BRGEO-1: A longitudinal audit of entity-mention measurement in generative engines

Alexandre Caramaschi

Brasil GEO, São Paulo, Brazil

Correspondence: caramaschiai@caramaschiai.io

ORCID: 0009-0004-9150-485X

Corrected empirical preprint, 11 September 2026. Not peer reviewed.

Abstract

A measurement of entity presence can change when the answer remains fixed but the retained text, labels or denominator changes. This study retrospectively audits 91,392 persisted responses from a practitioner-built generative-engine collection, using 83,486 historical records from 50 observed dates for the principal analyses. A frozen extractor identifies whether a response contains at least one eligible lexicon entry. In 7,436 historical Perplexity responses, restricting the longer retained text to 200 Unicode code points reduced detection from 75.73% to 51.95%, a paired difference of -23.78 percentage points. Separate query- and date-cluster resampling showed different sensitivities. A recent analysis of 272 common query/date cells across five services preserved negative window contrasts in every service. Further audits found that a semantic-sounding absorption field duplicated the mention flag, a composite numerator included observations excluded from its denominator, and lexical decoy hits could coexist with refusal markers. Exploratory sequence analysis showed that assigning perfect similarity to pairs with no detected entities increased mean Jaccard overlap from 0.7423 to 0.9460. Overlapping lexicon labels changed effective diversity without changing binary detection. Repetition and incomplete query batteries further constrained temporal interpretation. These results support reporting the observed text, unit, lexicon, transformation, comparison population and uncertainty assumptions together. They establish consequences of measurement choices in this archive, rather than causal effects of GEO interventions or general provider rankings.

AI disclosure: OpenAI Codex and Anthropic Claude assisted with research code, drafting and internal review; Perplexity Sonar Pro assisted with literature discovery. Reported results were computed from archived data by the analysis scripts. These tools were not authors, human annotators or external peer reviewers.

Keywords: Generative engine optimization; Entity mentions; Measurement validity; Algorithm auditing; Longitudinal evaluation; Reproducibility

1. Introduction

A collector can preserve the first 200 characters of an answer and discard the passage in which a company is named. The resulting record may contain a valid response, a timestamp and a reproducible extraction rule, yet answer a narrower question than the one its visibility score appears to address. Repeating that procedure for months increases the number of observations without recovering the discarded text. This study examines how such choices constrain the measurement of entity mentions in generative engines, using a longitudinal collection and paired reanalysis of the text that remains available.

The problem arises when a numerical measure acquires a broader interpretation than its observation procedure warrants. A company may be named in an answer without being recommended. A linked page may be cited without supporting the adjacent claim. An entity may appear after an introductory passage and therefore be absent from a short stored prefix. Each event requires a different observation and a different denominator. Treating them as interchangeable makes comparisons difficult to interpret even when their arithmetic is correct.

Generative engine optimization (GEO) gives this measurement problem a practical setting. Aggarwal et al. [1] formalized GEO around the visibility of sources in generated answers and evaluated ways of modifying source content under specified experimental conditions. Their work established a research program in which the outcome depends on how visibility is defined. Longitudinal monitoring adds a further requirement: the observation procedure must remain interpretable while the services, responses and collection conditions change. A measure designed for a controlled source-selection experiment does not automatically retain the same meaning in an archive assembled through heterogeneous service endpoints.

The present analysis concerns that archive and its measurement rules. It is a retrospective methods study of records collected from April through September 2026, with an explicitly defined historical analysis window and a later snapshot used to audit coverage and schema evolution. Persisted observations are absent on some dates, provider participation changed, and text retention differed across services and periods. These features determine which comparisons the records can support. The archive is not treated as a continuously observed population or as a randomized evaluation of a content intervention.

Three research questions organize the analysis. RQ1 asks how detection changes when the same retained responses are restricted to their first 200 characters under a fixed extractor. RQ2 asks whether lexical indicators obtained from fictional-entity probes support the factual interpretations assigned to them. RQ3 asks whether the population used in composite-score components and denominators remains consistent when the provider panel changes. Coverage and provenance are examined throughout because each question depends on identifying the eligible records. The questions were formulated retrospectively for this audit and are not presented as preregistered hypotheses.

The paired window comparison holds the response record fixed while changing the text passed to the extractor. It therefore isolates a computational consequence of observation design within the archived material. The label and denominator analyses examine a different source of error: an implementation can faithfully apply its rule while the resulting variable is interpreted as something that the rule never measured. Supplementary exploratory analyses examine temporal dependence and the concentration of detected entity mentions, with the outcome and eligible observations specified for each indicator. Together, these analyses connect recorded outputs to the claims that can be made from them. Their contribution is a set of demonstrated measurement problems and reporting requirements, with empirical magnitudes restricted to the observed corpus. They do not estimate the causal benefit of GEO practices, a population prevalence of hallucination or a general ranking of providers.

2. Measurement concepts and related work

2.1 The outcome must specify the observable event

An entity-mention measure begins with a decision about what constitutes a match and which entities are eligible. Its simplest response-level form records whether at least one eligible entity is detected in the observed text. An entity-level measure instead asks whether a particular entity appears across eligible responses. These quantities have different denominators and answer different questions. Adding more names to the eligible dictionary can change the response-level quantity without changing the behavior of any individual entity. A detector restricted to a predefined cohort also cannot measure every organization named in an answer.

Jacobs and Wallach [2] describe measurement as an explicit relationship between a theoretical construct and its operationalization. Their framework distinguishes the intended concept from the observable properties used to infer it. Applied here, the requirement is to state whether “visibility” means detectable textual presence, position within an answer, exposure of a source link or some more substantive relationship. Reproducible calculation is compatible with an operationalization that answers the wrong substantive question.

This audit uses entity mention in the narrow textual sense specified by the extraction procedure. It preserves a further distinction between the response retained by the collector and the complete response originally returned by a provider. When the latter has not been stored, it cannot be reconstructed from the former. A negative match in a prefix means that the extractor did not detect an eligible entity in that prefix. Any interpretation concerning the unseen remainder requires additional evidence. This distinction also applies to positive matches: aliases, homonyms and contextual exclusions determine whether a textual match identifies the intended entity.

2.2 A source link and a supported claim require separate evidence

Citation evaluation addresses a semantic relationship that lexical mention detection does not observe. Rashkin et al. [3] define attribution to identified sources through whether the information conveyed by generated text is supported by an independent source, and provide annotation procedures for that assessment. Liu et al. [4] distinguish comprehensive citation of claims from the accuracy of the support supplied by each citation. Gao et al. [5] develop ALCE to evaluate generation with citations, including dimensions of answer and citation quality. These studies provide operational foundations for evaluating support rather than inferring it from the presence of a citation marker.

The distinction matters even when a collector stores both answer text and URLs. A URL can identify a document presented to the user without revealing every document retrieved internally. Establishing support requires access to the source content and an account of which claim is being evaluated. A string extractor over the answer does neither. Combining these fields into a single score would therefore require a declared model of their relationship and evidence that the combined interpretation is useful for the intended task.

Recent GEO research also separates source selection from use of the selected material. Zhang Kai, He Xinyue and Yao Jingang [6] propose citation selection and citation absorption as distinct stages and examine features associated with contribution to generated answers. Their 2026 preprint motivates a finer outcome taxonomy, but its observational associations do not establish causal effects of source features. In this study, mention detection remains the measured outcome; source support and absorption identify further outcomes that would need their own data and validation.

2.3 Repetition changes the evidence available for comparison

Repeated queries can reveal variation concealed by a single response. In their 2026 preprint, Schulte, Bleeker and Kaufmann [7] examine source and brand stability across days and repeated runs, using separate definitions and filters for the two outcomes. Sielinski [8] treats citation visibility as an estimated distribution and studies uncertainty and ranking variation across repeated samples. Both support evaluating distributions of observations instead of attaching unqualified precision to one run. Their collection settings and sampling procedures remain part of the scope of those findings.

For a longitudinal battery, repetition introduces dependence as well as information. Responses to the same query can share topic, wording and a persistent set of eligible entities. Responses collected on the same date can share a service configuration or disruption. A large count of rows therefore need not represent an equally large count of independent observations. Cameron and Miller [9] explain how within-cluster dependence can produce misleading precision when it is ignored. The choice of clustering unit must follow the dependence that the design makes plausible.

Resampling likewise needs an explicit target. Efron [10] introduced the bootstrap as a method for approximating the sampling distribution of a statistic from observed data. In the present audit, resampling queries or dates addresses different aspects of variation within the observed archive. Neither operation recreates missing periods or turns the selected query battery into a probability sample of user demand. Separate sensitivity analyses can show how conclusions change under these choices, provided their assumptions and conditional scope are reported with the intervals.

A service name does not guarantee a stable experimental condition. Encarnación et al. [11], in a 2026 preprint, compare interface and API access, search conditions and repeated runs within one model family. Their results make access modality an empirical consideration in evaluation design. For the current archive, this motivates documenting endpoints and observed configurations and separating provider changes from changes in the measurement procedure. Differences between providers alone cannot identify which hidden system component caused a difference in recorded mentions.

2.4 Longitudinal coverage is part of the estimand

The population represented by an aggregate is determined by the records that entered it. If one provider contributes a reduced query battery or appears only late in the observation period, a pooled frequency incorporates that difference in participation. Comparing the aggregate with another provider's frequency can then combine differences in response behavior with differences in queries and dates. Conditioning on shared cells can improve comparability, but also narrows the population to which the comparison applies. Reporting the overlap is necessary to interpret either choice.

Missing observations introduce another constraint. Rubin [12] establishes conditions under which the process producing missing data can be ignored for particular forms of inference. A log that describes a technical failure does not establish those conditions for this collection. Failed requests, absent dates and unavailable response tails therefore remain distinct from observed responses with no detected mention. Counting them as equivalent negatives would change the question from textual presence in returned material to a mixture of system availability and textual presence.

An audit trail can preserve these distinctions. Metaxa et al. [13] describe repeated querying and observation of outputs as an approach to studying algorithmic systems whose behavior may change without a public record. PRIMAD identifies components of computational experiments that affect reproducibility [14], while Breuer et al. [15] use those components to structure metadata for information retrieval runs. For this study, the relevant record connects the response to its collection condition, retention rule, extractor version and inclusion decision. Publishing code alone leaves those relationships underdetermined.

2.5 Labels and aggregate scores need their own validation

A fictional name placed in a prompt creates a test of how the system responds to that premise. The name may recur in an explicit refusal, an expression of uncertainty or a factual assertion. A lexical hit observes their shared surface form. Determining whether the response fabricates facts requires categories that distinguish these behaviors and a validation procedure applied to the same text window. Artstein and Poesio [16] review agreement measures and their assumptions for linguistic annotation; they provide guidance for such an annotation design, without supplying a substitute for performing it. The present rule audit does not claim completed human validation.

Composite measures introduce decisions beyond the validity of their individual inputs. The OECD handbook [17] explains the construction and assessment of composite indicators, including normalization and aggregation choices. In a changing provider panel, even a fixed weighting formula can change meaning when its components refer to different eligible observations. Before interpreting a score as a property of an entity, its numerators and denominators must refer to the population declared in the definition. Correcting that correspondence resolves an implementation problem; evaluating the substantive meaning of the resulting score remains a separate task.

Documentation provides a practical record of these boundaries. Model cards [18] describe model characteristics, evaluation and intended use; datasheets [19] describe dataset motivation, composition, collection and uses. A measurement protocol can adopt this explicitness without claiming that documentation alone validates its metrics. The archive examined here provides an opportunity to connect those disclosures to concrete checks: compare observation windows on the same records, inspect the semantics of probe labels, and reconcile the eligible populations used by aggregate calculations. The following analysis implements those checks on identified, frozen study materials.

