How Much Measurement Does an AI Visibility Claim Need? Replicate Instability and Minimum Detectable Effects in Generative Search Audits

Muhammad Arshad

Independent researcher · arshad@telictech.com

6 September 2026

Abstract

A commercial category now sells businesses a measurement of how often generative search engines cite them, typically from a few dozen prompts sampled once. We ask what such a measurement can actually support. Using 4,320 observations in which identical queries were issued three times to each of four generative answer surfaces, we quantify three things that this literature does not report: the rate at which repeated identical queries disagree with themselves; how the width of a confidence interval on a citation rate falls with the number of prompts sampled; and the minimum difference two samples of a given size can distinguish from noise.

Instability is substantial and differs sharply by platform: from 1.4% to 40.3% of repeated prompt-platform cells return non-unanimous outcomes, meaning that on one widely used assistant a repeated identical question contradicts its own earlier answer two times in five. At twelve prompts — a plausible commercial configuration — the 95% interval on a platform citation rate spans 53 percentage points on the least precise platform, and the smallest difference resolvable between two samples ranges from 25 to 36 points. It follows that a monitoring product reporting month-over-month movement at that sample size is, for much of its range, reporting variation below its own detection threshold.

We give a sample-size table practitioners and researchers can use directly, and release the harness and observations so the thresholds can be recomputed as the platforms change. We do not claim the underlying products are without value; we claim that the precision at which their numbers are presented is not supported by the sampling behind them, and that this is straightforwardly fixable by reporting intervals and replicate counts.

1. Introduction

Generative search engines answer commercial questions by naming a few sources. Businesses have begun paying to know how often they are among them, and a category of "AI visibility" or "generative engine optimisation" monitoring has grown to serve that demand. Its outputs are typically point estimates: a percentage, a score, a trend line month over month.

Point estimates presuppose a measurement whose error is small relative to the differences being reported. That presupposition is rarely examined. Generative answers are not deterministic; the same question asked twice may cite different sources. If that variation is large relative to the changes being reported, then a visibility trend line is mostly resampling noise rendered as a line.

This paper measures the error term. We treat non-determinism not as a nuisance to be averaged away but as the quantity of interest, and derive from it the sample sizes at which claims of a given size become supportable.

Contributions.

1. **Replicate instability by platform.** The share of repeated identical queries whose citation outcome is not unanimous, measured on four commercial surfaces. This is a noise floor: no increase in prompt count reduces it, because it is variation in the system rather than in the sample.
2. **Precision as a function of sample size.** Confidence-interval width against the number of prompts, computed by a cluster bootstrap that resamples prompts rather than observations — replicates of one question not being independent questions.
3. **Minimum detectable effect.** The smallest difference two samples of a given size can separate from chance, which converts directly into guidance: to claim a gap of d, sample at least k.
4. **A worked demonstration of how selection manufactures an effect.** We show by resampling how often a small prompt set produces a large apparent language gap when the corpus-wide gap is near zero.
5. **Released instrumentation.** Harness and observations, so that the thresholds can be recomputed rather than cited from a paper that will date quickly.

2. Related work

Kumar (2026) establishes a large-scale baseline for brand visibility across AI search engines, reporting that visibility is measurable and differs by platform. It does not report replicates, so the stability of the quantity it measures is unknown; the present paper supplies that missing term and is best read alongside it rather than against it. Liu et al. (2023) evaluate whether generative search citations support the claims they are attached to — a question about citation *quality*, where ours is about citation *measurement*. Aggarwal et al. (2024) formalise optimisation for generative engines, which presumes a measurable objective; we characterise the noise in that objective. Allaham and Diakopoulos (2026) audit the same four engines with a different question, whether cited sources are themselves AI-generated. Our design follows the external-audit tradition of Sandvig et al. (2014) and Metaxa et al. (2021).

3. Data and method

3.1 Observations

We use 4,320 successful observations collected in September 2026 across four answer surfaces (Google AI Overviews, Gemini with search grounding, Perplexity, and OpenAI's web-search tool), six European languages, and five commercial verticals. Each of 360 buying-intent prompts was issued three times per platform. Prompts were authored natively per language and never translated. Full collection design is given in the companion audit paper and in the released codebook.

