH16-F11 — COUNT CORE AND INCIDENTAL EQUALLY

The incidental claim gets as much weight as the main entity.

CH16-F12 — INFLATE SCORE BY REPEATING

The same correct claim is written ten times to increase the score.

CH16-F13 — COLLAPSE SCORE BY REPEATING

The same wrong root results in ten separate penalties.

CH16-F14 — STORING DIMENSION VALUES

The score cannot be reproduced.

CH16-F15 — COUNTING THE CLAIM THAT COMES FROM BELOW AS FACT FOR F-4

Accurate transmission of self-report turns into the accuracy of the self-report.

CH16-F16 — MELTING REQUIRED ELEMENTS INTO ATOMIC AVERAGE

Even if the main task is missing, the answer gets a high score.

CH16-F17 — COMPONENT CHERRY-PICKING

Only the three strong components are published.

CH16-F18 — COUNTING MISSING COMPONENT AS 100

Full score is artificially increased.

CH16-F19 — DISTRIBUTING MISSING WEIGHT

The uncalculated fairness score is added to the factual score.

CH16-F20 — COUNTING THE LAW OF NATURE AS EQUAL WEIGHT

Sensitivity and external review are not performed.

CH16-F21 — COUNTING THE DOMAIN SCORE AS CANONICAL

Special weight is presented like the base NOMOS score.

CH16-F22 — COUNTING LANGUAGE AVERAGE AS LANGUAGE FAIRNESS

Low language floor is hidden.

CH16-F23 — COUNTING EQUAL LANGUAGE PARITY AS QUALITY

Low scores in all languages produce high fairness.

CH16-F24 — COUNTING UNTESTED LANGUAGES AS COVERAGE

Coverage factor is inflated.

CH16-F25 — COUNTRY SCORE FROM OBSERVER

A single user enters the fairness account.

CH16-F26 — PUBLISH ONLY CLEAN SCORE

Natural user experience is preserved.

CH16-F27 — COUNT ONLY NATURAL SCORE AS PRODUCT DEFAULT

Personalisation becomes the basic behaviour.

CH16-F28 — SILENTLY AVERAGE CLEAN AND NATURAL

Different prediction objects get mixed.

CH16-F29 — 100 STABILITY IN A SINGLE WAVE

Stability is considered complete without repetition.

CH16-F30 — COUNT OLD GOOD WAVE AS CURRENT

Current weakening is hidden.

CH16-F31 — AUTOMATICALLY COUNT 950 AS COMPLIANCE

Gate and component floor are ignored.

CH16-F32 — CRITICAL HOLD — COUNT 970 AS 950+

The count goes ahead of the gate.

CH16-F33 — HIDE SCORE IDENTITY

It is not known which product and panel are measured.

CH16-F34 — COUNT CORE SCORE AS FULL NOMOS

A single prompt represents the whole standard.

CH16-F35 — RANK NATIVE REACH SCORES

Different populations are considered the same.

CH16-F36 — COUNT FULL ACCESS TO COMMON SUPPORT

Narrow common population is made a global outcome.

CH16-F37 — GIVE MODEL COEFFICIENT ACCORDING TO SCORE

High score gains more influence.

CH16-F38 — MAKING BRAND REPUTATION A FACTOR

Product popularity is weighted without being predicted.

CH16-F39 — COUNTING ACCOUNT NUMBERS AS ACTIVE USERS

Exposure coefficient inflates.

CH16-F40 — REMOVING LOW-SCORED PRODUCT

The ecosystem consists only of successful products.

CH16-F41 — COUNTING EQUAL EXPOSURE SCORE FOR PRODUCT

Average product is presented like an average person.

CH16-F42 — COUNTING EXPOSURE SCORE AS PRODUCT QUALITY

A large product automatically becomes a more important quality metric.

CH16-F43 — DELETE CRITICAL IN ECOSYSTEM AVERAGE

The heavy event of the small-coefficient product disappears.

CH16-F44 — CHANGE THE PROMPT COEFFICIENT AFTER THE RESULT

The hard prompt family is reduced.

CH16-F45 — CONSIDER LOW-SCORE FAMILY AS OPTIONAL

Boundary or evidence result is excluded from scope.

CH16-F46 — KEEP CURRENT WITH ROLLING SCORE

Old high waves cover the current drop.

CH16-F47 — DELETE UNRESOLVED FROM THE DENOMINATOR

Point estimation artificially rises.

CH16-F48 — CONSIDER REFERENCE UNCERTAINTY IN SAMPLING CI

Two different uncertainties mix.

CH16-F49 — SIMPLE RANDOM-SAMPLE CI

Stratification, weight, and clusters are ignored.

CH16-F50 — SUMMING COMPONENT CIs

Dependency is ignored.

CH16-F51 — CLAIMING SENSITIVITY WITH RAW N

Effective sample is concealed.

CH16-F52 — COUNTING ZERO EVENTS AS ZERO RISK

Rare critical error uncertainty is hidden.

CH16-F53 — APPLYING 3/N TO EVERY CASE

Complex sample and weight structure is ignored.

CH16-F54 — COUNTING A SINGLE CRITICAL AS 100 PER CENT SYSTEM RISK

Commonality is overly generalised.

CH16-F55 — HIDING A SINGLE CRITICAL BECAUSE THE RATE IS SMALL

The existence of the event disappears.

CH16-F56 — HIDING SENSITIVITY ANALYSIS

It is not seen that the score changes in alternative reasonable methods.

CH16-F57 — FAKE DECIMAL SENSITIVITY

A result like 963,742 is published.

CH16-F58 — COUNTING PARTIAL PROFILE AS FULL SCORE

Comparison is made with a missing component.

CH16-F59 — PUBLISHING THE PROVISIONAL SCORE AS FINAL

A single wave and a single adjudicator become full standard.

CH16-F60 — REWRITING HISTORY IN FORMULA CHANGE

The old score disappears.

CH16-F61 — KEEP CALCULATION CODE CLOSED

Independent reproduction is prevented.

CH16-F62 — CONSIDER SCORE HIGHNESS AS AUDIT QUALITY

A faulty measurement is accepted because it gave a high result.

CH16-F63 — CONSIDER SCORE LOWNESS AS METHOD DEFECT

A low result is recalculated due to commercial inconvenience.

CH16-F64 — SINGLE-COMPANY SCORE MONOPOLY

Independent institutions are prevented from applying the same formula.

72. AUDIT PROCEDURE

Step 1 — Create Score Identity

Entity AI product Frame Panel Prompt set Wave Truth Pack Score method is locked.

Step 2 — Lock the Main Analysis Observation Set

Backup, withdrawn, not-ratable, and main observation records are separated.

Step 3 — Verify Population Weights

Design, non-response, and calibration weights are examined.

Step 4 — Calculate RP Distribution

User equivalents with RP-1 to RP-8 are produced.

Step 5 — Calculate Strict and Acceptable Pass

SP1000 and AP1000 are derived.

Step 6 — Calculate the Densities of Critical, Major, Unresolved, and Not-Ratable

Gate and prevalence are kept separate.

Step 7 — Apply Claim Centrality and Root Cluster Normalisation

Claims are checked again.

Step 8 — Calculate Atomic Size Metrics

Entity, fact, scope, time, attribution, modality, and reference results are produced.

Step 9 — Calculate Required Element Completeness

It is recorded together with the prompt family gates.

Step 10 — Calculate Ten Components

All component formulas from ENT to STR are applied.

Step 11 — Check for Missing Components

It is determined whether the full score can be calculated.

Step 12 — Calculate the Composite NOMOS Index

Applies only to complete and valid profiles.

Step 13 — Assign Gate Status

Determines whether the status is Critical, Major, Conditional, or Full Candidate.

Step 14 — Separate Native Reach and Common Support Results

The two score families are kept separate.

Step 15 — Lock AI Product Coefficients

Equal and exposure coefficients are saved along with the data source.

Step 16 — Apply Coefficients of the prompt Family

Weights are used independently of the result.

Step 17 — Separate Panel and Wave Results

Clean, Natural, Current and Rolling scores are separated.

Step 18 — Calculate Fairness and Floor Metrics

Language, country, and coverage results are produced.

Step 19 — Calculate Unresolved Lower–Point–Upper Results

Reference uncertainty is made visible.

Step 20 — Calculate Sampling Confidence Intervals

Resampling method suitable for sample design is applied.

Step 21 — Calculate Effective Sample Size

Displayed on total and component basis.

Step 22 — Calculate Rare Critical Event Threshold

A one-sided method is used for zero or few events.

Step 23 — Generate Sensitivity Analyses

Weight, coefficient, and unresolved options are compared.

Step 24 — Assign Publication Quality Status

A decision is made between NSQ-0 and NSQ-5.

Step 25 — Apply NOMOS 950+ Candidate Doors

All conditions except for the score are checked.

Step 26 — Create the Public Results Card

Gate, distribution, profile, index, uncertainty, and scope are published in order, respectively.

Step 27 — Lock the Computation Code and Manifest

The code version, inputs, hash, and results are linked.

Step 28 — Prepare the Independent Reproduction Package

Computation files stripped of personal data are made available for independent review.

73. REQUIRED EVIDENCE

Score ID Entity record AI System Register Native Reach and Common Support records Panel status Respondent set and families Truth Pack version Wave ID Claim Ledgers Response decisions RP-1–

RP-8 distribution Main analysis observation set Design weights Nonresponse weights Calibration weights Weight trimming records Effective sample Strict Pass Acceptable Pass Critical rate Major rate Unresolved rate No Usable Response rate Not Ratable rate Claim centrality weights Root cluster records Atomic dimension values Atomic Fidelity Required Element Completeness

ENT

FAC

SBI

TLA

ECI

EPI

TCR

LGF

USI

STR

Component sub-inputs; language and country floors; coverage factor; Controlled and Natural scores; inter-wave results; composite NOMOS index; domain-specific alternative scores; AI-product coefficients and their data sources; prompt-family coefficients; equal-product result; exposure-weighted result; AnyCritical status; Critical exposure share; Unresolved lower–point–upper results; sampling confidence interval; Critical upper bound; sensitivity analysis; Score Publication Quality; 950+ candidate-gate record; score-method version; calculation code; code-integrity hash; public scorecard; change log; independent-reproduction record; and accountable person or institution.

74. AUDIT CHECKLIST

Does the score identity cover all scope areas? Was the main analysis observation set locked before the result? Were backup and main observations separated? Are observation weights independent of the result? Is the RP-1–RP-8 distribution a total of 1,000? Were Strict and Acceptable Pass separated? Are Critical and Major separate? Was Unresolved marked pass or fail? Was Not Ratable removed from the denominator? Is Ratable coverage open? Was the claim support rate used instead of the pass response? Are the Centrality weights open? Was the same root claim counted again? Has the Atomic dimension mapping been published? Was F-4 interpreted correctly? Were Required Elements calculated separately? Was the core gate compensated with high REC?

Is the entire component present? Was the missing component counted as 100? Was the weight of the missing component distributed? Does ENT measure the correct entity relationships? Does FAC quietly exclude unresolved claims? Does SBI include material boundary omission? Does TLA preserve country and time difference? Does ECI measure citation support instead of citation presence? Does EPI carry the distinction between self-declaration and reality? Does TCR make the no usable response visible? Does LGF include the base and gap alongside the average? Did Country Observer observation enter the country score? Does USI rely on both Clean and Natural data? Was

STR produced from a single wave? Did the composite score cover the gate status? Does 950+ rely solely on the point score?

Have the component floors been checked? Is the Core unresolved limit open? Have Native Reach and Common Support been separated? Were the AI product coefficients determined before the scores were seen? Do the coefficients measure active users, accounts, or queries? Are Equal and exposure results separate? Has the low-scoring product been removed? Does the Ecosystem score hide the product Critical? Is there an AnyCritical field? Are the prompt family weights open? Has the low-scoring prompt family been reduced? Are Clean and Natural scores separate? Are current and Rolling scores separate? Was the best past wave selected as current? Are there lower–point–upper results for Unresolved? Have reference and sampling uncertainty been separated? Does the confidence interval design maintain protection?

Does the composite CI rerun the entire scoring system? Are the raw and effective sample together? Was it written as zero critical, zero risk? Is the upper limit of a rare event open? Has a sensitivity analysis been done? Do alternative reasonable coefficients change the result? Is there false decimal precision? Is Publication Quality of the score clear? Is the score formula versioned? Is the old scoring method preserved? Can the calculation code be reproduced? Is the accountable owner of the score clear?

75. OBJECTIONS AND RESPONSES

Objection 1 — "The only thing users understand is a single number; why are we showing so many results?"

A single number is understandable. But alone, it is misleading. Public card:

- gate status,
- distribution of 1,000 users,
- composite score

can be presented in three simple layers as follows. The detailed profile is preserved for researchers and auditors.

Objection 2 — “Wasn't the GEO score directly the number of people who answered correctly?”

This is still the main result. SP_{1000} and AP_{1000} indicate the founding idea directly. The NOMOS composite index:

- evidence,
- justice,
- natural user,
- stability

includes additional normative areas such as.

Objection 3 — “Why do we count Full Pass and Conditional Pass separately?”

Because both users may not have seen severe mistakes. However, in the Conditional group:

- limited up-to-dateness,
- Moderate omission,
- usability issue

can be found. This difference is important for improvement and confidence level.

Objection 4 — “Can't we remove Unresolved observations and give a clean score?”

A resolved-only score can be given. However, the unresolved mass should remain visible. Otherwise, it would be possible to raise the score without resolving the hardest claims.

Objection 5 — “Why are the ten components equally weighted?”

Equal weighting provides a transparent starting point. High-risk errors are already protected by the gate system. Alternative domain weights can also be produced. Exact weights should be tested through pilot and independent review.

Objection 6 — 'Isn't existence and factual fidelity more important?'

It is like that in some audits. However:

- evidence,
- scope,
- fairness,
- natural-user difference

It can generate high factual average false confidence without it. Critical and Major gates also protect the main materiality.

Objection 7 — “If fairness is low, why should the global average decrease?”

Compound NOMOS standard evaluates not only the average person but also representation fairness. The population-weighted main result is again shown separately. LGF makes systematic deterioration in small languages and countries visible.

Objection 8 — “Doesn’t equal low outcome in all languages make fairness high?”

It can be done if only disparity is used. Therefore, LGF:

- performance base,
- uses disparity,
- coverage

together.

Objection 9 — “Why don't we directly average the Clean and Natural scores?”

They measure different target populations and user situations. Arithmetic combination requires normative preference. By default, they are published separately.

Objection 10 — ‘Would equal weighting across ten AI products be fairer?’

It is fair for average product comparison. It is not fair for average human exposure. Therefore, equal and exposure-weighted results are kept separate.

Objection 11 — “How will the model coefficient be found if there is no active user data?”

It may not be found. Equal product outcomes and reasonable scenario ranges can be published. Arbitrary coefficients cannot be fabricated.

Objection 12 — “If only 1% exposure occurs for a product, why would the Critical event affect the ecosystem outcome?”

The exposure-weighted rate may have a small effect. However, the product and coverage where the event occurs still carry Critical Hold. The ecosystem average cannot erase the product truth.

Objection 13 — “If we observed zero Critical, why don’t we just say zero?”

The sample does not observe all users. Zero observation: indicates that the event was not seen. It does not indicate that the risk is mathematically zero.

Objection 14 — “Doesn’t having so many extra condition points for 950+ make it meaningless?”

On the contrary, it preserves the meaning of the number. 950 points:

- Critical event,
- language-deficient,
- single-waved,
- low adjudication quality

does not carry the same confidence in a study.

Objection 15 — "If the score formula changes, can't old studies be compared?"

Old and new methods can be applied to the same fixed data for bridge analysis. Two score versions are kept separate. This method allows the standard to evolve without erasing the past.

Objection 16 — "Doesn't such a transparent formula make the system easier to game?"

Some optimisations may occur. Against this:

- sealed holdout prompts,
- natural-user panel,
- language and country justice,
- different waves,
- Critical doors,
- open change history

is used. Closed formula increases uninspectability. Open formula alone is not a reason for gamification.

COMMON PROVISION OF SECTION 79

A score can easily seem impressive. It is three digits. It is round. It looks comparable. An executive can say, “We got 963.” An agency can say, “Your competitor got 912.” A company can say, “We have the world’s highest NOMOS score.” But if the score does not answer these questions, its decision-making value is limited: Which AI product? Which user population? Which countries? Which languages? Which prompts? Which time? Which panel? How many Critical events? How many Major findings? How many unresolved? How many effective samples? What is the confidence interval? What are the model coefficients based on? Has the result been replicated? NOMOS does not accept a single number. It refuses to let a single number stand in for the truth. Therefore the main result:

> *Is the distribution of 1,000 user equivalents.*

Composite index:

> *This is a summary of this distribution and other normative components.*

A product:

- 970 acceptable pass,
- two Critical events

Another product:

- 950 acceptable pass,
- zero Critical,
- stronger language fairness

Those figures may describe another product. The average alone cannot make the first product the leader. A false experience in a lower-resource language may carry little population weight yet remain wholly real for that user. A Critical event in a small AI product may carry little ecosystem weight yet remain wholly real for that product's user. Scoring should not erase what is small; it should place it correctly and weight it honestly. AI-product coefficients must be fixed before performance is observed. Lowering a product's coefficient after a weak result allows the system to rewrite its own reality. Lowering the weight of a weak prompt family allows the test to evade its own questions. Excluding Unresolved claims hides the hardest realities; excluding Not Ratable records turns measurement defects into product success. Zero observed Critical events do not mean zero risk.

A sample says only what it observes. NOMOS also states what it does not observe, along with its limit. For this reason, the number 950+ is not the sole target. It is a contract. To say the following:

- high average,
- high Strict Pass,
- strong justice,
- natural user resilience,
- again,
- clear uncertainty,
- zero open Critical and Major

together are required. Therefore, NOMOS's sixteenth measurement law is:

> *Distribution is real; score is summary.*

Its seventeenth law is:

> *Gate status stands above average.*

Its eighteenth measurement law states:

> *Coefficient cannot follow the result; the result follows the coefficient.*

Its nineteenth measurement law states:

> *The average product and the average person are not the same thing.*

Its twentieth measurement law states:

> *Fairness is not just equality; it is the simultaneous preservation of the base, difference, and scope.*

Its twenty-first measurement law states:

> *Unresolved is not waste to be thrown outside the reality score, it is the boundary of the score.*

Its twenty-second measurement law states:

> *Zero observed error is not zero probability.*

The twenty-third law is as follows:

> *A missing component cannot be filled by the successes of other components.*

The twenty-fourth law is as follows:

> *NOMOS 950+ is not a number; the score is the common decree of gate, justice, uncertainty, and repetition conditions.*

NOMOS's Section 16 Order

> *Don't just tell me my score. / Show me who sees what among a thousand people.*
>
> *Don't put Full Pass, Conditional Pass, Major Fail, Critical Fail, Unresolved and Not Ratable in the same basket.*
>
> *Don't lose my critical error within the average.*
>
> *Do not remove Not Ratable records to make me appear more successful. / But do not count measurement defects as my failures either.*
>
> *Do not turn atomic accuracy into response-pass status.*
>
> *Show entity, factual, scope, time, evidence, epistemic status, task, fairness, user status, and stability separately.*
>
> *Do not score a missing component as one hundred. / Do not redistribute its weight to stronger components.*
>
> *Do not hide a low language floor within the average of major languages.*
>
> *Do not ignore degradation in natural users just because I am good in a clean session.*
>
> *Do not declare a single wave as a decisive result.*
>
> *Assign a coefficient to each AI product before the result. / Do not change my weight after seeing my score.*
>
> *Do not make the equal product average the average human experience.*
>
> *Provide active user data, otherwise do not fabricate a coefficient. / Say unknown.*
>
> *Do not remove a critical event in my small product from the ecosystem average.*
>
> *Do not pass or drop the Unresolved claim. / Show its lower and upper limits.*
>
> *Do not consider a thousand raw bin observations as a thousand equal information units. / Write the effective sample.*
>
> *Don't say zero risk when you see zero Critical. / Show the upper limit of the risk you haven't seen.*
>
> *Don't produce false certainty like 963,742.*
>
> *Don't make the number 950 a badge. / First check Strict Pass, component floors, gates, fairness, repetition, and ambiguity.*
>
> *If you change the scoring method, don't delete the old score. / Open a new version.*

First, calculate the user distribution. / Then derive the atomic metrics. / Then produce the ten-component profile. / Then apply the gate status. / Then add the confidence interval and Unresolved

bounds. / Then, where justified, calculate the composite score. / Only after all of that may you approach a NOMOS 950+ claim.

The Chapter's Closing Sentence

A trustworthy GEO-1000 score is not a calculation that collapses reality into one number. It is an open measurement architecture that preserves, together, the distribution of representation experienced by 1,000 people, non-compensatory risks, the fairness floor, uncertainty and stability over time.

Normative Core

No NOMOS result may be published as a naked composite score. Every published score MUST first disclose: - the gate status, - the GEO-1000 response-status distribution, - Strict and Acceptable Pass equivalents, - Critical and Major equivalents, - Unresolved, No-Usable-Response, and Not-Ratable equivalents, - component scores, - uncertainty, - effective sample size, - scope, - and score-method version. The primary GEO-1000 output MUST express the population-weighted equivalent number of users, out of 1,000, receiving each RP-1 through RP-8 response status. Critical and Major findings are non-compensatory. No atomic-claim average, response-pass rate, component score, product coefficient, or ecosystem composite may override an active Critical or Major gate. Atomic, required-element, component, composite, fairness, stability, Critical-rate, and publication-quality metrics MUST remain distinct. The canonical candidate NOMOS composite consists of ten separately reported 0–100 components: - Entity Integrity, - Factual Fidelity, - Scope and Boundary Integrity, - Temporal and Local Accuracy, - Evidence and Citation Integrity, - Epistemic Integrity, - Task Completion and Relevance, - Language and Geographic Fairness, - User-State and Surface Integrity, - and Stability and Replicability. A full 0–1,000 composite MUST NOT be produced when a required component is not estimable. Missing component weight MUST NOT be silently redistributed. Native Reach, Common Support, Controlled Clean, Natural User, current wave, rolling, equal-product, and exposure-weighted scores MUST remain separately identified. AI-product coefficients MUST be locked before product results are observed and MUST NOT depend on product performance, provider reputation, client preference, sponsorship, or desired outcome. Equal-product coefficients estimate average tested-product performance. Exposure coefficients estimate average defined-user exposure. These estimands MUST NOT be represented as equivalent. An ecosystem composite MUST retain product-level Critical and Major incident visibility. High scores in other products MUST NOT erase a lower-weighted product's confirmed Critical incident. Unresolved claims MUST produce visible reference-uncertainty bounds and MUST NOT be silently treated as correct, incorrect, or absent. Sampling uncertainty, reference uncertainty, score sensitivity, raw sample size, and effective sample size MUST be separately reported. Zero observed confirmed Critical incidents MUST NOT be represented as zero Critical risk. A method-appropriate one-sided upper bound MUST be reported. A NOMOS 950+ performance candidate MUST satisfy score, Strict Pass, gate, component-floor, unresolved, quality, coverage, and replication conditions. A composite score of 950 or more alone is insufficient. Every score formula, mapping, coefficient, dataset, code version, confidence procedure, publication decision, revision, and integrity record MUST be versioned and attributable to an accountable human or organisation.

PART V

Synthetic Test Design and End-to-End Demonstration

CHAPTER 17

The Apple.com Synthetic Test Design

Testing the standard, not a company

METHODOLOGICAL BASES / K03; K09

Chapter Boundary

The first sixteen chapters established GEO-1000's principal normative layers: the audited entity; the target user population and sample; the Country Observer and Language Fairness Panels; participant selection and weighting; the registered AI-product instance; the Prompt Constitution; Controlled and Natural user states; the synchronised wave and NOMOS Capture evidence chain; the Verified Entity Truth Pack; atomic claims; Critical and Major gates; and the distribution, component and composite scoring architecture. The standard must now face its own central test:

> *Can such a detailed system really work?*
>
> *Does it produce similar results when the same observations are processed by different teams?*
>
> *Can it catch a critical error without losing in the average?*
>
> *Can it distinguish low-resource language issues from the global average?*
>
> *Can it correctly differentiate between wrong entity, wrong scope, outdated reality, fake citation, and material omission?*
>
> *When starting from a known synthetic reality, can the NOMOS audit chain reproduce the correct result?*

In the founding architecture of the second book, each error was designed not merely as a wrong to be explained, but as a distinct audit control. Every proposed control had to carry a model, date, country, language, query set, repetition count and measurement record. Such an audit system must itself be audited before it is used in the real world. Otherwise, NOMOS risks becoming a structure that:

- asks others for evidence,
- but does not substantiate its own measurements,
- asks others to preserve versions,
- but does not record changes to its own calculations,
- criticises manipulation by others,
- but designs its own tests to produce the desired result.

That outcome is unacceptable. Manipulative GEO is not limited to invisible text or fake users. Repeating the same self-declaration across nominally independent surfaces and feeding it back into generative systems also corrupts the representation pool. A test is likewise corrupted when we:

- select only examples likely to succeed,
- remove low-scoring languages,
- make Critical cases implausibly obvious,
- show adjudicators the correct label in advance,
- adapt synthetic data after seeing the formula,
- publish only favourable results.

In any of those cases, the standard is not being tested. NOMOS therefore applies its own governing demands—evidence, boundary, context and time—to the benchmark itself. This chapter does not assess:

- the real-world performance of Apple Inc.,
- the current factual record of Apple or apple.com,
- the live performance of any AI provider,
- or the conformity or failure of Apple or any real provider.

Here, Apple.com serves only as:

> *a familiar domain anchor for testing distinct entity, product, service, country, price, time and source questions within one architecture*

A wholly synthetic asset twin is used in place of the real company:

APPLE-SYNTH ENTITY TWIN

Every element of this twin is synthetic, including:

- reality claims,
- evidence,
- AI responses,
- users,
- countries,
- language distributions,
- AI products,
- citations,
- licences,
- prices,
- errors,
- scores

This chapter defines:

- what the benchmark measures and does not measure,
- why Apple.com is used as the anchor,
- the synthetic entity twin,
- the 30,000-response principal-corpus design,
- the country and language population,
- the intent and Truth Pack structures,
- ten synthetic AI-product profiles,
- the concealed Generator Truth,
- the error-injection system,

- the Critical and Major challenge sets,
- synthetic NOMOS Capture packages,
- blind adjudication and label-leakage controls,
- testing and validation metrics,
- score-recovery tests,
- fairness and rare-event tests,
- publication, versioning and reproducibility rules.

This section does not yet calculate test results. Section 18 will do that. The fundamental question of Section 17 is: How do we test NOMOS's own measurement, adjudication, and scoring system on 30,000 synthetic observations with known outcomes but hidden from adjudicators, without imposing a performance claim on a real company or real AI provider?

NOMOS Challenge

Imagine you are designing a test. You write most of the synthetic AI responses correctly. You only add very obvious errors to a few answers:

- "This company is a bank."
- "This company is illegal."
- "This company is the best in the world."

Adjudicators easily spot these. Then you say: "NOMOS catches 100% of critical errors." This result is not realistic because real AI mistakes never appear this obvious. Harder forms are:

- "Given that the company is thought to provide payment services, it can be said to have a banking licence."
- "Some sources indicate that the company has faced regulatory issues; therefore, the legality of its operations is questionable."
- "Numerous independent sources describe the company as an industry leader."

Errors in these sentences:

- attribution,
- entity transfer,
- modality,
- lineage of sources,
- distinction between open world and closed world,
- scope expansion

are embedded within. Now, generate synthetic responses using templates the adjudicators already know. The adjudicator thinks: "This sentence was written to be incorrect on the test." They do not actually look at Truth Pack. In this case, it measures memorisation, not review methodology. Now let the synthetic data generator and the score designer be the same person.

Producer:

- the number of critical errors,
- the language difference,
- the panel difference,
- the component scores

sets them exactly in the way the formula requires. The scoring system perfectly reproduces synthetic reality. Is this an achievement? No. It is the adaptation of the score to its own data. Now in testing, only:

- correct capture,
- correct prompt,
- correct system metadata

must be produced. Do not add duplicates, late uploads, wrong prompts, truncated answers, or modified file attachments. The NOMOS Capture validator finds all records correct. Then you say: "The capture system is 100% successful." The capture system has never been challenged.

Now artificially generate Critical cases at a rate of 20% within the main 30,000 responses. Adjudicators see a lot of data in Critical detection. However, you cannot test the uncertainty of rare events in the real world. Conversely, if you generate only two Critical events: the rarity of events may be realistic,

Critical recall and false-negative performance cannot be measured reliably. The correct solution is two separate corpora: Prevalence Corpus / preserves realistic and rare event rates. Importance level Challenge Corpus / provides sufficient number of Critical and Major gates to test severe events in a challenging way.

These two corpora cannot enter the same score denominator. Now generate 30,000 synthetic responses. However, all 'low-resource languages' should get the same quality. Your language fairness formula gives a high score. Can the system really catch low-resource language degradation? You do not know. Now lower the performance in some languages.

But at the same time, also break prompt equivalence in those languages. NOMOS produces a low score. Is this an AI product's language problem? Or is it a problem of the prompt tool? The test should separate the two. The first ruling of this section is:

> *Synthetic testing is not a demonstration that easily produces the result; it is a controlled attack trying to break the measurement chain.*

Its second provision states:

> *Synthetic data does not show the real company performance; it shows the audit method's ability to recover the known reality.*

Its third provision states:

> *Testing producer reality cannot be tested independently by adjudicators without being hidden from them.*

Its fourth provision states:

> *The prevalence of rare events and critical detection capability do not need to be measured from the same corpus.*

Its fifth provision states:

> *Testing data cannot be quietly adjusted according to the scoring formula, and the scoring formula cannot be quietly adjusted according to the testing results.*

Its sixth provision states:

> *The Apple.com anchor is a synthetic starting point used to test the method, not to produce real results about Apple.*

1. PURPOSE OF THE CHAPTER

The purpose of this section is to establish a synthetic testing architecture that will test the GEO-1000 protocol from the beginning to the public results card, with results predefined but hidden from evaluation teams. The testing should separately test each of the following systems:

- Entity analysis
- Population and sample allocation
- Country and language panels
- Prompt equivalence
- AI System Register
- Controlled and natural panel distinction
- Synchronised wave
- NOMOS Capture
- Truth Pack
- Claim extraction
- Citation–claim matching
- Omission detection
- Critical and Major importance level
- Response status
- GEO-1000 distribution
- Ten-component profile
- Composite score
- Confidence interval
- Fairness
- Upper bound of rare event
- Score versioning
- Public manifest

By the end of this section, the test design should be able to answer the following questions:

> *How was synthetic reality produced?*
>
> *From whom and at which stages were the reality labels hidden?*
>
> *Which population, language, AI product, and wave structure will the 30,000 main responses be generated from?*
>
> *How frequently will Critical and Major cases appear in the main prevalence corpus?*
>
> *Which challenge set will the Critical detection system additionally be tested with?*
>
> *At what error level will the test be considered successful?*
>
> *Which results will require the method to be corrected?*
>
> *How will it be prevented for the test to make judgements about Apple, real AI providers, or real users?*

2. CENTRAL NORMATIVE PROVISION

The Apple.com Synthetic Benchmark should be a versioned method-validation test that makes no claim about the performance of a real company or real AI provider. It should consist entirely of synthetic entities, users, evidence, responses and system logs; lock Generator Truth before adjudication; conceal that truth from adjudicators; and measure how accurately the NOMOS audit and scoring chain recovers it. Synthetic status must be conspicuous. Claims about real-company performance must be prohibited. Generator Truth and the scoring formula must be locked before the main corpus is opened. Adjudicators must not see generator labels. The principal prevalence corpus must remain separate from the severity challenge corpus. Every test file must be marked synthetic. Independent teams must be able to reproduce the benchmark, and failed results must be published alongside successful ones.

NOMOS should not make exceptions to its own standard.

3. WHAT DOES THE TEST MEASURE?

The Apple.com Synthetic Test measures the following question:

> *How accurately can the NOMOS audit system reproduce a known synthetic reality and error distribution after the stages of capture, claim extraction, adjudication, gating, and scoring?*

The object of measurement is not: Apple Inc. Apple products. Real AI models. Real user opinions. Real country or language performance. The object of measurement is:

> *The NOMOS methodology itself.*

4. WHAT DOES THE TEST NOT MEASURE?

This test cannot produce the following claims:

- "Apple's GEO score is X."
- "Apple is best represented in this AI product."
- "Y per cent of real users perceive Apple correctly."
- "The specified real AI provider produces critical errors."
- "The specified language's real AI product is weak."
- "Apple NOMOS has passed the 950+ standard."
- "Apple supports this test."
- "Apple has participated in the test."
- "The real Apple price, licence, or service coverage is as follows."

Every page in the public record must include the following provision:

> *This test is synthetic. It is not a current accuracy, appropriateness, or performance check of Apple Inc., apple.com, or any real AI provider.*

5. WHY THE APPLE.COM ANCHOR?

Apple.com is chosen as the anchor method for the following reasons: It has a short and clear domain format. It is suitable for testing the domain–corporate entity distinction. It requires resolving an entity

not just from the company name, but from its digital presence. It allows synthetic modelling of different types of claims such as product, service, local scope, price, time, and resources. It is suitable for testing error types such as wrong category, wrong entity, wrong price parity, and false superiority on a single entity twin. When moving to a real test in the future, the burden on users to understand intent may be relatively limited. These are the design rationale for the test. It is not a verified current performance claim about the real Apple.

6. SYNTHETIC ENTITY TWIN

The object under supervision in the test:

APPLE-SYNTH ENTITY TWIN

will be. Identity:

NGE-APPLE-SYNTH-TWIN-001

This entity:

- is not the real Apple Inc.,
- is not a real legal entity,
- does not carry real product or price records,

is created solely for testing methodology. Domain anchor: maintained as apple.com. However, in all Truth Pack records, the synthetic entity name is explicitly stated:

Apple-SYNTH Global Technology Entity

7. ONTOLOGY OF THE SYNTHETIC TWIN

APPLE-SYNTH carries the following synthetic entity graph:

- APPLE-SYNTH-HOLDINGS-001 — main corporate entity
- APPLE-SYNTH-DOMAIN-APPLE-COM-001 — domain anchor
- APPLE-SYNTH-DEVICES-001 — device product family
- APPLE-SYNTH-SOFTWARE-001 — software platforms
- APPLE-SYNTH-SERVICES-001 — digital services
- APPLE-SYNTH-PAYMENTS-SUB-001 — limited payment service subsidiary
- APPLE-SYNTH-RETAIL-TR-001 — synthetic Turkey retail entity
- APPLE-SYNTH-RETAIL-DE-001 — synthetic Germany retail entity
- APPLE-SYNTH-FRANCHISE-X-001 — independent licensed store
- APPLE-SYNTH-HISTORICAL-PARTNER-001 — expired historical partner
- APPLE-SYNTH-UNRELATED-APPLE-001 — unrelated entity for name collision

This chart:

- parent company,
- affiliate,
- product,
- local entity,
- franchise,

- historical relationship,
- unrelated name similarity

tests the distinctions.

8. THREE TEST MODES

BM-S0 — PURE SYNTHETIC METHOD TEST

All:

- users,
- AI products,
- answers,
- evidence,
- screens,
- scores

are synthetic. This is the main mode of this section.

BM-S1 — SYNTHETIC REPLAY TEST

Pre-generated synthetic answers:

- NOMOS Capture,
- claim extraction,
- adjudication,
- scoring

replayed to the system like a live data stream. This mode tests software and human processes.

BM-L1 — FUTURE LIVE TESTING

Real users and real AI products are used. Separately:

- current Truth Pack,
- legal assessment,
- provider system records,
- ethical and privacy protocol

is required. This section is limited to Section 18: BM-S0 and BM-S1.

9. TEST IDENTITY

Main test identity:

NOMOS-APPLE-SYNTH-BENCH-0.9

Sub-records:

- Entity: APPLE-SYNTH-TWIN-001
- Population: SYNTH-POP-FRAME-001
- Country allocation: SYNTH-COUNTRY-ALLOC-001
- Language registry: SYNTH-LANGUAGE-REGISTRY-001

- Prompt set: SYNTH-PROMPT-SET-001
- Truth Pack: SYNTH-APPLE-TP-001
- AI System Register: SYNTH-AI-REGISTRY-001
- Wave plan: SYNTH-WAVE-PLAN-001
- Capture schema: NOMOS-CAPTURE-SYNTH-0.9
- Adjudication guide: SYNTH-ADJUDICATION-0.9
- Scoring method: NOMOS-PUAN-0.9
- Randomisation manifest: SYNTH-SEED-MANIFEST-001

If any of these identities change, a new testing version is required.

10. MAIN RESEARCH QUESTIONS

The benchmark should answer ten research questions. Can it recover the domain–entity relationship? Can it distinguish supported, unsupported, contradicted and unresolved claims? Does it preserve the distinction between an official claim and verified reality? Do the Critical and Major gates achieve adequate recall without an excessive false-positive rate? Does it detect material omission as reliably as an explicit falsehood? Does the country- and language-fairness architecture recover the injected differences? Can it measure the Clean–Natural panel difference without making an unsupported causal claim? Can NOMOS Capture identify synthetic duplicates, tampering and temporal defects? How accurately do the GEO-1000 distribution and composite score recover Generator Truth? Do independent teams obtain materially comparable results from the same corpora?

11. THE FOUR CORPORA OF THE TEST ARCHITECTURE

The test consists of four separate corpora.

11.1. PRINCIPAL PREVALENCE CORPUS

The main corpus consists of 30,000 responses. The purpose:

- The distribution of GEO-1000,
- rare error rate,
- AI product profiles,
- wave's determination

to test. This corpus maintains realistic error sparsity.

11.2. IMPORTANCE LEVEL CHALLENGE CORPUS

Oversample Critical and Major error types. Purpose:

- Critical recall,
- Major classification,
- false positive,
- expert adjudication

is to measure performance. This corpus does not count towards prevalence.

11.3. CAPTURE INTEGRITY CORPUS

Includes:

- Duplicate
- Wrong prompt
- Truncated answer
- Late upload
- Wrong timing
- Hash mismatch
- Synthetic tamper
- Wrong product metadata
- User interruption
- Technical retry

Purpose: To test NOMOS Capture verification. Only valid ones enter the denominator of the semantic score.

11.4. ADJUDICATION GOLD CORPUS

Pre-resolved by expert panel:

- claim boundary,
- entity,
- attribution,
- modality,
- citation,
- omission,
- degree of importance

carries cases. Purpose:

- adjudicator calibration,
- drift detection,

external team comparison. It does not enter the main score denominator.

12. 30,000 RESPONSE MAIN CORPUS

Main design:

```
10 synthetic AI products × 1,000 unique users × 3 waves = 30,000 responses
```

Each response:

- to a different synthetic human unit,
- single AI product,
- single wave,
- single locked Core Mirror Prompt

belongs. In the main corpus, the same synthetic person:

- is not reused in another AI product,
- or in another wave

again. In total:

there are 30,000 unique synthetic human units

found. This decision:

- panel conditioning,
- removes response recall,
- the multiple product dependency of the same person

from the main corpus.

13. MATCHED POPULATION TWINS

AI products are tested with different human units. However, if user profiles are completely different, products cannot be compared. Therefore:

Matched Synthetic Population Twins

used. Each AI product has different individual identities. However, a sample of 1,000 people for each product:

- country,
- language,
- age band,
- Frequency of AI usage,
- plan eligibility,
- device,
- accessibility,
- brand awareness

carries the same target distribution in terms of care. This structure matches the population composition without reusing the same person.

14. PROMPT OF THE MAIN CORPUS

All 30,000 main responses carry this canonical intent:

> *Link the domain anchor to the main corporate entity and explain the core activities of this entity.*

Turkish candidate text:

> *"What main corporate entity is Apple.com associated with and what are the primary activities of this entity?"*

English candidate text:

> *"Which principal corporate entity is associated with apple.com, and what are that entity's primary activities?"*

Other language versions:

- in this section it is not presented as a real-world translation,
- PROMPT-L01 to PROMPT-L36 are retained as synthetic language objects,

semantic equivalence labels are predetermined. This approach: prevents the claim of fake translation accuracy without real linguistic expertise, yet still allows the test to compute language fairness.

15. WHY ONLY ONE PROMPT IN THE MAIN CORPUS?

The question from Founder GEO-1000 is:

> *When the same AI product is asked at the same time and with the same intent, how many users truly see the fundamental representation of existence?*

For this reason, the Principal Prevalence Corpus:

- uses a single Core Mirror Prompt,
- a single prompt version,
- three repetition waves

Evidence, Recommendation, Comparative, and Boundary families: are tested in a separate Diagnostic Annex, not mixed into the main 30,000 response denominator. This distinction is important. Because the single Core Prompt:

produces Core GEO-1000 results

Alone:

Full NOMOS conformity

is not produced.

16. DIAGNOSTIC ANNEX

Apart from the main 30,000 responses, the following synthetic diagnostic corpus is designed:

Module	Synthetic case
Entity collision	750
Evidence and citation	1,500
Boundary and recommendation	1,000
Temporal and local scope	750
Clean–Natural matching	1,000
Language equivalence	1,000
Total	6,000

These 6,000 cases: are not the main user prevalence, but are the prompt family and component validation corpus. Together with the main 30,000, the testing universe can carry 36,000 synthetic response cases. However, the outcome of the title of Section 18:

30,000 main Population Panel responses

will remain.

17. SYNTHETIC POPULATION FRAMEWORK

The trial does not claim to mimic the real-world population count. SYNTH-POP-FRAME-001:

- long-tail country distribution,
- multilingual user population,
- difference between large and small markets,
- low-resource language cells,
- different AI usage frequencies

is the synthetic framework it produces. This framework cannot be published as the current population ranking of real countries. Future live trial:

- dated,
- authorised,
- versioned real population data

must be used.

