Barkhausen AI

WHITEPAPER · BA-W-2026-01

Measuring AI visibility: statistical requirements and common failures

A brand's visibility in AI assistants is routinely 'verified' with a single screenshot or one daily query.

“Every leap begins as a crackle.”

DOCUMENT	BA-W-2026-01
SERIES	Whitepaper
DATE	2026
SUBJECTS	Measurement & statistics; Conventions & policy
LICENSE	CC-BY-4.0
AUTHOR	Barkhausen AI

ABSTRACT

A brand's visibility in AI assistants is routinely 'verified' with a single screenshot or one daily query. This paper argues such verification is not measurement. Answer engines are stochastic and their retrieval changes continuously: lightly rewording a query while holding its intent fixed cut the overlap of the brands an assistant recommended to a Jaccard similarity near 0.3 — far below the 0.50–0.61 overlap of a plain re-run — and an identical prompt re-issued a day later overlapped only 34–42% in cited sources and 45–59% in mentioned brands. A single observation of a moving distribution estimates nothing. The paper enumerates what voids a visibility claim — no sample size, no interval, no window, no engine version, one fixed phrasing, uncontrolled personalization, discarded refusals — and shows a three-sigma jump can be pure drift. It then states what valid measurement requires — repeated sampling to a declared precision, bounded intervals near the extremes, a phrasing distribution, partial pooling, explicit windows and versions, change-point monitoring, recorded refusals — specified in BA-C-2 and BA-C-3.

IN BRIEF

When someone says a company 'shows up' in an AI assistant like ChatGPT or Gemini, how do they know? Often the proof is a single screenshot, or one question asked once a day. This paper explains why that is not enough to trust. AI assistants do not give the same answer twice. Ask the same question a day later and studies find that up to half the brands mentioned change; reword the question only slightly — while still asking for the same thing — and roughly half the organizations an assistant recommends can change too. A single check captures one roll of the dice, not the real picture — so a result that looks like a large improvement can be nothing more than the normal day-to-day wobble. Measuring visibility honestly means asking each question many times, in many natural wordings, recording which assistant and version answered and when, reporting a margin of error, and treating a refusal to answer as a real result rather than throwing it away. The paper closes with llms.txt — a file many sites add in the hope that AI models read it, though no public evidence shows the major models do — as an example of adopting a tactic without proof it works.

KEY FINDINGS

1. Rewording a query while holding its underlying intent fixed sharply changed which brands an assistant recommended: across roughly 12,000 runs on two production models (one OpenAI, one Anthropic), a cosmetic rewording dropped the recommendation-set overlap to a Jaccard similarity of about 0.29 (95% CI 0.22–0.36) — roughly 55% of the recommended brands turned over — and a rewording that added a constraint dropped it to about 0.14, both far below the 0.50–0.61 overlap of the identical prompt simply re-run [1].
2. Re-issuing an identical, unchanged prompt on two consecutive days overlapped only 34–42% in cited sources and 45–59% in mentioned brands (the brand figure from the study's three qualifying verticals), measured across four answer engines and four verticals across 45–46 collection days (2026-01-24 → 2026-03-20), so a single query estimates a moving quantity rather than a fixed one [2].
3. A single query per term is one Bernoulli draw whose standard error reaches 0.5 on a 0–1 scale at p = 0.5; it therefore pins a mention probability to no useful precision, however many decimal places are reported.
4. Achieving a ±10% confidence interval on a mention probability requires about n ≈ 96 independent samples per term–engine–window–region cell, and ±5% requires about n ≈ 385 (worst case p = 0.5), from $n \approx 1.96^2 \cdot p(1-p)/E^2$ [3].
5. A period-over-period jump from 51% to 76%, measured with one query per term across 80 terms, computes to roughly 3σ under a naive two-proportion test yet is not attributable to the intervention: with one measurement day per period, day-to-day engine drift is perfectly confounded with the change and, at 45–59% same-prompt brand overlap, can produce a swing of that size on its own [2].
6. Near a mention probability of 0 or 1 the normal-approximation interval fails and can return an impossible negative bound, so bounded methods — the Wilson score interval or the Clopper–Pearson exact interval — are required [4][5].
7. There is no public evidence that major models fetch or consume llms.txt at answer time: a roughly 300,000-domain analysis found no statistically measurable association between llms.txt and AI citation (removing the variable improved model accuracy) [6], and independent server-log analysis of about 137,000 sites found that AI crawlers rarely request the file — and a request would not, in any case, establish that the file is used when an answer is generated [9].

