Where the Questions Come From: Prompt Provenance as a Bound on Generative-Search Visibility Measurement

Muhammad Arshad
Telic Technologies Pvt Ltd · MSc, KTH Royal Institute of Technology
arshad@telictech.com

Preprint, version 1 — 10 September 2026. Not peer reviewed.
https://doi.org/10.5281/zenodo.22694994

> **Competing interest, stated first.** The author builds and operates AnswerRange, a product in the category this paper examines. That is a direct commercial interest in the findings and the reader should weight the paper accordingly. Three things are done to make the work checkable rather than asking for trust: both experiments are run on the author's own instrument and the raw observations are released; the author's own product is subject to the same criticism as every other, and Experiment 2 is a failure of the author's own default configuration; and the vendor census in §7 is coded only from public documentation, against a protocol published in full so that any reader can reproduce it rather than take it on trust (§7.1 explains why the products are not named, and what that costs). Nothing here depends on a judgement the reader cannot re-make from the released material.

Abstract

Commercial products report how often generative search engines cite a business. Every such number is conditional on a question set, and that question set is a free parameter: the same brand, measured on the same platforms in the same week, can be assigned very different visibility depending on where its questions came from. We name this parameter **prompt provenance** and give it a three-class taxonomy — natively authored, machine-generated, and observed search queries.

We report two experiments. In the first, a single brand is measured twice on the same five generative-search surfaces with the same replicate structure, differing only in corpus provenance, in **four language markets** (English, German, Spanish, French; 864 observations). The machine-generated corpus returns the higher citation rate in **all four**, but the size of the gap varies more than thirteenfold: **−24.1 points in English ($p = 5.9 \times 10^{-4}$), −18.5 in Spanish ($p = 0.009$), −5.6 in German and −1.9 in French (both within noise)**. Pooled across the three non-English markets the difference is −8.6 points ($p = 0.034$).

That variability is the finding, not a weakness of it. A practitioner cannot know in advance whether switching corpus will move their headline number by two points or by twenty-four, which makes the parameter's disclosure more necessary rather than less. The direction is informative too: the machine-generated corpus was the more flattering one every time.

In the second, a leather-goods retailer is measured against the nearest available category corpus and returns **0 of 108 (95% CI 0–3%)** — a precise-looking zero produced by a corpus that asks about winter coats and merino base layers. Re-measured against a purpose-built leather corpus it returns **0 of 90 (95% CI 0–4%)**, which is a real finding. The two zeros are statistically indistinguishable: identical instrument,

comparable intervals, platforms answering normally. Only provenance separates the empty measurement from the informative one.

We then show that in a census of 16 commercial products, **none discloses the provenance of its prompts**, and five do not disclose the prompt count at all. Our claim is not that these products are inaccurate. It is that a visibility figure without a stated provenance is not interpretable, that the effect is large enough to exceed the period-over-period changes such products report, and that disclosure costs nothing.

Keywords: generative search, AI visibility, algorithm auditing, measurement validity, prompt provenance, construct validity

---

1. Introduction

A measurement of the form *"brand X is cited in N% of AI answers"* is never a property of brand X alone. It is a property of brand X **and a set of questions**. Change the questions and the number changes, without anything about the brand, the platforms, or the world having changed at all.

This is not a subtle point, and it is not new: it is ordinary construct validity. What is new is that a commercial category has grown up around producing exactly this number, at scale, for businesses that cannot see the question set and are not told where it came from.

The prior paper in this series (Arshad, 2026b) examined how *precisely* these figures are reported relative to their sample sizes. This paper examines something logically prior. Before asking whether a rate is precise, one must ask what it is a rate *of*. We argue that the answer depends on a parameter that is currently undisclosed across the entire category, and we show experimentally that the parameter is not small.

1.1 Contributions

1. A three-class taxonomy of **prompt provenance** (§3), with the different claims each class can support.
2. **Experiment 1** (§4): a within-brand, within-platform comparison of two corpora differing only in provenance, run in four language markets. The effect is directionally consistent (4/4) but varies from 1.9 to 24.1 points, with English and Spanish individually significant after correction.
3. **Experiment 2** (§5): a demonstration that a provenance mismatch can produce a confident zero that is arithmetically correct and substantively empty.
4. A census showing that **0 of 16** commercial products disclose provenance (§7), with the full coding sheet released.
5. A minimal disclosure standard that would make these figures interpretable (§9).