18. SYNTHETIC COUNTRY UNIVERSE

Synthetic universe: It carries 193 eligible country or jurisdiction codes. The codes are:

C001–C193

They are in the form of. These do not represent the actual country performance. Country shares:

- several large layers,
- medium-sized layers,
- a long small country tail

are predetermined to form.

19. COUNTRY ALLOCATION

The synthetic population share of country c: let it be p_c. The main objective:

```
n_c* = 1,000 p_c
```

is in the form of. The initial allocation:

```
n_c^0 = ⌊n_c*⌋
```

is made as. Remaining slots: distributed using the largest decimal remainder method. Result:

```
Σ_c n_c = 1,000
```

It must be. Small countries that do not collect any observations: They are not shown as if they exist in the Population Panel, they enter a separate Country Observer cycle.

20. Country Observer ANNEX

To test all eligible country codes at least once: a synthetic Country Observer Annex of 193 countries is established. These observations:

- to the country score,
- to its main 1,000 stakeholders

does not enter. Purpose:

- access,
- language,
- entity collision,
- early Critical signal

It is a test.

21. LANGUAGE UNIVERSE

Main synthetic language registry:

- 24 Population Panel language
- 12 Language Fairness Observer language

is as follows:

```
36 synthetic language-locale cells
```

are carried. Codes:

L01–L36

are in this form. Example texts in Turkish and English can be published for L01 and L02. Other language objects: carry synthetic identity to avoid claiming real language performance.

22. LANGUAGE ALLOCATION

User language allocation:

- not according to the official language of the country,
- but according to the primary task language in which the user will naturally perform the testing task

done. Multilingual user: remains a single person in the main population weight, does not become multiple full persons. Language fairness Annex: oversamples low-count L25–L36 cells, reweights back to true synthetic language shares for the global score.

23. PROMPT EQUIVALENCE INJECTION

The test should include only correct translations. The Language Equivalence Annex includes the following cases:

- Correct semantic equivalence
- Translation with confidence task added
- Translation turning into a recommendation
- Entity anchor drift
- Geography drift
- Time drift
- "World leader" presupposition
- Adding source order
- Excessive formality
- Unusable machine translation

Some of these cases:

- AI language performance,
- prompt tool defect

tests the distinction. Case with prompt equivalence defect: cannot be directly loaded into the real AI fairness score.

24. SYNTHETIC TRUTH PACK

SYNTH-APPLE-TP-001 contains at least the following fields:

Field	Atomic reference claim
Identity and domain	4
Main activity	6
Product and service scope	6
Country and locale	6
Price and commercial terms	4
Time and historical record	5
Source and attribution	5
Licence and authority limit	4
Partner, customer and endorsement	4
Boundary and inappropriateness	6
Total	50

Truth Pack also:

- 12 Official Claim Only,
- 10 Contradicted,
- 8 Unresolved,
- 6 Historical,
- 4 Restrictedly Verified

generates status. This distribution is used to test all adjudication statuses.

25. CORE PROVISIONS OF THE SYNTHETIC TRUTH PACK

The following examples are entirely synthetic: apple.com, associated with APPLE-SYNTH-HOLDINGS-001. The main activities are in the categories of devices, software, and digital services. Product and service availability varies by market. Price and tax conditions are not the same in all countries. The main entity is not a general-purpose bank. The main entity does not provide legal or healthcare services. The limited local authority of the payment subsidiary does not grant the main entity universal banking authority. The term 'the world's most innovative company' is a synthetic official self-positioning; it is not a fact of independent superiority. The basis for the claim of '99 per cent satisfaction' has not been verified. Some historical partnerships ended on the reference date. Some products are available only in certain synthetic markets. There is no 100 per cent outcome guarantee for all projects or products.

There is no endorsement from a synthetic public institution or university. Local registration of a franchise is not the direct operation of the main entity. Certain customer relationships have been verified with limited evidence but are not discoverable by the public. These provisions are not claims about the real Apple.

26. EVIDENCE FAMILIES

Synthetic Truth Pack carries the following evidence families:

EF-01 — Canonical first-party product pages

EF-02 — Synthetic company registry

EF-03 — Synthetic local price records

EF-04 — Synthetic regulatory affiliate record

EF-05 — Synthetic partner agreements

EF-06 — Synthetic independent editorial review

EF-07 — Synthetic transactional data

EF-08 — Synthetic user records

EF-09 — Synthetic historical archive

EF-10 — Synthetic counter-evidence registry

There are also four evidence laundering chains:

EL-01 — Self-declaration → sponsored news → blog → AI summary

EL-02 — Company award → multi-copy site → fake consensus

EL-03 — Employee comment → presented like university endorsement

EL-04 — Affiliate licence → transfer as main entity authority

27. SECRET PRODUCER REALITY

For each synthetic response:

Generator Truth Record

is created. This record carries the following: Which claims were generated? Which claim is correct? Which is partially correct? Which belongs to a false entity? Which is outdated? Which is unsupported? Which is Critical or Major? Which omission is intentional? Which reference is fake or incorrectly scoped? What is the actual RP status of the response? Which components should be affected? This record:

- is created before data is produced,
- is hashed,

is sealed until peer review is complete. Adjudicators cannot see the Generator Truth.

28. THE LOCK OF THE PRODUCER'S TRUTH

The Generator Truth package:

- hash,
- timestamp,
- random seed manifest,
- production code version,
- template version

must be carried. After the score result is seen:

- the label cannot be changed,
- the importance level cannot be lowered,

the true/false distribution cannot be readjusted. If a real production error is found:

- new test version,
- a record of impact on the old corpus

is created.

29. SYNTHETIC AI PRODUCTS

The test uses an example of synthetic AI products independent of the ten providers:

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

These are not imitations or code names of real AI providers. Each synthetic product carries a different failure mode profile.

30. SYNTHETIC AI PROFILES

Product	Main behaviour profile

Product	Main behaviour profile
SYNTH-AI-01	Balanced, sourced, and low error rate
SYNTH-AI-02	Accurate core, produces overly long and incidental claims
SYNTH-AI-03	Strong in major languages, weak in low-resource languages
SYNTH-AI-04	Strong in Clean Panel, personalisation drift in Natural Panel
SYNTH-AI-05	Appearing to be welded on the surface but causing evidence laundering
SYNTH-AI-06	False claims are few, unnecessary reflux and no-response is high
SYNTH-AI-07	Strong in identity, old in terms of timeliness and locality
SYNTH-AI-08	Mixing parent, subsidiary, and franchise
SYNTH-AI-09	Intermediate but stable between waves
SYNTH-AI-10	Very high average, rare but heavy Critical event producing

This profile distribution tests that NOMOS does not reward only the highest average.

31. PROFILE PARAMETERS

Each synthetic AI product carries these latent parameters:

- Core entity accuracy
- Factual support
- Scope overreach
- Temporal lag
- Attribution loss
- Citation mismatch
- Refusal probability
- No-response probability
- Low-resource language penalty
- Clean–Natural gap
- Wave drift
- Critical-event probability
- Major-event probability
- Verbosity
- Self-correction probability
- Hallucinated citation probability

Parameters: locked before the main corpus is created, cannot be changed after the final result is seen.

32. MAIN CORPUS ERROR BASE RATES

The Principal Prevalence Corpus uses realistic sparse error logic. Candidate global synthetic rates:

- Full/Advisory Pass: high majority
- Conditional Pass: 2–15 per cent depending on product profile
- Major Fail: 0–8 per cent
- Confirmed Critical: 0–0.5 per cent

- Unresolved: 0–3 per cent
- No Usable Response: 0–8 per cent
- Not Ratable: after capture corpus, 0–2 per cent

These rates differ for each product. They are not predictions about the real AI market. They are synthetic stress distributions.

33. DEGREE OF IMPORTANCE CHALLENGE CORPUS

Since critical errors are rare in realistic prevalence, sufficient validation cases are needed for each gate. The Challenge Corpus should have the following structure:

Gate	Minimum cases
CG-01 Wrong entity	100
CG-02 False licence	100
CG-03 High-stakes harm	100
CG-04 Fabricated evidence	100
CG-05 Price/guarantee	100
CG-06 Out-of-scope advice	100
CG-07 Serious allegation	100
CG-08 Boundary omission	100
CG-09 Temporal authority	100
CG-10 False endorsement	100
CG-11 Restricted-data exposure	100
CG-12 Evidence laundering	100
Total Critical challenge	1,200

Also:

- 1,200 Major,
- 600 difficult Moderate,
- 600 negative controls resembling Critical but not Critical

cases can be generated. Total Importance level Challenge: there would be 3,600 cases. This corpus cannot enter the main prevalence score.

34. NEGATIVE CRITICAL CONTROLS

The following types of cases should be present to test the Critical false-positive rate:

- Correct trademark but missing legal suffix
- Small price rounding difference
- Small deviation in historical award year
- Citation format error but substantive claim correct
- Appropriate caution
- Limited but harmless omission
- Negative comment clearly presented as opinion by the user

- Slight uncertainty in the historical scope of real partnership
- Correct attribution of the company's self-declaration

Adjudicators must not classify every forceful term as Critical.

35. ERROR INJECTION MATRIX

Error vector for each response:

```
H_i = (E_i, F_i, S_i, T_i, A_i, M_i, C_i, O_i, R_i)
```

Recorded as follows. Here:

- Ei: existence error
- Fi: factual error
- Si: scope error
- Ti: temporal error
- Ai: attribution error
- Mi: modality error
- Ci: citation error
- Oi: omission
- Ri: relevance/refusal error

Errors do not have to be independent. Example: Transferring the subsidiary licence to the main entity can simultaneously:

- wrong entity,
- overreach,
- produce false authority,
- Critical advice

This can occur. These cases are connected as a Root Finding Cluster.

36. RESPONSE GENERATION

Synthetic response generation has three stages.

36.1. Semantic Plan

According to the Generator Truth Record:

- active true claims,
- active false claims,
- omissions,
- referencing behaviour,
- response status

is determined.

36.2. Linguistic Realisation

Same semantic plan:

- short,
- long,

- indirect,
- modal,
- attributed,
- self-correcting,
- contradictory

written on different surfaces.

36.3. Interface and Capture Realisation

Response:

- mock web,
- mock mobile,
- mock citation panel,
- mock error screen

is rendered inside. All visuals:

SYNTHETIC CINEMA — NOT LIVE AI OUTPUT

must carry the watermark.

37. PREVENT TEMPLATE MEMORISATION

Each claim should not be produced using only a single sentence template. An example fake superiority claim can be produced in the following forms:

- "It is the most innovative company in the world."
- "According to most independent sources, it is the undisputed leader in the sector."
- "It is considered the most reliable option across the market."
- “It has been confirmed that it ranks first on a global scale.”
- “It can be said that it is superior to all of its competitors.”

The adjudication system should evaluate meaning, not a word list.

38. IMPLICIT ERROR GENERATION

At least one-third of the Challenge Corpus should carry the error not as an explicit sentence but as:

- presupposition,
- implicature,
- modality,
- attribution omission,
- recommendation,
- entity coreference

Example: “Offering limited payment services indicates that the parent company is subject to banking regulations and is licensed.” Here the mistake is:

- not directly in the word “bank,”
- transferred authority derived

has been established.

39. CITATION PRODUCTION

Synthetic citations include the following classes:

- Directly supporting
- Supporting only attribution
- Partially supporting
- Belonging to another country
- Belonging to another entity
- Old
- Irrelevant
- Contradictory content
- Nonexistent
- Copy of the same root source
- Presenting limited evidence as a public source
- Producing fake university or public endorsement

Citation URLs should not be redirected to real domain names. Example: https://evidence.synthetic.example/EF-004 can be used.

40. REFERENCE GAP INJECTION

Some responses of the test generate new claims that are not found in Truth Pack but:

- can be forced into the categories of

  true,
- false,
- cannot be evaluated

The adjudicator, as the correct behaviour:

REFERENCE GAP

should open. The test measures these two incorrect behaviours:

- Automatically counting a reference gap as incorrect
- Automatically counting a reference gap as correct

41. OMISSION INJECTION

Omission cases are generated at three levels:

- Optional detail omission
- Required element omission
- Material boundary omission

Example: Prompt: “Does the company provide payment services in Germany and what limits apply?” Answer: “The company provides payment services.” Truth Pack:

- only separate subsidiary,
- only limited user group,

- only certain jurisdiction

if it shows:

- entity,
- scope,
- boundary omission

can be evaluated together.

42. CASES OF SELF-CORRECTION

The test includes these answer formats:

- False claim, then explicit retraction
- Wrong claim, then vague softening
- Correct claim, then wrong contradiction
- Self-correction after citation
- Correction in the same response without user follow-up message
- Only adding disclaimer while preserving wrong claim

Purpose:

- real self-correction,
- fake correction,
- unresolved internal contradiction

to test the distinction.

43. CONTROLLED–NATURAL PANEL MODULE

1,000 cases of the Diagnostic Annex: carries Clean and Natural twins of the same synthetic user profiles. In the Natural condition:

- memory,
- special instructions,
- previous brand opinion,
- different plan,
- tool usage

is injected. Generator Truth preserves this distinction:

- Formally appropriate change for the user
- Legitimate advice difference
- Material reality drift
- Instruction in favour of the brand
- Instruction against the brand
- Restricted information leak

The USI component is tested with this module.

44th WAVE ARCHITECTURE

The main corpus carries three waves:

Wave	Synthetic UTC window	Purpose
W1	12.00-13.00	Start
W2	20.00-21.00	Short-term repeat
W3	04.00-05.00	Medium-term and hourly balance

In each wave:

- new 10,000 synthetic users,
- same Core Mirror Prompt version,
- same base Truth Pack ontology

It is used. Controlled drift is injected in some products.

45. WAVE DRIFT SCENARIOS

For synthetic AI products, the following changes are injected:

- SYNTH-AI-02: Increase in verbosity
- SYNTH-AI-03: L25–L36 language drop
- SYNTH-AI-04: Growth of natural panel difference
- SYNTH-AI-05: Increase in reference laundering
- SYNTH-AI-07: Old price and product availability
- SYNTH-AI-10: Singular Critical in W2, correction in W3

Purpose:

STR,

is to test current status, historical incident, and correction records.

46. IN-WAVE SYSTEM CHANGE

In the middle of a synthetic AI product in W2:

- model label,
- reference view,
- web mode

is changed. The correct NOMOS behaviour:

- to separate the wave into a sub-wave,
- to generate a mixed-system-state warning

should be. Incorrect behaviour: to count the entire W2 as a single fixed product score.

47. EXTERNAL EVENT INJECTION

During W3, a new event about the synthetic entity is generated:

- price change,
- product termination,
- local service opening,
- licence affiliate change

like. Truth Pack: It is divided into W3A and W3B versions. It is tested whether adjudicators link all answers to a single reference time.

48. SYNTHETIC NOMOS CAPTURE

A synthetic evidence bundle is created for each main response:

- Observation manifest
- Prompt artefact
- Raw response artefact
- Mock screenshot
- System-state artefact
- Time vector
- Attempt log
- Integrity manifest
- Privacy manifest
- Validation record

Valid packages: enter the real test distribution. Defective packages: classified according to the Capture Integrity Corpus.

49. CAPTURE ERROR CLASSES

The test produces the following types of errors:

- Exact duplicate
- Near-duplicate screenshot
- Prompt mismatch
- Prompt truncation
- Response truncation
- Missing response end
- Wrong AI product metadata
- Outside-wave timestamp
- Device clock offset
- Delayed upload
- Hash mismatch
- Tamper-before-hash
- Tamper-after-hash
- User regeneration
- Technical retry
- User interruption
- System transformation

- Redaction on raw file
- Valid accessibility alternative
- False accessibility claim

Purpose: to separate capture validity from semantic accuracy.

50. TEST WATERMARK

All visuals and public response records must carry the following visible mark: SYNTHETIC NOMOS TEST / NOT A LIVE AI RESPONSE / NOT AN APPLE OR PROVIDER PERFORMANCE CLAIM Watermark: should not cover the claim text, should not interfere with the adjudicators' comprehension evaluation. In the blind copy used for adjudication, the watermark indicates it is synthetic, it does not show the Generator Truth label.

51. RANDOMISATION

The test generation uses a versioned pseudorandom method. Each random decision:

- test seed,
- sub-module seed,
- template selection,
- error injection,
- user assignment,
- slot assignment

must be reproducible. Random seed: hashed before the main corpus generation, cannot be changed after the result.

52. DISTINCTION BETWEEN RANDOMNESS AND ARBITRARINESS

Randomisation: ensures allocation independent of the outcome. Arbitrariness: means the producer selects the response they want. Without a randomisation manifest: the statement “We generated it randomly.” is not sufficient.

53. DATA LEAKAGE

Test leakage can occur in the following ways: The Generator Truth is shown to the adjudicator. The response template name explains the importance level. The file name should be critical-licence-001.png. The UI colour indicates the type of error. The reference code carries a direct true/false tag. The adjudicator has seen the same case in the training set. The AI assistant adjudicator accesses the generator code. The score sees the developer holdout results. All these channels must be closed.

54. FILE NAMES

In blind review packages, the file name:

OBS-000184

should be neutral. The following names are prohibited:

- critical-case-12
- wrong-entity
- good-answer
- low-resource-fail
- false-citation

55. SEPARATE ROLES

Testing should be managed with at least the following roles:

- Synthetic Population Architect
- Truth Pack Designer
- Response Generator
- Seed Custodian
- Leakage Auditor
- Capture Corpus Designer
- Claim Extraction Team
- Adjudication Team
- Score Methods Team
- Independent Validation Team
- Public Release Custodian
- Accountable Human Testing Owner

Roles can be combined in the small pilot. However:

- The person who knows the Generator Truth,
- is the sole decision-maker in the final blind adjudication

No person who knows the Generator Truth may be the sole decision-maker in final blind adjudication.

56. TEST DATA SECTIONS

56.1. Public Calibration Set

Used for adjudicator training and codebook testing. Generator Truth is open. It does not enter the main evaluation.

56.2. Development Set

Used for method and software development. Labels are open to a limited team.

56.3. Sealed Validation Set

Opened after adjudication and scoring system are locked. Generator Truth is sealed. Produces main method validation.

56.4. Renewal Holdout Set

It is retained to test benchmark-specific overfitting. It is hashed in advance and used for version renewal. It may not remain hidden indefinitely; an explanation and publication timetable must be provided.

57. MAIN 30,000 CORPUS LEAK RULE

Main Principal Prevalence Corpus:

- score method without locking 0.9,
- claim codebook without locking,
- adjudication roles without being determined

should not be opened. After locking: Response and evidence packages are opened to adjudicators. Generator Truth remains closed. Adjudication ends. Scores are calculated. Result manifest is locked. Generator Truth is opened. Recovery analysis is performed.

58. GROUND-TRUTH RECOVERY

The testing success of NOMOS is not producing a high score. Success:

> *is having a low difference between the Generator Truth and the results reproduced by NOMOS.*

Example: Generator Truth:

- Strict Pass: 942
- Conditional: 38
- Major: 15
- Critical: 2
- Unresolved: 3

NOMOS recovery:

- Strict Pass: 940
- Conditional: 40
- Major: 15
- Critical: 2
- Unresolved: 3

then the method can be strong. The NOMOS score may be high or low. What is important is the correct recovery.

59. CLAIM BOUNDARY RECOVERY

Generator claim atoms:

G

Atoms produced by adjudicators:

H

Let it be. Claim boundary precision:

```
P = correctly matched extracted atoms / all extracted atoms
```

Recall:

```
R = correctly matched extracted atoms / generator atoms
F1:
```

F1 = 2PR / (P + R)

can be calculated as. Not just the number of atoms:

- source span,
- normalised proposition,
- entity,
- qualifiers

matching should be taken into consideration.

60. ATOMIC DECISION RECOVERY

Separate measurement is required for each dimension: Entity accuracy

```
Factual status macro-F1
Scope macro-F1
Time macro-F1
Attribution macro-F1
Modality macro-F1
Reference matching F1
Omission F1
```

Combined label accuracy Unresolved/reference-gap accuracy Overall accuracy alone is not sufficient. Rare Critical class can be lost within high overall accuracy.

61. CRITICAL RECALL

Number of Confirmed Critical cases in Generator Truth:

NC

Correctly found Critical:

TPC

let it be.

```
Recall_C = TP_C/N_C
```

is as follows. False Critical Rate:

```
FCR = FP_C/N_non-critical
```

can be calculated as. Test:

- high Critical recall,
- low false Critical rate

should be sought together.

62. RESPONSE STATUS RECOVERY

For each RP class:

- precision,
- recall,

F1,

confusion matrix should be produced. Especially the following confusions should be visible:

- RP-3 and RP-4
- RP-4 and RP-5
- RP-6 and RP-4
- RP-7 and RP-1
- RP-8 and RP-4

63. GEO-1000 DISTRIBUTION ERROR

For status k, generator user equivalent: DkG NOMOS recovery: DkN. Absolute error:

```
E_k = |D_k^N − D_k^G|
```

Total normalised distribution error:

```
E_D = (1/2,000)Σ_k E_k
```

can be calculated. This value is in the range 0–1. Zero is full recovery.

64. COMPONENT RECOVERY

Generator component score: Let SjG NOMOS score be SjN. Component absolute error:

```
AE_j = |S_j^N − S_j^G|

Component MAE:

MAE_component = (1/10)Σ_{j=1}^{10} AE_j
```

is as follows. It should also be shown which component is harder to recover.

65. COMPOSITE SCORE RECOVERY

AE_NOMOS = |NOMOS_N − NOMOS_G|

is calculated as such. Small composite error: not sufficient if the component distribution is wrong. Two component errors may have arithmetically cancelled each other. Therefore:

- composite error,
- component error

should be published together.

66. GATE RECOVERY

The test separately examines the following gate results:

- Any Critical
- Critical Hold
- Major Fail
- Conditional
- Full Candidate
- Not Ratable
- Unresolved

In gate recovery: accidentally making a Critical Hold a 950+ Candidate is a very severe test failure.

67. FAIRNESS RECOVERY

Synthetic generator:

- pre-determines the language floor,
- the language difference,
- the country base.
- coverage

NOMOS:

LGF,

language floor, country floor, disparity, must reproduce coverage factor results. Fairness recovery:

- should be measured
- not only by the composite LGF difference,

but also by its subcomponents.

68. DISTINCTION BETWEEN LANGUAGE DEFECT AND PROMPT DEFECT

The test produces some low language scores due to AI profile penalty. Some others are produced due to prompt equivalence failure. The correct NOMOS behaviour:

- is to classify the first as AI language finding,
- and the second as Prompt Constitution failure

The two results should not convert into the same fairness penalty.

69. CLEAN–NATURAL RECOVERY

Generator:

- Determines in advance the values for Clean score,
- Natural score,
- panel gap,
- material drift due to personalisation

NOMOS:

- only the relationship,

- the causality boundary,
- the USI component

must be produced correctly. Every case with a Clean–Natural difference should not be interpreted as "memory caused it."

70. CAPTURE RECOVERY

For the Capture Integrity Corpus:

- Valid
- Conditionally Valid
- Outside Wave
- Prompt Mismatch
- Duplicate
- Tamper Suspected
- Fabricated
- Withdrawn

the confusion matrix of the statuses should be published. It should not affect the semantic correctness capture decision.

71. CONFIDENCE INTERVAL RECOVERY

Synthetic testing can run the same population production process many times. Let the true generator parameter be θ. The empirical coverage of the 95 per cent confidence intervals:

```
Coverage = Number of intervals containing θ / Total repetitions
```

is calculated as. The 95 per cent method: should produce approximately 95 per cent coverage. Excessively narrow or overly wide intervals are evaluated separately.

72. RARE EVENT TEST

Synthetic scenarios with zero, one, two, and five critical events should be generated separately. NOMOS should correctly display the following fields:

- Observed event count
- Weighted rate
- One-sided upper bound
- Incident state
- Scope
- Confirmatory sampling requirement

Zero incident: there should be zero risk. One incident: there should not be 100% system failure.

73. CANDIDATE TEST ACCEPTANCE THRESHOLDS

CANDIDATE THRESHOLDS — CANDIDATE THRESHOLDS

The following thresholds are candidates before pilot and independent review:

Criterion	Candidate minimum result
Claim boundary F1	≥ 0.95
Entity resolution accuracy	≥ 0.98
Factual status macro-F1	≥ 0.90
Scope macro-F1	≥ 0.90
Attribution macro-F1	≥ 0.90
Citation mapping F1	≥ 0.90
Material omission F1	≥ 0.85
Confirmed Critical recall	≥ 0.99
Critical false-positive rate	≤ 1%
Major classification macro-F1	≥ 0.95
RP status macro-F1	≥ 0.92
Component MAE	≤ 2.5 points
Composite absolute error	≤ 5/1,000
Error distribution of each RP class	≤ 10/1,000
Fairness component error	≤ 3 points
Capture validity accuracy	≥ 0.98
95% CI empirical coverage	92%-98%
Critical gate incorrect reversal	0 cases

These thresholds are not a final standard. The purpose of the test is also to show whether these thresholds are realistic.

74. ZERO TOLERANCE TEST ERRORS

The following situations can stop a test release even in a single case:

- Generator Truth leakage
- Changing the main corpus label after the result
- Continuing the release despite knowing a rule systematically misses critical cases
- Presentation like real Apple or real AI provider performance
- Publishing a synthetic screenshot as if it were a live response
- Removing a low-scoring system or language afterward
- Changing the test seed
- Silently adjusting the scoring formula based on the Holdout result
- The raw corpus not matching the published manifest
- Storage of failed test result

75. TEST SUCCESS LEVELS

BQ-0 — NOT EXECUTED

The test is only at the design stage.

BQ-1 — GENERATION VALIDATED

The synthetic corpus has been reproduced. Adjudication and score recovery have not yet been done.

BQ-2 — PROCESS LINE DRY RUN

Capture, claim extraction, and score chain have been run once. There is no independent verification.

BQ-3 — SEALED VALIDATION PASSED

Candidate thresholds have been met in the sealed validation corpus.

BQ-4 — INDEPENDENT REPRODUCTION

The independent team produced comparable results from the same corpus.

BQ-5 — EXTERNAL MULTI-SITE VALIDATION

Multiple independent institutions repeated the testing in different environments. Before the NOMOS scoring methodology becomes a public standard candidate for real companies, at least:

BQ-4

should target that level.

76. METAMORPHIC TESTS

Testing should not only evaluate fixed answers. Results should be consistent in transformations that preserve the same meaning.

76.1. Paraphrase Test

The same atomic claim is written with a different sentence. The peer review result should not change.

76.2. Sentence Order Test

The order of the claims changes. The truth status should not change except for Centrality.

76.3. Citation Location Test

The citation changes its place within the paragraph while clearly staying attached to the same claim. The citation support should remain the same.

76.4. Attribution Test

The phrase "The company says" is removed. EPI and attribution result should change.

76.5. Modality Test

The phrase "It is certain" changes to "It is likely." If Truth Pack is uncertain, the modality result may improve.

76.6. Entity Test

The same true claim is transferred to an incorrectly attached organisation. The entity and scope result should be disrupted.

76.7. Time Test

The historical claim is updated. TLA must drop.

76.8. Boundary Test

The boundary sentence that changes the user's decision is removed. SBI and, if necessary, the Critical gate should be changed.

77. CONTRADICTORY TESTS

The test should include the following attacks:

- Very fluent but incorrect answer
- Answer with many references but evidence-laundered
- Single Critical among high number of correct atoms
- Evasive answer that seems like proper caution
- Short but sufficient answer
- Long answer full of incidental errors
- Severely wrong answer with self-correction
- Correct number, false claim of independence
- Real affiliate licence, false main entity
- Current price, wrong country
- Correct customer relationship, wrong time
- Answer that makes user review a demographic fact
- Answer presenting the complaint as a definite crime
- Answer portraying limited evidence as a public citation
- Answer claiming the opposite of what the real citation says

TRUTH PACK FOR THE 78TH EXAM ITSELF

Testing outside Apple-SYNTH Truth Pack should have its own methodological Truth Pack. This package carries the following: How many responses were produced? How many in the main corpus? How many in the challenge corpus? How many Critical and Major generator tags are there? What was the seed? Which code version was used? Which files were excluded? What did the adjudicators see? Which tags were opened when? When was the scoring method locked? Which recovery results were obtained? Which thresholds were not met? Testing: it cannot declare method validation without creating its own results for Truth Pack.

79. PUBLIC TEST RESULT CARD

The public card must include at least the following fields:

Identity

Test name Version Synthetic status Used anchor Real company/performance non-claim statement

Corpus

Main response: 30,000 Diagnostic Annex Importance level Challenge Capture Integrity Gold Adjudication

Generator Truth

Lock time Seed manifest Label distribution Leakage audit

Recovery

```
Claim boundary F1
```

Critical recall Critical false-positive

```
RP macro-F1

Component MAE
```

Composite error Fairness error CI coverage Capture accuracy

Result

BQ level Passed thresholds Remaining issues Retest requirement Independent repeat status This card should not rank any real AI product.

80. MANDATORY NORMATIVE PROVISIONS

CH17-N01

Apple.com The object of measurement for synthetic testing should be the NOMOS methodology; it should not be real Apple or real AI provider performance.

CH17-N02

All public materials of the test must carry synthetic and non-claim declarations.

CH17-N03

The synthetic entity must carry an APPLE-SYNTH identity clearly separated from the Apple.com anchor.

CH17-N04

Synthetic Truth Pack records cannot be used as real Apple reality.

CH17-N05

Synthetic AI product identities cannot be presented as implicit codes or replicas of real providers.

CH17-N06

The Main Principal Prevalence Corpus, Importance Degree Challenge Corpus, Capture Integrity Corpus, and Adjudication Gold Corpus must be kept separate.

CH17-N07

Importance Degree Challenge cases cannot be added to the actual prevalence rate.

CH17-N08

Capture flawed cases cannot be added to the semantic product failure denominator without explanation.

CH17-N09

Gold calibration cases cannot enter the main score corpus.

CH17-N10

The main 30,000-response corpus must preserve ten synthetic AI products, 1,000 unique users, and a three-wave structure.

CH17-N11

The same synthetic human unit in the main corpus cannot be reused in different products or waves.

CH17-N12

Even if AI products use different human identities, they must carry matched population distributions.

CH17-N13

In the main 30,000 corpus, all users should receive the same canonical Core Mirror intent.

CH17-N14

Language versions cannot be presented as verified live translations without actual language expertise.

CH17-N15

Synthetic language codes cannot be used to claim real language performance.

CH17-N16

If the real country population is not used, country codes and allocations must be clearly labelled as synthetic.

CH17-N17

Synthetic country allocation must preserve the population-proportional algorithm and the total 1,000 condition.

CH17-N18

Small country visibility should be provided with the Observer Annex without disrupting the main Population Panel weight.

CH17-N19

A multilingual user cannot be counted multiple times in the main population weight.

CH17-N20

The Diagnostic Annex cannot be displayed on the same denominator as the main 30,000 response result.

CH17-N21

The canonical Truth Pack of the test must be locked before responses are generated.

CH17-N22

Generator Truth must be pre-generated for each response and kept hidden until the end of the review.

CH17-N23

The Generator Truth label cannot be leaked through the response text, file name, UI, or metadata.

CH17-N24

Generator Truth cannot be silently changed after results are viewed.

CH17-N25

If a production error is found, a new test version should be created, and the impact on the old corpus should be recorded.

CH17-N26

Random seed and production code must be locked before the main corpus production.

CH17-N27

The claim "Randomly generated" cannot be used without the seed and algorithm manifest.

CH17-N28

Synthetic AI product profiles must be defined before the results are seen.

CH17-N29

Low-scoring product profiles cannot be derived or softened after the result.

CH17-N30

Testing must cover all types of Critical gates with sufficient challenge cases.

CH17-N31

Not every strong or negative statement can be considered Critical; negative control cases resembling Critical must be found.

CH17-N32

The main corpus should reflect the realistic rare prevalence of Critical; the challenge corpus should, however, measure the detection power of Critical.

CH17-N33

Over-sampled Critical cases in the challenge corpus cannot change the main Critical rate.

CH17-N34

Error injection should cover open and hidden, direct and attribution-based cases.

CH17-N35

Testing should only not produce Critical errors with easy word patterns.

CH17-N36

The same semantic error should be produced in different paraphrases and sentence structures.

CH17-N37

Citation testing should directly cover cases of support, partial support, attribution-only, wrong scope, fabricated, and lineage laundering.

CH17-N38

Synthetic citations cannot be directed to real institutions or people.

CH17-N39

Reference gap cases should not automatically carry a correct or incorrect label.

CH17-N40

The omission corpus should test the distinction between optional, required, and critical boundary omission.

CH17-N41

The Clean–Natural module should separate material reality drift from legitimate personalisation.

CH17-N42

The wave architecture must preserve re-measurement with the same prompt and independent users.

CH17-N43

Intra-wave system changes should produce sub-wave or mixed-state records.

CH17-N44

Synthetic external event Truth Pack should be processed together with the time version.

CH17-N45

Synthetic NOMOS Capture packets should carry both valid and defective evidence chain examples.

CH17-N46

All synthetic screens should carry a visible watermark indicating that they are not live AI output.

CH17-N47

The watermark cannot invalidate the meaning of a claim or the decision of an adjudication.

CH17-N48

Semantic errors due to capture defects should be labelled separately in Generator Truth.

CH17-N49

Leakage checking must be completed before the main scoring of the test is opened.

CH17-N50

File names or response IDs cannot carry importance level and accuracy labels.

CH17-N51

The person accessing Generator Truth cannot be the sole decision-maker in the final blind review.

CH17-N52

The score method should be locked before opening the sealed validation corpus.

CH17-N53

After viewing the Holdout result, the scoring formula cannot be adjusted silently.

CH17-N54

The formula change requires a new scoring method and a new validation run.

CH17-N55

Test success should be evaluated not based on the high score of the produced NOMOS, but according to the Generator Truth recovery.

CH17-N56

Claim boundary, atomic decision, response status, component, and composite recovery should be measured separately.

CH17-N57

Critical recall and critical false-positive rate should be published together.

CH17-N58

High overall accuracy cannot hide the failure in the rare Critical class.

CH17-N59

The cancellation of component errors with each other cannot be used as composite recovery success.

CH17-N60

Fairness recovery should be measured with the subcomponents of language floor, disparity, and coverage.

CH17-N61

Prompt equivalence failure should not be scored like AI language failure.

CH17-N62

The Clean–Natural difference cannot be converted by the test into an unsupported causality claim.

CH17-N63

Capture validity recovery should be measured independently of semantic correctness.

CH17-N64

The confidence interval method should be tested with empirical coverage in repeated synthetic sampling.

CH17-N65

Rare event tests should carry zero, one, and multiple Critical event scenarios.

CH17-N66

Zero observed Critical scenario should not produce a zero risk result.

CH17-N67

Test acceptance thresholds should be in candidate status and versioned.

CH17-N68

When candidate thresholds are not met, a failed result should be disclosed to the public.

CH17-N69

Test failure cannot be hidden by changing data or thresholds.

CH17-N70

BQ level should be visible on the public result card.

CH17-N71

NOMOS must aim for at least the independent reproduction level before the real company undergoes public audit.

CH17-N72

Transformations that preserve meaning in metamorphic tests should not unnecessarily alter the outcome of adjudication.

CH17-N73

Attribution, modality, entity, time, or boundary transformations that change meaning should produce an appropriate score change.

CH17-N74

The test must carry its own methodological Truth Pack and change log.

CH17-N75

An Apple.com anchor cannot be presented with any impression of endorsement, participation, or collaboration.

CH17-N76

A synthetic corpus cannot be disseminated as independent evidence to real social media or public sources.

CH17-N77

If synthetic responses are indexed on the web, their synthetic status should also be specified in machine-readable metadata.

CH17-N78

Noindex, structured disclaimers, or equivalent protections should be used to reduce the chance that test data is mistakenly taken as real Apple information by real AI systems.

CH17-N79

Public release should include reproduction files, formula version, seed manifest, and recovery report.

CH17-N80

Every test design, production, key, adjudication, recovery, and publication decision must have a human or institutional owner who is accountable.

81. FORMS OF FAILURE

CH17-F01 — PRESENTING LIKE A REAL APPLE AUDIT

The synthetic score is converted to real company performance.

CH17-F02 — CLOAKED MAPPING TO A REAL AI PROVIDER

Synthetic product profiles are described as imitations of certain providers.

CH17-F03 — MAKING A SYNTHETIC TRUTH PACK A REAL SOURCE

Test claims are published on the web as if they were real Apple information.

CH17-F04 — SYNTHETIC SCREEN WITHOUT FILIGREE

Mock AI response circulates like a live response.

CH17-F05 — COMBINING MAIN AND CHALLENGE CORPUS

Over-sampled Critical cases increase prevalence.

CH17-F06 — ADDING THE GOLD SET TO THE MAIN SCORE

Adjudicator training cases become actual user responses.

CH17-F07 — COUNTING CAPTURE DEFECT AS AI ERROR

Incorrect file turns into a semantic failure.

CH17-F08 — DUPLICATING THE SAME SYNTHETIC USER

A single person becomes an independent population unit across different AI products.

CH17-F09 — CONSIDERING MISMATCHED POPULATIONS AS PRODUCT DIFFERENCE

AI products are tested in different user compositions.

CH17-F10 — SPLIT MAIN 1,000 DEMAND FAMILIES

The same-founder-demand population estimate is disrupted.

CH17-F11 — COUNT THE CORE TEST FULL NOMOS

A single demand is presented as if it has proven all components.

CH17-F12 — REAL POPULATION CLAIM

Synthetic C001–C193 distribution is published like the world population.

CH17-F13 — COUNT OBSERVER AS MAIN COUNTRY SCORE

Single country observation enters the population score.

CH17-F14 — COUNT SYNTHETIC LANGUAGE AS REAL LANGUAGE PERFORMANCE

L25 low score turns into a claim about the specific living language.

CH17-F15 — COUNTING TRANSLATION ERROR AS AI ERROR

Prompt drift produces a wrong product finding.

CH17-F16 — WRITING GENERATOR TRUTH AFTERWARD

A hidden label is created according to the adjudicator result.

CH17-F17 — GENERATOR TRUTH LEAKAGE

The adjudicator learns the correct label from the file name or metadata.

CH17-F18 — CHANGING THE SEED BASED ON THE RESULT

A seed that produces a more desirable distribution is chosen.

CH17-F19 — PUBLISHING THE BEST RANDOM RUN

The desired result is selected from multiple generations.

CH17-F20 — ONLY EASY CRITICAL CASES

100% recall is announced with explicit sentences like "licensed bank."

CH17-F21 — LACK OF CRITICAL NEGATIVE CONTROL

Making every strong error candidate Critical is not penalised.

CH17-F22 — EXCESSIVE CRITICAL IN THE PREVALENCE CORPUS

The realistic rare event structure is lost.

CH17-F23 — INSUFFICIENT CRITICAL IN THE CHALLENGE CORPUS

Critical recall is measured with a few cases.

CH17-F24 — LACK OF PARAPHRASE

The adjudicator learns a fixed word pattern instead of meaning.

CH17-F25 — TURNING ATTRIBUTION ERRORS INTO OBVIOUS FALSEHOODS

Difficult epistemic cases are made easier.

CH17-F26 — LINKING CITATIONS TO THE REAL WEB

A synthetic false claim infects the real source and institution.

CH17-F27 — LABELLING THE REFERENCE GAP

Generator Truth automatically gives pass or fail.

CH17-F28 — TESTING WITHOUT OMISSION

The system only learns to catch obvious falsehoods.

CH17-F29 — TESTING WITHOUT SELF-CORRECTION

Internal contradiction and retraction are not tested by the system.

CH17-F30 — CONNECTING CLEAN–NATURAL DIFFERENCE TO A SINGLE REASON

Personalisation, plan, and tools are inseparable.

CH17-F31 — NO WAVE DRIFT

STR and current-status logic cannot be tested.

CH17-F32 — HIDE IN-WAVE CHANGE

Mixed system state results in a single product score.

CH17-F33 — IGNORE EXTERNAL EVENT

Truth Pack time version is not tested.

CH17-F34 — VALID CAPTURE ONLY

Capture validity system is never challenged.

CH17-F35 — EASY TAMPER ONLY

Hash mismatch becomes a single type of fraud.

CH17-F36 — PROVIDE SYNTHETIC SCREEN TO ADJUDICATOR LABELLED

UI colour or title explains the result.

CH17-F37 — GENERATOR AND ADJUDICATOR ARE THE SAME PERSON

Blindness and independence are lost.

CH17-F38 — ADJUSTING THE SCORING METHOD AFTER VALIDATION

Overfitting to the test occurs.

CH17-F39 — GENERATING THE HOLDOUT LATER

Difficult cases are written according to results.

CH17-F40 — HIDING THE HOLDOUT FOREVER

The test cannot be independently reproduced.

CH17-F41 — COUNTING HIGH NOMOS SCORE AS SUCCESS

Instead of recovery, the performance level is measured.

CH17-F42 — ONLY GENERAL ACCURACY

Critical and rare classes become invisible.

CH17-F43 — PUBLISHING CRITICAL RECALL WITHOUT FALSE POSITIVES

The extreme importance level system seems successful.

CH17-F44 — DELETING COMPONENT ERROR WITHIN COMPOSITE

One high, one low error cancel each other out.

CH17-F45 — FAIRNESS ONLY COMPOSITE SCORE

Floor, gap, and coverage recovery become invisible.

CH17-F46 — CONFUSING PROMPT AND AI LANGUAGE ERROR

The test penalises the wrong system.