CONTENTS

1 The claim

The usual proof that a brand is visible in an AI assistant is a picture. Someone opens an answer engine, types a question, and the brand appears in the reply. A screenshot is taken; the claim is filed as verified. A more diligent version issues the same question once a day and plots the fraction of days the brand appeared. Both procedures share a premise: that a small number of observations of an answer engine, at a moment or a cadence of the observer's choosing, establish a fact about how visible the brand is.

This paper argues that the premise is false, and that most claims about AI visibility circulating today are, in the strict sense, unmeasured. The object being described — the tendency of an answer engine to mention a given brand for a given kind of question — is not a fact that a single observation can reveal. It is a probability attached to a distribution of possible questions and a distribution of possible answers, and that distribution moves. A screenshot samples it once. A daily query samples it once per day, at one fixed phrasing, on one engine build, from one session. Neither procedure produces the two quantities that turn an observation into an estimate: a sample size and an interval. Without them, a percentage is not a measurement of anything; it is an anecdote formatted to look like a statistic.

The argument proceeds in four moves. First, that the quantity of interest is a distribution, and that its instability is now documented in public research, not conjectured. Second, a taxonomy of the specific ways a single-shot verification misleads, with a worked arithmetic example in which a jump that looks like three standard deviations of improvement turns out to be indistinguishable from ordinary day-to-day drift. Third, what valid measurement requires instead — the sampling, interval, and monitoring machinery that separates a real change from noise, specified normatively in the companion conventions BA-C-2 and BA-C-3. Fourth, a cautionary case: llms.txt, a widely deployed tactic promoted as a visibility lever for which no public evidence of the claimed effect exists, offered as an illustration of the same discipline applied to a technique rather than a metric.

Throughout, one metric is central. **Visibility Probability** (**VP**) is the probability that an answer engine mentions a given brand in response to a given information need, estimated by repeated sampling and reported with a confidence interval. Formally, for a brand, an information need k, an engine e, a time window t, and a region g:

$$\mathrm{VP} = P(\text{brand mentioned} \mid \text{prompt} \sim Q_k,\ e,\ t,\ g)$$

The notation is worth reading slowly, because every failure catalogued below is a failure to respect one of its parts. The mention is conditioned on a *distribution* of prompts Q_k — the many ways a real user might ask for the same thing — not one sentence. It is conditioned on a specific engine e, which includes its version. It is defined over a window t, because the answer to "how visible is the brand" has a date. And it is a probability, which means it is estimated, never observed, and an estimate without a stated precision is not yet a number one can act on.

2 Visibility is a distribution

Answer engines are stochastic at more than one layer. Consumer products decode with a non-zero temperature, so the same model can word the same answer differently on two calls. The retrieval layer that feeds the model is itself variable: indexes are sharded and refreshed, results are re-ranked for recency and load, and on the fan-out architectures now common a language model generates the sub-queries, injecting further randomness before any document is retrieved. Above all this sits the slower motion of the index itself, and the occasional discontinuity of a silent model or system-prompt update. The consequence is

that "does the engine mention this brand for this question" does not have a single answer. It has a distribution of answers, and the practical question is what that distribution's mean — the mention probability — actually is.

Three independent lines of public evidence establish that the distribution is wide enough that a single draw reveals little. They are summarized below and then read in turn.

INSTABILITY	PUBLIC FINDING	EVIDENCE
Rewrite sensitivity	Across roughly 12,000 runs on two production models (one OpenAI, one Anthropic), rewording a query while holding its intent fixed dropped the overlap of the recommended-brand set to a Jaccard similarity of about 0.29 for a cosmetic rewrite (about 55% of brands turning over) and about 0.14 when the rewrite added a constraint — both far below the 0.50–0.61 overlap of the same prompt simply re-run.	[1]
Day-to-day drift	The identical prompt issued on two consecutive days overlapped only 34–42% in its cited sources and 45–59% in the brands it mentioned.	[2]
Monthly turnover	Single-source industry monitoring (not independently corroborated) reports monthly turnover of the cited-domain set of roughly 40–60%.	See text

Rewrite sensitivity is the sharpest of the three because it isolates phrasing from time. In the study, each query was reworded while its underlying intent was held fixed — the kind of difference between "best study-abroad agencies" and "which study-abroad agency is best" — and, crucially, the paraphrase results were compared against a same-prompt rerun baseline that measures ordinary run-to-run noise on its own [1]. A cosmetic rewrite dropped the overlap of the recommended-brand set to a Jaccard similarity of about 0.29 (95% CI 0.22–0.36) — roughly 55% of the recommended brands turned over — and a rewrite that merely added a constraint dropped it further, to about 0.14. Both sit far below the 0.50–0.61 overlap of the identical prompt re-run, which is the study's central point: most of the movement is caused by the rewording, not by the model's run-to-run randomness. Two lessons follow. The obvious one is that a single fixed phrasing is a fragile probe; the answer it elicits may be specific to that exact wording. The less obvious one is that the effect is large enough to swamp the real differences between entities, so a verification built on one phrasing measures the phrasing as much as the entity.