---

2. Background and related work

The audit-study designs that this work belongs to were set out for algorithmic systems by Sandvig et al. (2014), who adapted the social-scientific audit — long used to test for discrimination in housing and employment — to internet platforms, and who treat the queries an auditor issues as part of the audit design rather than as a neutral window onto the system.

That the query set materially changes what an audit sees is not a conjecture. Hannák et al. (2013), measuring personalisation in web search, found that on average 11.7% of results differed between users, and — the part that matters here — that the degree of difference **varied widely by query**. An auditor choosing a different query set would have measured a different amount of personalisation in the same system on the same day.

Sampling-frame effects are the older and more general version of the same problem: in survey methodology the frame a sample is drawn from bounds the population any result can describe, and coverage error is treated as a property of the instrument that must be reported rather than a nuisance to be absorbed (Groves et al., 2009).

We are not claiming novelty for the underlying idea. The idea is standard. What we claim is that a commercial category now sells a number that depends on this parameter, does not disclose it, and — as §4 shows — the dependence is large enough to matter.

What is specific to generative-search visibility measurement is the combination of three properties:

- the buyer of the measurement usually cannot see the question set;
- the vendor is commercially free to choose it; and
- the resulting number is sold as a property of the buyer's brand.

Under those conditions the question set is not merely a methodological detail. It is the part of the instrument with the most freedom and the least scrutiny.

---

3. A taxonomy of prompt provenance

We distinguish three classes by how the questions come into existence.

(P1) Natively authored. Questions written by a person for the market and category in question, in that market's language. Costly, does not scale, and carries the author's assumptions about what buyers ask. Its virtue is that someone can be asked to defend each question.

(P2) Machine-generated. Questions assembled from a keyword or search-volume source for a category, or generated by a language model from a category description. Scales freely and is the dominant approach where any method is visible at all. Its weakness is that it describes a *category*, not a *business*: two competitors in the same category receive the same questions regardless of what they actually sell.

(P3) Observed search queries. Questions taken from queries the business is demonstrably found for — for example, its own search-console performance data — with branded queries removed. Requires per-business data access. Its virtue is that the question set is anchored to real demand for that specific business.

The classes are not ranked by quality in the abstract. They support different claims:

Provenance	The rate is a rate of…	Supports the claim
P1 natively authored	questions a human judged representative	"…on questions we can defend"
P2 machine-generated	questions typical of a category	"…on category-typical questions"
P3 observed queries	questions this business is actually found for	"…on this business's own demand"

A vendor reporting P2 as though it were P3 is not lying about arithmetic. It is answering a different question from the one the buyer asked.

3.1 Why branded queries must be excluded from P3

A corpus built from a business's own search data will contain queries naming the business. An assistant that names a brand in response to that brand's own name demonstrates nothing about visibility, and including such queries inflates the rate — most for the businesses with the strongest brands, which is precisely backwards. All P3 corpora in this paper exclude branded queries, including misspellings within one edit of the brand token; the real corpus in Experiment 1 contained four such misspellings that exact matching missed.

---

4. Experiment 1 — the same brand under two provenances

4.1 Design

A single retailer (`decomica.com`, designer-furniture reproductions, English market) was measured twice on the same instrument, **on the same day, forty-five minutes apart**. The two arms differ in corpus provenance and in nothing else we control:

	Arm A	**Arm B**
Provenance	P2 machine-generated	P3 observed search queries
Source	category search volume	the site's own Search Console, 90 days, brand-stripped
Questions	12	12
Platforms	AI Overviews, AI Mode, Gemini, Perplexity, ChatGPT	same five
Replicates	2× (1× ChatGPT)	2× (1× ChatGPT)
Observations	108	108
Run at	9 Sep 2026, 18:21 UTC	9 Sep 2026, 17:36 UTC

The order was not randomised: Arm B ran first. With a 45-minute gap this is a minor concern, but it is not zero, and a randomised repeated design is the obvious next step.

4.2 Result

	Cited	**Rate**	**95% CI (Clopper–Pearson)**
Arm A — machine-generated	61/108	**56.5%**	47–66%
Arm B — observed queries	35/108	**32.4%**	24–42%

Difference: **24.1 percentage points**. Two-sided Fisher exact $\mathbf{p = 5.9 \times 10^{-4}}$.