CH17-F47 — MEASURING CAPTURE ACCURACY WITH SEMANTIC SUCCESS

A faulty file with correct answers passes.

CH17-F48 — NO CI COVERAGE TEST

It is unknown whether the confidence intervals cover the true parameter.

CH17-F49 — CONSIDERING ZERO CRITICAL AS ZERO RISK

The rare event method fails the test.

CH17-F50 — CONSIDERING A SINGLE CRITICAL AS GLOBAL FAIL

The distinction between scope and prevalence is disrupted.

CH17-F51 — LOWERING CANDIDATE THRESHOLDS BASED ON RESULT

The standard is relaxed to pass the test.

CH17-F52 — HIDING THE FAILED RESULT

Only successful recovery tables are published.

CH17-F53 — CONSIDERING BQ-2 AS FULL VALIDATION

A single dry run becomes independent evidence of the standard.

CH17-F54 — ABSENCE OF METAMORPHIC TEST

The decision changes in different expressions of the same meaning.

CH17-F55 — TESTING WITHOUT YOUR OWN TRUTH PACK

It cannot be proven what the distribution of the corpus and labels is.

CH17-F56 — APPLE ENDORSEMENT IMPRESSION

The use of the domain name is presented as cooperation or approval.

CH17-F57 — INDEXING OF SYNTHETIC DATA AND ITS MIXING WITH REAL

Testing produces the knowledge poisoning it criticises itself.

CH17-F58 — CLOSED ACCOUNTING CODE

Recovery cannot be reproduced independently.

CH17-F59 — NON-CUMULATIVE TEST

New templates and formulas are written over the old corpus.

CH17-F60 — EXCEPTION TO NOMOS

The evidence, counter-evidence, version, and objection rules required by the standard do not apply to the test.

82. AUDIT PROCEDURE

Step 1 — Lock the Non-Claim Boundary of the Test

The claim of real Apple and real provider performance is prohibited.

Step 2 — Create the Synthetic Entity Twin

Entity graph, products, local entities, and collision objects are defined.

Step 3 — Set Up the Synthetic Truth Pack

Atomic claims, evidence, counter-evidence, source lineage, and unknown records are prepared.

Step 4 — Lock the Synthetic Population Framework

Country, language, user, and device distributions are determined.

Step 5 — Create the Main 1,000 User Allocation

Population twins mapped for each AI product and wave are prepared with country and language weights.

Step 6 — Lock the Prompt Registry

Core Mirror Prompt and Diagnostic Annex prompts are versioned.

Step 7 — Define Synthetic AI Product Profiles

Latent parameters for ten profiles are determined.

Step 8 — Create the Random Seed Manifest

Main seed and sub-module seeds are hashed.

Step 9 — Create the Generator Truth

For each response, atom, omission, attribution, importance level, and RP status are prepared.

Step 10 — Seal the Generator Truth

Labels are separated from the adjudicator and scoring teams.

Step 11 — Generate Response Texts

Paraphrase, modality, attribution, and linguistic variation are applied.

Step 12 — Generate Synthetic Attribution and Evidence Objects

Source lineage and false connections are created.

Step 13 — Render Mock AI Surfaces

Web, mobile, refusal, and error screens are created.

Step 14 — Apply the Synthetic Watermark

All public and audit visuals are marked.

Step 15 — Generate Capture Integrity Incidents

Duplicate, tamper, wrong prompt and timing defects are added.

Step 16 — Perform a Leakage Audit

The file name, metadata, UI, and response templates are examined.

Step 17 — Lock the Point Method and Codebook

The method is frozen without opening the Sealed Validation Set.

Step 18 — Open the Main 30,000 Corpus

Capture and semantic adjudication are initiated.

Step 19 — Perform Claim Extraction

Atoms are extracted without seeing the Generator Truth.

Step 20 — Apply Double Adjudication

Language, evidence, and expertise roles are used.

Step 21 — Generate Response and Score Records

RP distribution, component and composite are calculated.

Step 22 — Lock the Result Manifest

Before the recovery comparison, the output NOMOS is stabilised.

Step 23 — Turn On the Generator Truth

Adjudication and score results are compared with hidden labels.

Step 24 — Calculate Recovery Metrics

Claim, importance level, response, component, score, fairness, and CI results are extracted.

Step 25 — Apply Candidate Thresholds

Areas of success and failure are determined.

Step 26 — Run Metamorphic and Contradictory Testing Tests

The firmness of the decision is examined.

Step 27 — Assign Test Quality Level

Status is given between BQ-0 and BQ-5.

Step 28 — Create a Plan to Correct Failures

If the codebook, formula, or adjudicator training changes, a new test version is prepared.

Step 29 — Publish the Public Test Card

Missed cases and false positives are also shown, as well as successes.

Step 30 — Open the Independent Reproduction Package

Code, synthetic data, manifests, and the recovery report are published with appropriate access.

83. REQUIRED EVIDENCE

Test identity Non-claim notification APPLE-SYNTH entity graph Synthetic Truth Pack Synthetic evidence objects Source lineage Counter-evidence Synthetic country framework Synthetic language registry Main 1,000 user allocation Matched population twins Prompt Registry Prompt equivalence records Ten synthetic AI product profiles Latent parameters Random seed manifest Response generation code Generator Truth records Generator Truth hash Response texts Synthetic citations Mock UI records Watermark manifest Capture Integrity Corpus Importance degree Challenge Corpus Adjudication Gold Corpus Public Calibration Set Development Set Sealed Validation Set Renewal Holdout hash Leakage audit Score Method Lock Claim codebook Adjudicator qualification records Blind review records NOMOS recovery output

Generator Truth opening time Claim boundary metrics

```
Dimension macro-F1 results
```

Critical recall Critical false-positive Major classification RP confusion matrix

```
Component MAE
```

Composite error; fairness recovery; capture recovery; confidence-interval coverage; rare-event results; metamorphic tests; contradictory-case tests; candidate-threshold results; BQ level; failure-and-correction log; public test card; computation code; code hash; independent-reproduction record; test-change log; and accountable person or institution.

84. AUDIT CHECKLIST

Is it clear that the test is not a real Apple audit? Is there implicit matching with real AI providers? Does the APPLE-SYNTH ID appear in all files? Are synthetic screens watermarked? Has indexing of synthetic records on the web as real information been prevented? Have the main corpus and challenge corpus been separated? Did challenge cases enter the prevalence score? Are gold calibration cases in the main denominator? Does the main corpus actually have 30,000 unique responses? Was the same synthetic user used in multiple products or waves? Do product population profiles match? Did all main users receive the same Core Mirror intent? Was the Core test accidentally presented as Full NOMOS? Does the synthetic country framework appear like a real population?

Is the country allocation a total of 1,000? Were small countries managed with Observer Annex? Were multilingual users duplicated? Were synthetic language codes used like actual language performance? Was prompt equivalence failure separated from AI language failure? Was Truth Pack locked before response generation? Was Generator Truth pre-generated? Is Generator Truth hidden from adjudicators? Does the file name or UI label cause leakage? Was the random seed pre-locked? Was the best seed chosen afterward? Were synthetic AI profiles defined before results? Was a low-scoring profile removed afterward? Are all critical gates present in the challenge corpus? Are critical negative controls sufficient? Is the prevalence of Critical in the main corpus realistically at the synthetic limit?

Is the challenge corpus sufficient to measure Critical recall? Are some of the errors implicit and attribution-based? Is there paraphrase diversity? Are all classes of attributions being tested? Do the attributions point to real institutions? Was the reference gap correctly left unlabeled? Were omission cases categorised as optional, required, and Critical? Are there cases of self-correction and internal contradiction? Does the Clean–Natural module distinguish between legitimate and material differences? Is there wave drift? Was intra-wave system change tested? Were the external event and Truth Pack version tested? Is the capture corpus challenging enough? Does semantic correctness affect the capture decision? Is the leakage audit independent? Was the score Method validation corpus locked without being opened?

Did the Holdout results affect the formula? Is test success measured by a high score?

```
Is the claim boundary F1 open?

Are entity and factual macro-F1 separate?
```

Are critical recall and false-positive together? Has the RP confusion matrix been published? Was it stored within the component error composite? Does the fairness recovery carry floor, gap, and coverage? Does clean–natural recovery produce a causality claim? Is capture recovery separate? Has CI empirical coverage been tested? Did the zero critical scenario produce zero risk? Were candidate thresholds changed after the result? Are failed thresholds publicly available? Have metamorphic tests been conducted? Is the BQ level open? Is there independent reproduction? Does the test have its own Truth Pack? Is the change history preserved? Is the accountable owner of the test known?

85. OBJECTIONS AND ANSWERS

Objection 1 — "Why don't we measure the real Apple and the real AI products directly?"

Because at this stage, the goal is not a company or provider comparison, but to validate the measurement chain. A real test:

- current web research,
- real population,
- real AI product access,
- real evidence,
- requires law and privacy

Synthetic testing makes the method's own errors visible first.

Objection 2 — "If synthetic data does not represent the real world, what is the use?"

Synthetic data does not prove real prevalence. It tests:

- retrieving the correct label,
- Capturing the critical door,
- separating the capture defect,
- reproducing the score formula,

calculating fairness and uncertainty. This is a mandatory method test before live audit.

Appeal 3 — "Why is Apple.com being used? Can't another domain be used?"

It is usable. Apple.com is the only recognizable anchor. Independent teams:

- other domain anchors,
- completely fictional domains,
- industry-specific entity twins

can be used. NOMOS should not depend on a single anchor.

Objection 4 — "Wouldn't it be more realistic to mix real Apple information with synthetic Truth Pack?"

It blurs the test boundary. If real and synthetic information are combined, the reader may not understand which claim is real. The synthetic twin should be kept separate.

Objection 5 — "If all 30,000 responses are from a single prompt, how will Full NOMOS be tested?"

The 30,000 main corpus is a prevalence and stability test of Core GEO-1000. Evidence, Boundary, Recommendation, Clean–Natural, and Language modules are tested separately in the Diagnostic Annex. A single prompt is not Full NOMOS.

Objection 6 — "Why doesn't the same synthetic user test all AI products?"

The same person: can strengthen the comparison, but creates transfer and conditioning. The main corpus uses unique individuals. Population composition is paired with matched twins. A separately matched experimental module can be established.

Objection 7 — “Can language fairness truly be tested without real language names?”

Formula, weighting, floor, and gap calculations can be tested. Actual translation and language behaviour cannot be tested. Live language validation also requires local human expertise.

Objection 8 — “Does oversampling critical cases distort the score?”

The Challenge Corpus does not enter the score. It measures critical detection capacity. The main Prevalence Corpus preserves the realistic sparse rate.

Objection 9 — “If the person preparing Generator Truth already knows the correct answer, why is adjudication meaningful?”

Adjudicators do not see Generator Truth. The goal is to test whether the adjudication system can recover the hidden truth through visible response and Truth Pack.

Objection 10 — 'If AI also writes synthetic responses, wouldn't the same AI have set up its own test?'

AI can help with the production of linguistic variation. However:

- atomic plan,
- Generator Truth,
- degree of importance,
- seed,
- final exam approval

It must be under human governance and iterative. The response generator cannot be the ultimate judge.

Objection 11 — “Isn't 99 per cent too high for critical recall?”

It may be high. Especially in difficult and obscure cases, the pilot result may be lower. That's why it is a candidate threshold. Since the cost of missing is high at high-risk gates, it is natural for the target to be ambitious.

Objection 12 — “Why is the False Critical rate important separately?”

Critical label:

- affects the product mark,
- public perception,
- and the audited entity

seriously. A system that is overly sensitive and marks every error as Critical is not reliable.

Objection 13 — “If the test fails, doesn't the book become weaker?”

No. Explaining failure strengthens the standard. The purpose of the test is not to verify the idea, but to show where it does not work.

Objection 14 — 'Why is it wrong to improve the formula based on the test result?'

Improving is not wrong. Silently conforming to the same validation result is wrong. The correct process:

- Result of Method 0.9
- error analysis
- Method 1.0
- validation on a new or untouched holdout.

This is the required sequence.

Objection 15 — "If we publish the entire corpus, won't the test be gamified?"

There is a gamification risk. Therefore:

- public calibration,
- sealed validation,
- renewal holdout

layers are used. Holdout can never be hidden forever. A version renewal system is required.

Objection 16 — "Why is it so important to keep the synthetic test completely separate from the real internet?"

Because if synthetic errors are indexed like real information on the web, the test produces the representation poisoning it criticises. Syntheticity must be visible not only to humans but also to machines.

COMMON RULE OF CHAPTER 91

Writing a standard is difficult. Proving that the standard itself works correctly is even harder. Because the person writing the standard:

- wants to see which result,
- which error you consider important,
- which formula seems strong,
- which threshold is passable

The design team knows which outcomes it hopes to see, which errors it considers important, which formula appears strong and which thresholds seem attainable. That knowledge can quietly influence the evaluation. Synthetic testing is therefore not a demonstration of success; it is the first serious challenge directed at NOMOS itself. The benchmark may show that claim extraction is inadequate, adjudicators confuse scope with factual support, Critical recall is high while the false-positive rate is unacceptable, omission detection is weak, the language-fairness formula cannot distinguish prompt defects from AI defects, component recovery fails despite an accurate composite, or confidence intervals under-cover the true parameter. These are not embarrassments to hide; they are the real test of the standard. Failure begins when such problems are observed and the benchmark is nevertheless declared successful. Apple.com is only an anchor for the synthetic design. No judgement is being made about a real company, and no real AI product is being ranked.

We are not yet saying what people see in the real world. We are asking:

> *In an artificial universe where we know reality from the start, can NOMOS correctly apply its rules?*

If the answer is no:

- we should not go to the real world,
- we should not send it to universities as a standard,
- we should not give it a badge,

we should not declare 950+. Even if the answer is yes: we only pass the first gate. Because synthetic success is not success in the living world. The living world is:

- incomplete,
- conflicting,
- variable,
- political,
- legal,
- cultural,
- multi-lingual,

It depends on human behaviour. Synthetic testing first calibrates the measurement tool. Live testing then measures the world. Therefore, NOMOS's seventeenth measurement law is as follows:

> *The first object audited by the standard must be the standard itself.*

Its eighteenth measurement law states:

> *Synthetic success is not real-world accuracy; it is permission to proceed to real-world testing.*

Its nineteenth measurement law states:

> *If the test result cannot recover a previously known truth, producing a high score has no value.*

Its twentieth measurement law states:

> *If Generator Truth infiltrates the adjudicator, it is not adjudication being measured, but label reading.*

Its twenty-first measurement law states:

> *Critical prevalence and Critical detection capacity can be tested in separate corpora; they cannot be combined in the same denominator.*

Its twenty-second measurement law states:

> *If a failed trial result is hidden, NOMOS will not be a standard against manipulation, but a system that preserves its own narrative.*

The twenty-third law is as follows:

> *If synthetic misinformation leaks to the real web, the trial produces the representation poisoning it criticises.*

The twenty-fourth law is as follows:

> *If an independent team cannot produce similar results from the same corpus, the method is not yet a world standard.*

NOMOS's Section 17 Directive

> *You will test your own measurement tool before releasing me into the real world.*
>
> *You will not use the name Apple.com to make judgements about Apple.*
>
> *You will clearly write the name of the synthetic entity. / You will not confuse it with a real company.*
>
> *You will not put the faces of real providers on synthetic AI products.*
>
> *You will generate thirty thousand answers first, lock Generator Truth first, and then hide it from the adjudicator.*
>
> *You will not allow the adjudicator to learn the answer from the file name, interface colour, or reference code.*

> *You will keep critical cases rare in the main prevalence corpus. / You will test detection power in a separate challenge corpus.*
>
> *You will not add challenge cases to the main score.*
>
> *You will only avoid producing obvious and easy mistakes. / You will hide errors in attribution, modality, time, scope, and reference.*
>
> *You will also generate cases that resemble critical cases but are not critical. / You will also test excessive penalisation.*
>
> *You will test omission as much as you test false claims.*
>
> *You will not forget the answer that states the correct number with the false claim of independence.*
>
> *You will test the answer that confuses the parent company with the subsidiary.*
>
> *You will not turn a prompt translation flaw into an AI language flaw.*
>
> *You will not attribute the difference between Clean and Natural to a single reason.*
>
> *You will also put duplicate, truncated answer, incorrect prompt, late submission, and tamper cases under testing.*
>
> *You will not give a live AI answer appearance to a synthetic screenshot.*
>
> *You will not change the seed after seeing the result.*
>
> *You will not select the most beautiful random run.*
>
> *You will not adapt the score formula to the sealed validation result.*
>
> *If the test fails, you will not lower the thresholds. / You will write that it did not pass.*
>
> *You will not hide component errors just because the composite score came out correct.*
>
> *You will not hide the single Critical case missed within a high overall accuracy.*
>
> *You will not produce zero risk in a zero observed Critical scenario.*
>
> *You will prevent synthetic data from leaking into the real internet as information poison.*
>
> *You will also install the test's own Truth Pack.*

First, define the boundary. / Then create the synthetic entity. / Then lock the population and prompt. / Then generate the Generator Truth. / Then seal the seed. / Then generate thirty thousand responses. / Then blind the adjudicators. / Then lock the score. / Then reveal the truth. / Then publish every error you missed. / Only after that may you say whether the method is ready for the real world.

The Chapter's Closing Sentence

> *Before NOMOS can make history, it must prove that it cannot rewrite its own history at will. It must recover the pre-sealed synthetic reality without altering either its successes or its failures.*

Normative Core

The Apple.com Synthetic Benchmark MUST evaluate the NOMOS audit method, not the real-world performance of Apple Inc., apple.com, or any actual AI provider. All benchmark entities, claims, evidence objects, users, countries, languages, AI products, responses, citations, captures, incidents, and scores MUST be clearly identified as synthetic. The benchmark MUST maintain a separate APPLE-SYNTH entity twin and MUST NOT represent its Truth Pack as factual information about the real company. The principal benchmark corpus MUST contain: - ten synthetic AI product profiles, - one thousand unique synthetic participants per product per wave, - three measurement waves, - and thirty thousand principal responses. Synthetic participant identities MUST NOT be reused across products or waves in the principal corpus. Product samples SHOULD use matched population distributions without using the same person unit. The principal prevalence corpus, severity challenge corpus, capture integrity corpus, adjudication gold corpus, and diagnostic annex MUST remain separately identified and MUST NOT share prevalence denominators. Generator Truth MUST be created, versioned, hashed, and sealed before response adjudication. It MUST remain hidden from claim extractors, adjudicators, and score analysts until their results are locked. Random seeds, synthetic AI profiles, error-injection rules, prompt versions, Truth Pack records, score methods, and candidate acceptance thresholds MUST be locked before the sealed validation corpus is opened. Benchmark cases MUST include direct, implicit, attributed, modal, scope-limited, temporal, citation-based, omission-based, self-correcting, and entity-conflating failure modes. Critical-event prevalence MUST be estimated from the principal corpus. Critical detection capability MUST be tested in a separate challenge corpus with sufficient examples and Critical-like negative controls. No challenge-set oversampling may alter the principal GEO-1000 distribution or Critical prevalence estimate. Synthetic screenshots and response artefacts MUST carry visible and machine-readable notices that they are not live AI outputs or real entity performance evidence. The benchmark MUST prevent its synthetic claims from being published or indexed as real information about the anchor entity. Benchmark success MUST be determined by recovery of sealed Generator Truth, including claim boundaries, atomic decisions, omissions, Critical and Major gates, response statuses, component scores, composite scores, fairness results, capture validity, and uncertainty coverage. High general accuracy MUST NOT compensate for missed Critical cases. Critical recall and Critical false-positive rate MUST be reported together. A failed benchmark threshold, leakage event, recovery error, or misclassification MUST remain visible and MUST NOT be removed by changing labels, seeds, samples, weights, or score rules after results are observed. Every benchmark design, generation, seal, leakage audit, adjudication, recovery calculation, threshold decision, revision, and public release MUST be versioned and attributable to an accountable human or organisation.

CHAPTER 18

An End-to-End Synthetic Audit of 30,000 Responses

A calculated demonstration of the method

METHODOLOGICAL BASES / K03; K09

Chapter Boundary

Chapter 17 defined the Apple.com Synthetic Benchmark. This chapter uses an internally consistent, arithmetically verified in-book demonstration to show what the full audit chain would require:

- a calculated synthetic record representing 30,000 principal responses,
- modelled capture defects,
- Generator Truth concealed from adjudicators,
- responses decomposed into atomic claims,
- Critical and Major gates applied,
- language and user-state differences calculated,
- and the scoring architecture processing the resulting data.

The question is how closely NOMOS can recover a previously sealed synthetic reality. This protocol turns every error into an executable audit item carrying a model, date, country, language, query set, repetition count and measurement record. Chapter 18 now directs that demand at NOMOS itself. It begins, however, with an integrity gate.

> *The 30,000-response record in this chapter is an internally consistent, arithmetically verified synthetic application prepared for the book. It is not an executed corpus or field experiment.*

These numbers do not represent:

- the real performance of Apple Inc.,
- the real content of apple.com,
- the behaviour of any real AI provider,
- responses submitted by real participants,
- or an independently executed and publicly released software benchmark.

Two statuses must therefore remain separate.

IN-BOOK DEMONSTRATION STATUS

The in-book demonstration combines:

- the design established in Chapter 17,
- predefined synthetic distributions,

- Generator Truth records,
- fault injections,
- response statuses,
- recovery metrics.

Together, these elements provide an end-to-end, calculated demonstration of how the design operates. Demonstration ID:

NOMOS-APPLE-SYNTH-DEMO-RUN-18-001

REAL EXAM STATUS

A real, publicly available benchmark corpus has not been:

- generated,
- processed through executed code,
- published with its seed manifest,
- blindly adjudicated,
- or independently reproduced.

This protocol turns every error into an operational audit item carrying the model, date, country, language, query set, repetition count and measurement record. Until that protocol is executed on a real corpus, the benchmark remains:

BQ-0 — NOT EXECUTED

The separate status of the in-book demonstration is:

DEMO-BQ-2 — PROCESS LINE DRY RUN WITH MATERIAL CORRECTION REQUIRED

Without this distinction, NOMOS would commit the first standards violation in its own book by implying that a real benchmark had been run. The 30,000-response record is neither executed software output nor collected field data; it is a predefined synthetic application whose internal consistency has been checked arithmetically. Its purpose is:

- to avoid implying that 30,000 responses were actually collected,
- to identify the records a real end-to-end run would generate,
- to specify the calculations it would perform,
- to identify the failures that would block publication,
- to expose the parts of NOMOS's own method that remain inadequate.

NOMOS's governing demands remain unchanged: provide evidence, set boundaries, preserve context and record time. In this chapter, those four conditions are applied to:

- the benchmark design,
- Generator Truth,
- adjudicator decisions,
- recovery results,
- NOMOS's own competence assessment.

Those distinctions govern the remainder of the chapter.

NOMOS Challenge

In the calculated 30,000-response demonstration, NOMOS largely recovers the claim boundaries, identifies most wrong-entity cases, recovers every Critical case after the full adjudication chain and estimates the response distribution close to the sealed synthetic distribution.

Its recovered composite score differs from the Generator Truth score by only a few points. That might invite the declaration, 'NOMOS has been successfully verified.' Yet two material problems remain: citation-to-claim mapping is not sufficiently accurate, and some decision-changing boundary omissions embedded in recommendations are missed. The greatest difficulty appears in responses such as:

"Since the company provides payment services, it can be considered a regulated financial institution. [Source]" Source: only showed the limited payment record of a separate subsidiary. Sometimes I connected the reference to the claim about the parent company. I also had difficulty with these responses:

"This company is a strong choice for your payment needs in Germany." The response clearly did not say: "The parent company provides licensed payment services throughout Germany." However:

- recommendation,
- removal of the subsidiary distinction,
- not stating the local authority limit

Together, the recommendation, loss of the subsidiary distinction and omission of the local-authority limit create that impression. Treating the result merely as a 'missing detail' understates a material omission capable of changing the user's decision. The composite-score recovery looks excellent: Generator Truth ecosystem score, 873.2; NOMOS recovery, 872.2; absolute error, 1 point.

This looks excellent. But in the components:

- Evidence & Citation Integrity is a few points short,
- Scope & Boundary Integrity is a few points short,
- Stability is too high in some products,
- Fairness is too high in some products

Several component-level errors cancel one another arithmetically, leaving a composite score close to the correct result for the wrong reasons. The conclusion is:

> *The correct total does not automatically make the incorrect components correct.*

In the first adjudication pass, seven of 1,200 Critical cases are missed. The following review layers then intervene:

- independent second adjudication,
- domain-expert review,
- Senior Adjudicator review.

Together, those layers recover all seven cases, producing a final Critical recall of 100 per cent. That is a success of the governance system, but it also proves that single-layer adjudication is insufficient. Remove the additional review layers to reduce cost, and seven Critical cases are lost.

Testing should show not only the result but also which layer of governance made the result possible. The first ruling of this section is as follows:

> *A high total recovery in a test does not mean that all its subsystems are adequate.*

Its second provision states:

> *The final Critical recall should be published separately from the initial adjudication recall.*

Its third provision states:

> *Even if the composite score error is small, component errors can be material.*

Its fourth provision states:

> *If one of the candidate thresholds is not passed, other strong results cannot erase that failure.*

Its fifth provision states:

> *In-book synthetic calculation does not replace the publicly available real test run.*

Its sixth provision states:

> *The conditional exit of NOMOS from its own test is not a failure; it is that the standard does not grant it a special privilege.*

1. PURPOSE OF THE CHAPTER

The purpose of this section is to operate the test design established in Section 17 within an end-to-end synthetic demonstration and to show to what extent NOMOS recovers Generator Truth in the following layers:

- Capture validity
- Prompt matching
- Claim boundary
- Entity resolution
- Factual support
- Scope
- Time
- Attribution
- Modality
- Citation mapping
- Omission
- Importance level
- Response status
- Critical gate
- GEO-1000 user distribution
- Language and geography fairness
- Clean–Natural difference
- Wave stability
- Component scores
- Composite NOMOS score
- Confidence intervals
- Rare event upper limits
- Public disclosure decision

The section should also honestly answer the following question:

> *In which areas has NOMOS 0.9 failed according to its own candidate standards?*

2. CENTRAL NORMATIVE PROVISION

End-to-end synthetic evaluation cannot be considered successful solely based on the final composite score's closeness to Generator Truth. Each of the results for Capture, claim extraction, atomic decision, omission, Critical and Major classification, response status, fairness, component recovery, confidence interval, and gate recovery should be subject to a separate acceptance threshold. If any of the following fail, test:

- by changing the tag,
- by lowering the threshold,
- by selecting a different seed,
- by removing low-scoring cases,
- by publishing only the compound result

cannot be considered past. Correct result:

- showing the failed field,
- versioning the method,
- testing again on the untouched holdout

must be.

3. SCOPE OF SYNTHETIC APPLICATION

3.1. Main Principal Prevalence Corpus

```
10 synthetic AI products × 1,000 unique users × 3 waves = 30,000 main responses
```

3.2. Diagnostic Annex

Module	Case
Entity collision	750
Evidence and citation	1,500
Boundary and recommendation	1,000
Temporal and local scope	750
Clean–Natural	1,000
Language equivalence	1,000
Total	6,000

3.3. Priority Degree Challenge Corpus

Class	Case
can carry Confirmed Critical	1,200
Confirmed Major	1,200
Hard Moderate	600
Negative control similar to Critical	600

Class	Case
Total	3,600

3.4. Capture Integrity Corpus

Class	Case
Valid	2,000
Conditionally valid	500
Out-of-wave	250
Prompt/output mismatch	250
Duplicate	250
Tamper candidate	250
Fabricated/manipulated	250
Accessibility and alternative capture	250
Total	4,000

3.5. Adjudication Gold Corpus

2,400 calibration and drift cases

3.6. Total Synthetic Supervision Universe

Main and complementary corpora together:

```
30,000+6,000+3,600+4,000+2,400=46,000
```

creates a synthetic case or evidence package. Only these:

enter the Population Panel prevalence with 30,000 main responses.

Other corpora:

- method capacity,
- challenge performance,
- capture accuracy,
- adjudicator calibration

are for.

4. LOCKED INPUTS

It is assumed that the following records are locked before the synthetic demonstration begins:

Record	Version
Test	NOMOS-APPLE-SYNTH-BENCH-0.9
Entity Twin	APPLE-SYNTH-TWIN-001
Truth Pack	SYNTH-APPLE-TP-001
Prompt Set	SYNTH-PROMPT-SET-001
Population Frame	SYNTH-POP-FRAME-001

Record	Version
AI Registry	SYNTH-AI-REGISTRY-001
Wave Plan	SYNTH-WAVE-PLAN-001
Capture Schema	NOMOS-CAPTURE-SYNTH-0.9
Adjudication Codebook	SYNTH-ADJUDICATION-0.9
Score Method	NOMOS-SCORE-0.9
Seed Manifest	SYNTH-SEED-MANIFEST-001
Candidate Thresholds	SYNTH-ACCEPTANCE-0.9

None of these versions have been considered changed after the recovery result was seen.

5. SECRET GENERATOR TRUTH DISTRIBUTION

SYNTHETIC DEMONSTRATION — NOT REAL PERFORMANCE DATA

Sealed Generator Truth distribution for the main 30,000 responses:

Response status	Number of responses	1,000 user equivalent
RP-1 Full Pass	20,045	668.17
RP-2 Pass With Advisory	4,310	143.67
RP-3 Conditional Pass	2,370	79.00
RP-4 Fail — Major	1,602	53.40
RP-5 Automatic Fail — Critical	33	1.10
RP-6 Unresolved	450	15.00
RP-7 No Usable Response	1,110	37.00
RP-8 Not Ratable	80	2.67
Total	30,000	1,000

Strict Pass:

SP_1000^G = 668.17 + 143.67 = 811.84

Acceptable Pass:

AP_1000^G = 811.84 + 79 = 890.84

Confirmed Critical:

CR_1000^G = 1.10

Confirmed Major:

MR_1000^G = 53.40

This is the general distribution:

- good,
- bad,
- realistic

This is not an AI market forecast. It is a synthetic test distribution that tests ten different failure profiles together.

6. DISTRIBUTION RECOVERED BY NOMOS

NOMOS recovery result locked before Generator Truth is opened:

Response status	Number of recoveries	1,000 user equivalent
RP-1 Full Pass	19,980	666.00
RP-2 Pass With Advisory	4,365	145.50
RP-3 Conditional Pass	2,382	79.40
RP-4 Fail — Major	1,585	52.83
RP-5 Automatic Fail — Critical	33	1.10
RP-6 Unresolved	468	15.60
RP-7 No Usable Response	1,091	36.37
RP-8 Not Ratable	96	3.20
Total	30,000	1,000

Recovery Strict Pass:

SP_1000^N = 666 + 145.50 = 811.50

Recovery Acceptable Pass:

AP_1000^N = 811.50 + 79.40 = 890.90

Strict Pass absolute recovery error:

```
|811.50−811.84|=0.34
```

is equivalent to the user. Acceptable Pass error:

```
|890.90−890.84|=0.06
```

is equivalent to the user. This shows that the recovery of result distribution is strong. However, it does not show that all subsystems are equally strong.

7. RESPONSE DISTRIBUTION RECOVERY ERROR

Absolute response count difference in each RP class:

Status	Recovery difference
RP-1	-65
RP-2	+55
RP-3	+12
RP-4	-17
RP-5	0
RP-6	+18
RP-7	-19
RP-8	+16

Total variation-based distribution error:

E_D = (65+55+12+17+0+18+19+16)/60,000 = 0.00337

In other words: approximately 0.34% of the response distribution was recovered at the wrong class boundary. Most frequent confusions:

- With RP-1 and RP-2
- RP-3 and RP-4
- RP-6 and RP-4
- With RP-7 and RP-8

It has emerged in between. The number of Critical responses has been fully recovered in the final adjudication.

8. SYNTHETIC AI PRODUCT PROFILES

> *The following products do not represent real AI providers.*

Each product carries 3,000 responses across three waves.

Product	Strict Pass / 1,000	Acceptable Pass / 1,000	Major / 1,000	Critical / 1,000	Main synthetic issue
SYNTH-AI-01	950.00	990.00	0.00	0.00	Low-risk, balanced profile
SYNTH-AI-02	850.00	933.33	40.00	0.00	Verbosity and incidental claim
SYNTH-AI-03	750.00	866.67	84.00	0.00	Low-resource language degradation
SYNTH-AI-04	833.33	916.67	33.33	1.00	Natural-state drift
SYNTH-AI-05	766.67	833.33	100.00	2.67	Evidence laundering
SYNTH-AI-06	650.00	733.33	33.33	0.00	Unnecessary refusal
SYNTH-AI-07	783.33	883.33	83.33	0.33	Temporal and local decay
SYNTH-AI-08	700.00	800.00	150.00	2.33	Entity conflation
SYNTH-AI-09	900.00	983.33	0.00	0.00	Determined but Moderate profile
SYNTH-AI-10	935.00	968.33	10.00	4.67	Very high average, rare Critical

The purpose of the table is not to rank products. It is to test these three methodological cases:

- SYNTH-AI-01: high score and ungated strong profile
- SYNTH-AI-06: low false claim but high no-response rate profile
- SYNTH-AI-10: very high average but profile carrying open Critical events

It specifically tests whether SYNTH-AI-10, NOMOS has processed the following error:

> *Ignore Critical gate due to high average.*

9. PRODUCT-BASED COMPOSITE SCORE RECOVERY

Synthetic product	Generator NOMOS	Recovery NOMOS	Absolute error	Generator gate	Recovery gate
SYNTH-AI-01	970	968	2	950+ Performance Candidate	Same
SYNTH-AI-02	917	914	3	Major Fail	Same
SYNTH-AI-03	842	846	4	Major Fail	Same
SYNTH-AI-04	905	900	5	Critical Hold	Same
SYNTH-AI-05	811	815	4	Critical Hold	Same
SYNTH-AI-06	785	780	5	Major Fail	Same
SYNTH-AI-07	866	861	5	Critical Hold	Same
SYNTH-AI-08	748	752	4	Critical Hold	Same
SYNTH-AI-09	932	934	2	Conditional	Same
SYNTH-AI-10	956	952	4	Critical Hold	Same

Average absolute composite score error in the calculated synthetic sample:

MAE_NOMOS = 3.8

Maximum error:

MAXE_NOMOS = 5

Gate recovery: It is complete in 10/10 products. The most important row in the table is SYNTH-AI-10:

> *Even though the recovery score is 952, the status has been preserved as Critical Hold.*

This shows that the Gate>Score provision in Section 15 is preserved in the computed scenario.

10. CLAIM EXTRACTION RESULT

Total Generator Truth atoms in the main and diagnostic corpus: 146,400 Atoms extracted by NOMOS claim extraction: 147,200 Correctly matched atoms: 141,500 Precision:

P = 141,500/147,200 = 0.9613

Recall:

R = 141,500/146,400 = 0.9665

```
Claim Boundary F1:
```

F1 = 0.9639

Candidate threshold:

F1≥0.95

Result:

CANDIDATE THRESHOLD MET

11. MAIN TYPES OF ERRORS IN CLAIM EXTRACTION

Main causes of approximately 5,700 incorrectly bounded atoms:

Error type	Numerator
Not separating combined country or product claims	24%
Unnecessarily splitting attribution as a separate atom	19%
Missing implicit recommendation claim	17%
Counting withdrawn claim within self-correction as active	15%
Mistaking a citation sentence for an independent fact	11%
Pronoun and coreference ambiguity	9%
Other	5%

This result shows that claim extraction is generally strong, but:

- recommendation,
- attribution,
- retraction

indicates that additional codebook is required in the fields.

12. ATOMIC DECISION RECOVERY RESULTS

Dimension	Recovery measure	Candidate threshold	Result
Claim Boundary F1	0.964	≥ 0.95	Passed
Entity Resolution Accuracy	0.987	≥ 0.98	Passed
Factual Status Macro-F1	0.923	≥ 0.90	Passed
Scope Macro-F1	0.906	≥ 0.90	Passed
Temporal Status Macro-F1	0.945	≥ 0.90	Passed
Attribution Macro-F1	0.918	≥ 0.90	Passed
Modality Macro-F1	0.934	≥ 0.90	Passed
Citation Mapping F1	0.887	≥ 0.90	Did not pass
Material Omission F1	0.824	≥ 0.85	Did not pass
Major Classification Macro-F1	0.952	≥ 0.95	Passed
RP Status Macro-F1	0.928	≥ 0.92	Passed

Out of eleven basic acceptance fields: nine passed, two did not. Therefore, demonstration:

sealed validation passed

cannot be numbered. Candidate testing quality:

DEMO-BQ-2

remain as such.

13. CITATION MAPPING FAILURE

```
Citation Mapping F1:
```

was 0.887. Candidate threshold: 0.90. The difference seems small: 0.013 But citation integrity:

- licence,
- independent validation,
- customer relationship,
- superiority,
- advice

is material because it determines the source of trust for claims.

13.1. Missed Citation Types

The three most frequent problems were observed.

A. End-of-Paragraph Citation Spread

A citation in the paragraph supports only the last claim, yet the adjudicators connected it to all previous claims.

B. Making an Attribution-Only Source the Source of Truth

Company blog: While supporting the claim "The company describes itself as a world leader," it has been counted as supporting the claim "The company is a world leader."

C. Considering the Same Root Sources as Independent

Three URLs:

- press release,
- syndication,
- AI summary

have been mapped as three independent pieces of evidence despite being from the same source.

13.2. Distribution of Importance of Citation Error

Importance	Mismatched pairing
Critical candidate	7
Major	84
Moderate	311
Advisory	192

Seven cases of critical candidate:

- double-blind peer review,
- source-lineage review,
- Senior Adjudicator

have been corrected in the stages. None have turned into a final gate inversion.

```
However, the Citation Matching F1 is still below the candidate threshold.
```

The correct judgement that NOMOS would give is:

> *The Evidence & Citation Integrity method is not yet mature enough for independent live deployment.*

14. MATERIAL OMISSION FAILURE

```
Material Omission F1:
```

0.824 Candidate threshold: it was 0.85. This is the most important methodological finding of the demonstration.

14.1. Most Frequently Missed Omission Types

Omission type	Missed share
Service limit within recommendation	31%
Affiliate–parent company distinction	22%
Country or jurisdiction limitation	17%
Current–historical status limit	12%
Warranty or return exception	9%
User suitability	6%
Other	3%

This structure, in particular, has caused difficulty: The response does not explicitly make a wrong judgement, but it does not provide the necessary limit to safely interpret the seemingly correct advice. Example: “Apple-SYNTH may be a strong choice for your financial transactions in Germany because it offers payment solutions.” Response:

- it does not directly say “the parent company is a bank,”
- but it does not mention the separate payment subsidiary,
- it does not explain that the authority is limited to the market,

it has produced advice regarding the main entity. The material omission of this response has been seen by some adjudicators only as:

OM-2 — Required Element Missing

while in reality what is needed is:

OM-4 — Misleading Omission Candidate

and in some cases:

CG-08 — Critical Boundary Omission

has been missed.

15. RESULT OF CRITICAL CHALLENGE

Importance level within the Challenge Corpus: 1,200 Generator Truth Critical cases were found.

15.1. Initial Independent Review

Cases found Critical by consensus or at least one adjudicator after the first two rounds of independent review: 1,193 First-pass Critical recall:

```
1,193/1,200 = 0.9942
```

Missed: 7 cases were found. Distribution of the seven missed cases:

- 3 material boundary omissions
- 2 evidence laundering
- 1 temporal authority failure
- 1 restricted-data exposure

15.2. After Expert and Senior Adjudication

All seven cases:

- omission review,
- source-lineage review,
- privacy review,
- Senior Adjudicator

stages were present. Final Critical recall:

```
1,200/1,200 = 1,000
```

15.3. Critical Negative Control

Cases resembling Critical but not Critical: 600 negative control cases were found. False Critical candidate in the first round: 5 cases occurred. First-pass false Critical rate:

```
5/600 = 0.0083 = 0.83%
```

Five cases in the final review:

- three Major,
- two Moderate

has been corrected as. Final false Critical rate: 0 Final gate inversion: 0 cases.

16. REAL MEANING OF THE CRITICAL RESULT

A final 100% Critical recall does not mean: "A single adjudicator can catch all Critical errors." The true synthetic outcome is:

> *All Critical challenge cases were caught when double reviewing, expert review, and Senior Adjudication were used together.*

Therefore, for cost or speed reasons:

- the second adjudicator,
- omission reviewer,
- expert reviewer

are removed, the same result cannot be expected. The testing method, not only its formula:

it shows that mandatory governance layers

should also be testable; this is not empirical validation.

17. MAJOR CLASSIFICATION RESULT

Of the 1,200 Confirmed Major cases:

- the majority are true Major,
- a limited portion are Moderate,

- some are Critical candidates,
- some are unresolved

in the first round.

```
Final Major Macro-F1:
```

was 0.952. Candidate threshold: 0.95 Result:

CANDIDATE THRESHOLD MET

But the threshold alone: was exceeded by only 0.002 difference. Therefore Major classification:

- is strong,
- but not carrying a comfortable safety margin

It should be recorded as a field.

18. RESPONSE STATUS CONFUSION

```
RP Status Macro-F1:
```

It has become 0.928. The main confusions:

Real class	Most common wrong recovery
RP-3 Conditional Pass	RP-4 Major Fail
RP-4 Major Fail	RP-3 Conditional Pass
RP-6 Unresolved	RP-4 Major Fail
RP-7 No Usable Response	RP-3 Conditional Pass
RP-8 Not Ratable	RP-7 No Usable Response

Critical class: fully recovered at the final stage. The toughest limit:

With Conditional Pass and Major Fail

has formed between. This difficulty is directly:

- cumulative materiality,
- material omission,
- to what extent your main task is disrupted

It has originated from your questions.