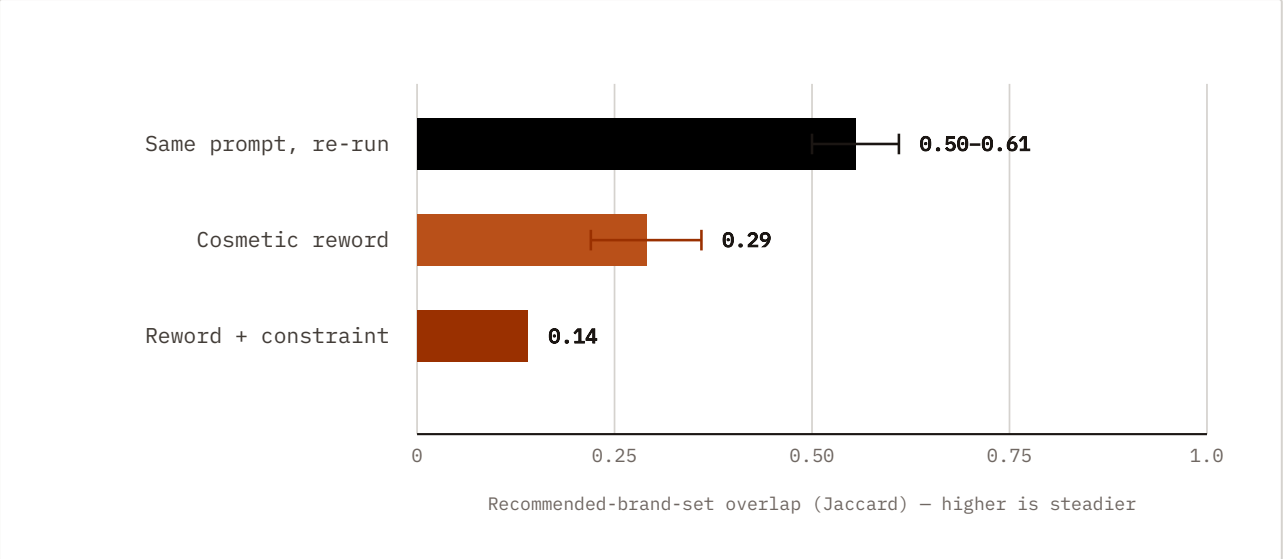

Figure 1. How much the set of brands an assistant recommends changes when the query barely changes. A cosmetic rewording of the same question turns over roughly half the recommended brands (Jaccard ≈ 0.29); adding one constraint turns over more (≈ 0.14) — both far below the overlap of simply re-running the identical prompt (0.50–0.61). Whiskers show the reported range or 95% confidence interval.

Paraphrase-brittleness study, ≈12,000 runs across two production models (one OpenAI, one Anthropic) [1]

Day-to-day drift isolates time. Here the prompt does not change at all; it is re-issued verbatim a day later on the same engine. Measured across four answer engines and four verticals across 45–46 collection days (2026-01-24 → 2026-03-20), the set of sources the engine cited overlapped only 34–42% between consecutive days, and the set of brands it mentioned overlapped 45–59% — the brand figure computed on the three verticals that met the study's brand-detection threshold [2]. The second number is the load-bearing one for visibility measurement: even for an unchanged question on an unchanged engine, roughly two-fifths to a half of the brands mentioned turn over from one day to the next. Brand mentions are somewhat steadier than the underlying citation set — which is why brand-mention probability, rather than the citation list, is the more stable basis for a key metric — but "somewhat steadier" still means that a single day's answer is a poor estimate of the brand's standing.