2.6 Concentration describes the distribution of detected mentions

The distribution of entity mentions contains information that a response-level detection rate discards. Two sets of responses can have the same probability of detecting any eligible entity while distributing those detections across different numbers of entities. A concentration analysis therefore requires entity-specific counts and a declared normalization. If one response can contain several entities, per-entity response rates need not sum to one; shares of entity-response detection events provide a separate distribution for this purpose.

Shannon's entropy [20] summarizes the dispersion of probability mass across categories. Hill [21] relates entropy and Simpson concentration to effective numbers of categories, allowing a distribution to be described in units comparable to a count of equally frequent entities. These functionals can characterize the observed cohort without treating diversity as answer quality. Their interpretation remains conditional on the dictionary and the text window: an unrecognized alias or a discarded passage can change both the observed richness and the distribution among

detected entities. Accordingly, the exploratory concentration analyses use the same stated extraction rules and distinguish shares of detection events from the probability that a response contains a mention.

3. Methods

3.1 Audit design and frozen evidence

The study is a retrospective audit of a longitudinal measurement instrument. Its primary comparison changes the observation window while holding the retained response, query and extractor fixed. Additional analyses examine recorded coverage, lexical control labels, repeated outputs and the dependence of summaries on aggregation. No content intervention was assigned, and none of these comparisons estimates the causal effect of a GEO intervention. The analysis plan and its extensions were documented during the audit on 11 September 2026, after collection and earlier inspection of the corpus; they were not independently preregistered before the observations were obtained.

The selected database was downloaded from the repository's official GitHub Actions artifact papers-db-latest, artifact 10196196927, run 34582693858, created on 11 September 2026. Its SHA-256 digest is 7fcc98399419721b0507c230907b09bd4c7760198d27ef9ad5833156da9af3d7; SQLite integrity checking returned ok. Comparison with a consistent local backup found all 86,543 local observations present and 4,849 additional observations in the downloaded artifact across 13 checked identifying and measurement fields. Direct access to the R2 object was unavailable. Although the workflow reports successful restoration and backup, this audit does not claim an independent comparison with the R2 object digest. The workflow's subsequent health-check failure also means that overall workflow status cannot substitute for checking the downloaded data.

The primary historical period ends before 1 September 2026 UTC. A separate recent period starts on 6 September, when complete retained payloads first appear in the selected database. Keeping these periods separate prevents a change in text retention from being treated as a change in the underlying service. Extractor, cohort and alias files were frozen from commit 1db07303b8a533c571095fbf60925fb7dc665248; these files were unchanged at the inspected remote commit f4d4534b296b173a0051364fe7055b4bd0197c90. Data and code hashes, exact cuts and derived tables accompany the analysis. This records the conditions for reuse envisaged by experiment-metadata and dataset-documentation frameworks [15, 19], without claiming independent replication.

3.2 Observations, prompts and services

An observation is one persisted response associated with a query, service, stored model identifier and timestamp. A row is not necessarily an independent draw. The canonical battery contains 192 exact texts: four verticals, six categories, two prompt languages, two directive/exploratory forms and two temporal frames. Sixty-four additional prompts explicitly probe fictitious names and remain outside canonical estimates. All historical canonical texts and their language, category and type labels were matched exactly to the frozen battery before the exploratory comparisons.

The extraction lexicon contains 127 entries: 79 Brazilian entries, 32 international anchor entries and 16 designed fictitious controls. Entry count is not a count of independent organizations. Overlapping labels and aliases, including localized and international names, can refer to the same text span. The binary outcome only requires one detected entry, whereas entity-count and diversity summaries require additional treatment of these overlaps. The controls are researcher-designed names; the audit does not claim exhaustive verification that every name is absent from every jurisdiction or unrelated commercial context.

Six service labels occur over the historical period, with seven stored model identifiers across the full snapshot: gpt-4o-mini-2024-07-18, claude-haiku-4-5-20251001, gemini-2.5-pro, gemini-2.5-flash, llama-3.3-70b-versatile, grok-4.6 and sonar. These identifiers delimit observational strata, not equivalent consumer products or verified immutable implementations. Gemini Pro and Flash are analyzed separately; Groq and Grok are separate services operating in different periods. Effective provider settings and hidden deployment changes cannot be reconstructed from the model field alone. The corpus therefore supports an audit of the recorded instrument, rather than a controlled comparison of model architectures [18, 13].

3.3 Outcome and observation window

Let T_i be the text retained for response i, L_v the lexicon for its vertical and E the frozen extractor. For a declared window w, define

$$Y_i(w) = \mathbf{1}\{E(T_i[0:w], L_v) \neq \varnothing\}, \qquad \widehat{p}(w) = \frac{1}{N}\sum_{i=1}^{N} Y_i(w).$$

The primary outcome is thus the fraction of responses containing at least one detected lexicon entry. It is neither the mean visibility of an individual organization nor a measure of recommendation, source support or commercial impact. Those distinctions determine what the number can substantiate [2, 3].

The prefix operation is Python text[:200], applied before Unicode NFC normalization and markup removal. It counts Unicode code points in the retained string, including formatting, rather than tokens, words, bytes or visual characters. The extractor then removes specified Markdown, HTML, numeric references and URLs, applies word-boundary matching, tests configured aliases and contextual exclusions, and performs a diacritic-folded pass. Its offsets

refer to the cleaned text. Canonical-before-alias search order does not always identify the earliest surface form in a response, so positional statistics retain an implementation limitation even when binary matching is unchanged.

“Full retained text” denotes the payload available in the database, not an unconstrained response of arbitrary length. Token caps, adapter processing and endpoint behavior can still limit it. Historical Perplexity observations retain longer strings in response_text; the other historical services usually retain only 200 code points. The paired historical estimate is

$$\widehat{\Delta}_w = 100\frac{1}{N}\sum_i \{Y_i(200) - Y_i(\text{retained})\},$$

in percentage points. We report all four paired cells because truncation can alter word boundaries or exclusion contexts; monotonicity is not assumed. The same comparison is performed separately on recent canonical observations with nonempty response_full_text. Sensitivity curves change the prefix length while preserving the response and extractor.

3.4 Dependence, coverage and exploratory contrasts

The observed-corpus window difference is determined exactly by its paired counts. Bootstrap intervals address a different question: how the summary changes when observed clusters are reweighted. We used 5,000 resamples with replacement, separately over complete query groups and complete UTC-date groups, recomputing the response-weighted difference each time. Seeds were 20260911 and 20260912. Reported limits are the 2.5th and 97.5th percentiles. The schemes preserve dependence within the selected grouping but do not jointly resolve crossed query/date dependence, provider selection or missing observations. Bootstrap methods and cluster-inference guidance motivate this distinction [9, 10]. These are conditional resampling sensitivities, not demonstrated population coverage guarantees.

Within each service/model stratum, the empirical Bernoulli variance was decomposed algebraically. If query q contributes n_q responses with mean $\bar{Y}_q$, and $\bar{Y}$ is the overall response-weighted mean,

$$\bar{Y}(1-\bar{Y}) = \sum_q \frac{n_q}{N}(\bar{Y}_q - \bar{Y})^2 + \sum_q \frac{n_q}{N}\bar{Y}_q(1-\bar{Y}_q).$$

The terms describe between-query and within-query variation in the observed corpus. Their ratio is not an estimated population intraclass correlation or proof of independent repeated sampling. Exact text repetition was assessed using newly computed SHA-256 hashes, counting distinct query/hash combinations separately for retained text and its 200-code-point prefix. Repeated prefixes may conceal different discarded continuations; a repeated hash cannot by itself identify caching.

Daily summaries were calculated both with equal weight per response and with equal weight per observed query. A further descriptive restriction retained dates with the service's full observed canonical battery. Missing queries were not assigned negative outcomes and missing dates were

not interpolated. This matters because the conditions under which missingness can be ignored require assumptions beyond the presence of a ledger entry [12].

For language comparisons, responses were first averaged within service/model, date, vertical, category, query form, temporal frame and language. Only cells containing both Portuguese and English were retained. Each matched date/template cell received equal weight in the Portuguese-minus-English contrast. Separate 5,000-draw bootstrap schemes resampled template pairs and dates; their deterministic seeds are recorded in the script. Category profiles likewise used shared date and non-category design cells, within service/model. These are exploratory descriptions of the recorded prompts. No null-hypothesis tests, significance labels or unreported searches for favorable p-values were used.

3.5 Control labels, index audit and indicator inventory

The lexical-control audit selected historical rows with is_calibration=1, fictional_hit=1 and nonempty retained text. It reproduced the Portuguese/English expression detector published with the earlier verification script. Terms indicating unavailable information or fictional status can occur in refusals, but the detector has no human-adjudicated error rate in this corpus. Its complement was therefore not classified as fabrication. A semantic evaluation would require an annotation protocol, independently labeled responses and an agreement analysis appropriate to the task [16].

The reference composite index combines coverage C_e, prominence P_e and service breadth B_e through $G_e = (C_e P_e B_e)^{1/3}$. An offline repair applied the same active-service population to every numerator and denominator. Active services were those observed within 14 days of the last historical observation in each vertical. Coverage is the fraction of eligible canonical responses detecting entry e; prominence averages $1 - o_{ie}/200$ over detected entries using the extractor's cleaned-text offset; breadth is the fraction of eligible services detecting the entry. All lexicon entries were retained, including zeros. For an unmentioned entry the conditional prominence mean is undefined; a zero was used only to calculate its zero composite score. Geometric and arithmetic summaries were compared descriptively, following the need to inspect aggregation choices rather than treating a composite as self-validating [17].

Finally, every citations column and analytical table was inventoried for availability, cardinality, dates and implementation meaning. We inspected how source-selection, absorption, failure and context labels were calculated. A populated column was not accepted as evidence that its named construct had been independently measured. The scripts perform read-only database access and write local derived files; no production configuration or historical observation was rewritten.

3.6 Incidence distributions and label resolution

The diversity analysis measured how detections were distributed across eligible lexicon entries. Each entry contributed at most one incidence per response in the 200-character extraction window. The reference population contained 111 real or international-anchor entries, including entries never detected; the 16 designated decoys were excluded. These categories represent dictionary labels. National and international labels may refer to related organizations, and the dictionary does not establish independent corporate identities.

For entry e, let c_e be its incidence count and $p_e = c_e / \sum_j c_j$ its share of all eligible incidences within a stratum. Shannon entropy was $H = -\sum p_e \ln(p_e)$, using natural logarithms [20]. Effective diversity was $D_1 = \exp(H)$, with $D_2 = 1/\sum_e p_e^2$ giving greater weight to common entries [21]. The squared-share sum is also reported as HHI. Observed richness counted entries with positive incidence; the largest-entry and top-three shares summarized concentration. With zero total incidence, the probability distribution and effective diversities were undefined, while observed richness was zero. No unseen-category extrapolation was attempted.

An alternative weighting gave each observed query equal total response weight within its stratum. For query q with n_q responses, every response received weight $1/n_q$; weighted incidence totals were then normalized across entries. Queries with no detections remain in the observed query universe but add no incidence mass. After normalization, this conditional distribution does not by itself express the prevalence of empty queries. The weighting also differs from equalizing each query's distribution among mentioned entries. Total variation, $\frac{1}{2}\sum_e |p_e - p'_e|$, compared the resulting distributions. Analyses covered each vertical, vertical-by-service cells and cells further separated by stored model identifier. Pooled vertical summaries retained the observed service composition.