Per platform, with Holm–Bonferroni correction across the five comparisons:

Platform	Arm A	Arm B	p	Holm-adjusted
Google AI Overviews	12/24	2/24	0.0034	**0.017**
ChatGPT	4/12	0/12	0.093	0.373
Gemini	19/24	14/24	0.212	0.637
Perplexity	22/24	18/24	0.245	0.637
Google AI Mode	4/24	1/24	0.348	0.637

Only Google AI Overviews survives correction individually; the effect is otherwise carried by the pooled comparison. A reader should not treat the five rows as five independent findings, which is why they are corrected as one family.

4.2.1 Prior run, three days earlier

An earlier pair of runs three days apart, on the four platforms then available, gave the same result at smaller magnitude: 50/84 (59.5%) for the machine-generated corpus against 34/84 (40.5%) for observed queries, p = 0.020. That comparison confounds corpus with three days of platform drift; it is reported here only because the same-day replication came out **larger**, which is evidence that drift was suppressing rather than producing the effect.

4.2.2 Three further language markets

The English comparison was repeated in German, Spanish and French on the same brand and the same five platforms, 12 questions per market per arm, 108 observations per market per arm (648 additional observations).

Market	Machine-generated	Observed queries	Difference	p	Holm
English	61/108 — 56.5%	35/108 — 32.4%	**−24.1pp**	5.9×10^{-4}	—
Spanish	69/108 — 63.9%	49/108 — 45.4%	**−18.5pp**	0.009	**0.028**
German	58/108 — 53.7%	52/108 — 48.1%	−5.6pp	0.496	0.993
French	53/108 — 49.1%	51/108 — 47.2%	−1.9pp	0.892	0.993

Holm–Bonferroni applied across the three non-English markets, which were run as one family. English was a separate prior experiment and is not part of that correction.

Pooled across German, Spanish and French: 180/324 (55.6%) against 152/324 (46.9%), a difference of −8.6 points, p = 0.034.

Two things follow, and they pull in different directions.

The direction is consistent. The machine-generated corpus produced the higher rate in all four markets. Under a null of no effect the probability of four arbitrary signs agreeing is 0.125 (two-tailed exact sign test) — suggestive on its own, and more persuasive taken with the two individually significant markets, but not conclusive from four observations.

The magnitude is not predictable. It ranges from 1.9 to 24.1 points, a thirteenfold spread, on one brand within one category. Whatever drives the size of the gap is not something a practitioner can read off in advance.

We regard the second as the more useful result. A fixed, known bias would be correctable by a footnote. An effect whose size varies by an order of magnitude between the markets of a single business cannot be corrected for — it can only be disclosed.

4.3 Interpretation

The machine-generated corpus produced the *higher* number. Inspection of the two question sets suggests why: the category corpus is weighted toward reproduction-specific long-tail phrasing on which this retailer ranks well, whereas its own observed demand includes head terms (e.g. a two-word product category on which it averages position 31.9 in classic search) where it does not. The category corpus contained a disproportionate share of the questions this business was already going to win.

The same inspection does not explain the *variation* between markets, and we do not have a mechanism to offer for it. The English and Spanish corpora diverged most from their category equivalents; the French and German ones least. Whether that reflects the maturity of the site's demand in each market, the structure of the category corpus in each language, or something about the platforms is beyond what four markets can settle.

What we do claim is narrower and survives the variation: **in the market where it mattered most, the choice of provenance moved the headline figure by 24 points — larger than most of the period-over-period changes these products are sold to detect, and larger than the minimum change their disclosed sample sizes could resolve at all — and a practitioner had no way of knowing in advance whether their market would be the 24-point case or the 1.9-point case.**

4.4 Threats to validity

Order effect, not temporal drift. The principal comparison is same-day and 45 minutes apart, so platform drift is largely controlled. What is not controlled is order: Arm B ran first. Generative-search surfaces are unstable between runs — our earlier work found one platform contradicting itself on roughly 40% of repeated identical questions — so a randomised, repeated-measures design is the correct next step.

Single brand, single category, single market. n = 1 business. This is a demonstration that the effect exists and is large, not an estimate of its typical size.

Both corpora built by the same authors. Arm A is our own category corpus, not a vendor's. We cannot claim it is representative of commercial P2 corpora generally.

Correlated observations. Repeated issues of one question are not independent draws, so the intervals above are, if anything, narrower than the truth.