3.2 Instability

Define a *cell* as one prompt issued to one platform in one market. For cells observed $r > 1$ times, the cell is *unstable* if the citation outcome is not unanimous across its replicates. We report the share of unstable cells per platform. This is deliberately a coarse measure: it asks only whether a practitioner running the same check twice would draw the same conclusion.

3.3 Precision

For a target sample size k we draw k prompt-clusters with replacement, compute the citation rate over all observations in those clusters, and repeat B times. The 2.5th and 97.5th percentiles give the interval. Resampling clusters rather than observations is essential: three replicates of one prompt carry roughly one prompt's worth of information, and treating them as three independent observations understates the interval by a factor that grows with the replicate count.

3.4 Minimum detectable effect

To ask what difference a study of size k could detect, we draw *two* independent samples of size k from the same platform and record the absolute difference in their citation rates. The 95th percentile of that distribution is the minimum detectable effect: an observed gap smaller than it is not distinguishable from having sampled the same population twice.

3.5 Population

All results are computed on the 864 observations covering the vertical in which the tracked brands trade. Restricting to that vertical matters: pooling in verticals where a brand sells nothing adds a population whose true citation rate is zero and whose replicates are therefore always unanimous, which deflates measured instability by roughly a factor of five. A practitioner would only ever measure on prompts relevant to the brand, so that is the population these figures describe.

4. Results

4.1 Instability differs by more than an order of magnitude

Platform	Citation rate	Non-unanimous	Instability
Gemini (grounded)	63.9%	29 / 72	40.3%
Google AI Overviews	16.2%	11 / 72	15.3%
ChatGPT (web search)	36.1%	4 / 72	5.6%
Perplexity	91.2%	1 / 72	1.4%

Table 1. Share of prompt-platform cells whose citation outcome was not unanimous across three identical repetitions.

Gemini disagrees with itself on two questions in five. Asking it once whether it cites a given brand yields an answer that a second identical ask would contradict 40% of the time. Perplexity, at the other extreme, is nearly deterministic on this measure. Any claim of the form "platform p cites brand b at rate r" carries a platform-specific reliability that ranges from near-certain to close to a coin flip on the disputed cases, and nothing in current practice reports it.

4.2 Precision against sample size

Platform	k=6	k=12	k=24	k=48	k=72
ChatGPT	±38.9	±26.4	±18.8	±13.5	±11.1
Gemini	±30.6	±20.8	±14.6	±10.8	±8.8
Google AI Overviews	±22.2	±18.1	±13.2	±9.0	±7.4
Perplexity	±16.7	±12.5	±10.4	±7.6	±6.5

Table 2. Half-width of the 95% interval on a platform citation rate, in percentage points, by number of distinct prompts sampled.

At twelve prompts — a plausible configuration for a low-cost monitoring tier — a ChatGPT citation rate is known to within ±26 points. Reported as "36%", it is consistent with anything from 10% to 62%.

4.3 Minimum detectable difference

Platform	k=6	k=12	k=24	k=48	k=72
ChatGPT	50.0	36.1	26.4	18.8	15.3
Gemini	44.4	30.6	20.8	15.3	12.5
Google AI Overviews	38.9	25.0	18.1	13.2	10.2
Perplexity	33.3	25.0	15.3	11.1	9.3

Table 3. Minimum detectable difference in percentage points: the 95th percentile of the absolute difference between two independent samples of size k drawn from the same platform.

The consequence is blunt. To distinguish a ten-point change — a large move by any commercial standard — requires on the order of 48 to 72 distinct prompts per platform, per market. Below roughly 24 prompts, only differences of 15 to 26 points are resolvable, and month-over-month movements of the size these products typically report fall entirely inside the noise.