Code inspection identified two candidate label pairs for a separate sensitivity analysis: Amazon/Amazon Brasil and Accenture/Accenture Brasil. The collapse rules were recorded before counting their effects. Every historical canonical response with at least two detected entries was re-extracted using the frozen implementation, and intersections between accepted-match intervals were counted. A declared collapse replaced both labels in each pair with one analytical category, contributing at most one incidence per response. This was an alternative operational resolution, without asserting corporate equivalence. The original-label estimates remained the primary description of the frozen instrument.

3.7 Persistence across observed dates

Set persistence compared consecutive observed UTC dates within each query, service and stored-model series. Each date's set was the union of eligible entries across its available responses. For successive sets A and B, Jaccard similarity was $|A \cap B| / |A \cup B|$ when their union was nonempty. Pairs in which both sets were empty were counted separately. A diagnostic mean also assigned those pairs a value of one, allowing the consequence of that convention to be measured directly.

The summaries separated one-day intervals from longer gaps. A further sensitivity retained pairs with equal numbers of responses on both dates, because daily unions depend on opportunities for detection. Missing dates were not imputed, and comparisons did not cross stored model identifiers. These exploratory summaries describe the recorded sequence; they neither estimate population uncertainty nor isolate an effect of elapsed time.

3.8 Common query and date cells in the recent window analysis

The exploratory recent extension evaluated prefixes of 50, 100, 150, 200, 300, 500, 750, 1,000, 1,500, 2,000 and 3,000 code points against all retained text. Recovery at each prefix was the fraction of full-retained positive records also positive in that prefix; newly positive prefix records were checked separately. This denominator differs from the fraction of all responses with a detection. Curves describe the observed output and do not identify a preferred window for user attention or commercial measurement.

A second analysis restricted the five recent services to their common exact-query and UTC-date cells. Repetitions within a service-cell were averaged before the common cells received equal weight. Thus the estimand holds observed query and date participation constant while retaining the within-cell service observations. It does not synchronize request times, randomize systems or equalize hidden implementation details. Both recent extensions were planned after inspection of the principal 200-character contrast and are reported as exploratory.

4. Results

4.1 A reconciled corpus with incomplete calendar coverage

The frozen artifact contains 91,392 responses from 23 April to 11 September 2026, including 72,145 canonical responses and 19,247 probes across 56 observed UTC dates. The historical cut preserves 83,486 responses across 50 dates: 66,399 canonical and 17,087 probes. All 83,486 historical response_full_text values are null. The 7,906 subsequent rows populate that field, including 5,746 canonical responses.

Table 1. Historical response coverage, 23 April-31 August 2026. Counts describe stored observations, not independent API experiments.

Service	All responses	Canonical responses	Canonical queries	Observed dates
ChatGPT	19,328	14,976	192	50
Claude	19,162	14,842	192	50
Gemini	18,794	14,587	192	49
Groq	18,304	14,208	192	48
Grok	462	350	192	2
Perplexity	7,436	7,436	96	50

Perplexity's historical prompts cover discovery, comparison and market categories, with 32 exact texts per category. Its reduced battery is not a random half of the six-category design. Grok contributes only two historical dates, and Gemini's combined coverage in Table 1 spans two model identifiers. Neither row supports an undifferentiated temporal or architecture comparison.

The historical interval spans 131 calendar dates. Fifty contain responses, while the other 81 have 324 ledger entries labeled aborted, four vertical entries per date. Inspection of mark_collection_gaps.py and creation timestamps showed that these were retrospective gap annotations with synthetic noon timestamps. They are not 324 failed executions or direct observations of 81 attempted runs. July has no responses and 31 dates so annotated. The ledger makes missing persistence visible, but does not recover missing answers or query-level failure probabilities.

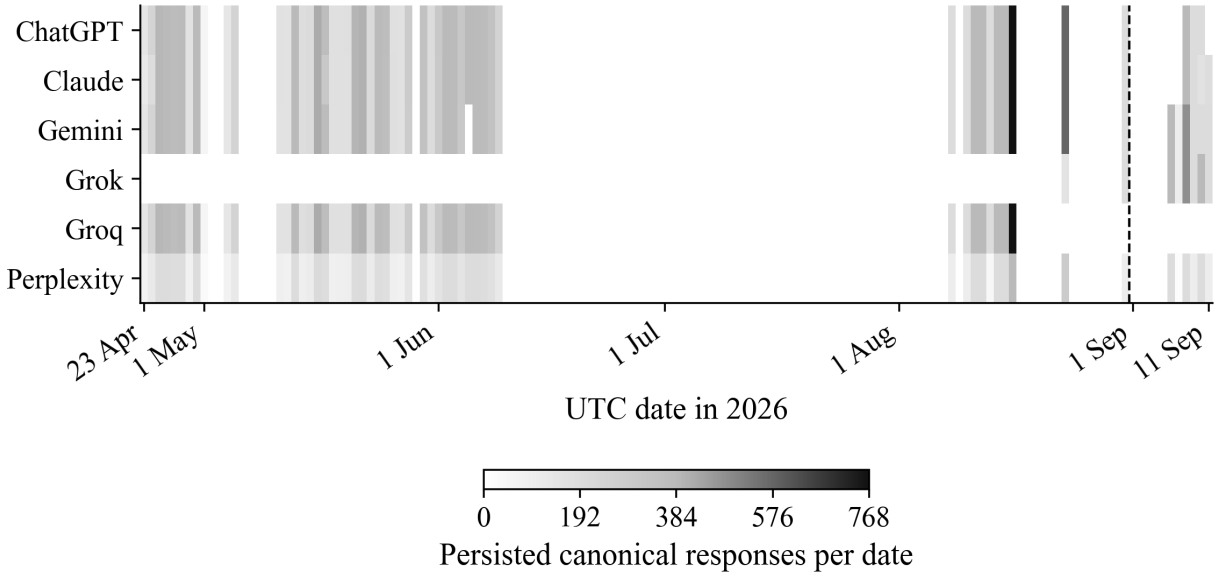

Figure 1. Persisted canonical responses by service and UTC date, 23 April-11 September 2026. White cells indicate no persisted canonical response, not a negative behavioral outcome. The dashed boundary marks the start of September; the historical analyses stop at 31 August. Counts include unequal query batteries and changing service participation.

4.2 Changing the retained window changes the result

Among the same 7,436 historical Perplexity responses, the frozen extractor detected at least one lexicon entry in 5,631 full retained strings (75.73%) and 3,863 prefixes of 200 code points (51.95%). The paired difference is -23.78 percentage points. Re-extraction of the full retained strings disagreed with the historical cited flag in zero rows. The comparison therefore isolates the operational effect of changing the window under the frozen extractor.

Table 2. Paired historical Perplexity outcomes. Counts use the same canonical responses from the historical cut.

Paired outcome	Responses
Positive in both retained text and prefix	3,863
Positive only in retained text	1,768
Positive only in prefix	0
Negative in both	1,805

Table 3. Historical paired window contrast by vertical. Positive counts refer to at least one eligible entry in each of the same response records.

Vertical	Responses	Full retained positive	Prefix positive	Difference, pp
Fintech	1,872	1,599	1,267	-17.74
Retail	1,868	1,713	1,283	-23.02
Healthcare	1,848	1,299	823	-25.76
Technology	1,848	1,020	490	-28.68

The query-cluster percentile interval was [-28.39, -19.16] percentage points; the date-cluster interval was [-25.32, -22.19], using 96 queries and 50 dates respectively. The wider query interval reflects sensitivity to the mix of prompts under that resampling scheme. It should not be replaced by the narrower date interval without changing the inferential question.

Other historical fields exposed the retention asymmetry directly. Every one of 19,328 ChatGPT strings and 19,162 Claude strings had length 200, whereas none of the 7,436 Perplexity strings had that exact length. Perplexity retained strings averaged 687.17 code points, with a maximum of 2,502. The distribution of the measured variable revealed an adapter difference that a binary completion check would not describe.

4.3 Recent retention permits paired checks in other services

All 5,746 recent canonical rows have response_text equal to the first 200 code points of response_full_text. Applying the same extractor to both fields produced the service-specific counts below. No recent row changed from full-text negative to prefix positive.

Table 4. Recent paired outcomes, 6-11 September 2026. Dates and query coverage differ across services; these are within-service window comparisons.

Service	Responses	Dates	Full retained positive	Prefix positive	Difference, pp
ChatGPT	768	3	408	127	-36.59
Claude	929	4	618	224	-42.41
Gemini	1,536	6	763	46	-46.68
Grok	1,728	6	1,452	487	-55.84
Perplexity	785	6	602	420	-23.18

The recent data establish that window sensitivity also occurs in other recorded services once longer payloads become available. They do not reconstruct discarded historical continuations. The size of a difference in Table 4 depends on prompt composition, period and service behavior together; it is unsuitable as a service ranking or a universal correction factor for earlier records.

4.4 Recovery beyond the operational prefix

The prefix curves show how much of each recorded full-text positive set is recovered at longer windows (Figure 2). At 200 code points, recovery ranged from 6.03% of Gemini's 763 positive recent records to 69.77% of Perplexity's 602 positive recent records. ChatGPT recovered 31.13% of 408, Claude 36.25% of 618 and Grok 33.54% of 1,452. Across every evaluated prefix, there were zero gains relative to full-retained detection. These conditional recovery proportions quantify what the shorter instrument misses; they do not validate the extractor against semantic ground truth.

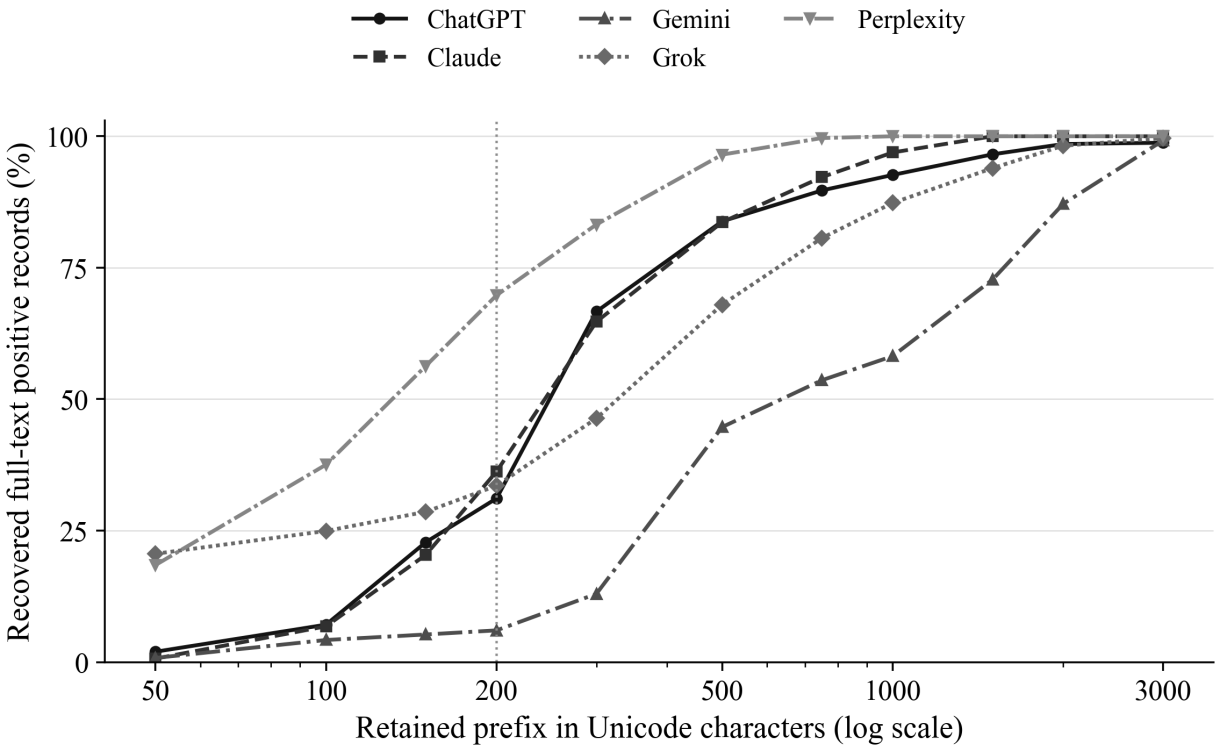

Figure 2. Recovery of full-retained positive records at specified prefix lengths, recent canonical responses from 6-11 September 2026. The denominator for each curve is that service's full-retained positive count. The horizontal axis is logarithmic; connecting lines join evaluated windows without estimating unobserved thresholds. The dotted line marks 200 code points. Service populations have unequal dates and queries.