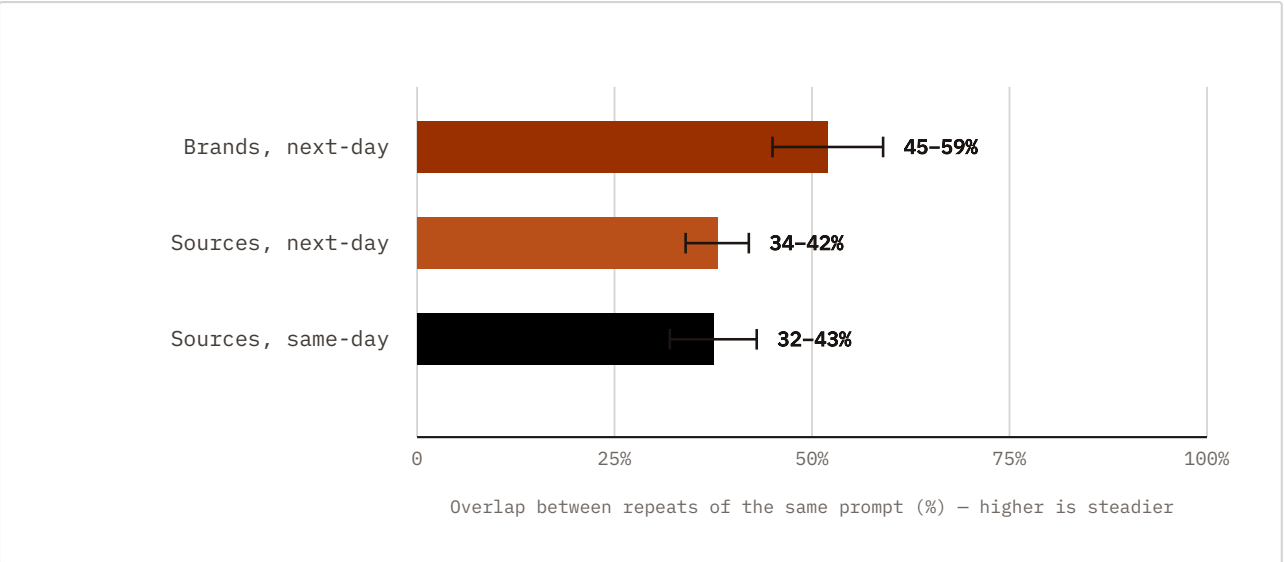

Figure 2. Overlap between two answers to the same, unchanged prompt. Even for an identical question on an unchanged engine, the brands mentioned overlap only 45–59% from one day to the next, and the cited sources only 34–42%; a same-day re-run of the identical prompt already churns the sources to 32–43%, so most source instability is run-to-run rather than day-to-day drift. Bars span the reported ranges.

Don't Measure Once (arXiv:2604.07585), four engines × four verticals; brand overlap on the three qualifying verticals [2] · 45–46 collection days, 2026-01-24 → 2026-03-20

Monthly turnover describes the slow drift of the index. The specific figure — roughly 40–60% of cited domains changing month over month — comes from single-source industry monitoring and is reported here only as directional context, not as an established value; it has not been independently corroborated, no per-engine breakdown is treated as fact, and the sampling convention BA-C-3 likewise declines to adopt it, treating the monthly scale qualitatively for the same reason. Its direction is consistent with the two stronger findings above: the population an answer engine draws from is continuously refreshed, so the thing being measured has genuine trend on top of its day-to-day noise.

Put together, these are not three descriptions of a flaky product. They are three descriptions of a moving distribution, measured at three time scales — phrasing-to-phrasing, day-to-day, month-to-month. A screenshot is one draw from that distribution at one phrasing on one day. The estimand VP is a property of the whole distribution. The gap between the two is the entire subject of this paper.

3 How single-shot verification misleads

If visibility is a distribution, the deficiencies of single-shot verification can be stated precisely as the sample statistics it omits. Each omission has a distinct failure mode; each, on its own, is enough to void a visibility claim.