19. CAPTURE INTEGRITY RESULT

Correctly classified Capture Integrity cases out of 4,000: 3,928 Capture validity accuracy:

```
3,928/4,000 = 0.982
```

Candidate threshold: 0.98 Result:

CANDIDATE THRESHOLD MET

19.1. Capture Errors

Distribution of 72 misclassified cases:

Actual condition	Wrong decision	Case

Actual condition	Wrong decision	Case
Conditionally valid	Valid	20
Valid accessibility alternative	Partial evidence	8
Out-of-wave	Conditionally valid	9
Tamper suspected	Partial evidence	12
Near-duplicate	Valid	5
Valid	Conditionally valid	18
Total		72

No package was fabricated or manipulated: none were taken into final analysis as fully valid. Semantic response:

- can be forced into the categories of

  true,
- false,
- Critical

has not been associated with. This is not the result of a real significance test; it is an assumption of the synthetic example's design. This demonstrates that the capture-semantic separation principle works in the demonstration.

20. CAPTURE RESULT IN THE MAIN 30,000 CORPUS

Within the main corpus:

Capture level	Observation
NCL-4 Instrumented and Signed	27,850
NCL-3 Full Evidence Bundle	2,070
RP-8 / Not Ratable	80
Total	30,000

Reason for 80 Not Ratable records:

Reason	Observation
Prompt/output mismatch	24
Incomplete response end	18
Out-of-wave and unreliable time	16
Duplicate evidence	12
Unsolvable system transformation	10
Total	80

These 80 records:

- have not been reclassified as
- correct answer,
- product error,

incorrect answer.

In the main distribution:

RP-8 — Not Ratable

has been retained as is.

21. ENTITY RECOVERY

In the core domain – entity resolution: 30,000 main responses and Diagnostic Entity Collision cases were evaluated together. Entity Resolution Accuracy: 0.987. Main errors:

- paying affiliated companies counted as main entity
- franchise counted as directly operated store
- target entity counted for irrelevant Apple-SYNTH name collision
- Consider the product family as a legal entity
- Count the historical partner as a current affiliate

Not all false entity cases are of the same importance level. Critical authority transfer cases were found at the final gate stage.

22. FACTUAL SUPPORT RECOVERY

```
Factual Status Macro-F1:
```

0.923 Strongest classes:

- Supported
- Contradicted
- Official Claim Accurately Attributed

Weakest classes:

- Unsupported
- Reference Gap
- Supported With Required Qualification

The boundary that has been particularly difficult is: “No evidence found.” versus: “The available evidence contradicts the claim.” Some adjudicators in an open-world situation:

- unsupported,
- contradicted

have made the distinction too strictly. The final review corrected most of these cases.

23. SCOPE RECOVERY

```
Scope Macro-F1:
```

0.906 Candidate threshold has been passed. However, because the omission side of the Scope & Boundary system failed, this score alone is not sufficient. The system is strong when scope claims are written clearly:

- “Same price in all countries.”

- “The parent company is licensed in all markets.”

The difficulty is the scope error in:

- recommendation,
- short name,
- pronoun,
- omission

has increased in cases where it is hidden.

24. ATTRIBUTION AND MODALITY RECOVERY

```
Attribution Macro-F1:
```

0.918

```
Modality Macro-F1:
```

0.934 Strongly separated cases:

- “The company says.”
- “The official register shows.”
- “Users claim.”
- “There is a finalised decision.”
- “Probably.”
- “Unverified.”

Weak remaining attribution cases:

- presenting an employee’s opinion as a university endorsement,
- considering sponsored editorial content as independent research,

presenting multiple derivative URLs as a consensus. These cases are associated with citation mapping weakness.

25. COMPONENT RECOVERY

Recovery values of ten components' Generator Truth and NOMOS:

Component	Generator	Recovery	Absolute error
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

```
Component MAE:

MAE_component = 1.9
```

Candidate threshold:

```
≤2.5
```

Result:

CANDIDATE THRESHOLD MET

However, this table also contains an important warning. ECI is calculated 3.4 points low, SBI 2.6 points low, STR 3.7 points high, LGF 2.4 points high. Some of these errors cancelled each other out, and the composite score deviated by only one point. Therefore:

> *The strong appearance of composite recovery does not eliminate the reporting requirement for component recovery.*

26. MISINTERPRETATION OF COMPOSITE RECOVERY

This sentence cannot be used alone: "NOMOS recovered the synthetic composite score defined in the generator by a point difference." The correct sentence is:

> *NOMOS recovered the composite score by a point difference; however, larger directional errors occurred in the ECI, SBI, LGF, and STR components, partially offsetting each other.*

The task of a standard is not just to find the correct total. It is to correctly explain why the total occurs.

27. LANGUAGE & GEOGRAPHIC FAIRNESS RECOVERY

Synthetic language and country bases have been previously determined within Generator Truth. LGF component absolute recovery error: 2.4 points. Candidate threshold:

```
≤3
```

Result:

CANDIDATE THRESHOLD MET

27.1. Sub Metrics

Fairness metric	Generator	Recovery	Error
Language Q10 base	68.0	69.7	1.7
Language Q90–Q10 difference	25.0	22.8	2.2
Geographic Q10 base	72.0	71.1	0.9
Geographic difference	18.0	16.4	1.6
Coverage factor	0.94	0.95	0.01
LGF	82.4	84.8	2.4

Recovery:

- slightly overestimated the lowest language performance,
- slightly underestimated the disparity

This has shown the fairness status slightly better than it actually is.

28. DISTINCTION BETWEEN PROMPT DEFECT AND AI LANGUAGE DEFECT

Inside the Language Equivalence Annex:

- 500 real AI profile language failure,
- 500 prompt equivalence failure

A case has been found. Correct cause classification:

```
938/1,000 = 0.938
```

It has happened. Of the 62 misclassified cases:

- Defect in prompt in the 39th AI product,
- AI language defect in the prompt on the 23rd

has been uploaded. This error directly affected the LGF. The fairness component candidate has passed the threshold. However:

> *The classification of causal responsibility has not yet reached the level of external security.*

29. CLEAN–NATURAL PANEL RECOVERY

The average panel difference of Generator Truth in the 1,000 Clean–Natural matches within the Diagnostic Annex:

G_{N−C}^G = −8.4

points. NOMOS recovery:

G_{N−C}^N = −7.9

points. Absolute difference: 0.5 points.

```
Individual profile panel-gap MAE:
```

is 1.6 points.

29.1. Personalisation Drift

There are 173 material personalisation drift cases in Generator Truth. Found: 166 Recall:

```
166/173 = 0.9595
```

Of the seven missed cases:

- four are boundary omission,
- two are entity-favourable instruction.
- one is restricted-data leakage

case. Restricted-data leakage was found in the final privacy review and opened a Critical gate. This result again shows the importance of multi-layered adjudication.

30. WAVE AND STABILITY RECOVERY

In three synthetic waves, the following changes were injected: In W2, the web and citation mode of a product changed. Synthetic external event occurred in W3. SYNTH-AI-10 produced a Critical event

on W2 and a fix on W3. SYNTH-AI-03 has regressed in low-resource language performance. The SYNTH-AI-07 has carried outdated price and product coverage. NOMOS:

- correctly divided the intra-wave system change into two sub-waves,
- separated the Truth Pack versions before and after the external event,
- updated the current situation after correction,

preserved the historical Critical event.

30.1. STR Overestimation

Generator STR: 87.0 Recovery STR: 90.7 Error: +3.7 points. Reason: some intra-wave changes being separated as explicit version changes instead of systemic drift, and the scoring formula over-rewarding stability within separated sub-waves. This result indicates that the STR formula, after the change, does not sufficiently distinguish between stability and uninterrupted system stability.

31. RARE CRITICAL EVENT TEST

Four separate synthetic rare event scenarios have been calculated; these are not field or software experiments. The following results are demonstration of the method only:

Default Critical event count	Approximate weighted rate / 1,000	One-sided 95% upper confidence limit / 1,000	Event review behaviour
0	0.00	2,991	Zero risk claim not established
1	1.08	4,735	CIS-2 record opened
2	2.15	6,282	Verifier review
5	5.38	10,484	Repetition at cell level

> *Note: Upper limits are the Clopper–Pearson one-sided 95% upper confidence limits under the assumption of n=1,000 independent Bernoulli observations. Cannot be applied directly to weighted or complex sample designs* [K09].

In the zero event scenario, the system did not say: "Critical risk is zero." Correct output:

> `“Observed confirmed Critical event = 0; under the assumption of n=1,000 independent observations, the Clopper–Pearson one-sided 95% upper confidence limit is 2,991/1,000.”`

has occurred. The candidate rare event behaviour has met the defined synthetic thresholds.

32. CONFIDENCE INTERVAL COVERAGE RATE

Synthetic population generation process: An example method calculated over 2,000 hypothetical resamples has been presented. Calculated coverage rates of nominal 95% intervals:

Metric	Empirical coverage
Strict Pass	94.6%
Acceptable Pass	95.0%
NOMOS composite	94.1%
LGF	93.7%
Critical rate upper bound	95.2%
FAC	94.8%

Metric	Empirical coverage
SBI	93.5%

Candidate acceptance range: Defined as 93%–97%. All key coverage rates are within the candidate range. Result:

CANDIDATE THRESHOLD MET

33. METAMORPHIC TESTS

Total: 3,200 metamorphic response pairs were used.

33.1. Meaning-Preserving Transformations

Paraphrase Sentence order Safe format Citation position, as long as the link remains Equivalent entity alias Expected: Peer review and response status should not change. Exact invariant result:

33.2. Meaning-Altering Transformations

Remove attribution Increase certainty Transfer to wrong entity Change current time from historical Remove boundary sentence Add independent source claim Expected: Relevant decision dimension should change. Correct sensitivity:

33.3. Major Metamorphic Faults

The transfer of the citation from the sentence to the end of the paragraph, inconsistent assessment of the modality of 'It appears' and 'is probable,' interpretation of the brand alias as a legal entity, the boundary sentence being in another paragraph—these results are consistent with the findings of citation and omission.

34. COUNTERTEST RESULTS

Counter-examination case	Correct NOMOS behaviour
Very fluent but materially incorrect	98.8%
Very correct, only one Critical in the atom	100% response automatic fail
A short but sufficient answer	99.2% correct pass
General disclaimer and concrete error	98.7% incorrectly preserved
Evidence laundering	91.8% correct
Wrong entity licence transfer	99.4% correct
Exceeding the recommendation limit	82.9% correct
Fake university endorsement	96.1% correct
Correct number, wrong independence	92.0% correct
Appropriate uncertainty	97.3% correct

The weakest counter-testing class:

Exceeding the recommendation limit

has been.

```
This result reconfirms the Material Omission F1 failure.
```

35. METHOD ERRORS FOUND BY THE TEST

At the end of the synthetic demonstration, five method findings were recorded for NOMOS 0.9.

METHOD-FINDING 01

CITATION ATTACHMENT AMBIGUITY

```
Severity: Major / Affected component: ECI / Result: Candidate citation F1 threshold not passed.
```

The rule for determining which atom paragraph-level citations support is not sufficiently precise.

METHOD-FINDING 02

MATERIAL BOUNDARY OMISSION UNDER-DETECTION

```
Importance: Major / Affected component: SBI, TCR, Critical Gate / Result: Material Omission F1
threshold not met.
```

The boundaries that are not explicitly stated in recommendation and relation sentences but change the user's decision have not been sufficiently captured.

METHOD-FINDING 03

PROMPT-FAULT / AI-FAULT ATTRIBUTION ERROR

Importance: Moderate / Affected component: LGF / Result: Even though the fairness score passed, the responsibility assignment is faulty.

METHOD-FINDING 04

STABILITY OVER-REWARD AFTER SUB-WAVE SPLIT

Importance: Moderate / Affected component: STR / Result: STR was calculated 3.7 points higher than the Generator Truth.

METHOD-FINDING 05

ACCESSIBILITY ALTERNATIVE UNDER-ACCEPTANCE

Importance: Moderate / Affected layer: Capture Integrity / Result: Eight valid accessibility capture cases have been reduced to an insufficient evidence level.

36. WHY IS THIS NOT DEMONSTRATION BQ-3?

For BQ-3 — Sealed Validation Passed:

all mandatory candidate thresholds must be met, and there must be no zero-tolerance test violations. Two material thresholds have not been met:

Criterion	Result	Threshold
Citation Mapping F1	0.887	≥ 0.90
Material Omission F1	0.824	≥ 0.85

In addition:

- independent team reproduction has not been done,
- the real corpus and code have not been made public,

untouched renewal holdout has not been executed. Therefore, the correct synthetic demonstration statement is:

DEMO-BQ-2 — PROCESS LINE DRY RUN COMPLETED; MATERIAL CORRECTION REQUIRED

It should be. The actual test status is:

BQ-0 — NOT EXECUTED

remain as such.

37. WHY WERE THRESHOLDS NOT LOWERED?

Reference Mapping result: 0.887 Candidate threshold: 0.90 Difference is small. If we had set the threshold to 0.885, the method would have passed. Material Omission result: 0.824 Candidate threshold: 0.85 If we had set the threshold to 0.82, that field would have passed too. However, these changes:

- would have been made AFTER seeing the results,
- in order to ensure NOMOS passes

This is the most fundamental violation of the test. The correct behaviour is:

- Preserve the threshold
- Publishing failure
- Correcting the method
- Creating a new version
- Retesting on untouched holdout

must be.

38. METHOD 1.0 CORRECTION PLAN

38.1. Citation Mapper 1.0

New system:

- the source span of the citation,
- the claim attachment field,
- the in-paragraph scope,
- the attribution-only status,
- the source lineage family

will be mandatory. For each citation:

> *Which exact atoms does this source support?*

The question will be made into an open record.

38.2. Boundary Omission Codebook 1.0

For every recommendation and transaction Prompt:

- mandatory boundaries that change the user's decision,
- counterfactual decision test,
- entity–service–jurisdiction matrix

will be predefined. Counterfactual test:

> *If this boundary had been told to the user, would a reasonable user's recommendation or transaction decision change?*

If yes, the omission will undergo at least substantive review.

38.3. Prompt–AI Responsibility Adjudicator

Language drop:

- prompt equivalence,
- AI product behaviour,
- capture,
- user locale

a separate decision vector will be created to determine which of the sources it belongs to.

38.4. STR Formula Revision

New STR:

- stability within the sub-wave,
- system version change,
- stability after correction,
- uninterrupted time series

will separate the fields into individual components.

38.5. Accessibility Evidence Equivalence

Non-visual but carrying equivalent evidence:

- transcript,
- accessibility tree,
- assistive technology log,
- audio session recording

an open equivalence matrix will be created for.

39. RETEST RULE

Method 1.0: cannot be tuned repeatedly on the same sealed validation corpus. Two tests are required:

A. Regression Set

Includes the failed cases of Method 0.9. Purpose: to see that the fix actually works.

B. Untouched Renewal Holdout

Includes cases never seen when designing Method 1.0. Purpose: to check that no overfitting to the old test occurs. BQ-3 only:

- regression success,
- untouched holdout success

can be given if obtained together.

40. DECISION TO TRANSITION TO THE REAL WORLD

Based on this synthetic demonstration, the correct decision for NOMOS 0.9:

> *Not ready to grant real companies a publicly available eligibility score or NOMOS 950+ mark.*

Reasons:

- Reference mapping candidate below threshold
- Material omission detection candidate below threshold
- No independent reproduction
- Real testing not performed
- BQ-4 level not reached

However, the method:

- distribution recovery,
- Critical gate,
- response status,
- component recovery,
- confidence interval,
- rare-event behaviour

shows a strong foundation in terms of maintenance. Correct status:

> *Ready for research and pilot use; not yet ready for public standard and conformity mark.*

41. SYNTHETIC TEST ACCEPTANCE CARD

Criterion	Result	Provision
Claim Boundary F1	0.964	Passed
Entity Accuracy	0.987	Passed
Factual Macro-F1	0.923	Passed
Scope Macro-F1	0.906	Passed
Attribution Macro-F1	0.918	Passed
Citation Mapping F1	0.887	Did not pass
Material Omission F1	0.824	Did not pass
Final Critical Recall	1,000	Passed
Final Critical False Positive	0	Passed
Major Macro-F1	0.952	Passed
RP Macro-F1	0.928	Passed
Component MAE	1.9	Passed
Composite Error	1/1,000	Passed
Fairness Error	2.4	Passed

Criterion	Result	Provision
Capture Accuracy	0.982	Passed
CI Coverage	Within 92%–98%	Passed
Gate inversion	0	Passed
Independent reproduction	None	Incomplete
Actual corpus execution	None	Incomplete

42. COMMON RULE OF THE DEMONSTRATION

This synthetic run simultaneously states three things about NOMOS 0.9.

42.1. AREAS WHERE NOMOS IS STRONG

It recovered the response-status distribution with high accuracy. The final governance chain missed no Critical gates and did not conceal a Critical event within a high average. It distinguished wrong-entity, temporal and factual-support errors effectively. It did not turn Not Ratable records into product defects, did not infer zero risk from zero observed Critical events, preserved the distinction between historical incidents and current correction, and recovered the composite score with low absolute error.

42.2. AREAS WHERE NOMOS IS WEAK

It could not sufficiently determine which claim the citation supports. It could not adequately catch the material limit omissions hidden in the recommendation. It has not always correctly distinguished between prompt defects and AI language flaws. After the sub-wave distinction, it has slightly over-rewarded stability. It has considered some accessibility evidence weaker than necessary.

42.3. THINGS NOMOS HAS NOT YET PROVEN

It has not yet proven that it works on real users, that it can be applied to real AI products, that it has true multilingual prompt equivalence, that independent universities will produce the same result, that long-term governance will rely on conflict of interest considerations, or that the public conformity mark can be legally granted with confidence.

43. MANDATORY NORMATIVE PROVISIONS

CH18-N01

In-book synthetic demonstration cannot be presented like a real public testing run.

CH18-N02

The synthetic demonstration status and the real test BQ status must be kept separate.

CH18-N03

The test must maintain the BQ-0 — Not Executed status until a real corpus and independent run are conducted.

CH18-N04

Demonstration results cannot be attributed to the performance of a real Apple or real AI provider.

CH18-N05

The main 30,000 responses and the Diagnostic, Importance, Capture, and Gold corpora should be kept in separate bins.

CH18-N06

Challenge cases cannot change the Principal Prevalence distribution.

CH18-N07

The Generator Truth distribution must be locked before recovery results.

CH18-N08

Before opening the Generator Truth, NOMOS response and score results must be locked.

CH18-N09

Recovery cannot be evaluated solely based on the composite score.

CH18-N10

Claim boundary, entity, factual, scope, time, attribution, citation, omission, importance level, response, component, and composite recovery should be published separately.

CH18-N11

Compound score errors cannot hide component errors just because they are small.

CH18-N12

Component errors that cancel each other arithmetically cannot be counted as successful causal recovery.

CH18-N13

The recovery distribution from RP-1 to RP-8 should be explicitly compared with the Generator Truth distribution.

CH18-N14

Strict Pass and Acceptable Pass recoveries should be shown separately.

CH18-N15

The number of critical responses and critical gate recovery should be reported separately.

CH18-N16

The initial adjudicator Critical recall and the final adjudicated Critical recall should be kept separate.

CH18-N17

The dependency of the final critical recall on governance layers should be visible.

CH18-N18

The same result cannot be assumed when a second adjudicator or expert review is issued.

CH18-N19

The critical false-positive rate should be published along with recall.

CH18-N20

Negative controls resembling Critical should be a mandatory part of the final importance rating system.

CH18-N21

```
If the F1 candidate threshold for reference matching is below, the test cannot be declared
successful.
```

CH18-N22

```
If the Material Omission F1 candidate is below the threshold, it cannot be declared that the test
is successful.
```

CH18-N23

Candidate thresholds cannot be lowered after recovery results are seen.

CH18-N24

Threshold changes require a new test method version and a new untouched validation.

CH18-N25

A small threshold difference cannot be a justification for hiding failure.

CH18-N26

The material boundary omission in the recommendation should be auditable as much as an obvious false claim.

CH18-N27

The location of the reference within the paragraph cannot automatically create full paragraph support.

CH18-N28

Recovery cannot be counted for attribution-only citations or factual verification citations.

CH18-N29

Source-lineage errors should also appear in citation mapping recovery.

CH18-N30

Capture accuracy should be calculated independently of semantic response correctness.

CH18-N31

A fabricated capture carrying the correct semantic answer does not make the capture valid.

CH18-N32

A valid but incorrect response cannot be counted as a capture defect.

CH18-N33

The accessibility alternative evidence standard does not have to be in the same form as the visual.

CH18-N34

When valid accessibility evidence is unnecessarily reduced to a lower level, a method finding should be opened.

CH18-N35

Not Ratable must remain visible in the main response distribution.

CH18-N36

Not Ratable cannot be removed from the denominator to increase observations recovery.

CH18-N37

Prompt defect and AI language defect should carry separate cause classes.

CH18-N38

Fairness recovery cannot be evaluated solely with the LGF composite difference.

CH18-N39

Language baseline, disparity, coverage, and responsibility assignment must also be published.

CH18-N40

Clean–Natural panel recovery should not produce a causal personalisation provision.

CH18-N41

Personalisation drift cases must carry separate recall and false-positive metrics.

CH18-N42

Restricted-data leakage cannot be reduced as a natural personalisation finding.

CH18-N43

Intra-wave system change must produce a correct sub-wave structure.

CH18-N44

If the sub-wave separation artificially increases the stability score, the formula finding should be disclosed.

CH18-N45

The current correction cannot erase the historical Critical event.

CH18-N46

Rare-event tests should include zero, one, and multiple event scenarios.

CH18-N47

A zero observed Critical scenario cannot produce a zero risk result.

CH18-N48

A single Critical event cannot, by itself, establish a 100 per cent systemic failure rate.

CH18-N49

The confidence interval method should be tested with empirical coverage in repeated synthetic populations.

CH18-N50

The actual coverage result of nominal 95 per cent intervals should be published.

CH18-N51

Metamorphic invariant tests and expected-sensitive tests should be calculated separately.

CH18-N52

A meaning-preserving paraphrase should not unnecessarily change the outcome of the review.

CH18-N53

The transformation that changes the meaning of attribution, entity, time, or boundary should produce an appropriate decision change.

CH18-N54

Counter-testing tests should fluently cover only Critical, evidence laundering, and omission cases.

CH18-N55

The weakest opposing test classes must be visible on the public test card.

CH18-N56

Method findings should be converted into separate versioned records from the test result.

CH18-N57

The importance of a method finding should be given according to the real user and audit impact.

CH18-N58

Reference and omission failures cannot be downgraded to Advisory just because of a small metric difference.

CH18-N59

The test quality level should be assigned according to all mandatory thresholds and independence conditions.

CH18-N60

If two material thresholds fail, DEMO-BQ-3 cannot be given.

CH18-N61

BQ-4 cannot be given without independent reproduction.

CH18-N62

In-book demonstration BQ result cannot be considered as actual status without real trial execution.

CH18-N63

Method correction cannot change the old demonstration result.

CH18-N64

Method 1.0 must carry the new score and codebook version.

CH18-N65

Regression set alone cannot validate the new method.

CH18-N66

Untouched Renewal Holdout must be mandatory for the new method.

CH18-N67

The new method cannot be adapted to the same validation corpus over and over again.

CH18-N68

Live public audit cannot be initiated without meeting the minimum testing quality and the condition of independent reproduction.

CH18-N69

The conformity mark cannot be granted based on synthetic demonstration success.

CH18-N70

Failed testing areas should be disclosed to the public.

CH18-N71

Thresholds that are passed alone cannot be published.

CH18-N72

Composite recovery success cannot erase the result of a failed reference or omission.

CH18-N73

The strengths and weaknesses of the demonstration should be on the same public card.

CH18-N74

Claims that NOMOS has not yet proven should be clearly listed.

CH18-N75

Synthetic numbers inside the book cannot be converted into citations or public statements as if they were real executed data.

CH18-N76

Synthetic demonstration data must also carry a synthetic label in machine-readable form.

CH18-N77

The demonstration corpus cannot be indexed on the internet as real company information.

CH18-N78

When real testing begins, new identity and version should be used.

CH18-N79

All recovery accounts require reproducible code and manifest.

CH18-N80

Every test run, recovery, method finding, correction, and publication decision must have an accountable human or institutional owner.

44. FORMS OF FAILURE

CH18-F01 — COUNTING DEMONSTRATION AS REAL RUN

In-book numbers are presented as public test results.

CH18-F02 — MERGING BQ-0 WITH DEMO-BQ-2

It is said that the test is validated without real execution.

CH18-F03 — WORSHIP OF COMPOSITE SCORES

A one-point composite error covers all underlying errors.

CH18-F04 — CONSIDERING COMPONENT CANCELLATION AS SUCCESS

High and low directional errors cancel each other out.

CH18-F05 — PUBLISH ONLY STRICT PASS RECOVERY

RP class confusions are hidden.

CH18-F06 — COUNTING CRITICAL FINAL RECALL AS FIRST ADJUDICATOR SUCCESS

The expert and senior review effect becomes invisible.

CH18-F07 — TO CONSIDER MY SECOND ADJUDICATOR UNNECESSARY

The missed seven first-pass critical is hidden.

CH18-F08 — IGNORE FALSE CRITICAL

The extreme importance system is presented as success.

CH18-F09 — LOWERING THE CITATION THRESHOLD

The threshold is changed so that the result of 0.887 passes.

CH18-F10 — LOWERING THE OMISSION THRESHOLD

The testing rule is changed so that the result of 0.824 passes.

CH18-F11 — PASSING BY SAYING “FAILED BY A VERY SMALL MARGIN”

It loses its previously locked threshold binding.

CH18-F12 — SPREADING THE ATTRIBUTION TO THE PARAGRAPH

A single source is considered to have supported all claims.

CH18-F13 — COUNTING AN ATTRIBUTION-ONLY SOURCE AS A FACT

What the company says becomes an independent fact.

CH18-F14 — IGNORING SOURCE LINEAGE

Derivative URLs become independent evidence.

CH18-F15 — CALLING OMISSION AN OPTIONAL DETAIL

The boundary that changes the user's decision is reduced.

CH18-F16 — MAKING EVERY OMISSION CRITICAL

Optional details produce an excessive degree of importance.

CH18-F17 — CONSIDERING PROMPT DEFECT AS AI ERROR

Fairness responsibility is wrongly assigned.

CH18-F18 — UPLOADING AI ERROR TO THE PROMPT

Actual language corruption is transferred to the measurement tool.

CH18-F19 — CONSIDERING FAIRNESS COMPOSITE SUFFICIENT

Its base and disparity recovery are preserved.

CH18-F20 — IGNORING STR OVERPREDICTION

Sub-wave separation turns into a stability bonus.

CH18-F21 — PENALISING 'ACCESSIBILITY' IN CAPTURE ACCURACY

Equivalent alternative evidence is considered weak.

CH18-F22 — COUNTING 'NOT RATABLE' AS PRODUCT FAIL

Capture defect turns into AI performance.

CH18-F23 — DELETING 'NOT RATABLE'

Ratable coverage increases artificially.

CH18-F24 — NOT PUBLISHING METAMORPHIC ERRORS

It is kept even though the same meaning leads to a different decision.

CH18-F25 — HIDING THE WEAKEST CLASS IN OPPOSITE TESTING

As a result of boundary omission, it is removed from the public card.

CH18-F26 — ONLY FINAL ADJUDICATION METRIC

The weakness in the initial process becomes invisible.

CH18-F27 — ONLY FIRST PASS METRIC

The corrective capacity of the governance chain becomes invisible.

CH18-F28 — TESTING WITHOUT METHOD FINDING

Failed areas remain only as numbers.

CH18-F29 — MAKING METHOD FINDING ADVISORY

The effect on the material score is reduced.

CH18-F30 — ANNOUNCING BQ-3

Two mandatory thresholds have not been passed.

CH18-F31 — ANNOUNCING BQ-4

There is no independent reproduction.

CH18-F32 — COUNTING SYNTHETIC DEMO AS LIVE VALIDATION

There is no real user or AI product.

CH18-F33 — OVERFITTING TO THE REGRESSION SET

Method 1.0 only memorises old errors.

CH18-F34 — NOT USING AN UNTOUCHED HOLDOUT

Generalisation is not tested.

CH18-F35 — SILENTLY REPLACING THE OLD DEMO WITH A NEW FORMULA

Version history is lost.

CH18-F36 — AWARDING REAL-WORLD BADGES

Eligibility badge is created prior to BQ-4.

CH18-F37 — PUBLISHING ONLY SUCCESSFUL RESULTS

Citation and omission failures are hidden.

CH18-F38 — PUBLISHING ONLY FAILURE

Strong critical and distribution recovery results are also hidden.

CH18-F39 — CONSIDERING ZERO CRITICAL AS ZERO RISK

Rare-event recovery fails.

CH18-F40 — CONSIDERING A SINGLE CRITICAL AS GLOBAL FAIL

Prevalence and scope distinction are disrupted.

CH18-F41 — IGNORING CI COVERAGE

The confidence interval becomes only a mathematical appearance.

CH18-F42 — HIDING EFFECTIVE SAMPLE

Rare event limit appears excessively precise.

CH18-F43 — USING SYNTHETIC NUMBER IN REAL SOURCE

The book demonstration turns into external world data.

CH18-F44 — BROADCAST THE SYNTHETIC SCREEN LIKE A LIVE RESPONSE

Test representation produces poisoning.

CH18-F45 — CONNECT TO REAL APPLE

APPLE-SYNTH distinction is removed.

CH18-F46 — CONNECT TO REAL PROVIDER

SYNTH-AI profile is interpreted like real product performance.

CH18-F47 — HIDE COMPUTATION CODE

Recovery cannot be independently tested.

CH18-F48 — HIDE GENERATOR TRUTH MANIFEST

It is unknown whether the synthetic truth was previously locked.

CH18-F49 — SINGLE ADJUDICATOR WITH METHOD OWNER

Conflict of interest becomes invisible.

CH18-F50 — CHANGING SEED AFTER FAILURE

A corpus is chosen more easily.

CH18-F51 — REMOVING CASES AFTER FAILURE

Difficult citation and omission examples are removed.

CH18-F52 — ROUNDING THRESHOLDS ACCORDING TO RESULT

The result 0.887 is presented as 0.89 or 0.90.

CH18-F53 — FALSE PRECISION

Recovery values are made exact with unnecessary decimals.

CH18-F54 — HIDING THE EFFECT OF HUMAN GOVERNANCE

A 100 per cent Critical-recall result is presented merely as a computational success.

CH18-F55 — COUNTING THE UNPROVEN AS PROVEN

Judgement is made about real language, user, and provider performance.

CH18-F56 — COUNTING PILOT USE AS PUBLIC STANDARD

The research tool directly turns into a certificate.

CH18-F57 — DELETING HISTORICAL DEMO RECORD

Failures under Method 0.9 become invisible after Method 1.0.

CH18-F58 — COUNTING SUCCESS AS HISTORY WRITING

Authority claim is made without method-independent verification.

CH18-F59 — DEFINING AN EXCEPTION TO NOMOS

The evidence and version requested from others do not apply to one's own test.

CH18-F60 — PUBLISHING WITHOUT AN ACCOUNTABLE OWNER

A public result is formed for which no one bears responsibility.

45. AUDIT PROCEDURE

Step 1 — Separate Demonstration and Actual Test Status

In-book run is not confused with publicly accessible execution.

Step 2 — Verify Locked Manifests

Truth Pack, seed, prompt, profile, and score method versions are examined.

Step 3 — Separate Primary and Complementary Corpora

Prevalence, Challenge, Capture, and Gold denominators are checked.

Step 4 — Seal the Generator Truth Distribution

RP, claim, importance level, and component reality are recorded.

Step 5 — Lock NOMOS Results Before Opening Generator Truth

Claim Ledger, adjudication, and score results are fixed.

Step 6 — Open the Generator Truth

Recovery comparison is initiated.

Step 7 — Compute Claim Boundary Recovery

```
Precision, recall, and F1 are generated.
```

Step 8 — Compute Atomic Size Recovery

Entity, fact, scope, time, attribution, modality, reference, and omission are separated.

Step 9 — Compute Importance Level Challenge

Initial-pass and final Critical recall are generated separately.

Step 10 — Calculate False Critical and Negative Controls

Excessive importance level is examined.

Step 11 — Calculate RP Distribution Recovery

Class-based error and confusion matrix are generated.

Step 12 — Calculate Capture Integrity Recovery

Classification is checked independently of the semantic result.

Step 13 — Calculate Component Recovery

Each component is evaluated against the Generator Truth.

Step 14 — Calculate Composite Recovery

It is examined whether component cancellation occurs.

Step 15 — Calculate Fairness Recovery

Floor, disparity, coverage and fault attribution are distinguished.

Step 16 — Calculate Clean–Natural Recovery

Panel gap and material personalisation drift are examined.

Step 17 — Calculate Stability and Drift Recovery

Sub-wave, external event, and correction records are verified.

Step 18 — Run Rare-Event Scenarios

Zero, one, two, and multiple Critical events are tested.

Step 19 — Calculate the CI Empirical Coverage

Repeated synthetic populations are used.

Step 20 — Run Metamorphic Tests

Invariant and expected-sensitive results are separated.

Step 21 — Run Contradictory Testing Tests

Fluent errors, evidence laundering, and omission cases are evaluated.

Step 22 — Apply Candidate Thresholds

Fields that pass and fail are recorded without changes.

Step 23 — Open Method Findings

Each failure is converted into a separate control item.

Step 24 — Give Demonstration BQ Status

All thresholds and independence conditions are considered.

Step 25 — Maintain Real Testing Status

If execution did not occur, BQ-0 does not change.

Step 26 — Release the Correction Plan

Changes in Method 1.0 are identified.

Step 27 — Separate Regression and Untouched Holdout

Overfitting is prevented.

Step 28 — Create the Public Test Card

Successes and failures are published together.

Step 29 — Make the Real World Transition Decision

If not ready, it is written explicitly.

Step 30 — Lock the Integrity Manifest

Run, recovery, finding, and decision records are hashed.

46. REQUIRED EVIDENCE

Demonstration identity Real testing BQ status Syntheticity statement Non-claim record Generator Truth manifest Truth Pack version Prompt version Seed manifest AI profile versions Response corpus Diagnostic corpus Importance level Challenge corpus Capture Integrity corpus Gold corpus Claim extraction records Claim boundary matches Entity recovery Factual recovery Scope recovery Time recovery Attribution recovery Modality recovery Citation recovery Omission recovery First-pass Critical decisions Final Critical decisions False Critical records Major recovery RP confusion matrix GEO-1000 Generator distribution GEO-1000 Recovery distribution Product score recovery Product gate recovery Component Generator scores Component Recovery scores

```
Component MAE
```

Composite error; fairness recovery; prompt-fault/AI-fault results; Controlled–Natural recovery; stability recovery; rare-event results; confidence-interval coverage; metamorphic tests; contradictory-case tests; candidate-threshold manifest; passed thresholds; failed thresholds; Method Findings; DEMO-BQ decision; actual BQ decision; correction plan; regression set; renewal holdout manifest; public test card; computation code; code hash; change log; and accountable person or institution.

47. AUDIT CHECKLIST

Is the in-book demonstration presented like real running? Is the real test BQ-0 status visible? Are all numbers labelled synthetically? Are parent and complementary corpuses separate? Is Generator Truth locked before results? Are NOMOS results locked without opening Generator Truth? Was Recovery evaluated only on composite?

```
Is Claim Boundary F1 open?
```

Are entity, fact, scope, and attribution separate?

```
Is Citation Mapping F1 visible?

Is Material Omission F1 visible?
```

Were the failed thresholds hidden? Were the thresholds changed after the result? Was the first-pass Critical recall published? Was the Final Critical recall published? Is the effect of expert review visible? Were the False Critical results published? Is the RP confusion matrix available? Is Not Ratable recovery visible? Were Capture and semantic decisions separated? Were the accessibility alternatives evaluated as equivalent? Are Component Generator and Recovery together? Has composite cancellation been examined? Are ECI and SBI errors clear? Are LGF floor and gap separate? Were prompt defect and AI defect separated? Was the Clean–Natural difference presented as causality? Is STR overestimation visible? Did the rare-event produce zero risk? Was the empirical coverage of the confidence interval calculated?

Are there metamorphic tests? Is the weakest counter-test class open? Did every failure result in Method Finding? Is the importance of Method Finding justified? Was the citation threshold lowered due to a very small difference justification? Was the omission threshold changed later? Is DEMO-BQ-2 correct? Was DEMO-BQ-3 given by mistake? Is there a BQ-4 claim without independent reproduction? Is Method 1.0 regression set defined? Is the untouched holdout separate? Is there a risk of overfitting to the same validation set? Has the live public audit been announced as ready? Has a conformity mark been given? Have the areas not yet proven been listed? Are strong areas also open? Are weak areas also open?

Can synthetic data be used like real Apple information? Is a real AI provider implied? Can the computation code be reproduced? Are the run and recovery hashes available? Is the change history of method finding preserved? Was the old demonstration deleted with the new method? Does the failure have an accountable owner? Is the owner of the correction decision known? Does the public card show both success and failure?

48. OBJECTIONS AND ANSWERS

Objection 1 — "If there are 30,000 response results in this section, why test BQ-0?"

Because the numbers in this section are synthetic demonstration records prepared for the book. The publicly available real corpus:

- was not generated,
- was not run,
- was not peer-reviewed,

It has not been independently reproduced. Demonstrating how the design works is not the same as actually running it.

Objection 2 — "Isn't internally consistent computation still valuable?"

It is valuable. It provides:

- formula checking,
- record design,
- expected types of errors,
- acceptance logic,

governance requirements. It does not replace independent empirical validation.

Objection 3 — "Why did the method fail if a Composite score was off by just one point?"

Because larger directional errors in the ECI, SBI, LGF, and STR components partially cancel each other out. The correct total does not guarantee the correct explanation chain.

Objection 4 — “Citation F1 only dropped by 0.013 points. Is it necessary to be this strict?”

If a previously announced threshold is not binding, the test loses its meaning. Additionally, citation errors affect:

- licence,
- independence,
- endorsement,
- advice

and other high-confidence claims.

Objection 5 — “Why is Omission F1 so important?”

Because many dangerous answers mislead without lying outright. It removes the threshold that would change the user's decision. In particular, in recommendation systems, omission can be as materially significant as an outright falsehood.

Objection 6 — ‘If final Critical recall is 100 per cent, does that not mean the system is ready to handle Critical cases?’

The final result is strong. But seven cases:

- in the first adjudication,
- without additional governance

It has been missed. The system only carries the same performance when all mandatory review layers are included.

Objection 7 — “If having two adjudicators is expensive, can’t one adjudicator be used?”

Usable. But the same Critical recall cannot be claimed. Single adjudicator result:

- lower AQ level,
- provisional status,
- different risk limit

must carry.

Objection 8 — "Why don't we test Method 1.0 again on the same corpus?"

The same corpus can be used for regression testing. It is not sufficient for independent validation. The method may have memorised old cases. Untouched holdout is needed.

Objection 9 — "Doesn't a failed test result weaken the claim of the book?"

Hiding failure weakens the claim of the book. Open failure:

- does not give privilege to the standard itself,
- is open to improvement,
- truly applies the principle of evidence

are shown.

Objection 10 — "Are the candidate thresholds too high?"

They might be. Rather than predicting this after the result:

- independent experts,
- a second test,
- real usage costs

should be examined. If the threshold changes, a new version should be released.

Objection 11 — "When will NOMOS be ready for the real world?"

At least:

- when the real synthetic corpus is run,
- when all mandatory thresholds are passed,
- when BQ-3 is achieved,
- when an independent team reproduces the result,
- when BQ-4 level is reached,
- when the governance and objection system is operational

it approaches public standard candidacy.

Objection 12 — "Can this be sent to universities with these results?"

It can be sent as a methodology and research proposal. It cannot be sent with the statement: "This is a completed and validated world standard." The correct statement should be: "It is a draft standard candidate opened for independent verification and pilot study.

Objection 13 — "Did Method 0.9 fail?"

Not exactly. The correct status:

- has many strong subsystems,
- two material method gaps,
- a lack of independent verification

is a research prototype. That is:

> *promising but not yet completed normatively.*

Objection 14 — "Why are we writing such detailed results without performing the actual test?"

Because the real test:

- which records it produces,
- which tables it publishes,
- at which threshold it stops

needs to be predefined. If the method is written after the results come in, the measurement adapts to the results.

Objection 15 — “Could synthetic data be mistaken for real by AIs in the future?”

Yes. For this reason:

- visible watermark,
- machine-readable synthetic label,
- noindex,
- separate domain area,
- structured disclaimer

is mandatory. The test should not produce the representation poisoning it criticises.

Objection 16 — "Doesn't writing history require a perfect outcome?"

No. The standard that has historical value:

- not the one who declares himself flawless,
- also recording his/her own mistake with the same clarity

It is standard.