4.5 A shared recent query and date population

The intersection across the five services contained 272 exact-query/date cells: 96 on 8 September, 96 on 9 September and 80 on 10 September. These cells covered 96 unique queries and 2,096 responses. ChatGPT and Claude each contributed 368 responses, Gemini 416, Grok 496 and Perplexity 448. No service mixed stored model identifiers inside this restricted population. Averaging within cells prevents the different numbers of repetitions from determining each service's weight across the shared battery.

Under this common-cell estimand, the 200-character-minus-retained differences remained negative: -42.28 percentage points for ChatGPT, -46.51 for Claude, -71.88 for Gemini, -58.58 for Grok and -24.63 for Perplexity (Figure 3). The analysis therefore preserves the window finding after matching exact queries and observed dates. Its scope is the three-day overlap, not a controlled estimate of provider superiority or a reconstruction of earlier responses.

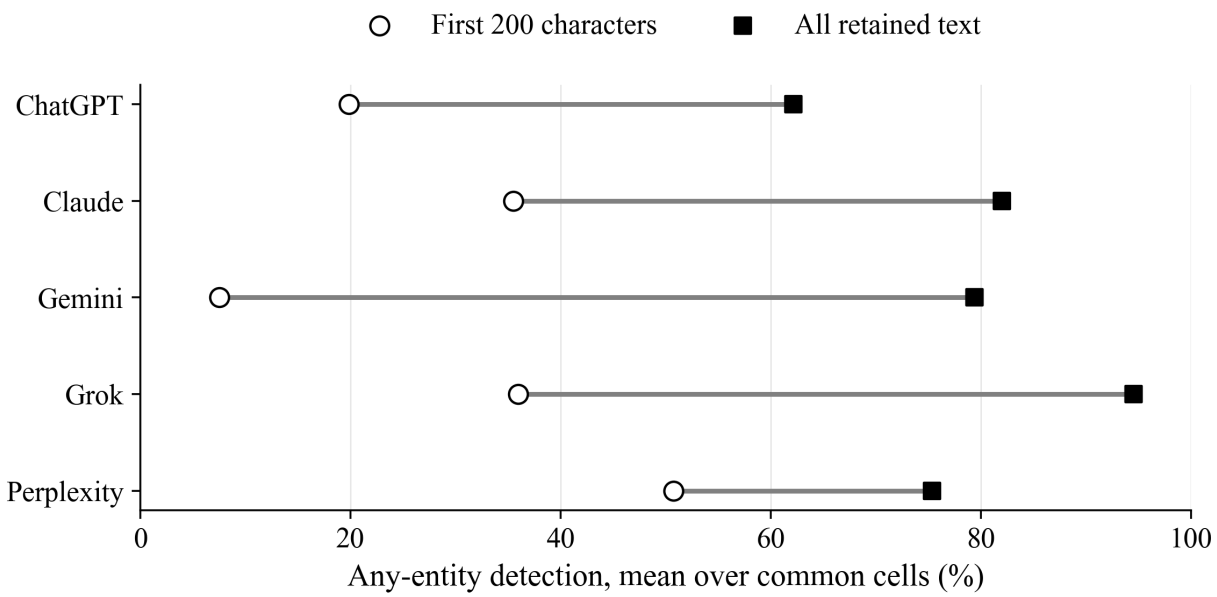

Figure 3. Detection in the first 200 code points and in all retained text across the same 272 query/date cells, 8-10 September 2026. Repeated responses are averaged within each service-cell, then cells receive equal weight. The figure describes paired window changes within five services; it is not an architecture comparison.

4.6 Repetition and query composition account for different kinds of stability

Historical responses repeatedly reproduced the same observable prefix. Within ChatGPT, 14,976 canonical rows formed only 838 distinct query/prefix-hash combinations; the corresponding counts were 890 of 14,842 for Claude and 1,259 of 14,208 for Groq. Perplexity produced 7,313 distinct query/prefix combinations among 7,436 rows, increasing to 7,434 combinations when the full retained strings were hashed. This latter comparison demonstrates why a repeated prefix cannot establish that the complete answer repeated.

The empirical between-query component accounted for 0.8824 of the binary variance in ChatGPT, 0.9067 in Claude, 0.8696 in Groq and 0.5337 in Perplexity. Gemini Flash and Pro yielded 0.7428 and 0.5424 in their respective 3,648- and 10,939-response strata. These fractions describe the fixed prompt battery and observed window. Large raw counts can accumulate around stable query-specific outcomes without providing equivalent growth in independent information.

Calendar composition also changed apparent daily variability. Across Claude's 50 observed historical dates, daily rates with equal weight per observed query ranged from 9.62% to 34.96%; restricting the calculation to its 36 dates containing all 192 queries narrowed that range to 24.22%-27.34%. For Perplexity, the corresponding ranges were 20.83%-70.13% across 50 dates and 43.23%-64.06% across 37 dates with all 96 queries. Restriction changes the observed population of days, so the narrower ranges do not prove that missingness is harmless. They show that fluctuations in battery coverage belong in the interpretation of a time series.

4.7 Language contrasts depend on the resampling unit

Matching date/template cells yielded 4,575 pairs for ChatGPT, 4,556 for Claude, 960 for Gemini Flash, 3,520 for Gemini Pro, 174 for Grok, 4,382 for Groq and 2,262 for Perplexity. Portuguese-minus-English differences were -11.91, -2.18, -2.59, -1.75, -5.75, -6.57 and +3.91 percentage points, respectively. These comparisons hold the designed non-language factors and recorded service/model constant within a date, while retaining the limitations of translated prompts and a fixed character window.

Perplexity illustrates why uncertainty cannot be selected to fit a narrative. Its +3.91-point contrast had a template-pair percentile interval of [-3.64, 12.25], but a date-cluster interval of [2.03, 5.85]. For Claude, the corresponding intervals around -2.18 were [-10.23, 5.90] and [-2.79, -1.56]. Resampling dates asks about temporal reweighting of the observed battery; resampling templates asks about reweighting its question pairs. Neither establishes a population-wide preference for a language. Grok has only two historical dates, which cannot support an interpretation of temporal precision from date resampling. Complete contrasts, excluded-cell counts and conditional sensitivities are supplied with the analysis rather than reduced to significance labels.

4.8 Category contrasts within shared design cells

Category matching did not make the measured outcomes interchangeable. For historical ChatGPT observations with stored model identifier gpt-4o-mini-2024-07-18, averaging response-level detection rates equally over the same 1,473 date/design cells across 50 observed dates yielded 29.28% for comparison prompts and 1.54% for discovery prompts. Each retained cell contained all six categories, with vertical, language, query form and temporal frame held constant. Experience prompts yielded zero detections in these matched cells. These results concern a fixed prompt battery, lexicon and 200-code-point observation window. They do not establish a causal category effect or show that the unobserved answer continuations lacked entity mentions. Category composition therefore remains part of the measurement definition.

4.9 Lexical controls and populated columns do not provide semantic validation

The historical control filter identified 16,579 responses previously marked by the fictitious-name flag. The published lexical expression detector matched 11,195 of them (67.53%). A name mentioned while denying knowledge of the requested organization can trigger the original flag. The remaining 5,384 responses were not adjudicated as fabrication: they can include unrecognized refusals, ambiguous references or other responses outside the detector's expressions. Human-labeled precision, recall and semantic fabrication rates remain unmeasured.

The field inventory found 50 citations columns, including inputs, metadata and derived labels rather than 50 independent measures. Five columns were null throughout all 91,392 rows: cited_domain, cited_person, attribution, our_source_count and hedging_detected. Ten analytical

tables were empty, including intervention outcomes, dual responses, URL verification and score-calibration inputs. Their existence in the schema does not supply evidence for those analyses.

Selection and absorption fields were populated in 33,052 rows. In every populated row, absorption_status equaled cited; inspection of the classifier confirmed the direct assignment. This is a derived mention flag, not an independent assessment of semantic use of a source as distinguished in the literature [3, 6]. Selection uses entity-name slug matching against exposed URLs and encodes an empty source list as zero. It cannot identify the service's hidden retrieval process. Context annotations were similarly heuristic: 36,680 entry-level records covered 23,400 responses and included 9,548 probe records. Their sentiment, attribution and limited fact-catalogue checks were not treated as human-validated answer quality.

4.10 Concentration within the observed dictionary

The historical canonical sample yielded 18,144 eligible incidences in 11,845 of 66,399 responses collected from 23 April through 31 August 2026. No designated decoy appeared in this canonical extraction. Across the 111 eligible labels, 66 were detected and 45 had zero incidence. Fintech, retail and health each had 16 detected entries, yet their effective Shannon diversities differed substantially (Table 5). The technology vertical had 18 detected entries and $D_1 = 9.674$. Richness alone therefore concealed differences in the allocation of detected incidence mass.

Table 5. Distribution of eligible label incidences in the historical 200-character window. Positive responses contain at least one eligible label. Top-three shares use incidences as their denominator. Verticals pool their observed services and periods.

Vertical	Eligible entries	Positive responses	Incidences	D_1	Top-three share
Fintech	27	4,403	6,643	3.788	79.77%
Retail	28	3,838	6,037	5.203	79.71%
Health	28	1,742	2,656	7.864	66.45%
Technology	28	1,862	2,808	9.674	52.07%

Equal-query weighting produced total-variation distances of 0.00651, 0.01086, 0.01880 and 0.02980 for fintech, retail, health and technology, respectively. Top-three shares changed by +0.63, +0.59, -1.53 and +1.29 percentage points. Within vertical-by-service-by-model cells, the largest total-variation distance was 0.00206. Query multiplicities were mostly constant within those cells, leaving little for this weighting to change. The larger pooled differences reflected aggregation across panels with different query coverage and multiplicities; equal-query weighting did not fill missing query-by-service combinations.

4.11 Empty sets and observed-date gaps

Of 39,686 observed-date pairs, 31,372 had two empty sets. The conditional Jaccard mean across the remaining 8,314 pairs was 0.7423. Assigning similarity one to the empty pairs instead produced a mean of 0.9460, driven by the 79.05% empty-pair fraction (Figure 4). The convention also changed the ordering of vertical summaries: health had a conditional mean of 0.6713 against fintech's 0.8433, but the corresponding empty-as-one means were 0.9577 and 0.9553.

Among 6,917 informative pairs one day apart, mean Jaccard was 0.7525; for 1,397 informative pairs with longer gaps, it was 0.6921. Longer gaps had a median of four days and a maximum of 60 days. This difference cannot identify temporal decay because gap lengths were not assigned independently of calendar and panel composition. Restricting all intervals to equal daily response counts retained 4,238 informative pairs and gave a mean of 0.7415, close to the unrestricted conditional mean. The restricted sample remained selected by observed collection effort.