---

5. Experiment 2 — when the provenance is simply wrong

5.1 Design and result

An independent leather-goods retailer (leather jackets, bags, belts, equestrian saddles; English market) was measured against the nearest available category corpus, `fashion-apparel`, on five platforms, 12 questions, 108 observations. The business is not named: it is a third party that did not ask to be measured, what is reported about it is a zero, and its identity carries no part of the argument, which is entirely about the two

corpora. Every observation is released, and both corpora with them, so the experiment is fully reproducible without the name.

Result: 0 of 108. 95% CI 0–3%.

The corpus asked, among others: *best men winter coat*, *best waterproof rain jacket women*, *best merino base layer*, *best denim jeans brand*.

5.2 Why this matters more than a null result usually does

The measurement is not broken. The assistants answered normally — 100% answer rates on four of five platforms — and cited 336 distinct pages, of which we retrieved 262 and found the retailer on none. Every number in the report is correct.

The report is nonetheless empty of information about this business, because the questions describe a different business. And the failure is *invisible in the statistics*: an observed 0/108 carries a **narrower** interval than any interior value, so the least informative measurement in this paper is also the one that looks most precise. At n = 108 the smallest non-zero rate distinguishable from zero is 6/108 (5.6%), so the interval is genuinely narrow — it is the construct, not the arithmetic, that has failed.

This is the sharp form of the paper's thesis. In Experiment 1, provenance moved the answer. In Experiment 2, provenance determined whether there was an answer at all, while leaving every visible quality indicator intact.

5.2.1 The same brand, two corpora, two zeros — one empty, one real

The obvious objection to §5.1 is that the retailer may simply not be cited. We tested it. A leather-goods corpus was built for the same market from observed search demand (10 questions: *best brand for leather jacket*, *best quality leather bag*, *best leather men wallet*, and so on) and the same brand was measured again on the same five platforms.

The result was again 0 — 0 of 90 (95% CI 0–4%).

The two zeros are not equivalent. The first was produced by a corpus about winter coats and merino base layers and carries no information about a leather retailer. The second was produced by a corpus about leather goods and is a real finding: across 595 citations to 210 distinct domains on leather buying questions, this retailer was named in none. That second measurement also supports an interpretation the first could not — the domains that *are* cited are overwhelmingly editorial and community sources (Reddit, YouTube, Forbes, John Lewis, specialist review sites) rather than retailers, which locates the route in as being covered by those publications rather than displacing them.

The two measurements are statistically indistinguishable. Both are zero, both have narrow intervals, both come from a correctly functioning instrument against platforms answering normally. Nothing in the output separates the meaningless zero from the meaningful one. Only the provenance of the question set does — which is precisely the parameter no product in §7 discloses.

We regard this as the strongest single argument in the paper. A disclosure requirement is not a matter of scholarly tidiness; without it, two reports that look identical can differ in whether they contain any information at all.

5.3 Disclosure

This was our own product's default behaviour, found by running it on a third-party site during ordinary testing. The instrument now lists the questions on every report and states, on a zero result, that the cause may be a corpus mismatch rather than absence. We report it here because a category-mismatch failure that is invisible in the output is exactly the kind of thing a vendor has every incentive not to publish.

---

6. Implications for interpretation

If provenance is undisclosed, then for any reported figure the reader cannot distinguish:

- a business with low visibility on its own demand, from
- a business measured against a category it does not occupy;

nor can they distinguish:

- a genuine period-over-period improvement, from
- a change in the corpus between periods.

The second is worth emphasising. A vendor that regenerates its category corpus — a routine maintenance action — may produce a change of the magnitude in §4.2 in a customer's headline number with nothing having happened in the world. Since corpora are not disclosed, neither the vendor nor the customer is in a position to rule this out after the fact.

---

7. What the category discloses

We coded 16 commercial products from their public documentation as of 8 September 2026. The coding sheet is released with this paper in de-identified form: products appear as Product A–Q, prices are banded, and the exact engine lists, source URLs and per-product verbatim notes are withheld because each is a fingerprint that would identify the row in seconds.

On provenance specifically: 0 of 16 disclose where their prompts come from. Some describe a prompt *count*; five do not disclose even that. None describes how the questions were produced, by whom, or from what source, and none offers the customer sight of the question set.

The census is deliberately narrow and is not the argument of this paper. It establishes one fact — that the parameter shown above to matter is, at present, universally undisclosed. We did not observe any vendor's output and make no claim about its accuracy.

7.1 Why the products are not named, and what that costs

Naming is the scholarly default and we have departed from it deliberately.