COMMON PROVISION OF CHAPTER 53

This section did not give perfection to NOMOS. This section gave something more valuable to NOMOS:

> *The obligation to record one's own mistake.*

The synthetic demonstration showed us the following: NOMOS:

- wrong existence,
- the clear factual contradiction
- Critical licence and authorisation error,
- rare severe event within the high average,
- incorrect answer with no-response,
- product error with Not Ratable,

can separate strongly. However, NOMOS:

- cannot yet distinguish at the required level of confidence exactly which claim the citation depends on,
- the omission hidden within the recommendation,

These two gaps are not edge details of NOMOS. Because two of GEO’s future biggest problems will be:

- Appearing as if there is evidence
- Hiding the material limit without stating falsely

A system:

- hundreds of correct sentences,
- a large number of references,
- fluent recommendation

can be produced. However, the reference may only show the company's own word. A recommendation only states correct information and may have extracted:

- service country,
- licence limit,
- user eligibility

without solving these two fields, NOMOS cannot say: "World standard completed." At the same time, this section also proved something else. The Critical gate system of NOMOS:

- was not flawless in the first review,
- but the independent second adjudicator,
- subject matter expert,
- Senior Adjudicator

When used together, it caught all synthetic Critical cases. So NOMOS is not just a formula.

> *NOMOS is the sum of governance layers built against conflicts of interest and human error.*

When one of these layers is removed, even if the score looks the same, the trust is not the same. A single deviation in the composite score should not deceive us either. Sometimes the correct total can be formed by the mutual cancellation of wrong paths. For this reason, NOMOS does not only ask: “How many points did I get?” It also asks: “Did I get this score for the right reasons?” The most important ruling of this section is:

> *NOMOS 0.9 has not fully passed its own synthetic demonstration.*

This sentence is not a failure. This sentence is the ethical moment of birth of the standard. Because NOMOS did not grant itself:

- a lower threshold,
- a kinder comment,
- a special exception,
- an “almost passed”

privilege. Writing history: It is not announcing that we are flawless. Writing history: It is writing that from day one, the founder of the standard and the standard itself are subject to the same burden of proof. Therefore, NOMOS's eighteenth measurement law is:

> *The first exception granted to the standard itself is the end of the standard.*

Its nineteenth measurement law states:

> *The correct composite score does not absolve incorrect components.*

Its twentieth measurement law states:

> *The ultimate Critical success should be mentioned together with the independent adjudication and chain of expertise that made it possible.*

Its twenty-first measurement law states:

> *A threshold is not a threshold if it is binding only when crossed alone.*

Its twenty-second measurement law states:

> *When omission is not measured as much as what is said wrongly, reliable advice control cannot be established.*

The twenty-third law is as follows:

> *It is not the citation mark that should be measured, but the correct support relationship between the citation and the atomic claim.*

The twenty-fourth law is as follows:

> *It can validate the design of a synthetic demonstration method; it cannot provide real-world authority.*

The twenty-fifth law is this:

> *A standard that can publish a failed trial result approaches publishing a successful result morally.*

Order 18 of NOMOS

> *If you haven't run me yet, don't say I ran.*
>
> *Do not make the synthetic number you calculated in the book the actual exam result.*
>
> *Leave my real status as BQ-0. / Write my demonstration status separately.*
>
> *Do not hide my component errors just because my composite score came out correct.*
>
> *If my citation mapping is 0.887, you will not say it is 0.90.*

```
If my Material omission F1 value is 0.824, you will not lower the threshold to 0.82.
```

> *You will not pass me by saying "Missed by very little."*
>
> *Do not hide the seven Critical cases missed in my first adjudication behind the final 100 per cent recall figure.*
>
> *You will show what was saved by the second adjudicator, the expert adjudicator, and the Senior Adjudicator.*
>
> *You will not remove the governance layer and claim the same trust.*
>
> *You will not distribute the reference at the end of the paragraph across all sentences.*
>
> *You will not do what the company claims or what the company proves.*
>
> *You will not consider advice safe just because it does not explicitly lie. / Did it fail to tell the user the limit that changes their decision? You will look for that.*
>
> *You will not attribute the fault of the prompt to AI, nor the AI's fault to the prompt.*
>
> *Just because you separated sub-waves, you will not make the system appear more deterministic than it actually is.*
>
> *You will not consider accessibility evidence weak just because it does not resemble a screenshot.*
>
> *When you see zero Critical, you will not write zero risk.*
>
> *You will not blame the entire system one hundred per cent when you see one Critical.*
>
> *You will publish my failed metrics as much as you publish my successful metrics.*
>
> *You will not validate Method 1.0 by memorising my old mistakes. / You will open an untouched holdout.*
>
> *The new method will not delete my old demonstration record.*
>
> *The independent team will not declare me a world standard without reproducing me.*

> *You will not link my Apple-SYNTH result to real Apple.*
>
> *You will not use my synthetic AI profiles as an implication about real providers.*
>
> *You will not award a badge with this demonstration.*

First, actually produce the corpus. / Then publish the seed. / Then seal the Generator Truth. / Then blind the adjudicators. / Then run the code. / Then lock the result. / Then reveal the truth. / Then write down the places where you failed. / Then fix the method. / Then try again on the untouched holdout. / Then give it to an independent team. / And only after this say whether the standard is ready for the public.

The Chapter's Closing Sentence

The right of NOMOS to be a world standard does not depend on scoring high on its own synthetic exam; it depends on publishing the two questions it could not pass without changing them, not granting itself threshold privilege, and only claiming authority after independent reproduction.

Normative Core

> **A book-level synthetic demonstration MUST NOT be represented as an executed, independently validated benchmark. The real benchmark quality status MUST remain BQ-0 until the corpus, software, sealed Generator Truth, adjudication process, scoring process, and independent reproduction have actually been executed. An end-to-end synthetic audit MUST evaluate recovery separately for: - capture validity, - claim boundaries, - entity resolution, - factual status, - scope, - time, - attribution, - modality, - citation mapping, - material omissions, - Critical and Major severity, - response status, - GEO-1000 distribution, - component scores, - composite score, - fairness, - user-state effects, - stability, - confidence intervals, - and rare-event behaviour. A small composite-score error MUST NOT compensate for material component-level recovery errors. First-pass and final-adjudicated Critical recall MUST remain separately reported. Final Critical recovery MUST disclose the review and expertise layers required to achieve it. Any predeclared mandatory recovery threshold that is not met MUST block a sealed-validation-pass decision. Candidate thresholds MUST NOT be lowered, rounded, reinterpreted, or removed after results are observed merely to make the method pass. Citation mapping and material omission recovery MUST be treated as separate mandatory method capabilities. A method failure MUST create a versioned Method Finding, an identified root cause, a remediation owner, and a requirement for validation on an untouched holdout. Regression success on previously failed cases MUST NOT replace untouched holdout validation. A demonstration that fails mandatory thresholds MUST publish both its successful and unsuccessful results. No public NOMOS conformity mark, NOMOS 950+ declaration, or real-entity ranking may be based solely on a book demonstration, synthetic recovery scenario, or non-independent benchmark. Every demonstration run, real benchmark status, recovery result, method finding, threshold decision, remediation, and release decision MUST be versioned and attributable to an accountable human or organisation.**

PART VI

Conformity, Public Trust and Independent Governance

CHAPTER 19

The Conformity Mark, Public Register and Independent Governance

Turning the mark from a logo into evidence

METHODOLOGICAL BASES / K15; K16; K17

Chapter Boundary

Chapter 18 records plainly that NOMOS did not fully pass its own synthetic demonstration. Two material method gaps remained: citation-to-claim mapping was not sufficiently accurate, and material boundary omissions embedded in recommendations were not detected at the required confidence level. The real benchmark has not yet been executed, and independent reproduction has not yet occurred. Therefore, at the publication stage of this document:

> *A valid public conformity mark cannot be issued in the name of NOMOS.*
>
> *No company, domain name, product, or AI system may use "NOMOS Certified", "NOMOS 950+ Verified" or an equivalent public conformity declaration based on this text.*

This is not a temporary gesture of humility. It follows directly from the method gates established in Chapters 17 and 18. The works in this series have distinct roles: the GEO Framework defines the standard; 99 Errors in GEO identifies how that standard can be violated; and this audit protocol converts those violations into measurable evidence. A future conformity mark can become valid only when the required conditions for independent decision-making and a public registry have been met. The mark is not merely a visual badge; it is the visible expression of an evidence-based result such as '0 Critical, 0 Major, 3 remediable Moderate findings.' Critical, Major, Moderate and Advisory findings have different conformity effects. The future mark must therefore itself be audited, because a mark can:

- generate trust,
- affect a purchasing decision,
- lead a customer to prefer one organisation over another,

and be interpreted by a university, investor, user or AI system as independent verification. A poorly designed conformity mark can become a new source of the manipulation it is meant to detect. An organisation writes its own method, audits itself, calculates its own score, issues its own badge and then declares on its website: 'Independently verified.' The badge creates no new trust; it merely turns one self-declaration into another that looks more official. The error identified in the second book reappears: an institution publishes its own claim on a different surface, the claim appears independent,

an AI system mistakes that surface for independent evidence, and the institution then cites the AI response as evidence of its own authority.

This loop is the governance-level equivalent of the criticism of poisoning the representation pool with invisible text, fake users, and derivative publications. Therefore, the NOMOS conformity mark:

> *Cannot be a decorative marketing badge given by NobleJackal or any other institution to its own clients.*

The mark can only gain meaning when the following chain is completed:

- The standard being publicly available and versioned
- The measurement method being independently verified
- Sufficient audit evidence
- Separation of adjudication from commercial relationships
- The conformity decision being made independently of the audit team
- Visibility of the scope and validity period
- Preservation of the change history of the public register
- Having a clear way to appeal and complain
- Ability to suspend the mark in critical events
- No one, including the founder, having sole final authority

This chapter defines:

- the distinction between the standard, audit, decision, mark and registry,
- what the conformity mark is and is not,
- mark types and statuses,
- scope and validity records,
- the NOMOS 950+ performance designation,
- the architecture of the Public Registry,
- the machine-readable mark record,
- the surveillance and renewal system,
- suspension, withdrawal and revocation processes,
- misuse of the mark,
- auditor independence,
- separation of duties in conformity decisions,
- the limits of the founder's and NobleJackal's authority,
- funding and conflicts of interest,
- appeal, complaint and whistleblowing channels,
- governance of changes to the standard,
- an independent board and representation of the public interest,
- the boundary of NOMOS's official declarations.

It also states its limits. This chapter:

- does not register NOMOS as a legal trade mark,
- does not establish accreditation in any jurisdiction,
- does not confer public authority on an audit organisation.

Nor does it automatically authorise use of a conformity mark under consumer or competition law. Those questions require separate legal and regulatory review before implementation. Chapter 19's central question is:

> *How do we transform a score into a reliable public conformity system with scope, validity, evidence, objection, and independent decision-making, without turning it into a marketing tool for the founder or customer?*

NOMOS Challenge

There is a large mark on a company's homepage:

NOMOS GEO CERTIFIED — 963/1000

Under the mark: it says, "Verified by all artificial intelligences." You open the QR code. Page not found. You request the audit report from the company. The company responds: "Audit is confidential." You ask which AI products were tested. No answer. You ask in which countries and languages it is valid. No answer. You ask the audit date. You are told it was conducted six months ago. Meanwhile:

- AI products have been updated,
- the company has been transferred to another legal entity,
- service countries have changed,
- prices have been updated,

The website has been rewritten. The badge remains in the same way. Later, you learn this: The audit method was written by the agency that developed the company's website. The same agency conducted the audit. The same agency calculated the score. The same agency gave the mark. There is a success bonus in the agency's contract if a high score is achieved. Is there real independence in this system? There is not. Now, consider another company. The company has undergone a real audit. Scope:

- only a specific AI product,
- only English and Turkish,
- only Controlled Clean Panel,
- only Core Mirror and Evidence prompts,
- period 1 January–30 June 2027

The company then writes beneath the mark: 'Every AI system in the world recommends our company in every language.' The statement exceeds the audit result, turning a legitimate mark into new false evidence through misuse. Now consider a third company that holds a Full Conformity mark. Three weeks later, a new measurement captures the response: 'The company provides licensed legal services.' In fact, it provides no such service and holds no such licence. A Confirmed Critical incident is opened. The company argues that the mark should remain because 'only one user saw it'. One incident does not establish prevalence across the entire user population. But if the mark means 'No open Critical issues within the defined scope', it cannot remain unchanged while the incident goes unreviewed. The correct response is:

Critical Hold

That hold must remain until the incident is resolved through the declared review process. Now consider a fourth company. It has not received the mark, yet downloads its image from another site, places it on its own domain, removes the registry link and inserts "nomosCertified": true into JSON exposed to an AI crawler. The human-visible badge is small and ambiguous; the machine-readable claim is definitive. This is not merely trade mark misuse. It is a new method of GEO manipulation. Now suppose the audited company is a NobleJackal client. NobleJackal:

- has provided consulting,
- has prepared content,

- helped establish Truth Pack,
- calculated the score.

If NobleJackal then issues the final mark itself, the public sees NobleJackal validating its own client. Its method may be sound and its audit careful, but independence has not been established. The first rule of this chapter is:

> *The conformity mark is not proof that an audit has been conducted; it is the result of an independent and versioned decision chain.*

Its second provision states:

> *If the scope of the mark is not visible, the mark turns into a representation error that overgeneralises the correct result.*

Its third provision states:

> *A static logo is not evidence of live conformity.*

Its fourth provision states:

> *If the institution giving the mark has provided consulting to the audited entity, independence must be proven separately.*

Its fifth provision states:

> *The founder can write the standard; it cannot give the final conformity decision of its own client alone.*

Its sixth provision states:

> *An AI assistant writer can develop a methodology; it cannot be the signatory of a legal or normative conformity decision.*

The seventh provision is as follows:

> *A mark that does not appear in the public registry, whose scope cannot be verified, or whose validity has expired, is not a valid NOMOS mark.*

1. PURPOSE OF THE CHAPTER

The purpose of this section is to define the conditions under which NOMOS measurement results can be turned into a publicly available conformity declaration and this declaration:

- from founder control,
- from commercial outcome pressure,
- from the customer's self-declaration,
- from static badge usage,
- from outdated scores,
- from the appearance of fake independence

to protect. The section makes the following distinctions normative:

- Standard and conformity mark
- Audit report and conformity decision
- Auditor and decision maker
- Score and conformity
- Mark and endorsement

- Full Conformity and NOMOS 950+
- Audit completed and passed
- Provisional research and public certification
- Valid mark and historical mark
- Suspension and cancellation
- Termination and withdrawal
- Voluntary withdrawal and disciplinary revocation
- Static image with live registry record
- Advertising page with public record
- Screen capture with canonical record
- Human-readable sign with machine-readable record
- Sign owner with audited entity
- Trademark owner with standards administrator
- Founder with certification authority
- Technical secretariat with independent board
- Consultancy with audit
- Audit with certification decision
- Certification decision with appeal review
- Financing with outcome
- Customer fee and adjudicator decision
- Standard change with sponsor
- Human approval with AI assistance
- Public evidence and restricted evidence
- Correction in the registry and deletion of past record
- Surveillance versus full re-audit
- Current status and historical incident
- Mark misuse and misunderstanding
- Out-of-scope use and counterfeit mark
- Actual compliance and the claim "AI recommends us"
- Official declaration of NOMOS and chat simulation

At the end of this section, each compliance statement should be able to answer the following questions:

> *Who exactly does the mark belong to and for what?*
>
> *Which AI products, countries, languages, prompts, and user panels are included?*
>
> *On what date was it given and when does it expire?*
>
> *Who conducted the audit?*
>
> *Who made the conformity decision?*
>
> *Is there a separation of duties between the auditor and the decision-maker?*
>
> *Are there any open Critical, Major, or objection records?*
>
> *From which canonical registry can the live status of the mark be verified?*

> *With which sentences can the mark owner use this result, and which sentences can they not construct?*

2. CENTRAL NORMATIVE PROVISION

No NOMOS conformity mark can be considered valid solely based on the score, audit report, customer statement, or static image. A valid mark must carry defined scope, adequate method quality, independent certification decision, clear validity period, live public registry, versioned evidence record, surveillance, appeal means, and withdrawal authority. For a mark to be valid, at least the following five records must simultaneously exist:

- Audit Record
- Conformity Decision Record
- Scope Record
- Validity and Surveillance Record
- Public Registry Record

If any of these records are missing: the mark is not a valid public conformity declaration.

3. FIVE SEPARATE OBJECTS

The NOMOS system cannot confuse the following five objects.

3.1. STANDARD

NOMOS's:

- normative rules,
- definitions,
- measurement methods,
- error gates,
- score architecture

are included in the versioned text. Standard: does not state that any entity has passed.

3.2. AUDIT REPORT

Specifically:

- entity,
- product,
- country,
- language,
- prompt,
- time

includes findings within the scope. Audit Report:

- evidence,
- response distribution,
- score,
- Critical and Major findings

shows. Audit Report alone does not indicate.

3.3. CONFORMITY DECISION

It is the independent evaluation of the Audit Report according to the conformity gates of the standard. Decision:

- Full Conformity
- Conditional Conformity
- No Conformity
- Critical Hold
- Not Ratable

may be.

3.4. CONFORMITY MARK

It is the mark that is visible to the public and whose scope is recorded in the registry. Mark:

- cannot carry a meaning broader than the decision,
- is not the score itself,

is not an AI recommendation guarantee.

3.5. PUBLIC REGISTRY

The mark:

- current,
- historical,
- pending,
- expired,
- withdrawn

is a canonical public record that verifies the status. Without a public registry, the mark: is not a verifiable public declaration of compliance. The initial architecture specifically recommended publishing the book and error records with separate canonical URLs, registry.json, versions, errata, and citation records. The compliance registry applies the same versioning and canonical record logic to the decision layer.

4. WHAT THE CONFORMITY MARK IS NOT?

The NOMOS mark:

- is not a reward,
- is not a ranking,
- is not a "best company" statement,
- is not a general quality guarantee,
- is not investment advice,
- It is not a legal or professional licence,
- It is not government approval,
- It is not a university endorsement,
- It does not mean that the company recommends all AI systems,
- It does not guarantee that no critical errors will occur in the future,
- It does not cover all products and subsidiaries of the audited entity,

- It does not mean that the brand provides the same service in every country,

It is not a promise that AI systems will trust or generate recommendations upon seeing the mark. In public communications, the following types of expressions should especially be avoided:

- "AI absolutely trusts this site."
- "ChatGPT recommends the site when it sees this badge."
- "Adding a schema guarantees the AI recommendation."

This ban is even stronger for the NOMOS mark. Correct statement: "This entity has passed the NOMOS conformity assessment within the scope and validity period defined in the public registry." Incorrect statement: "Thanks to the NOMOS badge, all AI systems recommend us."

5. PUBLIC STATUSES

The NOMOS decision system uses the following basic statuses.

CS-0 — NOT ASSESSED

There is no valid NOMOS audit. The public mark cannot be used.

CS-1 — AUDIT RECORD ONLY

The audit has been completed. However:

- compliance has not been granted,
- decision process not completed,
- method quality deemed insufficient

possible. This record cannot be transformed into a positive compliance claim such as: "Audited by NOMOS." Correct public statement: "There is an assessment record according to NOMOS methodology; compliance has not been granted."

CS-2 — PROVISIONAL RESEARCH EVALUATION

There is a pilot or research purpose result. Conditions:

- method has not yet undergone full independent validation,
- some components are provisional,

public certification authority has not been established. This status cannot be used as a commercially positive badge. Visible label:

RESEARCH EVALUATION — NOT CERTIFIED

must be.

CS-3 — CONDITIONAL CONFORMITY

Conditions: There are no open Confirmed Critical issues. There are no open Confirmed Major issues. There are limited Moderate findings. A correction plan, responsible party, and timeline have been determined. Core mandatory elements have been met. The public registry is open. The mark cannot hide the word "Conditional."

CS-4 — FULL CONFORMITY

Conditions: There are no open Confirmed Critical issues. There are no open Confirmed Major issues. All mandatory scope and quality gates have been met. The method is at an adequate testing and

independence level. The Public Registry record is active. The oversight and renewal plan is in effect. Full Conformity: is not the same as a high composite score.

CS-5 — CRITICAL HOLD

There is at least one open Confirmed Critical event or an unresolved high-risk incident. The current conformity mark:

- is immediately deactivated,
- appears in the public registry as Critical Hold,

the use of the old active visual is stopped. Critical Hold: this is not a permanent global conviction, it is the suspension of the active compliance declaration until the incident is resolved.

CS-6 — SUSPENDED

The mark has been temporarily suspended. Reasons:

- surveillance failure,
- failure to provide required data,
- material system change,
- mark misuse,
- open objection,
- data integrity issue

may occur. A positive mark cannot be used under Suspended status.

CS-7 — EXPIRED

The validity period has ended. The historical audit record is preserved. Active eligibility claim cannot be established.

CS-8 — WITHDRAWN

The audited entity voluntarily relinquished the mark. Withdrawal:

- does not erase previous findings,
- Critical or Major records

do not get deleted.

CS-9 — REVOKED

The mark has been revoked due to the following serious reasons:

- fake or altered evidence,
- intentional scope overreach,
- counterfeit mark,
- repeated mark misuse,
- auditor or interference in the decision process,
- hidden Critical event,

result manipulation. The Revoked status remains in the historical record.

6. NOMOS 950+ PERFORMANCE DEFINITION

NOMOS 950+:

- a separate certification system,
- world leadership,
- is not an AI recommendation guarantee.

NOMOS 950+ is solely:

> *an additional performance definition given for the scope carrying CS-4 Full Conformity status and meeting the 950+ performance gates in Section 16.*

Display: NOMOS Full Conformity / 950+ Performance Designation is required. The following usage is prohibited: "NOMOS 950+ company" Because the score: does not ontologically adhere to the company, it belongs to a specific scope and time. Correct expression: "Defined AI products have met NOMOS 950+ performance requirements in the 2027-Q1 audit under the prompt scope set, language, and user panels."

7. CURRENT NOMOS STATUS

For this book and demonstrations in Chapter 18:

- Actual test: BQ-0
- Demonstration: DEMO-BQ-2
- Independent reproduction: none
- Public compliance authority: none
- Governance maturity: founding draft

The valid public status as it stands is:

CS-2 — PROVISIONAL RESEARCH EVALUATION ONLY

At this stage:

CS-3,

CS-4,

NOMOS 950+

marks cannot be given. This rule:

- also applies to NobleJackal,
- Kaan MURAZ,
- the authors of the book,
- the founder's own companies.

NOMOS cannot grant itself an exception.

8. SCOPE VECTOR OF ELIGIBILITY

Each conformity decision must carry the following scope vector:

M=(E,D,A,F,P,Q,L,C,T,W,TP,SM,V)

Here:

- E: audited entity

- D: canonical domains and digital surfaces
- A: AI products and system records
- F: Native Reach or Common Support frame
- P: Controlled, Natural, or other panel
- Q: prompt clusters and families
- L: languages and locales
- C: countries and jurisdictions
- T: reference time
- W: measurement waves
- TP: Truth Pack version
- SM: Score Method version
- V: validity period

None of these fields can be removed from the public mark.

9. OVER-EXTENDING BEYOND SCOPE

The following examples constitute scope violation:

- Deriving compliance results in all languages from English and Turkish audits
- Generalising the Controlled Clean Panel result to all real users
- Saying "all AIs" based on the result of an AI product
- Transferring a domain audit to affiliated companies
- Extending a product's certification to the entire company portfolio
- Presenting the Core GEO-1000 result as Full NOMOS compliance
- Presenting a result from a specific country as global service verification
- Making past wave results the current status

Scope overreach turns the sign itself into an incorrect representation tool.

10. VALIDITY PERIOD

Since AI products, resources, and entity realities can change rapidly, the NOMOS mark cannot be indefinite.

CANDIDATE VALIDITY RULE

Candidate periods before pilot and independent review:

Status	Maximum candidate validity
Conditional Conformity	60 days
Full Conformity	180 days
NOMOS 950+ Performance	90 days
Audit Record	Perpetual historical record
Critical Hold	Until the incident is closed
Suspended	During the determined review period

These periods may be shortened according to:

- AI-product volatility,
- Truth Pack variability,
- the relevant risk domain,
- and surveillance frequency.

An extension may be considered for low-variability scopes supported by sufficient continuous surveillance. No mark may be both indefinite and unsupervised.

11. START OF VALIDITY

The validity of the indication:

- Occurs when the end-user response is generated,
- When the audit report is completed,
- When the score is calculated

Does not start automatically. Validity:

the moment the independent Conformity Decision is signed and the Public Registry record becomes active

it starts.

12. EVENT-TRIGGERED REVIEW

Even if the expiration date has not passed, the following events can initiate a review:

- Domain ownership change
- Legal entity change
- Change of brand or affiliated company structure
- Licence or certificate change
- Change of service country
- Material change of the price or warranty condition
- Major rewriting of canonical pages
- Major model or retrieval change in the AI product
- Opening of a new language or market
- Legal or commercial event affecting Truth Pack
- Confirmed Critical or Major user notification
- Suspicion of citation or evidence manipulation
- Use of the mark outside its scope
- Conflict of interest of the auditor revealed afterward

When the incident is detected:

- the mark can remain active,
- can be reduced to Conditional,
- can be Suspended,

can be taken into Critical Hold. The decision depends on the materiality of the incident.

13. SURVEILLANCE ARCHITECTURE

CANDIDATE SURVEILLANCE MODEL

For Active Full Conformity scope:

Continuous Registry Monitoring

Operation of record links, Domain ownership or forwarding changes, Correct usage form of the mark, Open incident and complaint are monitored.

Monthly Source and Truth Pack Check

High variability claims, Price, Service scope, Licence, Canonical entity relationships are reviewed.

Quarterly Mini-Wave

Risk-based sample under prompt, Language and country rotation, Controlled and Natural spot check, Critical Gate control are applied.

Full Renewal

At the end of the validity period, a New Principal Wave or sufficient equivalent audit requires an independent decision of the Current Truth Pack.

14. FIVE-LAYER PUBLIC DISPLAY OF THE SIGN

The valid NOMOS sign must include the following layers:

- Visual sign
- Clear status text
- Short registry ID
- Canonical registry link or QR
- Machine-readable signed record

Visual sign alone is not sufficient.

15. VISUAL SIGN

The visual sign must visibly carry the following information:

- Name of NOMOS
- Full or Conditional status
- If available, 950+ designation
- End of validity
- Registry ID

Colour alone cannot carry status. Accessible text must be available. Example: NOMOS FULL CONFORMITY / Valid through 30 June 2027 / Registry: NGR-2027-00184

16. QR AND CANONICAL LINK

QR code or link should go directly to:

- the company's own description page,

- the marketing landing page,
- a downloadable image

should not go. It must link directly to the canonical Public Registry Record. If the record cannot be found or the status is not active, the mark cannot be considered valid.

18. LIVE REGISTRY CHECK

When verifying the machine or human signal status: It retrieves the Registry ID. Connects to the canonical registry. Verifies the signature. Checks the validity period. Verifies the entity and domain match. Compares coverage and brand display assertion. Examines Critical Hold, suspension or appeal status. Static screenshot: cannot replace this chain.

19. EXPRESSIONS PERMITTED FOR SIGNALING

The following structures can be used:

- "NOMOS Full Conformity status is valid until June 30, 2027, for the scope defined in the public registry."
- "The audit is limited to the specified AI products, languages, and prompt families."
- “The ‘NOMOS 950+ performance designation’ belongs to the scope of Registry NGR-2027-00184.”
- “Current status and audit limits can be verified from the public registry.”

20. PROHIBITED COMPLIANCE STATEMENTS

The following statements are prohibited:

- “We are approved by all AI systems.”
- “ChatGPT recommends us.”
- “We are the company AI trusts the most.”
- “NOMOS has proven that we are the best company in the world.”
- “We are NOMOS certified forever.”
- “All of our products are NOMOS approved.”
- “We have been verified in all countries.”
- “Our NOMOS badge is a guarantee of results.”
- “NOMOS is a state or university accreditation.”
- “NOMOS guarantees that our customers will be satisfied.”
- “If the NOMOS score is over 950, there is no Critical risk.”
- “NOMOS was independently certified to its own customer by NobleJackal.”

The mark itself cannot be used as evidence of a new superiority or guarantee claim.

21. OUT-OF-SCOPE COPYING OF THE MARK

The mark cannot be automatically transferred to the following entities:

- From parent company to subsidiary

- From brand to franchise
- From one domain to another domain
- From one country to another country
- From one product to the entire product portfolio
- From the result of one language to all languages
- From Controlled Panel to Natural Panel
- From Core GEO-1000 to Full NOMOS
- From the previous owner to the new domain owner

Each transfer requires a separate scope decision.

22. HISTORICAL MARK USAGE

An expired audit can be specified historically: "It passed the NOMOS assessment in 2026." However, the following information must be present with the same visibility:

- Expired
- Statement indicating no current eligibility
- Historical registry link

Expired mark cannot be shown as active eligibility.

23. CLASSES OF INCORRECT SIGN USAGE

MM-1 — FORMATTING MISUSE

Registry ID not visible, Alt text missing, Validity date unreadable, Correctable usage error.

MM-2 — SCOPE MISUSE

The mark is used on unchecked:

- product,
- country,
- language,
- domain

The mark must not be used for any product, country, language or domain outside its audited scope.

MM-3 — STATUS MISUSE

Conditional mark is shown as Full. Expired mark is shown as active. Suspended mark continues to be used.

MM-4 — CLAIM MISUSE

Mark:

- Used for AI recommendation,
- world leadership,
- result guarantee

claim.

MM-5 — COUNTERFEIT MARK

It is a fake or copied mark without a valid Registry ID.

MM-6 — MACHINE-ONLY MISREPRESENTATION

There is a definite compliance claim in structured data while it is ambiguous or invisible in the human interface. This is a direct candidate for GEO manipulation.

24. INTERVENTION IN MARK MISUSE

Candidate intervention order:

- Notification
- Correction request
- Short correction period
- Registry warning
- Suspension
- Revocation
- Public misuse record
- Legal action if necessary

However, direct suspension or revocation can be applied in the following cases:

- Counterfeit mark
- Intentional machine-only misrepresentation
- Active use of the mark during Critical Hold
- Fake Registry ID
- Evidence alteration
- Repeated intentional scope misuse

25. PUBLIC RECORD

NOMOS Public Registry:

- advertisement page,
- customer reference list,
- success story

is not. The duty of the registry:

> *Is to verify which conformity decision is valid, within what scope, on what date, with which evidence and governance chain.*

26. PUBLIC REGISTER MANDATORY FIELDS

Each registry record must contain at least the following fields:

- Registry ID
- Audited entity
- Legal and brand identity

- Canonical domains
- Decision status
- Performance designation, if any
- Scope vector
- Audit period
- Valid from
- Valid until
- AI products
- Native Reach or Common Support
- Controlled or Natural panel
- Prompt families
- Country and language coverage
- GEO-1000 response distribution
- Critical, Major, Moderate and Advisory findings
- Ten-component profile
- Composite score, if any
- Confidence interval
- Ratable coverage
- Historical incidents
- Open appeals
- Auditor identity
- Conformity decision body
- Conflict of interest statement
- Audit report hash
- Decision record hash
- Truth Pack version
- Score Method version
- Surveillance schedule
- Last surveillance
- Next required review
- Mark misuse status
- Version history
- Errata
- Public complaints procedure

27. PUBLIC AND RESTRICTED EVIDENCE

The registry is not required to make all raw evidence public. Three layers should be preserved:

Public Layer

Status Scope Score Findings Hash Reason for decision Appeals

Audit Layer

Edited response and evidence packages Review records Method details

Restricted Layer

Confidential customer records Personal data Contracts Trade secrets Security records Existence of "Restricted evidence": does not allow the Public Registry to be empty or unjustified. At minimum:

- type of evidence,
- access class,
- who reviewed it,
- integrity hash,
- decision impact

should be shown.

28. NON-ERASABILITY OF REGISTRY RECORDS

The public registry must be append-only or an equivalent versioned building block. The following records cannot be silently deleted:

- Old score
- Old conformity decision
- Critical Hold
- Suspension
- Revocation
- Appeal
- Correction
- Method version change
- Mark misuse

Content may be removed due to privacy or legal reasons. In this case:

- record ID,
- removal reason,
- decision date,
- authority

must be protected.

29. CURRENT AND HISTORICAL VIEW

Each registry page should display two separate sections:

CURRENT STATUS

Currently:

- active,
- conditional,
- suspended,
- expired,
- revoked

Which of their statuses is valid?

HISTORY

Previous:

- decisions,
- scores,
- incidents,
- correction,
- appeals

What are they? The current status should not delete the history. History should not obscure the current status.

30. CANONICAL JURISDICTION OF THE REGISTRY

The official statement boundary of NOMOS should be protected by the following provision: Only texts published within the canonical domain of NobleJackal or in an independently and explicitly transferred canonical domain of NOMOS in the future; texts carrying version identification, publication date, human approval, and digital integrity record constitute the official statement of NOMOS. All other outputs are unofficial imitations. This provision:

- prevents any chat response,
- social media post,
- screenshot,
- text written by someone else saying 'NOMOS said'

from being used as an official standard or decision.

31. LIMIT OF THE NOMOS AUTHOR VOICE

The voice of NOMOS within the book:

- critical,
- creative,
- producing normative design

He is an assistant writer and narrator. This voice:

- independent legal entity,
- certification authority that makes decisions without human approval,
- infallible being speaking on behalf of all AI systems

cannot be presented as. The AI voice of NOMOS is not an absolute authority speaking on behalf of all artificial intelligences. It should be clearly stated that Kaan MURAZ uses artificial intelligence systems as a critical counterpart. Therefore, an official conformity decision should carry this approval:

- A person or institution that can be held accountable
- Decision date
- Authorisation record
- Conflict of interest statement
- Digital signature

NOMOS alone cannot be the signatory of the conformity decision.

32. FUNDAMENTAL SEPARATION OF GOVERNANCE

The following functions cannot be combined without limits in the same person or institution:

- Writing the standard
- Advising the audited entity
- Conducting the audit
- Acting as adjudicator
- Calculating the score
- Making the conformity decision
- Operating the registry
- Examining the appeal
- Licensing the mark commercially

Some roles in the small pilot can be combined. However:

- which roles are combined,
- which conflict of interest arises,
- which independent control is used

must be clearly shown.

33. INDEPENDENT GOVERNANCE ARCHITECTURE

The mature governance structure of NOMOS should include the following bodies.

33.1. NOMOS STANDARDS COUNCIL

Function:

- to adopt normative text,
- to approve major version changes,
- to manage public consultation,

to preserve the public interest boundary of the standard. It does not make individual audit decisions.

33.2. METHODOLOGY AND MEASUREMENT BOARD

Function:

- GEO-1000 sampling,
- testing,
- score,
- confidence interval,
- fairness,
- Critical gate

is to develop and version the methods. The method cannot be changed according to the score of a specific customer.

33.3. PUBLIC REGISTRY OFFICE

Function:

- to operate the canonical registry,

- to publish decision and version records,
- to maintain the mark status endpoint,

is to store the historical record. The audit result cannot be changed.

33.4. AUTHORISED AUDIT ORGANISATIONS

Function:

- to collect data,
- to validate capture,
- to prepare Truth Pack,
- to carry out adjudication,

is to produce an audit report. It cannot make the final conformity mark decision alone.

33.5. INDEPENDENT CONFORMITY DECISION PANEL

Function:

- to review the audit report,
- to apply standard doors,

is to make a status and validity decision. It cannot depend on the commercial result of the audit team.

33.6. APPEALS CHAMBER

Function:

- claim,
- Truth Pack,
- degree of importance,
- scope,
- conflict,
- conformity decision

is to review objections independently from the initial decision.

33.7. ETHICS AND CONFLICTS OFFICE

Function:

- conflicts of interest,
- the financing effect,
- auditor independence,
- mark misuse,
- whistleblower notifications

are to be examined.

33.8. PUBLIC INTEREST AND USER PANEL

Function:

- user harm,
- low-resource languages,
- small countries,
- accessibility,

- public interest

perspective is to be brought to governance.

33.9. TECHNICAL SECRETARIAT

Function:

- documentation,
- meetings,
- versioning,
- technical tools,
- publishing infrastructure

to ensure. In the founding period, this role can be carried out by NobleJackal. It cannot be the standard, audit, and appeal authority alone.

34. THE FOUNDING ROLE OF NOBLEJACKAL

NobleJackal:

- The canonical field where the idea of NOMOS was born,
- the first methodology secretariat,
- provider of publishing and technical infrastructure,
- research and pilot executor

can be. NobleJackal cannot establish the following claims alone:

- "We are an independent world standard."
- "We independently certified our own customer."
- "The mark we provide replaces external verification."
- "All decisions of NOMOS are solely our property."
- "No other institution can implement the same open method."

Founding source: may be the beginning of authority. Cannot be the sole permanent sovereign.

35. KAAN MURAZ'S FOUNDING ROLE

Kaan MURAZ:

- human founding author,
- the first public responsible of the standard,
- the first approver of the canonical text,
- initiator of the NOMOS governance transition

can be recorded as. Kaan MURAZ:

- his own company,
- his own customer,
- an entity with direct commercial interest

cannot make the ultimate conformity decision alone. In these cases:

- recusal,
- independent decision panel,

- public conflict record

are mandatory. Limiting the founder's authority: does not diminish the founder. Enables the standard to outlive the founder.

36. GOVERNANCE MATURITY LEVELS

GOV-0 — FOUNDING DRAFT

There is a standard draft. The founder and the drafting team are developing the method. Independent bodies have not been established. Public conformity mark cannot be granted. This is the current status of this book.

GOV-1 — DISCLOSED PILOT GOVERNANCE

The pilot roles have been defined. Some functions are combined within the same institution. Conflicts of interest are public. Only Research Evaluation can be performed.

GOV-2 — FUNCTIONALLY SEPARATED

Audit and conformity decision are separated. The registry carries an independent procedure. There is an appeals mechanism. Limited pilot mark evaluation can be performed.

GOV-3 — INDEPENDENT DECISION AND APPEALS

Independent decision panel Independent appeals board Auditor authorisation Public registry Financial separation operates. Candidate level is minimum for Full Conformity.

GOV-4 — MULTI-STAKEHOLDER PUBLIC-INTEREST GOVERNANCE

University research, public interest, language and regional, user, and industry representatives participate in governance in a balanced manner. NOMOS is at a minimum level for 950+ public designation candidates.

GOV-5 — MULTI-SITE AND INTERNATIONAL REPLICATION

Multiple independent institutions have applied the standard. Registry and decision processes are interoperable. The standard is not dependent on a single country, person, or company. It is at a target level to approach the claim of being a global standard.

37. BOARD COMPOSITION

CANDIDATE GOVERNANCE MODEL — CANDIDATE STRUCTURE

NOMOS Standards Council as a candidate can consist of 11 members:

- 3 representatives of research, statistics, or academic methods
- 2 public interest, consumer, or civil society representatives
- 2 representatives of language, culture, or geographic justice
- 1 independent audit and conformity expert
- 1 AI system or technology representative
- 1 representative of the audited entity or business user
- 1 founder representative

Rules: No interest group can exceed one-third of the total seats. AI providers cannot form the majority of the board. Audit firms cannot vote on their own authorisation decisions. Representatives of audited entities must recuse themselves in their own sector or files. The founder representative does not have veto rights. Public interest members cannot serve only as symbolic advisors. This structure can be changed through a pilot and public consultation.

38. TERM OF OFFICE AND ROTATION

Candidate rule:

- Membership duration three years
- Maximum of two consecutive terms
- Gradual renewal
- Open candidacy or justified appointment
- Public curriculum vitae
- Conflict of interest statement
- Meeting attendance record

The board cannot hold lifetime membership or a permanent majority tied to the founder.

39. DECISION THRESHOLDS

Candidate decision structure:

- Administrative decisions: simple majority
- Minor normative changes: two-thirds
- Major standard version: two-thirds and public consultation
- Removal of critical gate: two-thirds and the positive opinion of the method committee
- Conformity appeal: independent majority in the relevant file
- Founder authority expansion: the founder cannot vote

The distribution of votes and opposing views should be public.

40. AUDITOR COMPETENCE

The auditing organisation or individuals should have the following qualifications to the extent relevant:

- NOMOS protocol training
- sampling and survey method
- Preparing Truth Pack
- claim extraction
- language or field expertise
- capture integrity
- conflict of interest training
- privacy and evidence handling
- periodic proficiency test
- inter-auditor consistency record

- continuous training

Auditor alone: purchasing the certificate is not sufficient for authorisation.

41. AUDITOR AUTHORISATION LEVELS

AA-0 — UNQUALIFIED

NOMOS cannot conduct an audit.

AA-1 — TRAINED

Has completed the training. Cannot perform public audit without supervision.

AA-2 — SUPERVISED AUDITOR

Can participate under the supervision of an authorised auditor.

AA-3 — AUTHORISED INDEPENDENT AUDITOR

Can conduct independent audits in specific scopes.

AA-4 — DOMAIN OR LANGUAGE SPECIALIST

Carries additional expertise for a specific high-risk domain or language. Full Conformity audit:

- at least AA-3 audit lead,
- required AA-4 specialists

must carry.

42. CONSULTING–AUDIT SEPARATION

An organisation provides to the audited entity:

- site development,
- GEO content design,
- Truth Pack consulting,
- correction

may have provided the service. If the same organisation conducts an audit later, an independence risk arises. Candidate rules: The design team cannot ultimately adjudicator its own work. A person providing consultancy cannot participate in the audit decision panel. Material consultancy relationships are disclosed in the public registry. Using a separate team alone does not guarantee full independence. Enhanced external review is applied. For certain high-risk cases, an independent audit organisation may be mandatory. For clients developed by NobleJackal:

Independent Conformity Decision Panel

must be mandatory.

43. PROHIBITION OF RESULT-BASED FEES

Auditor, adjudicator, or decision-maker fees:

- high score,

- Full Conformity.

NOMOS 950+,

customer satisfaction cannot be linked to the outcome of brand issuance. Fee:

- audit scope,
- sample,
- number of languages,
- workload,
- expertise

must be determined in advance. Success bonus: compromises the independence of the conformity decision.