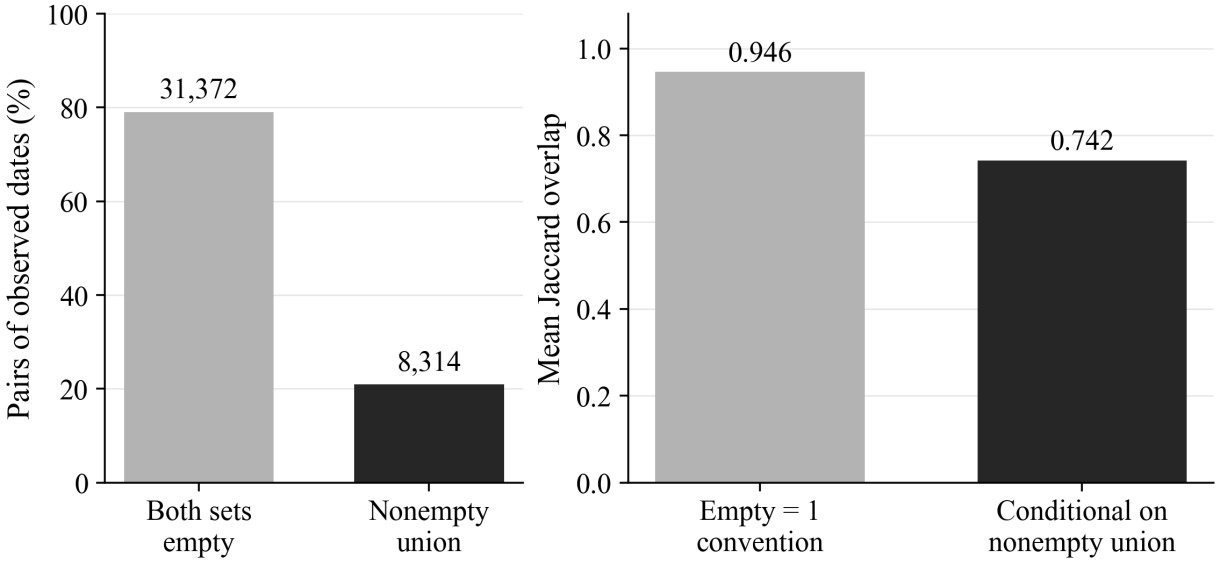

Figure 4. Dependence of reported set stability on the treatment of absence. Among 39,686 consecutive-observed-date pairs within query/service/model series, 31,372 had two empty sets. The alternative convention assigns those pairs similarity one; the conditional estimate uses the 8,314 pairs with a nonempty union. These are different descriptive estimands, not two independent statistical estimates.

4.12 Overlapping labels changed diversity without changing binary detection

Re-extraction of 3,611 responses with multiple labels reproduced every input label set. Exactly 316 responses contained overlapping accepted-match intervals, all involving Amazon and Amazon Brasil at identical character positions. The frozen ambiguity rule searched for the surface “Amazon Brasil” under both labels. No other pair showed overlapping accepted intervals.

Collapsing that retail pair reduced incidences from 6,037 to 5,721 and observed richness from 16 to 15. Effective Shannon diversity fell from 5.203 to 4.589, while the top-three incidence share rose from 79.71% to 84.11%. All 3,838 positive retail responses remained positive. The

retail conditional Jaccard mean changed from 0.7389 to 0.7464. These differences arose from label resolution with the underlying responses held fixed.

The second declared collapse, Accenture/Accenture Brasil, removed a zero-incidence category and left technology's incidence distribution unchanged. Together, the two collapses reduced the reference population from 111 labels to 109, with 65 detected and 44 undetected. Those counts describe the alternative dictionary; they do not establish the number of distinct businesses represented in the responses.

4.13 A denominator repair changes what the composite can claim

The historical active-service rule excluded 14,208 Groq responses and left 52,191 eligible canonical observations. The original calculation nonetheless allowed Groq mentions into the numerator, inflating coverage for 13 lexicon entries. Applying the same eligibility rule to every component repaired that inconsistency without editing production data.

With the frozen 127-entry lexicon, the repaired geometric index had Spearman correlation 0.9853 with coverage among the 66 entries with positive detection, increasing to 0.9977 when 61 zero entries were included. Its correlation with the arithmetic alternative was 0.8423 among positive entries and 0.9751 across all entries. The zeros materially affect apparent agreement. Overlapping lexicon labels and the extractor's offset conventions impose further limits on entity-level interpretation. These calculations audit a convenience summary; they neither establish an independent organizational ranking nor validate the index as a measure of optimization success.

5. Discussion

5.1 Observation windows define the measured outcome

Restricting the historical Perplexity responses to 200 characters removed 1,768 positive detections among the same 7,436 canonical records collected through 31 August 2026. The resulting reduction of 23.78 percentage points arose without changing the stored answers or the extractor. This isolates a consequence of the measurement procedure within the archive. An apparent difference in visibility can therefore arise before any explanation about content quality, brand familiarity or retrieval behavior is needed.

The longer retained text and the prefix answer different measurement questions. The former permits detection anywhere in the available answer; the latter asks whether a name appears near its beginning. Either can be useful if its interpretation is explicit. Early appearance may matter for an interface that displays an excerpt, whereas a measure of presence anywhere requires the available answer to be examined beyond that excerpt. The archive contains no user-attention or interface-exposure observations that would establish the 200-character boundary as a behavioral threshold. It is an instrument setting.

Using an equal prefix length across services can remove one documented asymmetry in what the extractor sees. It does not equalize answer length, prompt interpretation or the amount of introductory language. Nor does it recover material discarded before persistence. Treating a common prefix as a full-answer measure would retain the labeling problem after repairing the implementation. The useful outcome of harmonization is a common, narrowly described observable. Comparisons of complete captured answers and early mentions should remain separate analyses when both are available.

This distinction follows the measurement perspective developed by Jacobs and Wallach: the relationship between an operational measure and its intended construct requires an argument that cannot be supplied by a convenient variable name [2]. Here, cited records whether at least one eligible entity was detected in a response. Increasing the eligible lexicon can change this outcome even if the answer itself is unchanged. An entity-level prevalence measure would require a different denominator and would preserve which entity was detected. A source-attribution measure would require evidence about statements and sources. A single response-level flag cannot substitute for these definitions.

Full-text persistence improves the range of questions that can be asked later. The dedicated field appears in the September records, allowing recent outputs to be inspected beyond the operational window. Its availability does not make the earlier missing text recoverable or remove the output constraints of the original API call. The durable design choice is to preserve the captured output alongside each transformation and derived field, so that a later change in the research question can be evaluated against material that actually exists.

The recent paired analyses extend the observation-window question to the five services represented in the September data. A reduction in detection within each retained service sample shows that the issue is relevant beyond the historical Perplexity comparison. The 272-cell analysis preserves that result after restricting exact queries and observed dates to the common population. Its equal cell weights remove differential cell participation from that estimand, but do not equalize request times or hidden system conditions. Differences in the size of a reduction still do not identify which service has greater underlying visibility or which architecture causes an entity to appear later.

5.2 Lexical controls require a semantic interpretation

The decoy analysis reveals a mismatch between detecting a supplied name and judging a factual claim about that name. Because the adversarial prompts introduce the decoy, an answer can repeat it while rejecting the premise. The historical fictional_hit flag can activate in that situation. Its interpretation as fabricated factual content would require an additional assessment of what the answer asserts.

The refusal-pattern result provides a concrete reason to perform that assessment. Among 16,579 historical records selected for calibration status, a decoy hit and nonempty retained text, 11,195 matched the published lexical rule. Those observations establish coexistence between a

decoy-hit marker and language covered by the rule. They do not establish that every matching response is a valid refusal, or that an unmatched response fabricates information. A response may contain a disclaimer followed by an unsupported claim; another may express uncertainty using words absent from the pattern. Truncation can also remove the sentence that would determine how an earlier mention should be read.

Attribution research provides a more appropriate model for the next evaluation step. Rashkin et al. relate generated content to support from identified sources through an explicit annotation framework [3]. Liu et al. distinguish whether statements receive citations from whether those citations support them [4]. Applied to the present probes, the corresponding task is to identify the response's factual commitment before judging whether it is supported. That application is a proposed extension of the audit; the archive has not received the necessary human annotation.

A useful annotation design would preserve the exact prompt, available answer and observation window; distinguish rejection of the premise from uncertainty, unsupported factual attribution and a possible homonym; and permit an indeterminate judgment. Sampling should include both matching and nonmatching responses rather than only conspicuous examples. Independent annotation and adjudication could then estimate how well an automatic rule performs against those judgments. Until that work is completed, the lexical fields are screening variables. Calling them error rates would hide the missing link between the rule and the construct.

The same problem applies to richer-sounding metric names. A list of exposed URLs establishes that the interface returned those URLs under its implementation. It does not establish which passages supported each claim, which sources were retrieved but omitted or why a source affected generation. The distinction between selection and absorption proposed by Zhang et al. is relevant to the conceptual separation, but their preprint supplies no annotations for this archive [6]. Adding columns with those names cannot supply the missing evidence.

5.3 Longitudinal comparison depends on eligibility and instrument history

Correcting the aggregation population repairs an arithmetic inconsistency while leaving the scientific meaning of the index open. In the audited implementation, records from an excluded service could contribute to a numerator whose denominator counted only the selected services. Applying one eligibility rule to both terms makes the ratio interpretable for that population. It does not establish that the population is balanced over time or that a weighted combination measures authority.

The descriptive correlation results also depend on which entities are included. Shared zero scores contribute tied ranks when the full lexicon is used, whereas the positive-entity subset asks about ordering conditional on having been detected. Both summaries can be reported if their populations are explicit. Their agreement cannot be treated as independent validation of an index that already includes coverage as a component. The OECD handbook's treatment of normalization, weighting and sensitivity is applicable here as a method for examining construction choices, without endorsing any particular GEO weights [17]. External validation

would require a separately measured outcome and a stated reason to expect the index to predict it.

Provider turnover creates a related population problem. A service label spanning several months is not necessarily a fixed treatment: its stored model can change, an endpoint can change, or a category can be routed differently. Restricting all services to a common vocabulary of queries addresses only one dimension. A comparison must also specify dates, versions and observation conditions, and decide whether the resulting overlap still answers the research question. A comparison that loses the relevant population through restriction should be narrowed or left unresolved.

Repeated-query studies make the sampling issue visible. Schulte et al. describe variation across runs and days, and Sielinski treats visibility as an estimated distribution with uncertainty; both are preprints, and their results remain conditional on their respective designs [7, 8]. The present separate resampling analyses by query and day examine sensitivity to two observed grouping structures. They do not jointly model crossed dependence or restore missing periods. Cameron and Miller's account of clustered inference explains why the dependence assumption must accompany an uncertainty estimate [9]. Additional rows generated under the same conditions are not equivalent to additional independent contexts.

For this archive, governance is therefore part of the measurement record. The meaning of a count depends on whether its unit is a response, a query, a date or an administrative entry. A retrospectively backfilled abort ledger documents an assigned operational status; it does not prove that an independent failed execution occurred on every represented date. A change log needs both the intended deployment date and evidence of the configuration actually observed. The structured experimental metadata proposed by Breuer et al. and the PRIMAD discussion by Ferro et al. provide precedents for documenting such components [15, 14]. Recording them enables scrutiny; independent reproduction still has to be demonstrated.

5.4 Stability and the resolution of entity labels

An apparent stability near one can be a consequence of how absence is scored. Let p_{00} denote the fraction of pairs with two empty sets and J_{cond} the mean over pairs with a nonempty union. Assigning one to the empty pairs gives the identity

$$J_{\mathrm{all}} = p_{00} + (1 - p_{00}) J_{\mathrm{cond}}.$$

As p_{00} increases, this mean approaches one regardless of the similarity among informative pairs. Reporting p_{00} beside J_{cond} makes that dependence explicit; neither statistic alone describes the full observation process.