4.4 Selection effects at small samples

The same arithmetic explains how spurious cross-language findings arise. Drawing matched subsamples and computing the English–German difference under single-pass conditions: on Google AI Overviews, where the corpus-wide difference is 8.3 points, a study of eight prompts per language reports a difference of at least 20 points 16.4% of the time, with draws reaching 50 points; at four prompts, 33.5% of the time, reaching 75 points. On ChatGPT, where the corpus-wide difference is exactly zero, a four-prompt study reports a difference of at least 20 points in one direction or the other 12.5% of the time.

A published claim of a large language gap, derived from a handful of prompts sampled once, is therefore not evidence of a language gap. It is the expected output of the sampling procedure.

5. Guidance

From Table 3, the minimum prompt count needed to resolve a difference of a given size:

Claimed difference	Perplexity	Google AIO	Gemini / ChatGPT
25 points	12	12	24
15 points	24	24	48
10 points	48	72	>72
5 points	>72	>72	>72

Table 4. Approximate distinct prompts required per platform per market. Figures beyond the study's 72-prompt ceiling are marked as exceeding it rather than extrapolated.

Three recommendations follow, in descending order of how much they cost to adopt.

Report the replicate count. This costs nothing. A visibility figure given without it cannot be assessed at all, and the field has no convention requiring it.

Report an interval, not a point. Also nearly free: the bootstrap above runs in seconds on data already collected. A vendor unwilling to publish the interval is, in effect, declining to say how much of the number is real.

Do not report period-over-period change below the detection threshold. This one has a genuine commercial cost, because movement is what makes a dashboard feel alive. It is nonetheless the recommendation with the most substance: at the sample sizes in common use, most reported movement is resampling noise, and presenting it as a trend is not a rounding error but a category mistake.

6. Threats to validity

One corpus, one window. Instability is estimated on this prompt corpus in September 2026. It is a property of the platforms as configured then, not a constant. This is the reason the harness is released.

API versus consumer surfaces. Measurements describe provider APIs, which are not identical to the consumer products. If consumer interfaces add personalisation, their instability is plausibly higher than what we report, making our thresholds conservative.

Coarse outcome measure. We score whether a domain was cited, not its position or prominence within the answer. A finer measure would likely show more instability, not less.

Author interest. The author operates online stores in three of the markets studied and has a commercial interest in this measurement category. The instability and precision results do not depend on which brands are treated as tracked; the released data permits the analysis to be repeated with any target set. Tracked brands are pseudonymised in this release.

7. Data availability

The observation log accompanying this paper is released under CC BY 4.0, with tracked-brand domains pseudonymised as FOCAL-1 through FOCAL-4; every quantity reported here depends on the citation outcome rather than the identity of the brand, so the redaction does not affect reproducibility. The prompt corpus is released in full. The measurement harness is released under MIT. A full replication requires API credentials for the four providers and costs approximately $46.

References

Aggarwal, P., Murahari, V., Rajpurohit, T., Kalyan, A., Narasimhan, K., and Deshpande, A. (2024). GEO: Generative Engine Optimization. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

Allaham, M. and Diakopoulos, N. (2026). Synthetic Sources? Auditing Generative Search Engine Citations for Evidence of AI-Generated Sources. *arXiv preprint arXiv:2605.23684*.

Kumar, P. (2026). Generative Engine Optimization at Scale: Measuring Brand Visibility Across AI Search Engines. *arXiv preprint arXiv:2606.20065*.

Liu, N. F., Zhang, T., and Liang, P. (2023). Evaluating Verifiability in Generative Search Engines. In *Findings of the Association for Computational Linguistics: EMNLP 2023*.

Metaxa, D., Park, J. S., Robertson, R. E., Karahalios, K., Wilson, C., Hancock, J., and Sandvig, C. (2021). Auditing Algorithms: Understanding Algorithmic Systems from the Outside In. *Foundations and Trends in Human–Computer Interaction*, 14(4).

Sandvig, C., Hamilton, K., Karahalios, K., and Langbort, C. (2014). Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms. Presented at *Data and Discrimination*, a preconference at the 64th Annual Meeting of the International Communication Association.