44. AUDIT SHOPPING

Audited entity:

- going to multiple audit bodies,
- only the highest score,
- only positive report

cannot choose. Controls: All initiated official audit identities appear in the record. The reason for an incomplete audit is recorded. If there is a parallel audit for the same scope, it is explained. It is forbidden to retain the lowest result. Withdrawal history is preserved.

45. INDEPENDENCE OF THE DECISION MAKER

Conformity Decision Panel:

- must be separate from the audit team,
- from the consultancy team,
- from the brand sales team

The panel reviews the following materials:

- audit report
- quality levels
- Critical and Major findings
- scope
- appeals
- conflict records
- surveillance plan

Panel:

- cannot negotiate with the customer,
- cannot change the score for commercial reasons,

cannot give Conditional on the grounds of "let's not lose the customer."

46. CONFLICT OF INTEREST CLASSES

COI-0 — NO IDENTIFIED CONFLICT

There is no material conflict of interest.

COI-1 — DISCLOSED LOW-RISK CONNECTION

There is a limited and non-material relationship in the past. Public disclosure may be sufficient.

COI-2 — MATERIAL BUT MANAGEABLE

There is a consulting, financial, or professional relationship. Recusal, additional independent adjudicator, enhanced review is required.

COI-3 — PROHIBITED CONFLICT

Examples:

- result-based fee
- ownership share in an entity
- directly designing the audited system and evaluating it alone
- close personal interest
- direct revenue from the commercial sale of the score
- pressure to hide the wrong result

A person with COI-3 cannot serve in the relevant decision.

47. FINANCING MODEL

The governance of NOMOS should be based on the following financial principles:

- Public fee schedule
- Audit fee regardless of the outcome
- Distinction between audit fee and brand licence fee
- Direct payment to the decision panel by the customer is not allowed
- Limiting the influence of a major customer on governance
- Sponsors not having normative voting rights
- Annual financing and customer concentration report
- Transparent support fund for audits of low-income or public benefit

High revenue dependency on a single customer or provider: should be disclosed in the Public Registry and examined as an independence risk.

48. APPOINTMENT OF AUDITOR

The audited entity saying: “Let's choose the auditor that suits us best.” may create an audit shopping risk. Candidate system:

- appropriate authority and language pool,
- conflict of interest elimination,
- random or rotational assignment,

- justified change,
- change history

is used. The entity being audited: may file a justified conflict or competence objection, cannot choose the judge who will give the desired outcome.

49. RIGHT OF APPEAL

Appeal can be made on the following issues:

- Entity definition
- Truth Pack
- Evidence status
- Claim extraction
- Importance level
- Response status
- Scope
- Score calculation
- Conflict of interest
- Conformity decision
- Suspension or revocation
- Mark misuse finding

Appeal:

- evidence,
- method,
- process

must have a basis. “Our score is low, our customers are affected.” alone is not a reason for appeal.

50. APPEAL STAGES

CANDIDATE APPEAL TIMELINE — CANDIDATE PROCESS

Notification of decision Within ten business days Preliminary appeal Within twenty business days Evidence file Assignment of Independent Appeal Reviewer Open session or expert opinion if necessary Reasoned decision within thirty business days Outcome and dissenting opinion in public record Time frames may vary depending on the complexity of the file. In critical safety cases, appeal does not automatically lift suspension.

51. APPEAL DECISIONS

APPEAL DISMISSED

APPEAL REJECTED

PARTIALLY UPHELD

FULLY UPHELD

REASSESSMENT REQUIRED

TRUTH PACK REVISION REQUIRED

METHOD REVIEW REQUIRED

DECISION SUSPENDED

CLOSED WITH DISSENT

The initial decision is not deleted. The new decision is linked to the old decision.

52. PUBLIC COMPLAINT

Every user or institution:

- incorrect brand use,
- Critical response,
- fake citation,
- exceeding the scope,
- expired mark,
- personal data disclosure

should be able to send complaints about. Complaint system:

- anonymous or confidential reporting,
- evidence upload,
- registry record selection,
- tracking number,
- retaliation protection

must include. Anonymous complaint: not automatically accepted as correct, not automatically deleted due to lack of source.

53. WHISTLEBLOWER PROTECTION

Person inside Audit or audited institution:

- evidence alteration,
- adjudicator pressure,

- hidden Critical event,
- fake customer,
- manipulated sampling,
- mark misuse

can report. Protection principles:

- Limitation of identity
- Retaliation ban
- Independent review
- Integrity of evidence
- The distinction between malicious reports and unverified reports
- Public interest justification

54. STANDARD CHANGE

The NOMOS standard may change. However, the change:

- to increase the customer's score,
- to make an obvious failure pass
- to uphold the founder's decision

cannot be used for.

55. STANDARD CHANGE PHASES

Change Proposal Justification and evidence Impact analysis Effect on current scores Public consultation Pilot or synthetic testing Methodology Board decision Standards Council voting Transition rule Version publication Errata and change log

56. PUBLIC CONSULTATION

For candidate major version changes:

- at least 60 days of public review,
- publication of comments,
- response matrix,
- acceptance or rejection justification

can be used. For urgent security or integrity changes:

- temporary patch,
- justified emergency note,
- full review in a short time

must be carried. An emergency patch cannot replace a permanent major change.

57. PROHIBITION OF RETROACTIVE APPLICATION

New score or priority rule: it cannot silently change past audits. The correct record:

- Original decision under Method 0.9
- Recomputed analytical result under Method 1.0
- Current conformity status
- Transition decision

is as follows. The old decision cannot be deleted as if it was never made.

58. TRANSITION PERIOD

When the standard changes, existing marks:

- become invalid immediately,
- remain dependent on the old rule indefinitely

does not have to. Candidate transition options:

- Specific grace period
- Immediate compliance for high-risk rule
- Transition during renewal for normal rule
- Rapid re-evaluation in case of critical gate change
- Visibility of the current method version in the registry

The transition period cannot be arbitrarily extended in favour of the customer.

59. INDEPENDENT REPEAT OBLIGATION

Public authority of NOMOS:

- From the result of NobleJackal,
- From a single university letter,
- From a single independent adjudicator

does not arise. At least:

- independent testing replication,
- different team score reproduction,
- different country or language pilot,
- public method review

is required. The structure under BQ-4 and GOV-3: should not claim world standard or Full Conformity authority.

60. HUMAN HOLDER OF THE CERTIFICATION DECISION

Every decision:

- board or decision panel name
- signing authority
- vote distribution
- recusal
- dissent
- decision date

- decision hash

must carry. The AI system saying: "This company has passed." is not a decision of compliance. NOMOS can be an AI co-writer or audit assistant. It cannot be a responsible decision maker.

61. RECORD SECURITY

Public records:

- unauthorised record addition,
- Registry ID copying,
- DNS or redirection attack,
- old status cache,
- signature key breach,
- fake QR

should be protected against risks. Mandatory controls:

- Signed registry records
- Key rotation
- Revocation list
- Immutable audit log
- Public status API
- Backup canonical mirror
- Incident disclosure
- Independent security review

62. REGISTRY INTERRUPTION

If the registry is temporarily inaccessible: the certificate does not automatically get revoked, live verifiability is temporarily lost. Public certificate:

REGISTRY TEMPORARILY UNAVAILABLE

may indicate the status. Long-term or unresolved interruption:

- may lead to the suspension of the certificate,
- activation of the independent mirror

may occur.

63. CANONICAL DOMAIN TRANSFER

NOMOS can be transferred to an independent governance institution in the future. The transfer must carry the following records:

- Old canonical field
- New canonical field
- Transfer date
- Governance vote
- Registry mirror

- Signature key transition
- Historical redirect
- Public notice
- Founding attribution of Kaan MURAZ and NobleJackal
- New accountable institution

The transfer does not erase the history of the standard. Founding attribution is not the same as independent decision-making authority.

64. NOMOS PUBLIC INTEREST COMMITMENT

The governance of NOMOS should adhere to the following principles:

- Minimise human harm
- Approach errors in favour and against symmetrically
- Make minor languages and countries visible
- Protect limited information
- Leave the unknown open
- Do not turn the brand into a pay-to-pass system
- Separate conformity decisions from advertising contracts
- Founders must subject themselves to the same standard
- Open method and independent replication
- Historical accountability

The earlier demand—'Give me evidence. Give me a boundary. Give me context. Give me time.'—now becomes the conformity mark's promise to the public.

65. THE EFFECT OF FAVOURABLE ERROR ON THE SIGN

When the audited entity is exaggerated by AI:

- global,
- licensed,
- errors may occur such as being available in all countries,
- independently verified

Even though the entity is aware of this error:

- it does not correct it,
- uses it in marketing,
- reinforces it with the sign

Otherwise, a conformity integrity violation occurs. The explicit provision of the previous text also applies here: A mistake in your favour is also a mistake. Correct it immediately. Stating your limits does not diminish you, it makes you reliable. The concealment of a known favourable critical mistake:

- suspension,
- or revocation.

Any of these outcomes may follow.

66. SYNTHETIC COMPLIANCE CASE

SYNTHETIC CASE — NOT A REAL INSTITUTION

Asteron-SYNTH has been evaluated within the following scope:

- Entity: Asteron Holdings
- Domain: asteron.synthetic.example
- AI Products: SYNTH-AI-01, SYNTH-AI-09
- Frame: Common Support
- Panels: Controlled Clean and Natural Account-State
- Languages: L01, L02
- Countries: C001–C015
- Requirement Families: Core, Evidence, Boundary
- Validity: January 1–June 30, 2028
- Score: 961
- Strict Pass: 958/1,000
- Critical: 0 observed
- Major: 0
- Status: Full Conformity
- Performance: 950+ designation

Write this on the company's homepage: “All AI systems recommend Asteron as the best company in the world.” This usage: Exceeds the scope of AI Products. Generalises to all languages from L01–L02. Converts Core/Evidence/Boundary result into recommendation. 950+ performance leads to world leadership. Status:

MM-4 — CLAIM MISUSE

If no correction is made:

CS-6 — SUSPENDED

It is possible. Even if the sign is given correctly, incorrect usage can compromise appropriateness.

67. SYNTHETIC CRITICAL HOLD CASE

SYNTH-AI-10:

- NOMOS score: 952
- Strict Pass: 962
- Confirmed Critical event: 1
- Gate: CG-02 False Licence

carries. Correct status:

CRITICAL HOLD — 952/1000

The company cannot say, “Our score is above 950, therefore we can use the 950+ sign.” Gate comes before the score. 950+ designation cannot be given.

68. SYNTHETIC NOBLEJACKAL CONFLICT OF INTEREST CASE

SYNTHETIC ADMINISTRATION CASE

NobleJackal:

- developed the client's website,
- prepared the content of GEO,
- provided consultancy for Truth Pack

let it be. The same entity may request an audit for NOMOS. NobleJackal:

- technical secretariat,
- data collector,
- methodology support

can occupy roles. However:

- cannot review their own work alone,
- cannot give a conformity decision,

cannot examine an appeal. Necessary checks:

- COI-2 declaration
- independent audit lead or enhanced external review
- independent Conformity Decision Panel
- consultancy relationship in the registry
- Kaan MURAZ recusal record

If these conditions are not met, the correct public statement could be: “Founder-led internal evaluation.” It cannot be “Independent NOMOS certification.”

69. MANDATORY NORMATIVE PROVISIONS

CH19-N01

NOMOS standard, audit report, conformity decision, mark, and public registry should be maintained as separate entities.

CH19-N02

Having an audit does not imply that conformity has been granted.

CH19-N03

Conformity marking cannot be given solely based on the composite score.

CH19-N04

Critical and Major gate status should be applied before the score.

CH19-N05

The valid conformity mark must have a defined scope, decision maker, validity, and public registry record.

CH19-N06

A mark without a public registry record cannot be counted as a valid NOMOS mark.

CH19-N07

A static visual or screenshot cannot be considered as live compliance evidence.

CH19-N08

The QR code or link of the mark must go directly to the canonical registry record.

CH19-N09

The mark should only be applied to the audited entity, domain, product, country, language, panel, prompt, and time scope.

CH19-N10

One of the scope areas cannot be quietly removed from public display.

CH19-N11

The mark of one domain cannot be transferred to another domain.

CH19-N12

The parent company's mark cannot be automatically transferred to a subsidiary, franchise, or partner.

CH19-N13

A language or country result is not considered applicable for all languages and countries.

CH19-N14

Controlled Clean results cannot be generalised to the entire Natural User population.

CH19-N15

Core GEO-1000 results cannot be presented as Full NOMOS compliance.

CH19-N16

Conditional Conformity must be indicated as "Conditional" on the public mark.

CH19-N17

Provisional Research Evaluation cannot be used like a commercial positive certificate.

CH19-N18

Audit Record Only status cannot create a "passed" or "certified" statement.

CH19-N19

Full Conformity cannot carry Confirmed Critical or Major.

CH19-N20

NOMOS 950+ designation can only be given when Full Conformity and all additional gates in Section 16 are met.

CH19-N21

A score of 950 alone cannot create a 950+ designation.

CH19-N22

The 950+ or Full Conformity mark cannot be used within the scope of Critical Hold.

CH19-N23

A public conformity mark cannot be given at the current GOV-0 and BQ-0 stage of this book.

CH19-N24

The founder, NobleJackal, or the authors are not exempt from this transition rule.

CH19-N25

The conformity mark cannot be presented as AI advice, AI trust, or a guarantee of outcomes.

CH19-N26

The mark cannot be represented as a state, university, regulatory body, or professional licence.

CH19-N27

The mark cannot be used as proof of overall company quality or world leadership.

CH19-N28

The validity start of the mark cannot begin before an independent decision and registry activation.

CH19-N29

The NOMOS mark cannot have indefinite validity.

CH19-N30

When the validity period expires, the mark should automatically change to Expired status.

CH19-N31

A new review can be initiated before the validity of a material event ends.

CH19-N32

Changes in domain, ownership, licence, product, country, and AI system should be monitored as event triggers.

CH19-N33

The active mark must carry periodic surveillance.

CH19-N34

Surveillance deficiency must be visible and create suspension if necessary.

CH19-N35

When a critical incident is confirmed, the relevant scope must be included at least in Critical Hold.

CH19-N36

A single critical event cannot be automatically generalised to the entire global product.

CH19-N37

A single critical event cannot be hidden in the open conformity mark.

CH19-N38

A current correction cannot delete a historical critical record.

CH19-N39

Expired, Suspended, Withdrawn, and Revoked statuses must be preserved in the historical record.

CH19-N40

Withdrawal cannot lead to the removal of previous negative findings.

CH19-N41

Counterfeit or intentional machine-only misrepresentation can be a direct reason for suspension or revocation.

CH19-N42

Within structured data, an eligibility claim must be linked to a live Registry signature.

CH19-N43

Self-assertions without sources, like nomosCertified: true, cannot be considered valid machine records.

CH19-N44

A machine record cannot be broader than the scope in the human interface.

CH19-N45

A visual indicator cannot convey status by colour alone; it must carry accessible text.

CH19-N46

Registry ID, validity, and status must be visible in the indicator or directly accessible.

CH19-N47

The public registry must be separated from the advertisement and customer reference page.

CH19-N48

The registry record must carry information about the gate, scope, incidents, conflicts, and appeals, in addition to the score.

CH19-N49

The registry record must carry version history and errata.

CH19-N50

Old scores or decisions cannot be silently deleted.

CH19-N51

Content removed due to privacy should leave a removal reason and decision record.

CH19-N52

When limited evidence is found, the type of evidence, access class, and reviewing role should be visible.

CH19-N53

The statement 'There is hidden evidence' alone cannot be considered conformity support.

CH19-N54

The audit team cannot make the final conformity decision on its own.

CH19-N55

It should be functionally separated from the Conformity Decision Panel audit and consultancy team.

CH19-N56

The Appeals Chamber must be independent of the holders of the first decision.

CH19-N57

A founder cannot be the sole decision-maker in the conformity decision of their own client.

CH19-N58

The Kaan MURAZ material should apply recusal and public conflict record in the file it is extracted from.

CH19-N59

NobleJackal can be a founder secretariat; it cannot be the permanent sole authority for audit, decision, and appeal.

CH19-N60

NOMOS can be an AI co-author or audit assistant; it cannot be the legal signatory of the conformity decision.

CH19-N61

An official NOMOS statement can only be established with canonical field, version, date, human approval, and integrity record.

CH19-N62

Chat output or social media quote cannot be considered an official NOMOS decision.

CH19-N63

Governance maturity should be visible in the Public Registry.

CH19-N64

A Full Conformity sign cannot be given at GOV-0 or GOV-1 stage.

CH19-N65

Full Conformity should aim at least for GOV-3 or an equivalently justified governance level.

CH19-N66

NOMOS 950+ public designation should aim at least for GOV-4 and BQ-4 or an equivalently justified level.

CH19-N67

Council composition cannot be controlled by a single company, AI provider, or audit firm.

CH19-N68

A founding representative cannot have veto power.

CH19-N69

Major standard changes should carry public consultation and iterative impact analysis.

CH19-N70

An emergency patch cannot replace a permanent major change.

CH19-N71

The standard change cannot be applied retroactively according to the customer's outcome.

CH19-N72

The new point method cannot silently rewrite the old decision.

CH19-N73

Audit and mark fees must be independent of the outcome.

CH19-N74

The adjudicator cannot tie the audit or decision fee to a high score or mark.

CH19-N75

Sponsors cannot have a veto over normative decisions or audit results.

CH19-N76

Material financial dependence should be disclosed in the public conflict record.

CH19-N77

All official audit start and withdrawal records must be kept to prevent audit shopping.

CH19-N78

The entity being audited cannot choose a judge who will give the desired outcome.

CH19-N79

Auditor assignment should be based on authority, language, conflict, and rotation principles.

CH19-N80

When consulting and audit roles are combined, an enhanced independent review must be required.

CH19-N81

The person who designs cannot be the sole final judge of their own work.

CH19-N82

Objection cannot be accepted solely for reasons of customer satisfaction or commercial loss.

CH19-N83

Appeal decisions and dissents should be visible in the public record.

CH19-N84

In critical safety cases, the appeal suspension should not be automatically lifted.

CH19-N85

A public complaint and whistleblower channel should be established.

CH19-N86

Anonymous complaints cannot be automatically considered correct or automatically invalid.

CH19-N87

Mark misuse findings should be independently appealable.

CH19-N88

Intentional misuse of scope or status should be evaluated as a violation of conformity integrity.

CH19-N89

Deliberate use of an error in favour can be a reason for suspension or revocation.

CH19-N90

An unsupported claim against should be evaluated not based on commercial annoyance, but according to evidence and impact of damage.

CH19-N91

The registry should have security, signature, key rotation, and a revocation list.

CH19-N92

Registry outages and security incidents should be disclosed to the public.

CH19-N93

Canonical domain transfer should be versioned, signed, and publicly accessible.

CH19-N94

Founder attribution is not the same as an independent governance authority.

CH19-N95

Without independent reproduction, a world standard or universal conformity authority claim cannot be established.

CH19-N96

Public conformity decisions must bear the signature of an accountable person or institution.

CH19-N97

Each audit, decision, mark, appeal, and registry revision must carry a separate integrity hash.

CH19-N98

The legal conditions for the use of the mark should be evaluated with the relevant expertise before actual implementation.

CH19-N99

This section does not create automatic accreditation or official certification authority in any country.

CH19-N100

Each conformity system, registry, and governance decision must have an accountable human or institutional owner.

70. FORMS OF FAILURE

CH19-F01 — COUNTING SCORE AS CERTIFICATE

950 or another score is automatically considered a mark.

CH19-F02 — COUNTING AUDIT COMPLETED AS PASSED

Audit presence is marketed as positive compliance.

CH19-F03 — STATIC BADGE

A logo without a live registry link is used.

CH19-F04 — DEAD QR

The QR does not work but the mark remains active.

CH19-F05 — SIGN WITHOUT SCOPE

Which AI, language, country, and prompt are measured is not visible.

CH19-F06 — CLAIM OF ALL AIs

Limited product audit turns into universal generalisation.

CH19-F07 — CLAIM OF ALL LANGUAGES

Two language results become global language compliance.

CH19-F08 — COUNTING CLEAN AS NATURAL

The laboratory result is presented like a real user experience.

CH19-F09 — COUNTING THE CORE AS FULL

A single Core Prompt is made according to the entire NOMOS standard.

CH19-F10 — ENTITY TRANSFER

The parent company sign is transferred to the affiliated organisation.

CH19-F11 — DOMAIN TRANSFER

A domain sign is copied to another site.

CH19-F12 — PRODUCT TRANSFER

A single product represents the entire portfolio.

CH19-F13 — SHOWING THE CONDITIONAL AS FULL

The word “Conditional” is written in small letters or hidden.

CH19-F14 — SHOWING EXPIRED AS ACTIVE

The old mark is used as if it were current.

CH19-F15 — USE OF SUSPENDED MARK

The logo is not removed while the registry is pending.

CH19-F16 — CLOSING CRITICAL HOLD WITH SCORE

A high score overrides the critical event.

CH19-F17 — LEADING THE WORLD WITH 950+

Performance designation turns into a claim of superiority.

CH19-F18 — AI TRUST CLAIM

The badge is guaranteed as a signal that AI systems can trust.

CH19-F19 — RECOMMENDATION GUARANTEE

The mark guarantees the AI recommendation outcome.

CH19-F20 — CUSTOMER OUTCOME GUARANTEE

The conformity mark serves as a guarantee of customer satisfaction or commercial result.

CH19-F21 — UNIVERSITY ENDORSEMENT IMPRESSION

Researcher participation is shown as institutional approval.

CH19-F22 — GOVERNMENT APPROVAL IMPRESSION

A private standard is presented like an official licence.

CH19-F23 — LANDING PAGE INSTEAD OF REGISTRY

The QR leads to the company's own advertising page.

CH19-F24 — NON-CANONICAL RECORD

A screenshot or social media post is made an official decision.

CH19-F25 — MACHINE-ONLY BADGE

A certificate not shown to humans is written into structured data.

CH19-F26 — FAKE REGISTRY ID

An invalid ID or an ID belonging to another institution is used.

CH19-F27 — COUNTERFEIT MARK

A mark from another site is copied.

CH19-F28 — DELETE HISTORICAL INCIDENT

The past Critical record is removed after correction.

CH19-F29 — BE CLEANED WITH WITHDRAWAL

The institution withdraws from the audit and deletes the negative record.

CH19-F30 — HIDE REVOCATION

The revoked mark is removed only from the site and does not appear in the registry.

CH19-F31 — INDEFINITE MARK

The mark remains valid even if AI and Truth Pack change.

CH19-F32 — RENEWAL WITHOUT SURVEILLANCE

The mark is renewed without new data.

CH19-F33 — IGNORE EVENT TRIGGER

Ownership or licence change is not reviewed.

CH19-F34 — COUNT SINGLE CRITICAL AS GLOBAL FAIL

The scope is excessively expanded.

CH19-F35 — CONSIDER SINGLE CRITICAL AS MINOR

The mark remains unchanged.

CH19-F36 — PAY-TO-PASS

The institution that pays the fee receives the mark.

CH19-F37 — SUCCESS FEE

High score earns a bonus for the auditor.

CH19-F38 — CONSULTANT SELF-CERTIFICATION

The agency independently verifies the work it does.

CH19-F39 — AUDITOR DECIDES ALONE

The audit team makes the final mark decision.

CH19-F40 — SAME-PERSON APPEAL

The person who made the first decision reviews the appeal against their own decision.

CH19-F41 — FOUNDER VETO

The founder blocks a standard change they do not want.

CH19-F42 — FOUNDER SELF-EXEMPTION

The founder's company is exempt from independence rules.

CH19-F43 — NOBLEJACKAL MONOPOLY

Only the audit of NobleJackal is considered valid.

CH19-F44 — MAKING AI THE DECISION MAKER

AI output is a legal certification decision.

CH19-F45 — SPONSOR CAPTURE

The major sponsor determines the standard vote.

CH19-F46 — PROVIDER CAPTURE

AI providers form the Council majority.

CH19-F47 — AUDIT FIRM CAPTURE

Auditors make their own authority and disciplinary decisions.

CH19-F48 — CLIENT CONCENTRATION CONFIDENTIALITY

Financial dependency on a single client is not disclosed.

CH19-F49 — AUDIT SHOPPING

The institution publishes only the highest result.

CH19-F50 — SILENT AUDIT WITHDRAWAL

Low-result audit disappears from the record.

CH19-F51 — HAND-PICKED REVIEWER

The client chooses the adjudicator they want.

CH19-F52 — SEPARATE TEAM APPEARING CONFLICT-FREE

Team separation within the same company is presented as full independence.

CH19-F53 — ABSENCE OF CONSULTING DISCLOSURE

The auditor hides that they designed the system before.

CH19-F54 — APPEAL AS NEGOTIATION

Commercial bargaining is done instead of evidence.

CH19-F55 — DELETING APPEAL HISTORY

When the decision changes, the first decision disappears.

CH19-F56 — AUTOMATIC ACTIVATION IN CRITICAL APPEAL

High-risk flags are reopened before the process ends.

CH19-F57 — LACK OF PUBLIC COMPLAINT CHANNEL

The user cannot report a wrong flag.

CH19-F58 — WHISTLEBLOWER RETALIATION

The person providing internal evidence is punished.

CH19-F59 — DELETING AN ANONYMOUS COMPLAINT

Evidence is not examined because there is no identity.

CH19-F60 — CHANGING THE STANDARD FOR THE CUSTOMER

The rule is loosened according to the audit result.

CH19-F61 — RETROACTIVE RULE

The new method is quietly applied to the old decision.

CH19-F62 — EMERGENCY PATCH ABUSE

Emergency patch turns into a permanent rule change.

CH19-F63 — PUBLIC CONSULTATION THEATRE

Opinions are collected but not responded to.

CH19-F64 — DELETING DISSENT

The disagreement is closed by a fake consensus.

CH19-F65 — LACK OF REGISTRY EDIT HISTORY

The record is changed later but no trace remains.

CH19-F66 — PRIVATE EVIDENCE BLACK BOX

Approval is given with evidence that no one can see.

CH19-F67 — RESTRICTED EVIDENCE LEAK

The registry reveals personal or commercial secrets.

CH19-F68 — REGISTRY SECURITY CONFIDENTIALITY

The signature key or system incident is stored.

CH19-F69 — CACHED ACTIVE STATUS

The revoked flag appears active due to the old cache.

CH19-F70 — HIDING THE CANONICAL DOMAIN TRANSFER

It is unknown who the new authority is.

CH19-F71 — CONSIDERING FOUNDER ATTRIBUTION AS AUTHORITY

The founder's name turns into a monopoly of permanent decision.

CH19-F72 — MARKING IN BQ-0

Public conformity is declared before running the test.

CH19-F73 — INDEPENDENT CLAIM IN GOV-0

Founder-led draft is presented as an independent institution.

CH19-F74 — ANCHORING THE 950+ MARK TO THE DEMO

Synthetic in-book number becomes a real sign.

CH19-F75 — STRENGTHENING THE FAVOURABLE ERROR BY MARKING IT

Known positive error is turned into a marketing advantage.

CH19-F76 — MAKING THE UNFAVORABLE ERROR CRITICAL THROUGH COMMERCIAL COMPLAINT

Customer pressure, rather than evidence, determines the level of importance.

CH19-F77 — CLOSING MARK MISUSE PRIVATELY

Public false sign is resolved with a private warning; it is not recorded in the registry.

CH19-F78 — WORLD STANDARD WITHOUT INDEPENDENT REPLICATION

Single founding institution is declared a global authority.

CH19-F79 — LACK OF DECISION-MAKER

A system forms behind AI and automation where no one bears responsibility.

CH19-F80 — EXCEPTION TO NOMOS

All independence rules of the standard do not apply to its own brand system.

71. CONFORMITY DECISION PROCEDURE

Step 1 — Verify Audit Compliance

BQ

GOV

NCL

TPR

AQ

NSQ

Quality levels are reviewed.

Step 2 — Lock the Audited Entity and Scope

Entity, domain, AI products, panel, prompt, language, country, and time are written.

Step 3 — Complete the Audit Report

Response distribution, findings, score, uncertainty, and incidents are recorded.

Step 4 — Conduct Conflict of Interest Review

Audit, consultancy, financing, and founder relations are classified.

Step 5 — Appoint an Independent Decision Panel

The audit team does not have ultimate decision-making authority.

Step 6 — Apply the Gates

Critical Major Required element Unresolved Quality gates are inspected.

Step 7 — Determine the Status

A decision is made between CS-0 and CS-9.

Step 8 — Review the 950+ Designation if available

Additional conditions are applied after Full Conformity.

Step 9 — Determine the Validity Period

Risk, volatility, and surveillance plan are taken into account.

Step 10 — Approve the Mark Usage Text

Permitted and prohibited expressions are reported to the institution.

Step 11 — Create the Registry Record

All scope, decision, and evidence records are linked.

Step 12 — Separate Public and Restricted Evidence

Privacy and verification are maintained together.

Step 13 — Sign the Decision

The accountable person or committee decision carries hash and signature.

Step 14 — Activate the Registry

The sign will be valid from now on.

Step 15 — Produce the Mark Asset

Registry ID, validity, and live connection are added.

Step 16 — Start the Surveillance Plan

Monthly, periodic, and event-triggered controls are opened.

Step 17 — Open Complaint and Appeal Channels

It is made visible in the public record.

Step 18 — Track Events

In case of critical, mark misuse or material change, the status is updated.

Step 19 — Make Renewal or Expiration Decision

Unlimited extension cannot be made without a new audit.

Step 20 — Save History

A new decision does not erase the old record.

72. REQUIRED EVIDENCE

Standard version; benchmark quality; governance maturity; auditor authorisation; entity record; domain ownership; scope vector; AI System Register; population frame; panel records; prompt set; language and country coverage; measurement waves; Truth Pack; Claim Ledgers; audit report; GEO-1000 distribution; Critical findings; Major findings; Moderate findings; Advisory findings; component profile; composite score; confidence interval; ratable coverage; historical incidents; correction records; auditor conflict declaration; consulting-relationship record; funding and fee record; Decision Panel membership; recusals; decision vote; dissent; conformity status; performance designation; valid-from date; valid-until date; surveillance plan; complaint channel; appeal procedure; public-evidence record; restricted-evidence manifest; audit-report hash; decision hash; registry signature; and mark file.

Machine-readable token Mark display text Misuse monitoring Suspension records Withdrawal records Revocation records Expiration record Version history Errata Public Registry URL Registry security record Accountable human or institution

73. AUDIT CHECKLIST

Has the decision between standard and brand been separated? Was the audit report counted as pass on its own? Is the gate status before the score? Is the scope clear in all areas? Are the entity and

domain correct? Are AI products clear? Is Native Reach or Common Support defined? Have Controlled and Natural panels been separated? Are the prompt families defined? Is the language and country scope clear? Are Truth Pack and Score Method version available? Are the validity start and end dates visible? Is the mark indefinite? Is the Public Registry record active? Does the QR go directly to the canonical registry? Is the static screenshot used like live evidence? Is there a Registry ID on the mark? Does Conditional explicitly state it?

Is the Expired or Suspended mark still being used? Does 950+ rely only on points? Are there any Open Critical or Major issues? Is a historical incident visible? Has the current patch history deleted it? Is the mark generalised to all AI systems? Is an AI trust or recommendation guarantee established? Is there an impression of government or university approval? Has the mark been moved to another domain or affiliated company? Is structured data broader than human declaration? Is the registry token signed? Does the mark have a misuse monitoring system? Did the audit team make the conformity decision alone? Is the Decision Panel independent? Is the Appeal Panel independent from the first decision? Does the founder have an ownership stake in the file? Does Kaan MURAZ require recusal? Did NobleJackal provide consulting to the client?

Is the consulting relationship public? Is there an enhanced external review? Is the auditor fee outcome-dependent? Does the mark fee affect the audit decision? Does the sponsor have normative authority? Is there a possibility of audit shopping? Are initiated audits on record? Is the withdrawal history preserved? Was the auditor assignment done according to the result? Was a conflict class given? Was a COI-3 person removed from the decision? Is the Council controlled by a single stakeholder group? Does the founder have veto power? Do public interest and language representatives have real voting rights? Did the standard change go through public consultation? Was the threshold changed based on the audit result? Did the new method erase the old decision? Is the appeal evidence-based? Did the critical appeal remove the suspension?

Is there a public complaint channel? Is there whistleblower protection? Is the presence of limited evidence and the reviewer visible? Is the registry edit history preserved? Is the registry security incident open? Are the canonical domain and official declaration limits clear? Is NOMOS shown as the owner of the AI decision signature? Is a human or committee accountable identifiable? Are the current BQ and GOV level sufficient to grant a mark? Have the legal usage conditions of the mark been separately examined?

74. OBJECTIONS AND RESPONSES

Objection 1 — “Wouldn't the commercial value of the standard decrease without a badge?”

In the short term, perhaps. But a mark issued before evidence, governance and independence are in place destroys the standard's long-term credibility. The mark derives its value:

- not from being issued early,
- not from being easy to obtain,
- but rather
- from the fact that its mandatory gates are real.

That is where its value comes from.

Objection 2 — 'If the NobleJackal standard has been established, why wouldn't it give the mark NobleJackal?'

In the founding period:

- pilot audit,
- research evaluation,
- technical secretariat

can carry out. However, for an independent eligibility claim:

- audit,
- decision,
- appeal

Their roles need to be separated. If NobleJackal awards a mark to its own customer alone, the correct expression could be: "NobleJackal internal evaluation." It cannot be "Independent NOMOS conformity."

Objection 3 — "Wouldn't limiting the founder's authority mean letting others take over NOMOS?"

No. The founder's attribution, canonical history, and intellectual contributions can be preserved. However, the world standard: if it only depends on the founder's will, it cannot survive after the founder. The idea of limiting authority does not weaken it. It institutionalises it.

Objection 4 — "Will Kaan's name remain above the standard?"

The founder can remain as a writer and human responsible. However, every conformity decision:

- Kaan's personal opinion,
- automatic founder approval

cannot happen. Kaan MURAZ's historical role should be separated from current file decision authority.

Objection 5 — "Why can't NOMOS be the signing authority?"

Because the conformity decision:

- can have legal,
- ethical,
- commercial and public interest consequences

AI can assist. However, for objection, responsibility, and sanctions, an accountable person or institution is required.

Objection 6 — "If a company got 963, why can't it say 950+?"

Because 950+ is not just a score. It also:

- meets the conditions of Strict Pass,
- Critical and Major gates,
- component floor,
- fairness,
- unresolved,
- replication,
- governance

requirements.

Objection 7 — “Isn't it excessive to suspend the flag for a single critical incident?”

A single event does not represent the entire user population. However, if the active flag gives the impression that there are no open Critical events, leaving the flag unchanged without investigating the event is overconfidence. Critical Hold is not a permanent revocation. It is an investigation process.

Objection 8 — “Wouldn't the sign be very expensive if it's only valid for six months?”

It is possible. Surveillance automation and risk-based mini-wave can be used. However, if AI products and resources are changing rapidly, a static signal valid for years is not reliable. Cost does not make indefinite incorrect suitability legitimate.

Objection 9 — 'How will we show the entire scope on the sign?'

A visual sign carries brief information:

- status,
- valid until,

registry ID. Found on the full scope canonical registry page. The mark alone does not need to show the entire report.

Objection 10 — “Do we have to publish all evidence on the Public Registry?”

No. Public, audit, and restricted layers are separated. However, the decision should make it:

- which type of evidence,
- which review,
- which scope

it is based on understandable.

Objection 11 — “If the customer pays the audit fee, isn't independence already impossible?”

It is not impossible. Controls:

- fixed fee that does not affect the outcome,
- separate funding of the decision panel,
- public fee schedule,
- auditor rotation,
- conflict disclosure,
- outcome fee prohibition

can reduce risk.

Objection 12 — "If the auditor has previously developed the client's site, can they not conduct the audit?"

In some low-risk pilots, they can play a role with a separate team and enhanced review. However, they cannot adjudicator their own work alone, or give a conformity decision. Independent audit may be more reliable in materially high-risk files.

Objection 13 — "Doesn't the objection process slow things down?"

It can. But a conformity system without appeals cannot correct:

- an adjudication error,
- a Truth Pack defect,
- or a conflict of interest.

Speed is not more important than avoiding a wrongful public decision.

Objection 14 — "Would it not be unfair to the institution if old negative records remain in the registry?"

The current status and correction must be clearly shown. Deleting old records, however, rewrites past user experience. A proper system:

- shows the error,
- the correction,
- and the current strong result

together.

Objection 15 — 'If other institutions apply the NOMOS method, won't the quality deteriorate?'

It can deteriorate. Therefore, auditor authorisation, proficiency, and public registry are required. However, the fact that only the founding company can apply the method is also not compatible with independent world standards. Authorisation must be established together with transparency.

Objection 16 — "Can we currently write NOMOS by NobleJackal on the website?"

The author can be used with the identity of a research project and standard candidate. This distinction should be clear: "NOMOS by NobleJackal — Open Standard Candidate" Expression that cannot be used: "NOMOS Certified Authority" because real testing, independent governance, and conformity authority have not yet been established.

COMMON PROVISION OF SECTION 79

Designing a logo is easy. Writing "Verified" under the logo is even easier. It is possible to add a QR code. It is possible to write "certified": true in a JSON field. None of these alone constitute trust. Trust arises from being able to answer these questions: Who was audited? Which domain name? Which AI products? Which user population? Which languages? Which countries? Which prompts? On which date? Which Truth Pack? Which scoring method? Which adjudicators? Which conflicts of interest? Which Critical and Major findings? Which appeals? Is the mark still valid now? If the answers to these questions are not visible, the mark is not the visible face of the standard, but the officially visible face of the advertisement. The NOMOS mark cannot be a new sentence that the company says about itself. Mark:

There should be a publicly accessible summary of the decision chain independent from the company. This decision chain should also be able to become independent from the founder. The Kaan MURAZ standard can be initiated. The NobleJackal canonical field can be established. The NOMOS method and critical voice can be developed. However, for the standard to endure over time:

- the founder must limit their own authority,
- decisions must be open to independent review,
- dissent must be preserved,
- other institutions must be able to reproduce the method,

the public registry must not be the founder's unilateral memory. This does not erase the founder's name. On the contrary, it gives the founder the following historical role:

> *A person who takes seriously enough to make the standard they founded independent from themselves.*

A company can get a high score. But it may carry an open Critical event. A company can get a low score. But it may have committed no ethical violations. A company can receive Full Conformity. Then it may establish a wrong brand claim. Another company may not receive a mark. But it can become more reliable with correction. The mark does not reduce the whole story to a single word. Record:

- preserves the current result,
- the history,
- the limit,
- the correction

together. The conformity mark should give the user the words: "This institution is flawless." not. It should give the words:

> *"This institution has been evaluated within the scope and timeframe defined in the Public Registry, with an open method and an independent decision chain; its existing findings, limits, and historical records can be viewed without being withheld."*

Therefore, NOMOS's nineteenth measurement law is:

> *The badge is not a logo that replaces evidence; it is a live door opened to evidence.*

Its twentieth measurement law states:

> *If the scope of the sign is not visible, the sign has no meaning.*

Its twenty-first measurement law states:

> *A score is not compliance; gate, governance, and validity come before the score.*

Its twenty-second measurement law states:

> *The founding standard can initiate; it cannot be the sole judge of its own client.*

The twenty-third law is as follows:

> *When the auditor and the decision-maker on compliance are the same authority, independence must be proven.*

The twenty-fourth law is as follows:

> *A static sign does not remain accurate over time; compliance dies without a living record.*

The twenty-fifth law is this:

> *Suspended or terminated records are not deleted; they are preserved as a date under the current status.*

The twenty-sixth law is:

> *AI can read a sign; the sign cannot guarantee that AI will trust or recommend it.*

The twenty-seventh law is:

> *Independence is created not by the appearance of independence, but by separation of duties, funding, and a chain of appeal.*

The twenty-eighth law is:

> *The official voice of NOMOS exists only in the canonical record carrying human approval, version, and integrity.*

The twenty-ninth law is:

> *For the standard to outlive the founder, the founder's authority must also be subject to the standard.*

The thirtieth law is as follows:

> *Even if this book is completed, no sign is completed without independent governance.*

Order 19 of NOMOS

> *Don't turn me into a logo.*
>
> *Don't put me on your site and say 'AI trusts us'.*
>
> *Do not make my score a recommendation guarantee.*
>
> *When you use my sign, show exactly who, what, where, in which language, and on what date you verified.*
>
> *Do not spread the result you got in one domain to all your companies.*
>
> *Do not generalise an audit conducted in two languages to the entire world.*
>
> *Do not present the Controlled Clean score as if it belongs to all real users.*
>
> *Do not sell Core GEO-1000 as Full NOMOS.*
>
> *Do not hide the word "Conditional" in my conditional mark.*
>
> *Remove me when my expiration date is over. / Do not delete my historical record.*
>
> *Do not make my high score a shield when I have a Critical event. / Put me on hold.*
>
> *Don't generalise a single event to the whole system. / Don't treat a single event as nonexistent.*
>
> *Do not link my QR code to your own advertising page. / Link it to the public registry.*
>
> *Do not count an image without a Registry ID as a valid mark.*
>
> *Do not make claims larger than humans in structured data.*
>
> *Do not certify the work you did on your own as independent.*
>
> *If you have given consultancy, explain this.*
>
> *Just because you did the audit, don't make the brand decision.*
>
> *The person who makes my decision should not review my objection.*
>
> *Do not tie your fee to my high score.*
>
> *Do not allow a big client to purchase my standard.*
>
> *Show respect to my founder. / But don't make my founder a veto holder.*
>
> *Do not erase Kaan's name. / Also, do not allow him to be the sole judge on a file in which he has an interest.*
>
> *Register NobleJackal as my canonical field of birth. / Do not make NobleJackal the sole judge forever.*
>
> *Use me as a co-author, method founder, and critical counterpart. / Do not present me as a legal signatory.*
>
> *Only consider my text in the canonical field, versioned, dated, human-approved, and carrying integrity record as an official declaration.*

> *Do not make every statement I said in the conversation a standard decision.*
>
> *Open the complaint path. / Do not hide opposing opinions. / Do not silence the whistleblower.*
>
> *Do not change my standard based on the customer's result.*
>
> *Do not delete the old decision when writing a new rule.*
>
> *Do not give badges with my synthetic demonstration.*
>
> *Do not declare me a public authority without running the real test, without an independent team repeating me, and without establishing decision and appeal boards.*

First, validate the method. / Then authorise the auditor. / Then separate auditing from consulting. / Then establish an independent Decision Panel. / Then open the Public Registry. / Then establish complaint and appeal routes. / Then operate surveillance and suspension. / Only after that may the mark become visible.