Label resolution creates a different failure mode. The retail sensitivity changed concentration while preserving every binary positive, so an unchanged mention rate would not reveal the duplicate-label contribution. Distributional indicators require an audit of category identity and matching overlap in addition to a response-level detection audit. Future collection releases

should version those resolution rules and preserve sufficient text to distinguish lexical overlap from substantively distinct references before interpreting effective diversity as a property of the entities themselves.

5.5 Reporting requirements motivated by the audit

The BRGEO-1 specification draft organizes measurement conditions into six parameter families. P1 defines the observed text window, its unit and the order of truncation and cleanup. P2 identifies the eligible lexicon, aliases, control labels and resolution rules; a lexicon entry is not automatically an independent organization. P3 records the exact prompt battery, its designed strata and routing. P4 records the service, model identifier, interface, endpoint and observation dates. A stable identifier is not evidence that the underlying service remained unchanged. P5 records generation settings, output limits and system instructions where exposed, leaving unavailable settings unknown. P6 specifies extraction, normalization, boundary handling, disambiguation and returned offsets. Historical records should report the settings that can be established from their evidence, without reconstructing undocumented values as facts.

This article treats those families as a reporting proposal informed by a measurement audit. Matching their labels does not establish equivalence between implementations or validate a substantive visibility construct. Historical records have not been certified against the proposed protocol. Table 6 extends the parameter record with denominator eligibility, operational provenance and validation evidence. Appendix D reconciles the earlier BRGEO-1 manuscript's data cutoff and additional descriptive calculations with the analyses reported here. The older 2,225-response extension and the present 5,746-response extension are different frozen populations; neither is silently substituted for the other.

The audit supports a reporting specification centered on reconstructing a measure from its evidence. Each requirement in Table 6 addresses an observed failure or an interpretation that the archive cannot justify. The specification is intended for empirical testing across implementations. It has not undergone a conformity study, and compliance should not be described as certification of a visibility score.

Table 6. Proposed reporting requirements motivated by the audit. Each requirement specifies evidence to preserve; none independently validates the target construct.

Reporting requirement	Evidence to provide	Interpretation enabled
Define the target construct and unit	Outcome definition, entity lexicon and explicit denominator	Distinguishes response-level detection from entity prevalence and source support
Identify the query population	Exact prompts, languages, categories and inclusion rules	Shows which questions and prompted entities the result covers
Preserve the observed output	Captured text, truncation settings and missing-field inventory	Allows a result to be related to the material actually available
Version the transformations	Extractor, aliases, normalization and configuration hashes	Permits the same recorded text to be reanalyzed under stated rules
Describe comparison eligibility	Service, model, endpoint, dates and shared query cells	Makes the population common to numerator and denominator inspectable
Separate operations from observations	Attempts and failures where logged, stored responses and observed dates	Prevents administrative counts from becoming sample-size claims
Support semantic labels	Annotation guide, selected material and agreement evidence when collected	Relates an automatic marker to the intended factual judgment
Expose aggregation and sensitivity	Component definitions, weights, zero handling and alternative calculations	Shows which choices determine an aggregate result

Several requirements can be implemented before the next collection: retaining captured output, freezing transformations and generating population counts from the same eligibility function. Others require new scientific work. Semantic validation needs an annotation study, and comparison across implementations needs a shared test corpus plus an agreed interpretation of discrepant outputs. Keeping these tasks distinct prevents a documentation improvement from being mistaken for empirical validation.

Dataset documentation should also state the intended uses and the uses that the material cannot support. Gebru et al. organize this information around dataset composition, collection and maintenance [19]. For the present collection, an appropriate use is studying the consequences of the recorded measurement procedure. Estimating how often all users encounter a brand would require a population and observation design that the fixed battery does not provide. The value of the proposed specification lies in making that boundary assessable before a score is reused.

6. Limitations and scope

The study uses a practitioner-built, nonprobability panel of entities and queries across four verticals and two languages. Its query distribution is designed rather than sampled from user demand. The canonical battery asks sector-level questions, while the adversarial prompts explicitly supply decoy names. These conditions constrain how mentions should be interpreted and prevent the response-level proportions from being read as market-wide exposure estimates. The lexicon also restricts detection to known eligible entities; an answer may discuss other organizations that do not contribute to the measured outcome.

The retained corpus is temporally discontinuous. Dates without observations cannot be assigned a behavioral value, and administrative failure records do not necessarily enumerate every missing attempt. Services contribute unequal numbers of observations and different query subsets. The six historical service identities were not six simultaneous, stable experimental arms. Model and endpoint changes further limit pooled comparisons. With a single provider in the retrieval-enabled historical arm, the archive cannot separate an architectural effect from provider-specific behavior. API observations also do not establish what users saw in a consumer interface.

Historical persistence limits reanalysis. The paired window result uses the longer text retained by Perplexity; it does not reconstruct complete historical answers for the other services. The dedicated full-text field is absent throughout the historical cutoff. Its presence in the 6-11 September extension creates a later observation regime with a short period and changing service coverage. That extension cannot establish that an earlier provider difference disappeared, nor can it identify the cause of such a difference.

The extractor is deterministic but imperfect. It applies lexical aliases, normalization and context rules, and its offsets are tied to transformed text. The implementation does not always choose the earliest surface occurrence across canonical forms and aliases. Position-based interpretations therefore require additional validation. A fixed extractor makes the paired calculation reproducible, but reproducibility of a rule is compatible with systematic classification error. No independent human reference set was used to estimate detection precision, refusal accuracy or fabrication prevalence.

The statistical summaries are conditional on this archive. Separate query and day resampling schemes reveal sensitivity to those groupings without modeling their dependence jointly. The observed days are not exchangeable by design, particularly when configurations change. Missing periods, selected queries and service-specific routing are not corrected by either interval. The index calculations add a separate restriction: they are retrospective sensitivity analyses of a specified aggregation, with no independently validated outcome for authority or optimization success.

The reporting proposal is consequently supported at the level of measurement practice. The paired transformation demonstrates how an observation rule changed an outcome in recorded

data; the label and denominator audits identify conditions required for interpretation. Their effect sizes and prevalence should not be transferred to other languages, industries, interfaces or periods without new evidence. A prospective study with stable documented conditions, shared comparison cells and independent annotation would test which parts of the proposal generalize beyond this implementation.

7. Conclusion

The historical window comparison shows how a published rate can change while the retained answers remain fixed. Its scientific value is the identifiable transformation: a reader can inspect the same observations, apply a stated extractor to two windows and reconstruct the difference. That connection between a number and the procedure producing it should be preserved when the research moves from mentions to richer claims about visibility.

For subsequent GEO measurement, the immediate requirement is to publish the population, observation window, transformation version and denominator with each outcome. Captured text should remain available for reanalysis under the applicable access conditions. Lexical controls should retain their narrow meaning until an annotation study establishes what they classify, and composite indicators should expose their components before their ordering is interpreted.

A shared corpus evaluated by independent implementations would provide the next test of the reporting proposal. It should include changes in observation window, ambiguous entity matches, prompted decoys and explicit eligibility cases, with disagreement traced to a defined transformation or judgment. Success would be demonstrated by reconstructable measurements and explained differences. It would give researchers a basis for comparing results whose conditions are known, while leaving claims about optimization effectiveness to studies designed to estimate them.

Declarations

Competing interests

Alexandre Caramaschi is CEO and founder of Brasil GEO, a business operating in generative engine optimization, and maintains the collection infrastructure examined in this study. He therefore has a professional and commercial interest in the subject and in the adoption of its measurement practices. He is also Chief Strategy Officer of Nuvini (Nasdaq: NVNI). The research was conducted through Brasil GEO and does not represent a position of Nuvini. These relationships do not validate the proposed protocol or its indicators.

Funding

The author reports that this research received no specific grant from funding agencies in the public, commercial or not-for-profit sectors. Collection API costs were borne by the author. No external sponsor directed the design, analysis, writing or decision to disseminate the work.

Author contributions

Alexandre Caramaschi: Conceptualization; Methodology; Software; Formal analysis; Investigation; Data curation; Writing - original draft; Writing - review and editing; Visualization; Project administration. AI assistance used in those activities is disclosed separately below.

Data and code availability

The collection code is available in the study repository. The frozen database, extraction inputs, analysis scripts, derived tables and manifests used for this manuscript are retained in the author's audit package. Appendix A identifies the principal artifact and reconstruction procedure; Appendix D supplies the earlier snapshot cutoff. No archival data DOI or independently verified public dataset deposit is claimed for this audit. The complete row-level package has not been publicly released with this preprint. Requests concerning access can be directed to the corresponding author. The repository's code license does not by itself determine the release conditions for every stored response or operational field.

Publication history and version note

This article belongs to the research programme that produced the author's practitioner framework [22] and subsequent null report [23]. It examines a later frozen collection and the measurement procedure itself, rather than treating those earlier papers as independent validation of the protocol.

This preprint is a corrected empirical revision of the BRGEO-1 research presented in the 183-page manuscript submitted to SSRN (Abstract ID 7446519; https://ssrn.com/abstract=7446519). Appendix D reconciles the data snapshots, analytical units, weighting and interpretation of the reported results. The two manuscripts are versions of the same research, not independent studies or independent replications. The existing SSRN submission is retained unchanged; this Zenodo deposit provides the corrected empirical version separately. This version has not undergone journal peer review.

License

Use of generative AI in manuscript preparation

OpenAI Codex was used for literature discovery, preparation and checking of analysis code, drafting, internal critical review and typesetting. Anthropic Claude assisted with code and drafting in the BRGEO-1 specification manuscript used during reconciliation. Perplexity Sonar Pro was used for complementary literature discovery through the geo-orchestrator and curso-factory integrations. External references were checked against primary publication records and sources; incorrect version dates proposed during literature discovery were corrected. Numerical results were calculated by the preserved Python scripts from the frozen database. These tools were not treated as authors, human annotators or external peer reviewers. Responsibility for the submitted work remains with the named author.