The reason is not doubt about the coding. Every cell was read from a page the vendor published about itself, and we stand behind each one.

It is proportionality. The argument of this paper is about a category-wide practice, and it is carried entirely by the aggregate: whether *any* product discloses provenance. No step in it depends on which product occupies which row. Attaching names would add adversarial weight to a claim that does not need it, from an

author who competes with the products being coded — a position from which naming is easy to read as a competitive act whatever the intent. Withheld names cost the argument nothing and remove that reading entirely.

We state the cost plainly. A de-identified census is weaker than a named one. A reader cannot audit our coding of any particular product, and must either take the aggregate on trust or reproduce it. That is a genuine loss and we will not describe it as anything else — a paper arguing for disclosure has no business being coy about its own.

What we offer instead is reproducibility. The coding protocol in Appendix B states the criteria and the decision rules in enough detail that a reader can apply them to any product in the category in about ten minutes each. The claim in this section is not "trust our table"; it is "run this on five products of your choosing and see whether you find a provenance statement". We expect a reader who does so to find what we found, and the claim should be discarded if they do not.

Two consequences follow and we accept both. The paper cannot support any statement about an individual product, and none is made. And no right of reply was sought, because there is no allegation against a named party to reply to.

---

8. Limitations

Beyond §4.4:

- **Arm order was not randomised**, and each market pair was run once. The point estimates should be treated as provisional until a randomised repeated design is run across several brands.
- **One brand, four markets.** Varying market while holding the brand constant isolates the corpus cleanly, but it cannot separate a market effect from a brand-in-that-market effect. Four markets is also too few to characterise the distribution of the gap; we can say it varies, not how.
- **The non-English arms ran a day after their controls**, not 45 minutes. Platform drift is therefore not excluded for those three, and the English pair remains the tightest comparison in the paper.
- **P3 requires data access that not every measurement context has.** We are not arguing that all measurement should be P3; we are arguing that provenance should be stated.
- **Our own instrument is the measuring device for both experiments.** A replication on a second instrument would strengthen the result considerably.
- **Branded-query removal is our definition, not a standard one.** Different exclusion rules would yield different P3 corpora.

8.1 P3 has a floor, and it excludes the businesses most likely to need it

Observed-query provenance requires that a business *has* observed queries. We checked two further properties under the same ownership as the Experiment 1 brand — a Spanish and a German store in the same category — intending to run the experiment in three markets. Neither could support a P3 corpus at all. Over 90 days, no query on either site cleared a 30-impression floor, and most sat beyond position 40.

A corpus could have been built from what was there: queries with two impressions in ninety days. We declined, because a rate measured against such a corpus would carry the same precise-looking interval as any other and would be an artefact of whichever half-dozen queries happened to register. That is Experiment 2 by another route.

This is a real limitation of the class, not of one implementation:

- **P3 is unavailable to small sites**, which are disproportionately the businesses buying a visibility measurement in the first place.
- Any vendor offering per-customer corpora therefore has a silent **eligibility threshold**, and a customer below it will receive a P2 measurement while believing they are receiving a P3 one — unless the vendor discloses which they got.
- The threshold is a judgement, not a constant. Ours is deliberately conservative. A vendor with a lower one is not wrong, but the customer cannot compare two products without knowing both.

This strengthens rather than weakens the paper's recommendation. It is a further reason the provenance class must travel with the figure: it can differ between two customers of the same product, and between two periods for the same customer, without either being told.

9. A minimal disclosure standard

We propose four lines, all cheap, none requiring a vendor to reveal a corpus:

1. **Provenance class** of the prompt set (P1 / P2 / P3, or a stated hybrid).
2. **Prompt count**, and whether the set changed since the previous period.
3. **Replicate count** per prompt per period.
4. **Whether branded queries are included.**

A figure accompanied by those four facts is interpretable. Without the first, the rest does not rescue it.

---

10. Conclusion

The question set is part of the instrument. Where it comes from changes the answer by an amount comparable to the changes these products exist to detect, and in the mismatch case determines whether the answer means anything at all. The category currently discloses none of this. Disclosure is four lines long.

References

Arshad, M. (2026a). *Cited or Invisible? A Multilingual, Multi-Platform Audit of Brand Citation in Generative Search.* Zenodo. https://doi.org/10.5281/zenodo.22491737

Arshad, M. (2026b). *How Much Measurement Does an AI Visibility Claim Need? Replicate Instability and Minimum Detectable Effects in Generative Search Audits.* Zenodo. https://doi.org/10.5281/zenodo.22480733

Groves, R. M., Fowler, F. J. Jr., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey Methodology* (2nd ed.). Hoboken, NJ: Wiley. ISBN 978-0-470-46546-2.