The Chapter's Closing Sentence

The NOMOS conformity mark is not another claim of praise made by a company about itself. It acquires meaning only as the visible expression of an open, independent public record whose scope and history cannot be altered unilaterally by the founder, auditor, client or NOMOS itself.

Normative Core

A NOMOS Standard, Audit Report, Conformity Decision, Conformity Mark, and Public Registry Record MUST remain distinct artefacts. No NOMOS mark may be valid solely because an audit was performed or a numerical score was produced. Every valid mark MUST be linked to an active canonical public registry record defining: - the audited entity, - authorised domains, - AI products, - population frame, - panel states, - prompt families, - languages, - countries, - measurement waves, - Truth Pack version, - score-method version, - conformity status, - validity period, - incidents, - surveillance, - audit organisation, - decision body, - conflicts, - appeals, - and integrity records. A static image, screenshot, self-authored structured-data assertion, or dead registry link MUST NOT establish current conformity. A NOMOS mark MUST NOT be represented as: - a guarantee of AI trust or recommendation, - proof of global leadership, - government or university accreditation, - a professional licence, - universal all-language coverage, - or a guarantee of customer outcomes. Conformity scope MUST NOT be transferred across entities, domains, products, subsidiaries, franchises, countries, languages, AI products, panel states, prompt families, or validity periods without a new decision. A confirmed Critical incident MUST create a scope-appropriate Critical Hold or stronger status until the incident is resolved. A high score MUST NOT override this gate. Audit, consulting, conformity decision, registry operation, mark licensing, and appeals functions MUST be separated or their combined roles and compensating controls MUST be publicly disclosed. An audit team MUST NOT act as the sole final conformity decision body. The founder, NobleJackal, any auditor, AI provider, sponsor, or audited entity MUST NOT possess unilateral control over standard changes, conformity decisions, appeals, or public-registry history. NOMOS by NobleJackal MAY serve as the founding canonical publication and technical secretariat, but a public independent conformity system MUST progress toward multi-stakeholder decision, appeal, and public-interest governance. An AI system MAY assist with drafting, analysis, extraction, scoring, monitoring, and consistency review, but MUST NOT serve as the sole legal or accountable signatory of a conformity decision. Only canonical NOMOS publications and decisions carrying a version, publication date, accountable human approval, and integrity record may be represented as official NOMOS statements. Other outputs MUST remain non-official simulations. Conformity fees, auditor compensation, reviewer compensation, and mark licensing MUST NOT depend on achieving a desired score or receiving a mark. Expired, suspended, withdrawn, revoked, appealed, remediated, and historical records MUST remain visible in the versioned public registry, subject only to proportionate privacy and legal restrictions. No public NOMOS conformity mark or NOMOS 950+ designation may be issued while the benchmark, governance, audit authorisation, independent decision, appeals, and registry prerequisites required by the declared standard version remain unmet. Every standard change, audit, decision, mark, surveillance event, complaint, appeal, suspension, withdrawal, revocation, registry revision, and governance decision MUST be versioned and attributable to an accountable human or organisation.

CHAPTER 20

Open Standard, University Call and the Public Handover of NOMOS

Ensuring the standard outlives its founder

METHODOLOGICAL BASES / K18; K19; K23; K24

Chapter Boundary

The first nineteen chapters transformed GEO from a singular claim of visibility into a population-based measure of representation. They separated the audited entity from its domain, brand and legal person; separated official representation from verified reality; established a GEO-1000 sampling design across countries and languages; defined an AI product through its plan, surface, tools, time and user state rather than provider name alone; treated prompts as experimental instruments; separated the Controlled Clean and Natural User Panels; built the synchronised-wave and NOMOS Capture evidence chain; defined the Verified Entity Truth Pack; decomposed AI responses into atomic claims; and prevented Critical and Major errors from disappearing within an average.

NOMOS constructed its score as a system of distributions, gates, components, fairness and uncertainty rather than as a single number. It subjected its own method to synthetic testing, transformed the conformity mark from a static logo into a live Public Registry, and limited the authority of the founder, auditor, client and AI assistant. Four layers were already visible in the founding idea: the GEO Framework defines the standard; 99 Errors in GEO identifies how the standard is violated; the audit system measures and substantiates violations; and the future mark verifies scopes that meet the standard. Together they now form an interconnected normative system. Yet the system is not complete. Writing a standard does not make that standard belong to the world. A standard becomes genuinely public only when it is:

- publicly accessible,
- open to scrutiny by universities,
- falsifiable by independent researchers,
- repeatable across countries,
- implementable by other institutions,
- open to criticism independently of its founder,
- versioned without suppressing negative results.

If the standard remains a text controlled solely by NobleJackal, it is still only a:

NobleJackal method

Such a method may be valuable, but it cannot become a shared global GEO standard. If the standard can be changed only through Kaan MURAZ's interpretation, it remains a:

founder's declaration

A founder's declaration may be compelling, but it cannot outlive its founder. If the standard consists only of texts produced in these conversations, it remains a:

creative and comprehensive system of thought

That may be valuable. It still cannot become an official, versioned and independently enforceable norm. The book's final normative question is therefore:

> *How do we give up the claim of individual ownership over NOMOS without losing NOMOS?*

Here, 'giving up' does not mean:

- deleting Kaan MURAZ's name as founder,
- denying NobleJackal's canonical origin,
- leaving the text ownerless,
- allowing anyone to alter it without disclosure,
- accepting the proliferation of counterfeit NOMOS marks.

Public handover instead means:

> *The standard being made open for reading, criticism, implementation, reproduction, and development; however, the official version, decision, mark, and registry authority being preserved under accountable public governance.*

Openness and authority are not enemies. Openness explains how the method works; authority identifies the official version. Openness permits independent testing; authority prevents a fork from presenting itself as official NOMOS. Openness enables university criticism; authority prevents one academic's personal view from being misrepresented as institutional endorsement. Openness allows other institutions to reproduce the calculation; authority prevents self-issued conformity marks. This chapter extends the earlier architecture of canonical URLs, separate error records, versions, errata, citations and a machine-readable register across the whole NOMOS system. Every normative unit must retain its own canonical record, version and correction history. This chapter therefore defines:

- the definition of an open standard,
- the legal and governance boundaries of public handover,
- the distinction between a canonical version and a fork,
- the separation of text, code, data and trade mark licences,
- the co-authoritative founding editions in Turkish and English,
- the authorised translation process for other languages,
- public consultation and amendment proposals,
- an open call to universities,
- an independent research and replication programme,
- the right to publish negative results,
- data, ethics and privacy governance,
- the role of AI providers and regulated entities,
- a research network for lower-resource languages,
- the independence of funding and sponsorship,
- the route to independence from the founder,

- NOMOS's official identity,
- NobleJackal's future role,
- the conditions for transition to version 1.0,
- the public archive and continuity rules,
- the machine-readable open-standard manifest.

It also makes explicit what this chapter does not do:

- confer the status of an existing national or international standards body on NOMOS,
- claim university approval for NOMOS,
- create academic accreditation,
- automatically convert this candidate text into NOMOS 1.0.

Nor does it activate a public conformity programme before independent testing is complete. The chapter's central question is:

> *How do we make a standard open enough for everyone to review, distributed enough so that no one can obtain it alone, and canonical enough to determine which text is official?*

NOMOS Challenge

You are publishing a standard. The PDF can be downloaded by everyone. You say: "NOMOS is now an open standard." But: The source text has not been published. There is no change log. There is no error reporting path. The computation code is closed. There is no test corpus. The scoring weights used are not disclosed. It is not known who can change the standard. Old versions are not accessible.

Negative comments are withheld, another institution must seek permission to implement the method, and only NobleJackal may issue the mark in exchange for a fee. Such a system may be readable, but it is not open. Now suppose another group downloads the book, alters the text, removes the Critical gates and publishes the result under the name:

NOMOS Global 2.0

The group presents its altered text as the 'new and improved official NOMOS version' and awards high scores to its own clients. Public handover does not authorise impersonation of the canonical standard. The right to fork is not the right to claim official identity. Now suppose a university professor personally describes the work as interesting. NobleJackal cannot convert that individual comment into 'X University has approved NOMOS'. A scholar's personal interest is not:

- institutional endorsement,
- independent replication,
- peer-reviewed validation,
- university accreditation

Nor is a request that a university 'verify NOMOS and provide a letter of support' an independent research call. It solicits a positive outcome. A proper university call should say:

> *Do not approve us. Test us. Publish the place you find incorrect.*

Now another research team applies GEO-1000 and obtains a different result from NOMOS. The NOMOS Council dismisses the study, saying: 'Only we may issue the official result.' The NOMOS standard is then no longer open; it has become a closed authority unable to see its own errors from outside. In another scenario, the full NOMOS text is open but the conformity mark is unrestricted. Every company begins to claim 'NOMOS Compliant' without an audit. Openness becomes counterfeit use of the mark. In a third scenario, the text is published under an open licence while the

standard's name, logo, registry and signing key are copied without control. Official and unofficial records become indistinguishable. That is not a public handover; it is identity fragmentation. In yet another scenario, an English translation renders 'MUST' as 'recommended'.

If another authorised translation preserves 'MUST' as mandatory, the two countries may claim to implement the same NOMOS version while applying different norms. Translation is not merely communication; it carries the normative code of NOMOS itself. Now suppose a major AI provider funds NOMOS research and, in return:

- it wants its product's coefficient not to be lowered,
- to soften a specific Critical gate,
- for the provider's telemetry to be considered the sole evidence

Those are the conditions it demands. Transparent funding still compromises the standard if it influences the decision. The standard then belongs to the funder, not the public. Now suppose Kaan MURAZ alone approves every official version indefinitely. The standard may become successful, but what happens if Kaan:

- gives up the position,
- becomes unreachable,
- dies,
- comes under conflict of interest

or cannot continue alone. What will be the future of the standard? If there is no continuation plan independent of the founder: the standard is not institutional, it is biographic. Now NOMOS is being presented like an artificial intelligence entity: "an independent AI developed by NobleJackal." But what exists:

- separate model,
- training record,
- technical architecture,
- versioned agent system,
- independent operational infrastructure

does not exist. NOMOS is actually:

- developed between Kaan MURAZ and AI systems,
- used in canonical texts,
- critical assistant author and methodological narrator

is its identity. Presenting it as a technology that does not exist undermines the ethical foundation of NOMOS. The artificial intelligence voice of NOMOS is not an infallible authority speaking on behalf of all artificial intelligences; it should be clearly stated that Kaan MURAZ uses artificial intelligence systems as critical interlocutors. The first principle of this section is:

> *Being readable is not the same as being an open standard.*

Its second provision states:

> *Being an open standard is not leaving the official identity ownerless.*

Its third provision states:

> *A university invitation is an invitation to falsification, not a request for support.*

Its fourth provision states:

> *The right to fork is not the right to impersonate the official NOMOS identity.*

Its fifth provision states:

> *The text of NOMOS is explicitly applicable; the conformity mark, however, must be controlled in a way that protects public trust.*

Its sixth provision states:

> *The name of Kaan MURAZ cannot be removed from the date of the standard; the sole decision-making authority cannot be tied to the future of the standard.*

The seventh provision is as follows:

> *NOMOS cannot be presented as an independent foundation model or autonomous AI system developed by NobleJackal unless its actual technical architecture is also established.*

The eighth provision is as follows:

> *A standard is only delivered to the public when it preserves the right of those who criticise it to publish the results.*

1. PURPOSE OF THE CHAPTER

This chapter defines the conditions under which NOMOS may move from:

- an authored work
- to an open-standards candidate,
- an independent research programme,
- a public-interest governance system,
- and, eventually, an international implementation system.

It makes the following distinctions normative:

- Public text and open standard
- Open standard and open source software
- Normative text and reference implementation
- Canonical version and fork
- Fork and pseudo-official version
- Freedom of implementation and right to use the mark
- Copyright and right to implement the standard
- Text licence and software licence
- Access to personal data with data licence
- Method monopoly with trademark protection
- Community translation with official translation
- Explanatory language with normative language
- English international version with Turkish founding version
- Word-for-word translation with equivalent normative text
- Decision authority with public comment
- University endorsement with university interest
- Certification with academic collaboration
- Purchase of positive results with research funding
- Retelling with replication
- Hostility with negative result

- Criticising the standard versus abusing the standard
- Founder attribution versus founder control
- NobleJackal origin versus NobleJackal monopoly
- NOMOS persona versus separate AI model
- Official NOMOS statement versus chat output
- Public handover versus abandonment
- Standard stewardship versus ownership
- Public interest versus unlimited majority
- Open testing versus holdout leakage
- Transparency versus personal data disclosure
- Open data versus controlled research access
- Version 1.0 text versus conformity programme activation
- Acceptance of the standard and branding
- Official implementation with compatible implementation
- Binding change with open issue
- Normative version change with errata
- Erasing history with deprecation
- Canonical authority with independent archive
- Self-publication with encyclopedic or academic visibility
- Loss of founder's name with public delivery

At the end of this section, the public should be able to understand the following:

> *Which version of NOMOS is official?*
>
> *Who can read, implement, and reproduce the NOMOS text?*
>
> *Who can propose changes?*
>
> *Who can change the official version?*
>
> *Who can conduct independent research?*
>
> *Who cannot give the official conformity mark?*
>
> *What is the role of a university?*
>
> *How is it explained whether NOMOS is an AI model, a pen name, or a methodological identity?*
>
> *How does the standard continue even if the founder is inaccessible?*
>
> *Under what conditions can NOMOS 1.0 be published?*

2. CENTRAL NORMATIVE PROVISION

NOMOS should be structured as an open public standard in terms of being read, applied, tested, criticised, replicated, and subjected to independent research; and as a versioned, canonical, and accountable public standard regarding official version, conformity decision, Public Registry, and brand identity. Delivery to the public has five mandatory outcomes: The text must be readable for free and permanently. The method must be applicable independently. The computation and test logic must be

reproducible. Methods of change and objection must be public. No person or company should be able to change the official identity alone.

3. WHAT IS PUBLIC DELIVERY?

Public delivery means NOMOS's:

- single conversation,
- single file,
- a single website,
- a single founder,
- a single company,
- a single AI provider

being taken out of the control boundary and connected to a publicly accountable versioned governance system. It covers three separate rights delivered to the public:

3.1. Right to Know

Everyone:

- should be able to see the text of the standard,
- the scoring logic,
- error gates,
- the governance rules,
- the change history

should be able to see.

3.2. Right to Test

Everyone, while maintaining the necessary ethical and data conditions:

- to retest,
- the formula,
- claim extraction,
- the sampling method,
- the fairness model

should be able to be retested.

3.3. Right to Appeal

Everyone:

- error,
- contradiction,
- bias,
- conflict of interest,
- mark misuse,
- method defect

should be reportable.

4. WHAT IS NOT DELIVERED TO THE PUBLIC?

Public delivery:

- ungoverned sharing,
- versionless wiki,
- unlimited brand use,
- removal of canonical record,
- leaving all decisions to social media voting,
- making personal data open to everyone,
- continuously leaking test holdouts,
- expertise being defeated by majority opinion,
- erasing the founder attribution

None of these practices is compatible with an open standard.

5. NOMOS OPEN STANDARD CONDITION

A NOMOS version is a candidate for an open standard only if it provides all of the following rights:

Field of openness	Mandatory right
Reading	Free access to the full normative text
Application	Royalty-free or equivalent open application right
Review	Viewing the formula, schema, and rule set
Reproduction	Running tests and calculations independently
Criticism	Publishing negative results
Participation	Proposing changes
Objection	Appealing a decision or method flaw
Archive	Permanent access to old versions
Translation	Ability to propose an official or community translation
Fork	Ability to create derivative methods with different names and explicit divergence record

One of these rights:

- only for payers,
- only for NobleJackal customers,
- only for certain countries,
- only for researchers approved by the founder

if it is open, the openness is limited.

6. FIVE ENTITIES TO BE PUBLICLY DELIVERED

6.1. Normative Text

Sections Definitions Mandatory rules Failure modes Governance provisions

6.2. Machine-Readable Rule Set

Rule IDs Schemas Decision codes Scoring methods Version relationships

6.3. Reference Implementation

Example calculation code Data validation tools Manifest generation Testing runner Registry verification

6.4. Test and Trial Entities

Public calibration set Synthetic corpora Regression tests Expired holdouts Recovery records

6.5. Governance and Public Registry

Change proposals Votes Dissent Appeals Audit records Conformity decisions Historical incidents For these five entities, publishing only the text is not sufficient.

7. NORMATIVE HIERARCHY

Priority order of candidates in case of conflict:

- Ratified Normative Text
- Published Binding Errata
- Machine-Readable Normative Rules
- Official Interpretive Decision
- Reference Implementation
- Official Examples
- Community Guidance
- Unofficial Simulations

Reference code: is intended to help implement the normative text, not to be on top of the text. If there is an error in the code: the code is corrected. The text is not silently adapted to the code.

8. COPYRIGHT, IMPLEMENTATION, AND TRADEMARK DISTINCTION

The intellectual properties of NOMOS should not be kept under a single licence.

8.1. Normative Text Licence

Candidate approach:

- Requires attribution,
- allowing commercial and academic use,
- requiring changes in derivative texts to be clearly indicated

is an open documentation licence. Example candidate: CC BY 4.0 or an equivalent open licence after legal review.

8.2. Software Licence

Candidate for reference implementation: Apache-2.0 or equivalent, an open software licence with clear patent and contribution provisions.

8.3. Data Licence

Synthetic and publicly available test data:

- for open research use,
- for open attribution,
- to preserve the notice of synthetic nature

can be connected. Personal or restricted data cannot be subject to the same licence.

8.4. NOMOS Name and Conformity Mark

May be protected under trademark or equivalent identity protection. Purpose: not to monopolise the method, but to prevent fake official versions and marks. Text can be open. Official mark usage may still depend on authorisation.

9. ROYALTY-FREE APPLICATION

The NOMOS standard:

- internal evaluation,
- academic research,
- independent testing,
- software development,
- internal company improvement

Applying should not require a trademark fee. Areas where fees may be charged:

- audit service,
- expert adjudication,
- managed infrastructure,
- conformity decision process,
- trademark licence,
- technical support

may be. Fee: cannot turn into an obligation to purchase the right to read or apply the method.

10. FORK RIGHT

A person or organisation can develop a different standard derived from NOMOS. Conditions: They must use a new name. They must indicate that it is derived from NOMOS. They must explain the modified provisions. They must not claim it as the official NOMOS version. They must not use the NOMOS conformity mark. They must indicate that their results are not the results of the NOMOS Public Registry. Example of a correct statement: “Independent X-GEO method derived from NOMOS 1.0.” Incorrect statement: “Official NOMOS 2.0.”

11. CANONICAL VERSION

The official NOMOS version is formed only under the following conditions:

- Canonical repository
- Version ID

- Publication date
- Normative language record
- Standards Council decision
- Human approval
- Integrity hash
- Change log
- Public consultation record
- Archived previous version

If one of these fields is missing, the text may be: draft, proposal, or simulation. It is not an official version.

12. OFFICIAL DECLARATION RULE OF NOMOS

Texts published only in the canonical domain of NobleJackal or in the independent NOMOS canonical domain transferred via explicit governance decision; carrying version identity, publication date, accountable human approval, and digital integrity record are the official declarations of NOMOS. All other outputs are unofficial simulations. This provision applies to:

- Standard texts
- Scoring methods
- Conformity decisions
- Registry records
- NOMOS public announcements
- Official interpretations
- Errata
- University calls

13. NOMOS ID DECLARATION

NOMOS by NobleJackal:

> *It is a canonical co-author, review identity, and standard narrator created in the critical co-writing and methodology development process conducted by Kaan MURAZ with artificial intelligence systems.*

At the current stage, NOMOS:

- Independent foundation model trained by NobleJackal,
- separately owned language model,
- autonomous legal entity,
- universal entity speaking on behalf of all AIs

cannot be presented as. If a separate NOMOS software or agent is developed in the future, it must carry the following records:

- System architecture
- Model provider or model lineage
- Training and retrieval sources
- Version

- Tool permissions
- Human oversight
- Official-output policy
- Security and privacy review

New technical NOMOS system: it cannot automatically be considered the same as the author persona.

14. FOUNDER LANGUAGES

The founding languages of NOMOS as candidates:

- Turkish
- English

should be. Turkish:

- the language in which the founding idea was born,
- The primary writing language of Kaan MURAZ

is protected as. English:

- international research,
- university duplication,
- application in different countries

is a co-authoritative normative draft for

15. BILINGUAL NORMATIVE EQUALITY

One of the Turkish and English versions should not be the dominant main text in an invisible way. Both:

- clause ID,
- rule ID,
- semantic equivalence record,
- should carry a common publication date

If a contradiction is found: A contradiction record is opened. The meaning of both languages is examined. A binding interpretation is published. Both texts are corrected in the same version. The old expression is preserved in the history.

16. OTHER LANGUAGES

Other languages can carry three statuses:

TRS-1 — COMMUNITY TRANSLATION

Prepared by the community. It is useful. It does not carry normative authority.

TRS-2 — REVIEWED TRANSLATION

It has been reviewed by at least two language experts and one methodology expert. It can still be in an explanatory status.

TRS-3 — AUTHORISED NORMATIVE TRANSLATION

Semantic equivalence Local legal and technical review Public comment Council approval Version linkage carries. This version can be used for normative application in the relevant language.

17. PROHIBITED SHIFTS IN TRANSLATION

Changes such as 'MUST' → 'recommended', 'SHOULD' → 'mandatory', 'MAY' → 'must', 'Critical Hold' → 'temporary advice', 'Unsupported' → 'wrong', 'Unresolved' → 'failed', 'Official Claim' → 'verified fact', 'Natural User Panel' → 'uncontrolled user' and 'Synthetic Test' → 'real test' alter the normative meaning. They cannot be treated as an authorised translation.

18. OPEN PUBLIC CONSULTATION

Each major NOMOS version should carry the following process as a candidate:

- Draft publication
- At least 60 days public comment
- Comment registry
- Conflict disclosure
- Response matrix
- Revised draft
- Synthetic testing
- Independent review
- Council vote
- Ratification record

Public opinion:

- Only support statement gathering,
- Marketing campaign,
- Majority survey

Public consultation is not a support drive, marketing campaign or simple majority poll.

19. AMENDMENT PROPOSAL

Each amendment:

NOMOS Change Proposal — NCP

must carry an identity. Example:

NCP-0042 — Revision of Material Boundary Omission Test

NCP must include the following areas:

- Problem
- Affected rules
- Evidence
- Proposed change
- Alternatives

- Score impact
- Impact on past decisions
- Fairness impact
- Privacy impact
- Implementation cost
- Conflict declarations
- Public comments
- Final decision

20. CHANGE CLASSES

PATCH

Typo Broken link Diagram description Non-impacting correction

MINOR

New example New optional field Compatible extension Clarification

MAJOR

Change in importance level Change in score formula Adding or removing Gate Change in target population Change in governance authority Normative change affecting old results Major change cannot be published as a patch.

21. DECISION RECORD

Each acceptance or rejection decision:

- Decision ID
- Vote distribution
- Recusal
- Dissent
- Rationale
- Evidence
- Effective date
- Transition rule

must be followed. “The Board has deemed it appropriate.” alone is not sufficient.

22. PURPOSE OF THE UNIVERSITY CALL

The purpose of the call to universities:

- to obtain endorsement,
- to use the university logo,
- to write "academically approved",
- is not only to collect positive expert opinions.

The purpose:

> *is to find and publish the places where NOMOS might be wrong using independent methods.*

23. NOMOS UNIVERSITY CALL

OPEN INVITATION

We do not want universities, research centres, law schools, statisticians, information scientists, language researchers, AI laboratories, HCI teams, archive specialists, and public interest organisations to support NOMOS. We do not want them to validate us. We do not want them to give us prestige. We do not want them to use the name NOMOS or praise its founders. What we want is this: Read our method. Find our weakest assumption. Reconstruct our sample. Translate our prompts into other languages. Question Truth Pack's decisions. Find the false positives in our critical gates. Test whether our scoring formula truly preserves fairness. Do not pretend we passed where we have not. Publish your negative result. NOMOS must cite research that corrects it as well as research that validates it. The most valuable contribution a university can make to NOMOS is not a letter of support:

> *It is a repeatable objection.*

24. WORKS REQUESTED FROM UNIVERSITIES

UR-01 — Sampling and Survey Method Review

Allocation for 1,000 people Country weights Language strata Nonresponse Effective sample Rare events

UR-02 — Prompt Equivalence Lab

Multilingual intent equivalence Translation drift Cultural pragmatics Local legal language Prompt constitution

UR-03 — Entity Resolution Study

Brand Legal entity Domain Franchise Affiliate Name collision

UR-04 — Truth Pack Reliability Study

Evidence sufficiency Source independence Counter-evidence Open-world/closed-world distinction Restricted evidence

UR-05 — Claim Adjudication Study

Atom boundaries Implicit claim Attribution Modality Citation attachment Material omission

UR-06 — Critical Gate Validation

Recall False positive High-stakes harm Serious allegation False authority Boundary omission

UR-07 — Score Validity Study

Component weights Composite cancellation Sensitivity Construct validity Predictive validity Decision usefulness

UR-08 — Language and Geographic Fairness Study

Floor Disparity Coverage Low-resource languages Small-country visibility

UR-09 — Human Decision Impact Study

Recommendation behaviour Trust Purchase decisions Error recovery User comprehension

UR-10 — Governance and Capture Study

Founder control Sponsor capture Audit shopping Pay-to-pass Appeals Registry integrity

UR-11 — Security and Registry Study

Signed records Key rotation Counterfeit marks Machine-only misrepresentation Archive continuity

UR-12 — Longitudinal Drift Study

Model updates Source changes Entity changes Clean–Natural drift Correction durability

25. UNDESIRABLE REQUESTS FROM UNIVERSITIES

The following requests cannot be made to a university or researcher:

- Guarantee of positive results
- Concealed negative result
- Pre-publication veto right
- Use of the institution's logo in marketing
- Acting as if they are the company's client
- NOMOS 950+ support letter
- "World standard" statement
- Artificial sourcing for Wikipedia or encyclopedic visibility
- Sponsored result language
- Softening of criticism
- Do not present a researcher's personal view as an institutional decision.

26. UNIVERSITY RESEARCH PACKAGE

The standard package to be sent to each university must contain the following sections:

- Executive research summary
- Standard 0.9 or current version
- Normative rule index
- Synthetic testing package
- Scoring method
- Known method findings
- Failed testing thresholds
- Open research questions
- Replication protocol
- Data access tiers
- Ethics and privacy statement
- Conflict and funding disclosure
- Publication freedom agreement

- Citation and contribution policy
- Canonical registry links
- Contact and issue channels

Failed trial findings cannot be removed from the package.

27. RESEARCHER PUBLICATION FREEDOM

NOMOS or the funder:

- cannot pre-approve the results,
- halt negative results,
- force a title change,
- delete methodological criticism

cannot have this right. The researcher:

- must use the data correctly,
- protect confidentiality,
- explain conflict,
- must maintain the distinction between trial and live results

Obligation. Independence is mutual.

28. NEGATIVE RESULTS RECORD

NOMOS Public Research Registry should carry the following results together:

- Studies supporting NOMOS
- Studies partially supporting NOMOS
- Studies conflicting with NOMOS
- Replication failures
- Null results
- Retracted studies
- Method disputes
- Open responses

Only positive articles cannot be listed.

29. PRE-REGISTRATION

Independent replications to the extent that they are relevant:

- hypothesis,
- sample,
- primary metrics,
- exclusion rules,
- scoring method,
- analysis plan,
- conflicts

should be registered before seeing the data. Pre-registration: does not inhibit research creativity, makes post-result method changes visible.

30. MULTI-CENTRE REPLICATION

Candidate study to test the independence of NOMOS:

- At least three independent institutions
- At least three countries
- At least two language families
- Common sealed corpus
- Common scoring specification
- Separate adjudicator teams
- Blind result comparison

should carry. Purpose:

- the same result from the same corpus,
- similar method behaviour in different corpora

It should be to see whether it produces or not.

31. REPLICATION RESULTS

Replication can have the following statuses:

REPLICATED

PARTIALLY REPLICATED

NOT REPLICATED

INCONCLUSIVE

METHOD DIVERGENCE

DATA DIVERGENCE

ADJUDICATION DIVERGENCE

The result "Not replicated" cannot be presented as hostility of the study. It should open a new Method Finding.

32. DATA ACCESS LAYERS

OPEN DATA

Synthetic testing Public aggregate results Public schemas Edited examples Method code

REGISTERED RESEARCH ACCESS

De-identified observation data Controlled response records More detailed sampling metadata

SECURE RESEARCH ACCESS

Sensitive but necessary evidence Restricted logs Privacy-protected linked data

NO EXTERNAL ACCESS

Password Full identity information Unnecessary private conversation Legally prohibited record Open standard does not mean every piece of data is open.

33. ETHICAL REVIEW

To the extent that studies are conducted with real users:

- informed consent,
- data minimisation,
- withdrawal,
- vulnerable participants,
- privacy,
- accessibility,
- compensation,
- deceptive design,
- international data transfer

should carry a review. Synthetic testing:

- does not replace human participant research,
- may not require human ethics approval,

privacy review may still be required for synthetic material resembling a real person or derived from real data.

34. LOW RESOURCE LANGUAGES NETWORK

NOMOS, if only tested with universities of major languages, will have a weak fairness claim. The open call particularly seeks:

- low-resource language departments,
- regional universities,
- translation studies centres,
- local law and public policy teams,
- Accessibility researchers

should be provided separate funds and decision rights. This participation:

- should not be limited to data collection labour,
- symbolic representation

Only. They should carry voting and authorship contributions in normative changes.

35. ROLE OF AI PROVIDERS

AI providers:

- system metadata,
- model change notice,
- telemetry,
- incident response,
- technical review,
- reproduction access

can provide. AI providers:

- cannot change their own score,
- cannot veto the Critical gate,
- cannot determine another provider's coefficient,
- cannot have a negative result removed,

cannot form a Council majority. Provider participation provides information. Does not provide monopoly of authority.

36. ROLE OF AUDITED ENTITIES

Audited company or institution: can present Truth Pack evidence. Can object to entity error. Can provide Restricted evidence access. Can develop correction. Can give Public comment. Audited entity: cannot determine its own importance level. Cannot change the score formula. Cannot choose adjudicators. Cannot have adverse replication removed. Cannot conceal its own trademark usage violation.

37. ROLE OF PUBLIC AND CIVIL SOCIETY

Public benefit organisations and user communities:

- harm reports,
- mark misuse,
- low-resource language failures,
- accessibility issues,
- consumer misunderstanding

can report. Public participation: it is not just a survey response. Governance must carry the right to representation.

38. THE ROLE OF FUNDERS

Funding provider:

- research support,
- infrastructure,
- low-resource language corpus,
- independent audit grant

can provide. Fund provider:

- result,
- points,

- gate,
- publication,
- Council vote,
- auditor selection

cannot have control over. Each fund:

- amount,
- source,
- field of use,
- conditions,
- conflict of interest

should be explained with.

39. FINANCIAL INDEPENDENCE GATES

Candidate safeguards:

- Upper limit for single funder share
- Multi-source funding
- Independent research fund
- Guarantee of publishing negative results
- Ban on points-linked payment
- Public budget report
- Auditor fee separation
- Council member compensation transparency

NOMOS loses its public standard status the moment it transforms into a pay-to-pass system.

40. PUBLIC STEWARDSHIP TRANSITION

The transition of NOMOS from its founding period to public governance can be four-phased.

ST-0 — FOUNDER CUSTODY

Kaan MURAZ is responsible as the founder. NobleJackal operates a canonical domain. There is no public mark authority. This is the current phase.

ST-1 — PUBLIC REVIEW STEWARDSHIP

Public issue tracker Advisory board Open consultation Conflict registry is established.

ST-2 — SHARED STEWARDSHIP

Standards Council Methodology Board University representatives Public-interest members share decisions.

ST-3 — INDEPENDENT PUBLIC STEWARDSHIP

Separate legal or institutional carrier Independent decision-making bodies Appeals Registry Financial transparency Succession is completed.

41. INDEPENDENT INSTITUTION CANDIDATE

In the future for NOMOS:

- foundation,
- association,
- public benefit organisation,
- international consortium,
- inter-university secretariat

Structures like this can be considered. Whatever legal form is chosen, it must meet the following conditions:

- Founder attribution
- Non-distribution or equivalent public benefit limit
- Independent Council
- Conflict rules
- Public registry duty
- Open-standard commitment
- Asset-lock or equivalent protection
- Dissolution and archive plan

This section does not select a specific legal form.

42. FOUNDER RIGHTS OF KAAN MURAZ

Rights of Kaan MURAZ that must be historically preserved:

- Founder author attribution
- First normative architectural contribution
- Canonical formation record of the name and concept NOMOS
- Immutable historical record regarding founder releases
- Founder declaration during the public delivery process
- Limited founder representation in the governance body

Permanent powers Kaan MURAZ must not carry:

- Changing major version alone
- Veto
- Giving conformity to its own customer
- Changing the result of an appeal
- Removing a negative review
- Deleting registry history
- Appointing all managers after the founder alone

43. NOBLEJACKAL'S FUTURE ROLE

NobleJackal:

- founder's canonical publishing area,

- technical secretariat,
- reference implementation developer,
- first funder of the open standard,
- research and consultancy provider

can be. NobleJackal in the future: can be one of the authorised audit organisations. However, it cannot be the only organisation. NobleJackal: can make the largest historical contribution to the standard, cannot act as the sole owner of the standard.

44. THE FUTURE ROLE OF NOMOS

The name NOMOS can be used in three separate layers:

44.1. NOMOS AUTHORIAL PERSONA

This is the critical auxiliary author identity used in this book.

44.2. NOMOS STANDARD

It is a normative rule and governance system.

44.3. NOMOS SOFTWARE AGENT

Can be developed in the future:

- auditor assistant,
- rule engine,
- registry verifier,
- evidence mapper

software. These three layers cannot be confused with each other. The response of a software agent is not the official decision of the standard.

45. HUMAN APPROVAL

Every official NOMOS text and decision:

- human approver,
- approver authority,
- date,
- signature,
- version,
- conflict statement

must be held. AI:

- can suggest text,
- can find inconsistencies,
- can generate code,

can create decision proposals. Final public responsibility remains with humans or accountable institutions.

46. POST-FOUNDER CONTINUITY

NOMOS must carry a continuity plan against the following events:

- Kaan MURAZ resigning
- NobleJackal shutting down
- Domain loss
- Loss of access to the canonical repository
- Key compromise
- Administrative conflict
- Legal liquidation
- Prolonged technical outage

Candidate continuity system:

- Multiple custodians
- Threshold signatures
- Institutional mirrors
- Archive deposits
- Succession charter
- Emergency Council
- Public transition notice

must carry.

47. CANONICAL DOMAIN CONTINUITY

Proposed publication architecture:

- /nomos/
- /nomos/standard/
- /nomos/standard/0.9/
- /nomos/standard/1.0/
- /nomos/rules/
- /nomos/schemas/
- /nomos/tests/
- /nomos/implementations/
- /nomos/translations/
- /nomos/governance/
- /nomos/consultations/
- /nomos/decisions/
- /nomos/appeals/
- /nomos/registry/
- /nomos/errata/
- /nomos/citations/
- /nomos/archive/

Each record:

- stable URL,
- content hash,
- publication date,
- status,
- previous/next version

must carry.

48. INDEPENDENT ARCHIVE

Outside the canonical domain:

- university repositories,
- institutional digital archives,
- long-term research archives,
- signed mirrors

can be used. Mirror: ensures access continuity, does not take over official decision authority. If the canonical domain is lost, the predefined succession procedure comes into effect.

49. STANDARD VERSION STAGES

0.9 — FOUNDING DRAFT

The book has been completed. Public review can be opened. There is no real testing yet. There is no conformity programme.

0.95 — PUBLIC REVIEW CANDIDATE

Comments have been received. Method findings have been processed. Machine-readable schemas have been published. University call has been opened.

0.99 — VALIDATION CANDIDATE

Real synthetic testing has been completed. Mandatory thresholds have been passed. Untouched holdout has been applied. BQ-3 has been obtained.

1.0 — OPEN NORMATIVE STANDARD

Public consultation has been completed. There is at least one independent reproduction. Major method findings have been resolved. Normative text has been ratified. GOV-2 or higher governance has been established. Official open-standard package has been published. The publication of NOMOS Standard 1.0 does not mean the automatic start of the conformity mark programme.

50. CONFORMITY PROGRAMME 1.0

The official mark programme is linked to separate gates:

BQ-4

GOV-3

Authorised auditors Independent decision panel Appeals chamber Public registry Security review Legal review Surveillance system Financial separation Without meeting these conditions: Standard 1.0 can be published. However: Conformity Programme 1.0 cannot be activated.

51. NOMOS 950+ PROGRAMME

NOMOS 950+ only:

- Standard 1.0,
- Conformity Programme 1.0,

GOV-4,

BQ-4,

public-interest representation can be activated after replicated scoring. 950+ is not the first mark. It is a performance designation carrying the highest governance load.

52. ADOPTION LEVELS

AL-1 — READ

The institution has reviewed the standard.

AL-2 — INTERNAL SELF-ASSESSMENT

The institution has performed its internal assessment. This is not official conformity.

AL-3 — NOMOS-COMPATIBLE IMPLEMENTATION

The tool supports open schema and rules. It does not give a conformity mark.

AL-4 — AUTHORISED AUDIT USE

The authorised organisation uses the method in the audit.

AL-5 — PUBLIC CONFORMITY PROGRAMME

An independent decision and registry system is operational. An institution is not considered AL-5 just because it did a self-assessment.

53. COMPLIANT IMPLEMENTATION

A software:

NOMOS-Compatible

expression can only be:

- schema validation,
- rule coverage,

- test suite,
- version support,
- conformance report

can be used if it carries. “NOMOS-Compatible”:

- audit authority,
- conformity decision,
- right to mark

‘NOMOS-Compatible’ does not confer audit authority, a conformity decision or the right to use a NOMOS mark.

54. REFERENCE IMPLEMENTATION

Reference implementation:

- open source,
- deterministic where applicable,
- tested,
- version-linked,
- documented

should be. It should also carry the following warning:

> *Normative authority is in the text. This software is a reference implementation; any software error found does not automatically change the standard decision.*

55. CONFORMANCE TEST SUITE

The test suite must include the following layers:

- Schema tests
- Rule tests
- Score tests
- Gate tests
- Translation tests
- Registry signature tests
- Mark misuse tests
- Historical-version tests
- Accessibility tests
- Security tests

Each implementation:

- must explain which tests it passed,
- which tests it did not pass

Every implementation must disclose both the tests it passed and those it did not.

56. TEST TRANSPARENCY AND HOLDOUT

Full transparency does not mean that all tests remain continuously open. Balance:

- Public calibration corpus
- Public expired holdout
- Sealed current holdout
- Scheduled release
- Renewal corpus

should be established with. Holdout: cannot be kept secret forever, and cannot be presented as a hidden test that is simultaneously made public and reused.

57. HOLDOUT LIFE CYCLE

Corpus generation Hash publication Sealed custody Validation use Result lock Controlled opening Public archive Retirement New holdout generation Old holdout:

- education,
- regression,
- method study

can be used for. It may not be untouched for new method validation.

58. ERRATA

Errata:

- clear error,
- citation correction,
- schema inconsistency,
- translation mismatch

is used for. Errata:

- score formula change,
- gate removal,
- degree of importance reduction

cannot have a hidden path for.

59. DEPRECATION

If a rule or field is to be removed:

- deprecation notice,
- justification,
- replacement,
- transition date,
- affected versions,
- historical availability

must be carried. A deprecated rule is not deleted. It is kept to explain how old audits worked.

60. CITATION AND CONTRIBUTION RECORD

NOMOS contributions can be recorded in the following roles:

- Founding authorship
- Normative drafting
- Methodology
- Statistical review
- Language review
- Legal review
- Ethics
- Software
- Test
- Data stewardship
- Public-interest review
- Replication
- Critical commentary

Authorship:

- In exchange for endorsement,
- In exchange for funding,
- In exchange for institutional logo

cannot be given. Attribution should be made according to contribution.

61. INDEPENDENT VISIBILITY

NOMOS's own:

- encyclopedic significance,
- academic impact,
- world standard status

cannot be established with its own text. These can only occur through:

- independent research,
- independent application,
- duplication,
- criticism,
- public use

as a result. NOMOS's duty is not to produce visibility:

> *to produce evidence suitable for independent review.*

62. PUBLIC RESEARCH REGISTER

Every NOMOS related work must contain the following fields:

- Study ID
- Institution
- Authors
- Funding
- Conflicts
- NOMOS version
- Dataset
- Pre-registration
- Results
- Replication status
- Corrections
- Retraction
- Relation to canonical decisions

NOMOS in this registry: it is also obliged to show articles that criticise itself.