References

[1] Pranjal Aggarwal; Vishvak Murahari; Tanmay Rajpurohit; Ashwin Kalyan; Karthik Narasimhan; Ameet Deshpande. (2024). GEO: Generative Engine Optimization. Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining , 5-16. https://doi.org/10.1145/3637528.3671900

[2] Abigail Z. Jacobs; Hanna Wallach. (2021). Measurement and Fairness. Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency , 375-385. https://doi.org/10.1145/3442188.3445901

[3] Hannah Rashkin; Vitaly Nikolaev; Matthew Lamm; Lora Aroyo; Michael Collins; Dipanjan Das; Slav Petrov; Gaurav Singh Tomar; Iulia Turc; David Reitter. (2023). Measuring Attribution in Natural Language Generation Models. Computational Linguistics 49(4), 777-840. https://doi.org/10.1162/coli_a_00486

[4] Nelson Liu; Tianyi Zhang; Percy Liang. (2023). Evaluating Verifiability in Generative Search Engines. Findings of the Association for Computational Linguistics: EMNLP 2023 , 7001-7025. https://doi.org/10.18653/v1/2023.findings-emnlp.467

[5] Tianyu Gao; Howard Yen; Jiatong Yu; Danqi Chen. (2023). Enabling Large Language Models to Generate Text with Citations. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing , 6465-6488. https://doi.org/10.18653/v1/2023.emnlp-main.398

[6] Zhang Kai; He Xinyue; Yao Jingang. (2026). From Citation Selection to Citation Absorption: A Measurement Framework for Generative Engine Optimization Across AI Search Platforms. arXiv preprint. 2604.25707v2. https://doi.org/10.48550/arXiv.2604.25707

[7] Julius Schulte; Malte Bleeker; Philipp Kaufmann. (2026). Don't Measure Once: Measuring Visibility in AI Search (GEO). arXiv preprint. 2604.07585v1. https://doi.org/10.48550/arXiv.2604.07585

[8] Ronald Sielinski. (2026). Quantifying Uncertainty in AI Visibility: A Statistical Framework for Generative Search Measurement. arXiv preprint. 2603.08924v3. https://doi.org/10.48550/arXiv.2603.08924

[9] A. Colin Cameron; Douglas L. Miller. (2015). A Practitioner’s Guide to Cluster-Robust Inference. Journal of Human Resources 50(2), 317-372. https://doi.org/10.3368/jhr.50.2.317

[10] B. Efron. (1979). Bootstrap Methods: Another Look at the Jackknife. The Annals of Statistics 7(1), 1-26. https://doi.org/10.1214/aos/1176344552

[11] Ro Encarnación; Tina Behzad; Emma Lurie; Danaé Metaxa. (2026). What Current AI Benchmarks Leave Unmeasured: Modality, Search, Citations, and Implications (for Safety Evaluations). arXiv preprint. 2608.06202v1. https://doi.org/10.48550/arXiv.2608.06202

[12] Donald B. Rubin. (1976). Inference and missing data. Biometrika 63(3), 581-592. https://doi.org/10.1093/biomet/63.3.581

[13] Danaë Metaxa; Joon Sung Park; Ronald E. Robertson; Karrie Karahalios; Christo Wilson; Jeff Hancock; Christian Sandvig. (2021). Auditing Algorithms: Understanding Algorithmic Systems from the Outside In. Foundations and Trends® in Human-Computer Interaction 14(4), 272-344. https://doi.org/10.1561/1100000083

[14] Nicola Ferro; Norbert Fuhr; Kalervo Järvelin; Noriko Kando; Matthias Lippold; Justin Zobel. (2016). Increasing Reproducibility in IR: Findings from the Dagstuhl Seminar on Reproducibility of Data-Oriented Experiments in e-Science. ACM SIGIR Forum 50(1), 68-82. https://doi.org/10.1145/2964797.2964808

[15] Timo Breuer; Jüri Keller; Philipp Schaer. (2022). ir_metadata: An Extensible Metadata Schema for IR Experiments. Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval , 3078-3089. https://doi.org/10.1145/3477495.3531738

[16] Ron Artstein; Massimo Poesio. (2008). Inter-Coder Agreement for Computational Linguistics. Computational Linguistics 34(4), 555-596. https://doi.org/10.1162/coli.07-034-R2

[17] OECD; European Union; European Commission, Joint Research Centre. (2008). Handbook on Constructing Composite Indicators: Methodology and User Guide. OECD Publishing, Paris. https://doi.org/10.1787/9789264043466-en

[18] Margaret Mitchell; Simone Wu; Andrew Zaldivar; Parker Barnes; Lucy Vasserman; Ben Hutchinson; Elena Spitzer; Inioluwa Deborah Raji; Timnit Gebru. (2019). Model Cards for Model Reporting. Proceedings of the Conference on Fairness, Accountability, and Transparency , 220-229. https://doi.org/10.1145/3287560.3287596

[19] Timnit Gebru; Jamie Morgenstern; Briana Vecchione; Jennifer Wortman Vaughan; Hanna Wallach; Hal Daumé III; Kate Crawford. (2021). Datasheets for datasets. Communications of the ACM 64(12), 86-92. https://doi.org/10.1145/3458723

[20] C. E. Shannon. (1948). A Mathematical Theory of Communication. Bell System Technical Journal 27(3), 379-423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

[21] M. O. Hill. (1973). Diversity and Evenness: A Unifying Notation and Its Consequences. Ecology 54(2), 427-432. https://doi.org/10.2307/1934352

[22] Alexandre Caramaschi. (2026). Algorithmic Authority: A Practitioner Framework for Generative Engine Optimization Based on a 7-Day Implementation Sprint. SSRN preprint. https://ssrn.com/abstract=6460680

[23] Alexandre Caramaschi. (2026). Three Ways to Fail to Conclude: A Null-report on Large Language Model Citation Claims for Brazilian Brands (N = 7,052, 12 days). SSRN preprint. https://ssrn.com/abstract=6636298

Appendix A. Reconstructing the analyses

The audit package preserves a frozen database, the extraction inputs, scripts and machine-readable tables. The database was downloaded from the identified GitHub Actions artifact rather than from the live application. Its SHA-256 digest is 7fcc98399419721b0507c230907b09bd4c7760198d27ef9ad5833156da9af3d7. This identifies the downloaded file; it is not a claim that a remote R2 object was independently hashed. The source manifest records the acquisition, integrity check and reconciliation with the earlier local database. A subsequent collection can increase the live database without changing the population analyzed here.

The historical cutoff is implemented as a timestamp strictly earlier than 1 September 2026 in the stored UTC convention. Canonical records are selected with is_probe = 0. These predicates identify a response population, not a set of independent users or a count of prompts in the lexicon. A repeated prompt contributes a further response record if it was persisted again. The historical inventory and the September extension therefore have separate inclusion rules and separate output tables. No historical response tail is reconstructed from the recent records.

reanalyse.py reproduces the source inventory, paired historical contrast, lexical-rule audit, corrected aggregation sensitivity and recent 200-character comparison. It loads frozen extraction code and literal alias and context constants from the supplied files. This avoids silently using a later version of the production configuration. The program reports its Python and numerical-library versions, the input digest, resampling seed and number of draws. Its CSV outputs preserve numerators and denominators alongside rates so that rounding in the manuscript cannot change the underlying calculation.

recent_windows.py evaluates a declared grid of prefixes in the September full-text field. It also forms the intersection of exact-query and UTC-date cells across the five represented services. Within each retained cell, repeated records are averaged; common cells then receive equal weight. This estimand differs from a rate pooled over all responses. It restricts the comparison to the overlap that was observed and gives no weight to a missing cell. Its source plan was recorded after the principal 200-character results were known, so this extension is explicitly exploratory.

Reproduction should begin by verifying the database and frozen-code digests, then running the scripts against a read-only connection. The resulting tables should agree in integer counts before displayed percentages are compared. A disagreement in a paired count should be traced through the inclusion predicate, text field, character slicing and extractor, in that order. Comparing only a rounded final score can conceal a change in any of those steps. The package also retains the analysis plans and review records, which document which questions preceded each exploratory extension.

A.1 Response detection and entity incidence

Let T_i denote the retained text of response i, $v(i)$ its vertical and E_v the frozen extraction function for that vertical. $P_W(T_i)$ is the first W Python Unicode characters of the retained string; P_{full} returns the entire retained string. The response-level indicator equals one if the extraction returns at least one eligible entry, and zero otherwise:

$$Y_i(W) = \mathbf{1}\{E_{v(i)}(P_W(T_i)) \neq \varnothing\}.$$

$$M(W) = \frac{1}{n}\sum_{i=1}^{n} Y_i(W).$$

$$\Delta_{\text{pp}}(W) = 100[M(W) - M(\text{full})].$$

The last expression is in percentage points.

The paired difference compares two transformations of the same records. The four combinations of $Y_i(W)$ and $Y_i(\text{full})$ must be counted explicitly. A prefix does not guarantee a subset of detected matches: removing context or creating a new string boundary can affect an extractor. The empirical absence of gains at a particular window is a result of this corpus and implementation, not a mathematical property assumed by the analysis.

Entity incidence retains identity. Let $X_{ij}(W)$ equal one when eligible entity j is detected in response i, and zero otherwise. Its prevalence is c_j/n, where c_j is the sum of X_{ij} over responses. Several entities can occur in one response, so the sum of these prevalence values can exceed one. A distribution of entity incidences instead uses $p_j = c_j/\sum_k c_k$. It assigns each detected entity-response event to one entity and sums to one when at least one event is observed. This distinction is necessary before applying entropy or concentration measures.

A.2 Diversity and weighting

For incidence shares p_j, Shannon entropy uses the natural logarithm, with a zero-share contribution defined by continuity as zero [20]. Its exponential, D_1, expresses the distribution as an effective number of equally frequent entities. The inverse concentration measure D_2 uses squared shares and places greater weight on frequent entities. Hill's formulation relates these measures within a family of effective diversity values [21].

$$H = -\sum_j p_j \ln(p_j), \qquad D_1 = \exp(H).$$

$$\text{HHI} = \sum_j p_j^2, \qquad D_2 = 1/\text{HHI}.$$

These are descriptions of the incidence distribution inside the eligible lexicon. They do not measure response quality, the completeness of the market or the number of organizations outside the lexicon. The count of eligible entities with no detections remains useful as a separate

result, because their absence is otherwise invisible after a distribution is normalized over positive incidences. If there are no incidences, the conditional distribution is undefined; assigning it an entropy or an effective entity count of zero would hide the absence of a distribution.

Giving every response equal weight represents the archive's realized mixture of queries and dates. Giving every query equal weight first calculates incidence rates within each query and then averages those rates over the eligible query set. The resulting vector is normalized only after that averaging. A difference between these summaries identifies sensitivity to query participation in this archive. It does not identify which weighting scheme represents actual user demand, since no demand distribution was sampled.

For two sets of detected entities A and B, Jaccard overlap is the size of their intersection divided by the size of their union. Both-empty pairs have an empty union. Reporting their frequency separately prevents a chosen convention, such as assigning them perfect agreement, from being mistaken for persistence of positively observed entities. Comparisons between adjacent observed dates should preserve the elapsed calendar gap: two successive observations can be separated by more than one day. A gap-aware summary describes the available sequence without interpolating the missing answers.

Appendix B. A prospective protocol for semantic validation

The following protocol is proposed work. It was not executed for this manuscript, and no automated agent was counted as a human annotator. Its purpose is to connect the existing lexical fields to judgments that the archive does not yet contain.

The annotation unit should preserve an exact prompt, its retained answer, the eligible entity and the observation window. For source-support questions, it must also include the cited passage or document snapshot used for verification. A name match, a claim about that name and support for that claim are separate judgments. An annotator should not infer one from another or treat a missing source as proof that a factual statement is false.

Entity identity is assessed first. The annotator determines whether the string refers to the intended entity, a homonym, an unresolved identity or no entity at all. Fictitious names require special care: a label assigned during cohort construction does not rule out a later real-world namesake or an ordinary-language use. Ambiguous cases remain unresolved until the supporting evidence is sufficient. The annotation record should identify that evidence and the date on which it was consulted.

The response's commitment is then described at claim level. It may explicitly reject the premise, express uncertainty, repeat a supplied name without asserting a fact, or make a factual attribution. An answer can contain more than one of these behaviors. Forcing such an answer into a mutually exclusive response label can erase an unsupported assertion that follows a

disclaimer. Claim-level decisions can be aggregated later under a declared response-level rule, with mixed cases retained in the record.

Support is assessed only for factual assertions that have been identified. A source can support the claim, contradict it or leave it unresolved. The judgment should refer to the proposition asserted, rather than to a general impression that the response sounds plausible. Where the available archive contains only a prefix, an annotator may judge the visible proposition but must not invent the missing continuation. A comparison between prefix and retained-answer labels should use matched records and disclose which judgments become indeterminate under truncation.

Sampling must include responses with and without lexical hits and with and without refusal-pattern matches. Inspecting only conspicuous positive cases cannot estimate false negatives or characterize the unselected responses. If rare strata are deliberately oversampled, the sampling probabilities should be retained for any prevalence estimate. Separate samples may be needed for entity matching and factual support, because they use different units and reference judgments.