Hannák, A., Sapieżyński, P., Molavi Kakhki, A., Krishnamurthy, B., Lazer, D., Mislove, A., & Wilson, C. (2013). Measuring personalization of web search. In *Proceedings of the 22nd International Conference on World Wide Web (WWW '13)*, 527–538. https://doi.org/10.1145/2488388.2488435

Sandvig, C., Hamilton, K., Karahalios, K., & Langbort, C. (2014). Auditing algorithms: research methods for detecting discrimination on internet platforms. Paper presented at *Data and Discrimination: Converting Critical Concerns into Productive Inquiry*, a preconference of the 64th Annual Meeting of the International Communication Association, Seattle, WA.

Data availability

All 1,062 observations behind both experiments, the de-identified coding sheet and the analysis script are released under CC BY 4.0 at https://doi.org/10.5281/zenodo.22694994 (10.5281/zenodo.22694994).

Each observation is one line of JSON carrying the question as it was put to the platform, the market, the platform, the domains cited in the answer, and whether the focal brand was among them. The questions are released in full — a paper arguing that prompt provenance must be disclosed cannot withhold its own prompts. `README.md` in the deposit is the codebook.

`provenance-analysis.js` recomputes every figure in §4 and §5 from those files alone, with no dependencies and no network access: `node provenance-analysis.js`. Nothing in it is hard-coded, so if a figure here and a figure it prints ever disagree, the data are right and the paper is wrong.

One redaction, and what it costs. The questions in the P2 arms are released in full. The questions in the P3 arms are not: they are one live business's own search demand, and releasing them would publish its competitive position to serve a methodological point. For each P3 observation we release the stable prompt id and the word count — which is what the head-term reading in §4.3 rests on — and every outcome field, so all figures recompute exactly.

We state the cost. A reader cannot inspect the P3 questions and judge for themselves whether they look like head terms. But note what the release policy does *not* change: a P3 arm is not reproducible by a third party under any policy, because reproducing it requires that business's Search Console access. That is a property of the provenance class, not of this deposit, and it is the same asymmetry §8.1 describes. It is also why the disclosure standard in §9 asks a vendor to state the provenance class rather than publish the corpus.

Appendix A — Analysis

`provenance-analysis.js` reproduces every figure in §4 and §5, reading only the deposited observations. Exact Clopper–Pearson intervals throughout; two-sided Fisher exact for every comparison; Holm–Bonferroni within each family. No normal approximation is used anywhere.

Two decisions in it are worth stating here. Observations whose request failed in transport are excluded. Observations where the platform *declined to answer* are **not** excluded: a refusal to answer is a real observation of that platform's behaviour, and dropping those rows would flatter every rate in the paper.

Appendix B — Coding protocol and de-identified sheet

Released as `coding-sheet-anonymised.csv`: one row per product (Product A–Q), one column per coded criterion, with the access date for every row.

Sampling frame. Products marketed as measuring brand visibility, citation or mention in generative search engines, discoverable from public listings and search as of 8 September 2026, that publish a price and a feature list without requiring a sales call. Products behind an enterprise-only gate are excluded, and that exclusion is a limitation: they may disclose more.

Coding rules. Each criterion is coded from the vendor's own public pages only — pricing page, documentation, changelog. No inference from output, no third-party reviews, no sales conversations.

- *replicates_stated* — **yes** only if the vendor states how many times a prompt is issued per period. A statement of *frequency* ("daily") is not a replicate count and is coded **no**.
- *uncertainty_stated* — **yes** only if an interval, margin, confidence or sample-size qualifier appears alongside a reported figure. A denominator mentioned elsewhere in the documentation does not qualify.
- *sampling_method_stated* — **yes** if the vendor states how prompts are selected or sampled.
- *prompt_provenance* — **yes** if the vendor states where its prompts come from: who or what produced them, from what source. A prompt *count* is not provenance. "AI-generated prompts" with no source is coded **yes, P2**; silence is coded **not assessed**.
- *engine_type* — whether the engines measured are AI search products end users use, raw model APIs, or a mix.

To reproduce. Take any five products in the category. For each, open the pricing page and the documentation and apply the five rules above. The claim under test is that none states prompt provenance.