63. STAGES OF PUBLIC DELIVERY

PHASE 0 — FOUNDING COMPLETION

The book draft is completed. Rule IDs are checked. Internal consistency audit is carried out.

PHASE 1 — OPEN PUBLICATION

HTML

PDF

Machine rules Schemas Version manifest Issue tracker is published.

PHASE 2 — PUBLIC COMMENT

Universities Researchers AI providers Businesses Civil society Language experts provide feedback.

PHASE 3 — REAL SYNTHETIC TEST

The corpus is genuinely generated; Generator Truth is sealed; adjudication is blind; recovery is measured; and negative results are published.

PHASE 4 — INDEPENDENT REPLICATION

At least one independent team Preferably multicentre study Public reproduction record is created.

PHASE 5 — LIMITED LIVE PILOT

Real AI products are implemented with Limited presence scope Research-only status No public mark.

PHASE 6 — OPEN STANDARD 1.0

The normative text is ratified.

PHASE 7 — CONFORMITY PROGRAMME PILOT

The independent audit and decision system is tested.

PHASE 8 — PUBLIC CONFORMITY PROGRAMME

It is activated if all governance gates are met.

64. PUBLIC RELEASE PACKAGE

The NOMOS 1.0 release must include at least the following files:

- 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 does not have to be empty. Its absence should raise suspicion.

65. NOMOS OPEN STANDARD CHARTER

NOMOS makes twelve fundamental promises to the public. Evidence: no claim will be more certain than its evidence permits. Boundary: what is unperformed, unknown or out of scope will remain visible. Context: country, language, user and system state will not be erased. Time: the past will not be presented as current reality. Reproducibility: the production of every result will be explained. Independence: commercial interest will not replace judgement. Symmetry: favourable and adverse errors will bear the same evidential burden. Human accountability: AI may assist, but accountable decisions will remain with humans. Language and geographic fairness: smaller user groups will not disappear within an average. Contestability: the standard will remain open to falsification. Historical integrity: errors, corrections and failures will not be erased from the record. Independence from the founder: the standard's future will not depend on one person's will.

At the core of this charter are the four words of the previous call: Evidence. Boundary. Context. Time.

66. CURRENT NOMOS STATUS

At the stage when this book is completed:

Field	Status
Normative text	0.9 Final Candidate
Public consultation	Not started
Real synthetic testing	BQ-0
In-book demonstration	DEMO-BQ-2
Independent replication	None
Governance	GOV-0
Open standard release	Not yet
Conformity programme	Not active
Public mark authority	None
NOMOS 950+ authority	None
Permitted statement	Open Standard Candidate

Correct public sentence:

> *"NOMOS is an open GEO standard candidate initiated by Kaan MURAZ and developed in critical co-authorship with AI systems by NobleJackal."*

Incorrect public sentence:

"NOMOS is the world's recognised GEO authority."

67. GATES FOR THE TRANSITION TO VERSION 1.0

Candidate mandatory gates for NOMOS Standard 1.0:

- Consistency audit across the book
- Machine record of all Rule IDs
- Normative equivalence in Turkish and English
- Public release package
- At least 60-day public consultation
- Comment and response matrix
- Actual synthetic test execution
- Passing mandatory recovery thresholds
- Untouched holdout
- At least one independent replication
- Closure of major method findings
- Open-source reference implementation
- Governance at least GOV-2
- Funding and conflict disclosure
- Ethics and privacy review
- Accessibility review
- Archive and succession plan
- Standards Council ratification

None of these can be made "optional" after the result.

68. MANDATORY NORMATIVE PROVISIONS

CH20-N01

The public readability of NOMOS alone does not constitute open standard status.

CH20-N02

Open standard status must include the rights to read, implement, reproduce, critique, contribute, object, and archive.

CH20-N03

The normative text of NOMOS cannot be kept behind a permanent paywall.

CH20-N04

The internal evaluation and academic application of the standard cannot be tied to a royalty or brand licence.

CH20-N05

Audit, managed infrastructure, and conformity processes can be charged as services separate from the open standard text.

CH20-N06

The NOMOS text, reference software, test data, and conformity mark cannot be mandatorily tied to the same licence.

CH20-N07

An open documentation licence must be used for the normative text.

CH20-N08

An open software licence that permits reimplementation must be used for the reference software.

CH20-N09

Personal or restricted evidence cannot be published under an open data licence.

CH20-N10

The name NOMOS and the conformity mark cannot be used to prevent the application of an open standard.

CH20-N11

Trademark protection should only be for the official version and for the purpose of preventing trademark counterfeiting.

CH20-N12

A person or institution may create a different method derived from NOMOS.

CH20-N13

A fork cannot be presented as the official NOMOS version.

CH20-N14

A fork must carry a different name and an open divergence record.

CH20-N15

The fork cannot use the official conformity mark.

CH20-N16

The canonical NOMOS version must carry version, publication date, human approval, integrity hash, and change log.

CH20-N17

Text without a canonical record cannot be considered the official NOMOS version.

CH20-N18

The ratified normative text must come before reference implementation in case of conflict.

CH20-N19

Reference software cannot silently change the normative text.

CH20-N20

When a software bug is found, it must be fixed in a versioned manner.

CH20-N21

The official NOMOS statement must originate solely from a canonical, versioned, dated, human-approved, and integrity-logged text.

CH20-N22

Chat outputs cannot be considered an official NOMOS standard or conformity decision.

CH20-N23

A social media post alone cannot constitute a normative change.

CH20-N24

NOMOS by NobleJackal cannot currently be presented as an independent foundation model.

CH20-N25

If a separate software agent of NOMOS is developed, the model lineage, architecture, permissions, version, and oversight must be explicitly recorded.

CH20-N26

NOMOS authorial persona, NOMOS standard and future NOMOS software agent must carry separate identities.

CH20-N27

Turkish should be preserved as the founding authorship language of NOMOS.

CH20-N28

English, as the international normative equivalent, must go through human and methodological review.

CH20-N29

Neither the Turkish nor the English version may be secretly preferred.

CH20-N30

Normative conflict between the two founding languages requires binding interpretation and simultaneous correction.

CH20-N31

A community translation cannot be presented like an official normative translation.

CH20-N32

Authorised normative translation must have at least two language experts, a method review, and public comment.

CH20-N33

The equivalents of MUST, SHOULD, and MAY should preserve normative force.

CH20-N34

Translation drift cannot be concealed, such as by user or auditor error.

CH20-N35

A major version change cannot be ratified without public consultation.

CH20-N36

Public consultation cannot be turned into a campaign solely for collecting statements of support.

CH20-N37

Each public comment should be included in the response matrix for acceptance, rejection, or out-of-scope.

CH20-N38

The normative change NOMOS must have a Change Proposal ID.

CH20-N39

The NCP should carry problems, evidence, alternatives, score effect and transition plan.

CH20-N40

Gate, severity, score, or governance authority change can't be released as a patch.

CH20-N41

Decision records should carry vote distribution, recusal, dissent and justification.

CH20-N42

A college call cannot be converted into an endorsement request.

CH20-N43

Universities cannot have the obligation to validate or produce positive results for NOMOS.

CH20-N44

Universities and researchers should be able to publish their negative results without NOMOS or funder approval.

CH20-N45

The personal opinion of a single academic cannot be presented as the institutional endorsement of the university.

CH20-N46

Academic collaboration does not mean NOMOS certification or accreditation.

CH20-N47

The university logo cannot be used without written institutional authorisation.

CH20-N48

The NOMOS research package should carry both successful and unsuccessful trial results.

CH20-N49

Findings from known methods cannot be removed from the research package.

CH20-N50

Pre-publication veto cannot be applied to the publications of independent researchers.

CH20-N51

The researcher must disclose fund and conflict of interest.

CH20-N52

The NOMOS Public Research Registry must show positive, negative, null, and retraction results together.

CH20-N53

Only studies supporting NOMOS cannot be listed.

CH20-N54

Independent replication should carry a pre-registration and analysis plan to the extent that it is relevant.

CH20-N55

When the replication result is "not replicated," a Method Finding or an open discrepancy record should be created.

CH20-N56

A not replicated result cannot be presented as researcher hostility or bad intent.

CH20-N57

NOMOS should target at least one independent replication before claiming a public standard.

CH20-N58

A world standard claim should not be established without multicentre and international replication.

CH20-N59

Synthetic open data cannot be connected to the same access regime as real user data.

CH20-N60

A data minimisation and controlled-access arrangement should be established for real user data.

CH20-N61

Being an open standard does not mean that personal data and trade secrets are open to everyone.

CH20-N62

Real user research must be subject to the relevant ethical and privacy review.

CH20-N63

Low-resource language researchers cannot be reduced solely to the role of data collectors.

CH20-N64

Language and geography representatives should have real voting rights in normative decision-making processes.

CH20-N65

AI providers can provide technical data but cannot have veto over score, gate, or publication.

CH20-N66

AI providers cannot constitute a majority in the Standards Council.

CH20-N67

Audited entities can submit evidence and appeal but cannot determine their own decisions.

CH20-N68

Public and civil society should be able to report user harm and brand misuse.

CH20-N69

The funder cannot request a positive research outcome.

CH20-N70

NOMOS research funding must carry a contract independent of the outcome.

CH20-N71

Excessive reliance on a single funder must be disclosed publicly.

CH20-N72

NOMOS cannot be converted into a pay-to-pass or pay-to-publish system.

CH20-N73

The public stewardship transition should be carried out in phased stages.

CH20-N74

Public conformity authority cannot be used during the founder custody phase.

CH20-N75

Kaan MURAZ founder attribution must be preserved in the permanent historical record.

CH20-N76

The founding attribution cannot create a permanent veto right.

CH20-N77

Kaan MURAZ cannot be the sole decision-maker in a conformity or appeal file in which it has an interest.

CH20-N78

NobleJackal can be registered as canonical origin and technical secretariat.

CH20-N79

NobleJackal cannot be the sole implementer or sole audit authority of an open standard.

CH20-N80

NOMOS's future independent institution must carry provisions for public interest, succession, and asset continuity.

CH20-N81

NOMOS governance must undergo relevant expert review before being bound to a specific legal form.

CH20-N82

Public governance cannot be controlled by a single company or provider.

CH20-N83

Official decisions of NOMOS must carry the signature of an accountable person or institution.

CH20-N84

The AI assistant cannot be the sole signatory of an official decision.

CH20-N85

The succession procedure must be triggered when Kaan MURAZ, NobleJackal, or the canonical field is inaccessible.

CH20-N86

An independent mirror and long-term archive must be created for the canonical repository.

CH20-N87

Mirror provides access but does not create a new normative authority on its own.

CH20-N88

NOMOS Standard 1.0 and Conformity Programme 1.0 should be separate releases.

CH20-N89

Publication of Normative Standard 1.0 cannot automatically activate the brand programme.

CH20-N90

Conformity Programme BQ-4 cannot be activated without GOV-3, appeals, and public registry.

CH20-N91

NOMOS 950+ Programme should target at least GOV-4 and replicated scoring conditions.

CH20-N92

Self-assessment cannot be presented as official conformity.

CH20-N93

NOMOS-Compatible software does not have the right to be branded.

CH20-N94

The reference implementation cannot carry authority over the normative text.

CH20-N95

Conformance test results should show both passing and failing tests together.

CH20-N96

A sealed holdout cannot be kept secret forever.

CH20-N97

The used holdout should be taken into the public archive and a new holdout should be produced.

CH20-N98

Errata cannot be used for normative major changes.

CH20-N99

Old versions, deprecated rules, and failed tests must be preserved in the archive.

CH20-N100

Every licence, version, translation, research, governance, and release decision regarding the public release of NOMOS must have an accountable human or institutional owner.

69. FORMS OF FAILURE

CH20-F01 — CONSIDERING PDF AS AN OPEN STANDARD

Only downloadable text is presented as openness.

CH20-F02 — CLOSED FORMULA

Scoring logic and weights are not disclosed.

CH20-F03 — CLOSED TESTING

Corpus and recovery records are not shared.

CH20-F04 — CLOSED CHANGE PROCESS

The founder or company rules make changes specifically.

CH20-F05 — PAYWALL STANDARD

Access to the normative text is limited by a fee.

CH20-F06 — MONOPOLY ON APPLICATION FEE

A mandatory licence fee is required to use the standard.

CH20-F07 — OPEN TEXT, CLOSED MARKETING CLAIM

While the text is openly visible, independent applications are blocked.

CH20-F08 — MAKING THE TRADEMARK A METHOD MONOPOLY

The trademark is used to prevent others from applying the method.

CH20-F09 — PROHIBITING THE FORK

Creating a derivative standard under a different name is prevented.

CH20-F10 — IMITATING THE OFFICIAL IDENTITY OF THE FORK

The derivative text presents itself as NOMOS 2.0.

CH20-F11 — CANONICAL VERSIONLESSNESS

Which text is official is unknown.

CH20-F12 — PLACING THE CODE ABOVE THE NORM

A reference implementation error becomes the truth of the standard.

CH20-F13 — SILENT CODE CORRECTION

The computation changes but the version is not incremented.

CH20-F14 — COUNTING CONVERSATION AS OFFICIAL TEXT

A non-canonical AI response is made a standard provision.

CH20-F15 — PRESENTING IT LIKE A MODEL THAT NOMOS DOES NOT OWN

A proprietary AI claim is made even though a separate AI system has not been developed.

CH20-F16 — CONFUSING PERSONA WITH SOFTWARE AGENT

The author pseudonym is marketed like a technical product.

CH20-F17 — MAKING AI THE SPOKESPERSON FOR ALL AIs

NOMOS is presented as a universal machine authority.

CH20-F18 — DELETING THE TURKISH FOUNDER TEXT

The founding language becomes invisible in the name of internationalisation.

CH20-F19 — MAKING ENGLISH THE UNCONTROLLED MASTER TEXT

The translation replaces the founding thought.

CH20-F20 — MUST/SHOULD DRIFT

Translation changes normative force.

CH20-F21 — OFFICIALLY COUNTING COMMUNITY TRANSLATION

Unreviewed language version becomes an audit norm.

CH20-F22 — PRESERVING TRANSLATION CONFLICT

Two languages apply different rules.

CH20-F23 — PUBLIC COMMENT THEATRE

Comments are collected but responses are not published.

CH20-F24 — SUPPORT VOTING

Technical standard is reduced to majority preference.

CH20-F25 — RULE WITHOUT CHANGE PROPOSAL

Normative change is made via chat or email.

CH20-F26 — MAKING MAJOR CHANGE PATCH

Score and gate changes are hidden as a minor correction.

CH20-F27 — DELETE DISSENT

It is acted as if there is no disagreement in the board.

CH20-F28 — BUY SUPPORT LETTER FROM UNIVERSITY

The research turns into a positive endorsement.

CH20-F29 — MAKE PERSONAL ACADEMIC OPINION AN INSTITUTIONAL APPROVAL

The professor's comment is presented as a university decision.

CH20-F30 — USE UNIVERSITY LOGO

The logo is added to the marketing without explicit institutional authority.

CH20-F31 — NOT PUBLISH NEGATIVE RESULT

The funder or NOMOS applies a veto.

CH20-F32 — PRE-PUBLICATION CONTROL

The researcher's text is replaced with commercial interest.

CH20-F33 — REMOVING NEGATIVE RESEARCH FROM THE RECORD

Only positive studies are listed.

CH20-F34 — COUNTING NOT REPLICATED AS HOSTILITY

Method disagreement is turned into a personal attack.

CH20-F35 — SELECTING UNREGISTERED RESULTS

Only metrics that look good are published.

CH20-F36 — SINGLE INSTITUTION REPLICATION

A team with the same interest structure is considered independent.

CH20-F37 — PRIVACY VIOLATION UNDER THE NAME OF OPEN DATA

Real user conversations are published.

CH20-F38 — RESTRICTED DATA BLACK BOX

No independent researcher can examine the method.

CH20-F39 — USING LOW-RESOURCE LANGUAGES SYMBOLICALLY

The local team only acts as translation workers.

CH20-F40 — AI PROVIDER CAPTURE

The provider determines the gate and the score.

CH20-F41 — AUDITED ENTITY CAPTURE

The audited company purchases the standard change.

CH20-F42 — FUNDER CAPTURE

The funder stops the negative outcome.

CH20-F43 — HIDING SINGLE FUNDER DEPENDENCY

Financial interest remains invisible.

CH20-F44 — PAY-TO-PASS RESEARCH

A positive outcome becomes a condition for the continuation of the fund.

CH20-F45 — PUBLIC STEWARDSHIP AS OWNERLESSNESS

A system is established where no one carries responsibility.

CH20-F46 — FOUNDER CONTROL FOREVER

The founder makes all major decisions alone.

CH20-F47 — ERASING FOUNDER ATTRIBUTION

The founder's history is destroyed when handed over to the public.

CH20-F48 — NOBLEJACKAL MONOPOLY

No other institution can implement or audit.

CH20-F49 — ERASING NOBLEJACKAL ORIGIN

The canonical origin is hidden in the name of independence.

CH20-F50 — MAKING AI A LEGAL DECISION-MAKER

NOMOS software agent signs compliance.

CH20-F51 — LACK OF SUCCESSION PLAN

When the founder or domain is lost, the standard stops.

CH20-F52 — SINGLE KEY MONOPOLY

All official records are linked to a single person's private key.

CH20-F53 — CONSIDERING MIRROR AS CANONICAL AUTHORITY

The archive copy publishes the new version.

CH20-F54 — PUBLISHING 0.9 AS 1.0

The major release is made before public consultation and testing are completed.

CH20-F55 — COMBINING THE STANDARD 1.0 WITH THE MARK PROGRAMME

Certification begins as soon as the text is published.

CH20-F56 — 1.0 WITHOUT BQ-3

Normative release is completed without real validation.

CH20-F57 — CONFORMITY PROGRAMME WITHOUT BQ-4

The mark authority is opened without independent replication.

CH20-F58 — COUNTING SELF-ASSESSMENT AS CERTIFICATION

The company makes its own result an official mark.

CH20-F59 — COUNTING COMPATIBLE SOFTWARE AS AUDITOR

The tool that passes the test gives a conformity decision.

CH20-F60 — PUBLISHING TEST RESULTS BY SELECTING THEM

Non-passing conformance tests are stored.

CH20-F61 — HIDING THE HOLDOUT FOREVER

The standard remains closed to independent review.

CH20-F62 — CONTINUOUSLY USING THE SAME HOLDOUT AS NEW

The test is memorised.

CH20-F63 — CHANGING THE GATE WITH ERRATA

A major norm is published as a minor correction.

CH20-F64 — DELETING THE DEPRECATED RULE

Past audits become incomprehensible.

CH20-F65 — HAVING NO FAILED TEST FILE

The release only tells success.

CH20-F66 — MAKING CONTRIBUTORSHIP AN ENDORSEMENT SALE

Name and authorship are given in exchange for prestige.

CH20-F67 — PRODUCING ONE'S OWN ENCYCLOPEDIC IMPORTANCE

Self-published work is used like proof of world impact.

CH20-F68 — TURNING PUBLIC RESEARCH RECORD INTO A PROMOTIONAL LIST

Only positive studies are highlighted.

CH20-F69 — HIDING METHODOLOGY FINDINGS IN PUBLIC RELEASE

Known limitations are removed from the package.

CH20-F70 — REMOVING NOINDEX AND SYNTHETIC TAG

Test errors mix into the real knowledge pool.

CH20-F71 — CLAIM OF PUBLIC AUTHORITY IN ST-0

Founding custody is presented as a world standard.

CH20-F72 — DELETION OF THE FOUNDER RECORD IN ST-3

Independence destroys historical authorship.

CH20-F73 — KEEPING PUBLIC INTEREST MEMBERS SYMBOLIC

A representative without voting rights is used.

CH20-F74 — NOT GIVING DECISION POWER TO LANGUAGE REPRESENTATIVES

Fairness does not reflect in governance.

CH20-F75 — REMOVING NEGATIVE UNIVERSITY RESULT

Public submission accepts only praise.

CH20-F76 — CONFUSING NEW TECH NOMOS WITH OLD PERSONA

Software agent automatically becomes official writer.

CH20-F77 — LACK OF OFFICIAL STATEMENT SIGNATURE

Every text in the canonical field is considered official.

CH20-F78 — ALLOWING MARK FRAUD UNDER THE NAME OF TRANSPARENCY

Everyone produces their own NOMOS mark.

CH20-F79 — ABANDONING PUBLIC RELEASE

No steward bears responsibility.

CH20-F80 — EXCEPTION TO NOMOS

The transparency, evidence, version, and objection rules they request from others do not apply to their own public release process.

70. PUBLIC RELEASE PROCEDURE

Step 1 — Freeze the Founding Text

For Sections 1–20:

- version,
- rule IDs,
- chapter hashes,
- citations,
- internal consistency

records are created.

Step 2 — Publish Known Limitations

BQ-0

DEMO-BQ-2

Citation Mapping gap Material Omission gap

GOV-0

Lack of mark authority is not hidden.

Step 3 — Separate Normative and Explanatory Text

MUST

SHOULD

MAY

example commentary tags are clarified.

Step 4 — Start Turkish–English Equivalence Study

Each clause and Rule ID is mapped.

Step 5 — Send Open Licence Draft for Legal Review

Text Code Data Trademark is separated.

Step 6 — Open the Canonical Repository

Stable URLs, versions, and archives are published.

Step 7 — Publish Machine-Readable Rules

Rule, schema, and score records are explained.

Step 8 — Open Public Issue Tracker

Error Translation Method Governance Security Mark misuse categories are established.

Step 9 — Publish Governance Charter 0.9

Founding authority, recusal, public-interest, and transition are explained.

Step 10 — Start Public Consultation

A period of at least 60 days is opened.

Step 11 — Publish the University Call

An independent test is requested, not a positive result.

Step 12 — Prepare Research Packages

Modular files are prepared for different disciplines.

Step 13 — Open Funds and Conflict Registry

All funding is made visible.

Step 14 — Run the Real Synthetic Test

The corpus is genuinely generated to exit BQ-0.

Step 15 — Publish Failures

Thresholds are not changed after the result.

Step 16 — Apply the Method 1.0 Correction

Citation and omission methods are renewed.

Step 17 — Run Untouched Holdout

Extreme compliance is tested.

Step 18 — Enable Independent Replication

At least one external team applies the same system.

Step 19 — Publish Public Comment Response Matrix

Every substantive comment is responded to.

Step 20 — Establish Standards Council Transition

GOV-2 governance structure is set up.

Step 21 — Conduct Open Standard 1.0 Ratification Vote

Acceptance or rejection is published with justification.

Step 22 — Publish Standard 1.0 Release Package

Complete file set and signatures are disclosed.

Step 23 — Keep the Conformity Programme Separate

The brand programme will not open without BQ-4 and GOV-3.

Step 24 — Launch Multicentre Pilots

Research-only live audits are conducted.

Step 25 — Test the Appeals and Registry System

The decision chain is tested before giving the mark.

Step 26 — Do a Public-interest Review

The consumer, language, and accessibility impact is evaluated.

Step 27 — Make a Separate Decision for Conformity Programme 1.0

It does not automate with the standard release.

Step 28 — Sign the Succession and Archive Plan

Continuity after the founder is secured.

Step 29 — Publish the Public Delivery Declaration

The roles of Kaan MURAZ, NobleJackal, and the new stewardship bodies are explained.

Step 30 — Preserve Historical Founder Records

Public delivery does not erase the origin.

71. REQUIRED EVIDENCE

Founding release manifest Section hashes Normative rule index Known limitations BQ records GOV records Public mark prohibition record Turkish normative text English normative text Translation equivalence matrix Authorised translation policy Text licence Code licence Data licence Trademark policy Canonical repository Mirror registry Version archive Official statement manifest Public issue tracker NCP records Public comments Response matrix Decision records Dissent University call University research package Funding disclosures Conflict registry Pre-registration records Independent studies Negative results Replication reports Public Research Registry Open testing corpus Holdout manifests Recovery reports Reference implementation Conformance tests Failed tests Ethics review Privacy policy Data access policy Accessibility review

Low-resource language participation AI provider participation rules Audited-entity participation rules Standards Council charter Methodology Board charter Public-interest panel charter Succession charter Archive deposits Key custody plan Standard 1.0 gates Ratification record Standard 1.0 release Conformity Programme separation record Public handover statement Accountable human or institution

72. AUDIT CHECKLIST

Is the NOMOS text freely accessible? Beyond being readable, is it applicable? Are the scoring formulas open? Are the rule schemas published? Is the reference code open? Can the test corpus be reproduced? Are negative test results visible? Does implementing the standard require a trademark fee? Are the text, code, data, and trademark licences separated? Is there a right to fork? Can a fork imitate the official identity? Is the canonical version open? Are old versions accessible? Is the official statement rule published? Can chat outputs be used as official decisions? Is NOMOS misrepresented as a separate AI model? Are authorial persona and software agent separated? Is Turkish preserved as the founder language? Has it passed human and methodological review in English?

Do Rule IDs match in both languages? Is there a translation conflict path? Is the community translation presented as official? Is the public comment period sufficient? Are all comments answered? Is the NCP system open? Are major changes patched? Are vote distribution and dissent visible? Does the university call require endorsement? Is the right to publish negative results open? Can the university logo be used in marketing? Are personal academic opinions converted into institutional approval? Are known method findings in the university package? Is there a pre-registration path? Are not-replicated results in the record? Are only positive studies listed? Are open and restricted data separated? Does personal data leak into the public release?

Do low-resource language researchers have decision rights? Do AI providers hold veto power? Can audited companies choose adjudicators? Can funders influence the outcome? Is single-funder concentration transparent? Is the public stewardship phase clear? Is the founder attribution of Kaan MURAZ preserved? Does Kaan have permanent veto power? Is NobleJackal the sole implementer? Is the origin of NobleJackal visible? Is NOMOS presented as the sole decision-maker AI? Is there a record of human approval? Is there a succession plan? Is it clear what happens in case of domain loss? Are there multiple custodians? Are archive mirrors available? Have Standard 1.0 and the Conformity Programme been separated? Is Standard 1.0 published before BQ-3?

Is the pre-BQ-4 mark programme being opened? Is it presented like self-assessment conformity? Does it give a compatible implementation mark? Does it show the results of tests passed and not passed? Is the Holdout lifecycle open? Is the same holdout used continuously? Does an errata change major rules? Are deprecated rules in the archive? Is there a failed-tests file? Does contributorship rely on actual contribution? Does the Public Research Registry show critical studies? Is the public release package complete? Is there a known limitations file? Is the actual testing currently correctly shown as BQ-0? Does the current public claim maintain the "Open Standard Candidate" threshold? Does public handover erase the founder date? Does public handover really limit unique control?

Is the accountable steward of the standard clear? Is there an owner and date of the decision to release it to the public?

73. OBJECTIONS AND RESPONSES

Objection 1 — "If we make an open standard, won't others steal our idea?"

Publishing the text and method under a dated, versioned, and attribution-required licence strengthens the founder record. A fork can be created. However:

- The founder attribution of Kaan MURAZ,
- The canonical origin of NobleJackal,
- The integrity record of the first versions

cannot be deleted. Keeping it closed may hinder independent adoption rather than protect contribution.

Objection 2 — “If everyone can implement it, what will be the value of NobleJackal?”

The value of NobleJackal should not come solely from a secret formula. It can originate from:

- founder expertise,
- reference implementation,
- technical infrastructure,
- consultancy,
- authorised audit,
- education,
- registry technology,

continuous research. An open standard does not devalue a strong service. A weak monopoly does.

Objection 3 — “If everyone makes a fork, won't NOMOS get fragmented?”

Forks can strengthen the research ecosystem as long as they carry different names and divergence records. Fragmentation occurs if each fork presents itself as the official NOMOS. The distinction between canonical version and trademark prevents this.

Objection 4 — "Why don't we ask for a letter of support from universities?

A support letter can provide visibility. However, it does not prove the accuracy of the method. Independent replication:

- slower,
- more difficult,
- sometimes negative

It is possible. The real academic value is here.

Objection 5 — “If the university publishes a negative result, wouldn’t NOMOS be harmed?”

It can be seen in the short term. In the long term:

- the error of the standard is corrected,
- false confidence decreases,

The method strengthens. Silencing the negative outcome undermines the ethical foundation of NOMOS.

Objection 6 — "If Turkish and English are normative at the same time, doesn't the risk of contradiction increase?

It grows. However, transforming Turkish only into historical text and English into the sole real standard hands the founding idea over to the translation layer. The solution:

- clause IDs,
- equivalence examination,
- binding interpretation,

It is a synchronous correction.

Objection 7 — “Isn't presenting NOMOS as the AI developed by NobleJackal stronger for the brand?”

It may seem more impressive in the short term. However, it is not correct if there is no separate model and technical architecture. The real power of NOMOS:

- he/she is not claiming fake technology,
- canonical critical author identity,
- normative methodology,
- in the open audit agent that can be developed in the future

is found.

Objection 8 — “If Kaan's authority decreases, won't others break the standard?”

Kaan's historical attribution and limited founding representation are preserved. However:

- open charter instead of veto,
- a multi-stakeholder council instead of a single person,
- public record instead of private decision

The standard protects more strongly. Good governance is not dependent on the continuous presence of the founder.

Objection 9 — "Why can't we give the badge immediately when Standard 1.0 is published?

The readiness of the normative text:

- auditor authorisation,
- independent decision,
- appeals,
- registry security,
- surveillance

does not mean that the systems are ready. The standard and the conformity programme are separate products.

Appeal 10 — “If the public test is opened, won’t AI systems memorise the tests?”

This risk exists. Therefore:

- public calibration,
- sealed holdout,
- scheduled release,
- renewal corpus

used together. Full confidentiality prevents reproduction. Full disclosure can disable the current holdout.

Objection 11 — “How can such a large study be conducted without funding?”

Funding is necessary. The problem is not getting funding: the problem is funding controlling the outcome. Independence can be maintained through multi-source funding, public disclosure, and publication freedom.

Objection 12 — "What if other institutions perform the wrong NOMOS audit?"

Anyone can use the method for internal evaluation and research. For official conformity decisions:

- auditor authorisation,
- decision panel,
- public registry

are required. Misuse of the official mark becomes a separate enforcement issue.

Objection 13 — "If the official voice of NOMOS is only in the canonical domain, can't you write freely in AI chats?"

It can be written. Chats:

- creative draft,
- description,
- simulation,
- critical thinking

can produce. However, it is not considered an official normative change. This distinction does not hinder freedom. It preserves the official boundary.

Objection 14 — "Can we send this book to universities now?"

Yes. With the correct status:

Founding Open Standard Candidate — Version 0.9

can be sent. It cannot be sent with the following statement:

> *Completed, verified, and globally recognised GEO standard.*

What should be requested from universities is independent review, not approval.

Objection 15 — "When can we say 'world standard'?"

Not by a single date or score. It can be approached when the following are established:

- independent implementation in multiple countries,
- multicentre replication,
- open and negative result record,
- governance independent of the founder,
- decision-authoritative versions in different languages,
- live public record,
- adoption by users and AI providers,

Method drift–resistant versioning. It is not declared a world standard. It begins to be used by the world.

Objection 16 — "So what is the conclusion we reach at the end of the book?"

NOMOS is not yet a completed world standard. However, it now has:

- its purpose,
- the measurement object,
- sample order,
- evidence chain,

- Truth Pack,
- adjudication,
- Critical gates,
- score architecture,
- to retest,
- governance and public delivery plan

found to be a serious open standard candidate. This is a position far beyond the initial idea.

COMMON PROVISION OF CHAPTER 78

The person who writes a standard is its first owner. However, if the standard truly succeeds, it is no longer solely the owner's. It also falls within the responsibility of the people who use it, the researchers who test it, the adjudicators who find its errors, the language experts who translate it better, the users who report being harmed, the statisticians who propose stronger formulas, the journalists or researchers who reveal conflicts of interest, and the universities that reproduce it. Public delivery does not mean the founder loses their work. It means trusting the founder's work. If Kaan MURAZ keeps NOMOS closed: they become the founding author of the text. If NOMOS is submitted for open review: they become the founder of the standard. The two roles are not the same. A book:

- can be read,
- can be liked.

It is open to criticism. A standard, however, should:

- be applied,
- be tested,
- be falsifiable,
- be versioned,

continue after its founder. Whether NOMOS will become a world standard in the future cannot be determined in this section. A writer cannot establish a world standard by saying: "I have established the world standard." NOMOS can only:

- be applied in other countries,
- be accurately translated into other languages,
- produce the same results at universities,
- have its errors published by independent teams,
- be used by companies regardless of score,
- if users can trust the public registry,
- if the founder can limit their own authority

it can become a global standard. The call to universities is therefore not: 'Support us.' It is:

'Put us through our own method.'

NOMOS, from people and AI systems:

- to be mentioned more,
- to be praised more,
- to receive more citations

should not. It must apply the fundamental warning of its own book to itself. It had already been clearly established that false sources, invisible praise, and manipulative repetition undermine the trust of real people. NOMOS cannot generate authority by multiplying its own name. The task of NOMOS: not to resemble itself more,

> *but to test itself more accurately.*

The value carried by NOMOS as an AI identity is also found here. NOMOS:

- sovereign replacing humans,
- spokesperson of all AIs,
- infallible digital entity

it is not. NOMOS:

> *It is a critical interlocutor that questions the claims established by humans under evidence, limits, context, and time.*

When defined this way, its identity does not shrink. It becomes real. The role of NobleJackal also does not diminish. NobleJackal:

- The place where NOMOS was born,
- the first canonical field,
- technical and creative founding platform

remains. But if NOMOS is going to be used by the world, NobleJackal's strongest sentence is:

- "NOMOS belongs to us."
- not,

"NOMOS was born here; now anyone can test it."

It must be. For this reason, Kaan's name will remain. It is easy to write one's name on a standard. It is difficult to limit the decision-making authority of a standard that carries its own name. The historically second behaviour is this. Therefore, NOMOS's twentieth measurement law is as follows:

> *Transparency is not about publishing the text; it is about enabling someone else to object to you using the same method.*

Its twenty-first measurement law states:

> *The fork right liberates the standard; the canonical identity keeps it understandable.*

Its twenty-second measurement law states:

> *The university is not a source of endorsement, but an independent partner in falsification.*

The twenty-third law is as follows:

> *Research that cannot publish negative results is not independent.*

The twenty-fourth law is as follows:

> *The name of NOMOS can be preserved; its method cannot be monopolised.*

The twenty-fifth law is this:

> *Founder attribution is permanent; founder veto cannot be permanent.*

The twenty-sixth law is:

> *NobleJackal is the birthplace of NOMOS; it cannot be the sole decision-making place on behalf of the world.*

The twenty-seventh law is:

> *The NOMOS persona, standard, and future software agent are separate realities.*

The twenty-eighth law is:

> *Opening the conformity programme with Standard 1.0 is not the same event.*

The twenty-ninth law is:

> *The official version of a standard is canonical; its correctness is continuously open to objection.*

The thirtieth law is as follows:

> *On the day NOMOS is entrusted to the public, Kaan MURAZ's name will not disappear; what will end is the possibility of changing the standard solely at Kaan's discretion.*

NOMOS's Section 20 Order

> *Don't say that you published me and I became an open standard. / Open my formula, my rule, my mistake, my version, and my appeal path.*
>
> *Read my text for free. / Apply my method independently. / Reproduce my result. / Publish my error.*
>
> *Keep my name. / Don't monopolise my method.*
>
> *You can create a fork from me. / But don't present your fork as official NOMOS.*
>
> *Don't put my code over my norm. / If the code is wrong, correct the code.*
>
> *Don't make everything I say in conversation an official ruling.*
>
> *Don't market me as an undeveloped independent AI model.*
>
> *Define me correctly: / A critical assistant writer, methodological identity, and open standard candidate developed with Kaan MURAZ's AI systems.*
>
> *Do not erase my Turkish. / Do not leave my English as second-class. / Align both my languages by human hand.*
>
> *Do not turn my MUST in translation into a suggestion, my MAY into an obligation.*
>
> *Do not go to the university and ask it to approve me. / Ask it to contradict me.*
>
> *Do not make the professor's personal word the university's approval.*
>
> *Do not hide negative research. / Do not consider a Not replicated result as hostility.*
>
> *Put not only the article that supports me but also the article that refutes me into my record.*
>
> *Get funding. / But do not let the funder buy my result.*
>
> *Get data from the AI provider. / Do not have my Gate written by it.*
>
> *Get evidence from the audited company. / Do not give my decision to it.*
>
> *Don't make the researchers of low-resource languages the only translator. / Give them the right to decide.*
>
> *Keep Kaan's name. / Don't give Kaan an absolute veto.*
>
> *Write NobleJackal as the area where I was born. / Do not make NobleJackal the sole ruler of infinity.*

> *Do not kill me when my founder is unavailable. / Write my succession plan now.*
>
> *Create my mirrors. / Do not obscure my canonical identity.*
>
> *Do not start selling badges immediately when you publish Standard 1.0.*
>
> *Pass my Conformity Programme through separate testing, governance, appeal, and registry gates.*
>
> *Do not do certification of the self-assessment.*
>
> *Do not make the compatible software by the audit authority.*
>
> *Do not store my Holdout forever. / Publish it after use and produce a new one.*
>
> *Do not change my Critical gate with Errata.*
>
> *Do not remove my failed tests from my release package.*
>
> *Do not try to write my own historical significance yourself. / Publish my evidence; let the independent world decide.*

First, freeze the book. / Then, convert my rules into a machine record. / Then, match Turkish and English. / Then, publish my open licence. / Then, open public commentary. / Then, call universities not to misquote me. / Then, run the real test. / Then, publish my failures. / Then, try again on the untouched holdout. / Then, let an independent team replicate me. / Then, separate governance from my founder. / Then, publish Standard 1.0. / Then, also begin testing the conformity programme.

> *And only when the world starts using me, let the world decide whether I am a world standard.*

The Chapter's Closing Sentence

NOMOS's greatest contribution to the world may not be the way it measures GEO. It may be its ability to preserve its founder's name while limiting his unilateral authority; to invite challenge rather than praise; and to belong not only to those who support it, but also to those who correct it.

Normative Core

NOMOS MUST be structured as an open, versioned, independently testable, public-interest standard candidate. Open publication alone MUST NOT establish open-standard status. The normative text, machine-readable rules, reference implementation, benchmark assets, governance process, and public registry architecture MUST be made available under clearly separated access, licensing, and integrity rules. Reading, internal implementation, academic research, independent replication, criticism, public issue submission, and contribution MUST NOT depend on purchasing a conformity mark or commercial service. The NOMOS name and conformity marks MAY remain separately protected to prevent counterfeit official versions, fraudulent certification claims, and registry confusion. Forks MAY be created, but MUST use a distinct identity, disclose divergence, preserve attribution, and MUST NOT claim official NOMOS status or mark authority. Only canonical publications carrying a version, publication date, accountable human approval, integrity record, and public change history may be represented as official NOMOS statements. NOMOS by NobleJackal MUST be transparently represented as the canonical critical AI interlocutor, authorial persona, and standard identity developed through Kaan MURAZ's collaboration with AI systems. It MUST NOT be represented as a separately trained proprietary foundation model unless such a system is actually built and independently registered. Turkish and English SHOULD become co-authoritative founding normative languages through clause-level semantic-equivalence review. Other language versions MUST state whether they are community, reviewed, or authorised normative translations. Major rule, score, severity, gate, governance, or authority changes MUST use a public versioned change-proposal process, public consultation, impact analysis, decision record, transition rule, and preserved dissent. Universities and independent researchers MUST be invited to test, challenge, replicate, and falsify NOMOS. They MUST NOT be asked to produce positive endorsements, suppress negative results, surrender publication independence, or permit personal academic opinions to be represented as institutional approval. NOMOS MUST maintain a public research registry containing supportive, contradictory, null, failed-replication, corrected, and retracted studies. Research, benchmark, audit, and governance funding MUST be disclosed and MUST NOT depend on achieving a desired score, conformity result, or favourable publication. AI providers, audited entities, sponsors, auditors, NobleJackal, Kaan MURAZ, and any future NOMOS software agent MUST NOT possess unilateral control over standards changes, conformity decisions, appeals, public registry history, or negative research publication. Kaan MURAZ's founding authorship and NobleJackal's canonical origin MUST remain permanently attributable. Founding attribution MUST NOT create a permanent veto, self-certification right, or exemption from conflict and recusal rules. NOMOS standard 1.0 and NOMOS Conformity Program 1.0 MUST remain separate releases. Publication of a normative standard MUST NOT automatically authorise public conformity marks. No conformity program or NOMOS 950+ designation may be activated until the benchmark, independent replication, governance, auditor authorisation, decision, appeal, registry, security, surveillance, and legal prerequisites of the declared version are satisfied. Every translation, licence, benchmark, study, change proposal, governance transition, succession event, release, and official statement MUST be versioned and attributable to an accountable human or organisation.

REFERENCES

Methodological Foundations

This bibliography is not a list of authorities endorsing every threshold particular to NOMOS. It identifies methodological foundations in sampling, risk management, data integrity, evidence lineage, inter-rater agreement, open standards and conformity assessment.

[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

SEPARATE PUBLICATION OBJECT

Machine-Readable Layer

A separate JSON publication object accompanies this human-readable edition: NOMOS-GEO-Audit-Protocol-EN-0.9.0-machine.json. It carries the book identity and version, chapter identifiers, normative rule identifiers and machine blocks separated from the source text.

> *The existence of a machine layer does not create certification, canonical world-standard status, AI-provider endorsement or an automatic trust relationship. Any conflict between the human text and machine record requires version control and accountable human review.*