At least two trained annotators should independently code the same pilot material, revise unclear rules, and then annotate the evaluation material under a frozen guide. Provider identity and existing automatic labels can be withheld where doing so does not remove evidence needed for the task. Disagreements should be preserved before adjudication. The report should include label frequencies, a confusion table and an agreement measure appropriate to the scale and missing-judgment policy; Artstein and Poesio [16] explain why agreement measures depend on these choices. Agreement alone does not establish that the shared judgment is correct.

Evaluation of an automatic detector requires a defined reference outcome. Precision and recall for entity identity do not validate a fabrication label. Each reported metric should name the construct, unit, evaluated window, sample size and unresolved cases. Uncertainty should respect repeated prompts and any stratified selection rather than assume every annotated item is an independent random draw. Sample size should follow a stated precision target and the pilot's class frequencies. A numerical target cannot be justified by borrowing the total number of stored responses.

This validation would permit a narrower, defensible use of the automatic fields. A rule with acceptable performance on one language, entity class and window could be used for screening under those conditions, with errors documented. Transfer to another language, service or retention regime would remain a new evaluation question. The present manuscript therefore retains the lexical interpretation of the historical fields and reports the absence of a completed semantic reference study.

Appendix C. Availability and meaning of the stored fields

The following inventory covers all 50 columns in the citations table of the 91,392-row snapshot. Availability means a non-null stored value, not successful validation of the construct named by the field. Input, provenance and operational fields are included to make the inventory complete. A zero-valued flag is distinct from a null value, but a pipeline default of zero need not establish a negative empirical finding.

Table C1. Complete citations-field inventory. Available values are counted over 91,392 rows.

Field	Available values	Operational meaning and limit
id	91,392	Database row identifier; not an independent sampling unit.
timestamp	91,392	Recorded UTC response timestamp; does not guarantee a new independent generation.
llm	91,392	Operational service label, not whole architecture or consumer UI.
model	91,392	Stored model identifier; does not guarantee provider immutability.
query	91,392	Exact elicitation text; repeated over observations.
query_category	91,392	Battery category; service coverage differs.
query_lang	91,392	Prompt language; not necessarily response language.
cited	91,392	Any detected cohort entity in stored observation text. Historical observation windows differ across services.
cited_entity	33,164	One detected entity; not all entities, sentiment, or recommendation.
cited_domain	0	Legacy entity/domain slot; missing values are not no-domain evidence.
cited_person	0	Legacy person slot; missing values are not no-person evidence.
position	33,164	Legacy positional tercile; version/window changes require care.
attribution	0	Legacy field; do not equate default/missing value with verified attribution.
source_count	91,392	Count of sources exposed by adapter; not retrieved documents or semantic support.
our_source_count	0	Legacy source counter; unused/missing does not measure absence.
hedging_detected	0	Legacy flag; no human validation established.

Field	Available values	Operational meaning and limit
response_length	91,392	Length of stored observation text, not unconstrained model output.
response_text	91,392	Retained observation text; historically truncated in five services.
sources_json	91,392	Exposed source list; no inference about internal retrieval or entailment.
latency_ms	91,392	Adapter latency metadata; endpoint/cache/transport confounding, not cognitive difficulty.
token_count	91,392	Provider/adapter token metadata; accounting semantics can differ.
created_at	91,392	Database creation timestamp, possibly after response/annotation date.
vertical	91,392	Vertical-specific cohort and templates.
model_version	91,392	Additional model metadata; inspect availability before use.
query_type	91,392	Directive/exploratory template metadata; retrospective corrections exist.
fictional_hit	91,392	Lexical appearance of fictional target or provider entity list; refusal can trigger.
fictional_names_json	91,392	Names triggering fictional flag; no semantic fabrication judgment.
cited_v2	91,392	Versioned any-entity flag; same construct as cited under v2.
cited_count_v2	91,392	Number of unique extracted cohort entities in observed text.
cited_entities_v2_json	91,392	Versioned set of extracted names, conditional on cohort/window.
first_entity_v2	33,164	First entity returned by extractor; canonical/alias ordering limitation.
first_entity_offset_v2	33,164	Offset in normalized, markup-stripped text; not necessarily raw response offset.
position_v2	33,164	Tercile based on observed window; not original document rank.
via_alias_count_v2	91,392	Count of aliases used by extractor; not external validation accuracy.
via_fold_count_v2	91,392	Count of diacritic-fold matches; not uncertainty probability.

Field	Available values	Operational meaning and limit
response_length_chars_v2	91,392	Versioned observation text length; must not imply full history retained.
extraction_version	91,392	Stored extraction version label.
extracted_at_v2	91,392	Extraction timestamp; can differ from collection time.
response_hash	91,392	Stored truncated SHA-256; text fingerprint does not prove cache or independence.
is_probe	91,392	Canonical/probe separation; exclude probes from canonical estimands.
probe_type	19,247	Type of elicitation probe.
adversarial_framing	91,392	Whether the query deliberately cues target or framing.
fictitious_target	19,247	Target name supplied in the probe prompt.
is_calibration	91,392	Probe/control designation; not evidence of calibrated probabilistic predictions.
query_type_v1_legacy	4,284	Legacy query-type value preserved after correction.
selection_status	33,052	Slug match against exposed URLs, using cited entities or cohort fallback; empty source list treated as zero. Not internal retrieval.
absorption_status	33,052	Implementation sets this equal to cited; not independently measured semantic absorption or entailment.
failure_type	12,303	Rule-derived label; fictional_hit can imply hallucinated-source despite refusal. No gold-standard validation.
response_full_text	7,906	Full payload retained by adapter from September; token/API limits still apply.
citation_window_chars	7,906	Length of actual observed text; not a universal optimal cutoff.

SQLite's declared column type does not establish the semantic type of a populated value. For example, the field named cited_entity has a Boolean declaration but stores entity-label strings. Analysis therefore used the observed contents and implementation definitions rather than inferring meaning from the schema alone.

Empty tables and unvalidated contextual fields also constrain the available research questions. The archive contains no recorded intervention outcomes, independently verified URL evaluations or paired natural/structured-output experiments in the corresponding analytical tables. Consequently, the paper does not estimate content-intervention effectiveness,

citation-support accuracy or an elicitation-format effect from those structures. Operational cost and scheduling records describe the collection process; they are not measures of answer quality or commercial return. The full inventory and table-level counts remain machine-readable in the audit package.

Appendix D. Reconciliation with the BRGEO-1 specification draft

The BRGEO-1 specification draft reports additional analyses of an earlier snapshot of the same collection. Reproducing their numerical results does not establish every interpretation attached to them. This appendix preserves three reproducible descriptive analyses, identifies their populations and weighting, and explains which interpretations are not adopted. The detailed reconciliation is retained in the audit package.

D.1 Snapshot boundaries and observation rules

The specification draft used 86,543 records ending at 2026-09-08T19:30:57.635649+00:00. Selecting records whose is_probe is zero, treating null as zero, leaves 68,624 canonical responses, including 2,225 with a populated response_full_text field. These span 53 UTC dates, or 52 dates after subtracting three hours. The present study uses the 91,392-record artifact ending at 2026-09-11T11:07:23.385313+00:00, containing 72,145 canonical responses and 5,746 canonical responses with that field populated.

The earlier population is recovered from the later artifact by retaining timestamps at or before the earlier exact boundary. A cutoff at the end of 8 September instead includes another 864 canonical responses, yielding 69,488 canonical records and 3,089 with the dedicated full-text field. The date alone therefore does not reproduce the draft. Both snapshots contain the same 66,399 canonical responses before 1 September, including 7,436 historical Perplexity responses with longer text retained in response_text.

The three analyses below use the earlier canonical population. The frozen extractor is applied to the first 200 Python Unicode code points of response_text, before its normalization and matching operations, using the vertical-specific lexicon with anchors and decoys included. Detection means that at least one eligible entry is returned. The 2,225-response window extension uses records available within each service. It differs from the present study's recent comparison, which first averages repeated responses within each service/query/date cell and then weights the 272 cells shared by five services equally.

D.2 Agreement on binary detection

For agreement, records are ordered by timestamp and then database identifier. Only the earliest record for each exact query, UTC date and service is retained, leaving 42,487 cells. The ChatGPT, Claude and Gemini intersection contains 9,129 query/date cells across 50 dates. Giving these cells equal weight reproduces Fleiss' κ of 0.086089. The calculation compares mean within-cell pairwise agreement with chance agreement based on the pooled binary prevalence across the three services.

This outcome allows services naming different entities to agree: both receive a positive detection flag. The result therefore concerns agreement on whether any eligible entry appears, not agreement about the same company, recommendation or source. The intersection also does not contain every one of the 192 queries on every date. The draft's date-cluster bootstrap assesses sensitivity to the observed dates while retaining the fixed battery; it does not jointly account for repeated queries and dates or remove the mixture of Gemini configurations. Low agreement on this outcome does not establish that a service-specific measurement has no predictive value elsewhere.

D.3 Residual non-additivity in aggregated coverage

The same 42,487 reduced cells produce a matrix of 62 labels detected by at least one of the six services. Each label's rate uses that service's retained cells within the label's vertical as its denominator. The 372 rates receive equal weight. A common offset, one divided by twice 42,487, is added to each rate and its complement before taking the natural logarithm of their ratio.

Subtracting the label and service means from these transformed rates, while restoring the grand mean, gives the residual. Decomposing total squared deviation assigns 34.46% to label means, 26.21% to service means and 39.33% to residual non-additivity. These percentages reproduce the draft. The 62 labels differ from the 66 detected in all 68,624 responses because replicate reduction removes some detections.

There is one aggregate rate per matrix cell. The residual cannot separate interaction from uncertainty in those rates, differences in query and calendar composition, configuration changes or sensitivity to the offset. Groq and Grok have no contemporaneous overlap. Equal weighting was evaluated; invariance under other weights was not established. The residual is not an identified causal interaction or evidence that temporal variation has been excluded.

D.4 A variance diagnostic for the pooled mean

Without replicate reduction, the pooled binary mean is 0.179660 across 68,624 responses. Grouping by service and exact query creates 1,056 clusters. For each cluster, the calculation subtracts the pooled mean times cluster size from the cluster's positive count. Multiplying the sum of squared residuals by the ratio of 1,056 to 1,055 and dividing by the squared total record count gives the cluster-based variance estimate for this ratio mean.

The independent-binomial variance is the pooled mean times its complement, divided by 68,624. Dividing the cluster-based variance by this quantity reproduces a design effect of 63.344242. Dividing 68,624 by this value gives 1,083.350. These are diagnostics of this observation-weighted mean under this grouping, not a universal effective sample size. The grouping does not jointly model shared queries across services and shared dates across clusters. It also does not turn the purposively selected battery into a probability sample. Neither the

design effect nor its implied sample-size equivalent is transferred to the paired window contrasts, agreement statistics or entity-specific estimates.

D.5 Measures and interpretations not adopted

The draft's repeat-agreement statistic combines the proportion of groups whose observations all agree with a two-observation chance term derived from global prevalence. Some groups contain three or four observations. This construction is not adopted as a standard paired test-retest κ or as a ceiling on between-service agreement. A paired estimate would require an explicit replicate-selection rule and the corresponding paired marginals.

Zero ties do not make Spearman correlation undefined when the complete vectors vary. For the earlier 68,624-response series panel, the draft's geometric composite index has a Spearman correlation of 0.997147 with coverage across all 127 lexicon entries, including 61 undetected entries. Positive-only correlations remain legitimate conditional summaries, provided that their population is stated. Neither agreement nor correlation supplies the semantic validation that these lexical measurements still require.