#	FAILURE	DIAGNOSIS	HOW IT MISLEADS
1	No sample size	A percentage with no n behind it is not an estimate. A single 0/1 observation of a mention has a standard error as large as 0.5 on the 0–1 scale.	″Mentioned in 60% of answers,″ derived from one query per term, is a coin flip reported to two significant figures.
2	No confidence interval	Without an interval no reader can separate a real move from sampling noise, and point estimates invite over-reading of small differences.	62% and 55%, each from $n = 100$, have overlapping ±10-point intervals; reporting them as a “7-point decline” reports noise.
3	No time window	Engines drift day to day and month to month, so an undated number describes an instant, not a standing state, and cannot be compared to anything.	A figure captured “in 2026” cannot be placed against a later measurement; the comparison is undefined.
4	No engine and version	Engines differ from one another, and a silent model or system-prompt update changes retrieval, so results are not portable across engines or across builds of one engine.	″Visible on the leading assistant” conflates distinct engines and undated builds into one ungrounded claim.
5	One fixed phrasing	A single sentence is a degenerate stand-in for the distribution Q_k of ways users actually ask; rewrite sensitivity makes any one wording an unrepresentative probe.	Rewording a query while holding its intent fixed turned over roughly 55% of the recommended-brand set (Jaccard ≈ 0.29), far below the same prompt re-run [1]; the “verified” wording may be the only one that works.
6	No personalization or region control	Logged-in history, stored memory, and region change the answer, so an uncontrolled sample silently mixes populations.	A result gathered from a personalized, logged-in session in one country cannot be generalized to anyone else.
7	Refusal treated as missing data	A refusal is an outcome — the term is, for now, unavailable on that engine — not a gap to be dropped; discarding refusals biases the reported rate upward.	Dropping the refusals turns “mentioned in 3 of 3 answers the engine gave” into “100% visible” while hiding that it refused the other 7 of 10 attempts.

Failures 1 and 2 are the ones a worked example makes vivid, so consider a concrete, neutral one. It is a constructed illustration, not an audit of any specific report.

A worked example. A brand is tracked across a fixed set of $N = 80$ terms — each a query in the sense of BA-C-3, one information need expressed through a distribution of real phrasings — on one answer engine. “Verification” consists of issuing each term once and recording whether the brand appears. In an early period the brand appears for 41 of 80 terms, a rate of $\hat{p}_1 = 0.51$; a round of optimization work is done; in a later period it appears for 61 of 80, a rate of $\hat{p}_2 = 0.76$. The report presents a 25-point improvement and attributes it to the work.

Treat the 80 terms as independent replicates, as the naive reading implicitly does, and the improvement looks decisive. The standard error of the difference between two proportions is

$$\mathrm{SE} = \sqrt{\tfrac{p_1(1-p_1)}{N} + \tfrac{p_2(1-p_2)}{N}} = \sqrt{\tfrac{0.51 \cdot 0.49}{80} + \tfrac{0.76 \cdot 0.24}{80}} \approx 0.073$$

At the conservative worst case $p = 0.5$ this is $\sqrt{0.25/80 + 0.25/80} \approx 0.079$. Either way the observed 0.25 difference is roughly $0.25/0.079 \approx 3.2$ standard errors — about 3σ, nominally significant beyond the 1% level. On its face, proof.

It is not proof, for two reasons the design cannot escape. The first is that each term–period cell holds exactly one observation. A single Bernoulli draw estimates that term's true mention probability with a standard error of up to 0.5 — it is a coin flip. The 0.073 figure is a *between-term* quantity; it describes how much the portfolio average would wobble *if* each term's single draw equalled that term's true probability and the terms were independent replicates of one stable process. That "if" is exactly what is not established. The arithmetic of significance has been run on an assumption the data cannot support.

The second reason is fatal to the causal claim regardless of sample size. The two periods are measured on two different days. Everything that moves between those days — the engine's index refresh, a silent model or system-prompt update, load-dependent re-ranking, time of day, and any drift in the exact phrasings used — is perfectly aliased with "the optimization." Write the period difference as its parts:

$$\hat{p}_2 - \hat{p}_1 = \Delta + (D_2 - D_1) + \varepsilon$$

where Δ is the true effect of the work, D_t is the engine's day-level state, and ε is sampling noise. With one measurement day per period, $D_2 - D_1$ is a single unobserved draw that cannot be separated from Δ. And the drift evidence tells us $D_2 - D_1$ is not small: re-issuing an identical, unchanged prompt on two consecutive days overlaps only 45–59% in the brands mentioned [2], so with zero intervention a large share of the 80 per-term indicators would flip between two days on their own. A net movement of 20 terms out of 80 sits comfortably inside that envelope. The 3σ statistic, honestly read, certifies only that *something changed between the two measurement days*. It does not, and under this design cannot, attribute that change to the work. In the language of the framework, the claim fails the **Barkhausen Criterion** — the rule BA-C-3 defines, in four conditions, for when a visibility change counts as real: statistically significant against a properly estimated interval, sustained across the engine's drift, clear of any coincident engine update, and controlled for the multiplicity of testing many terms at once. A single jump measured once meets none of them.

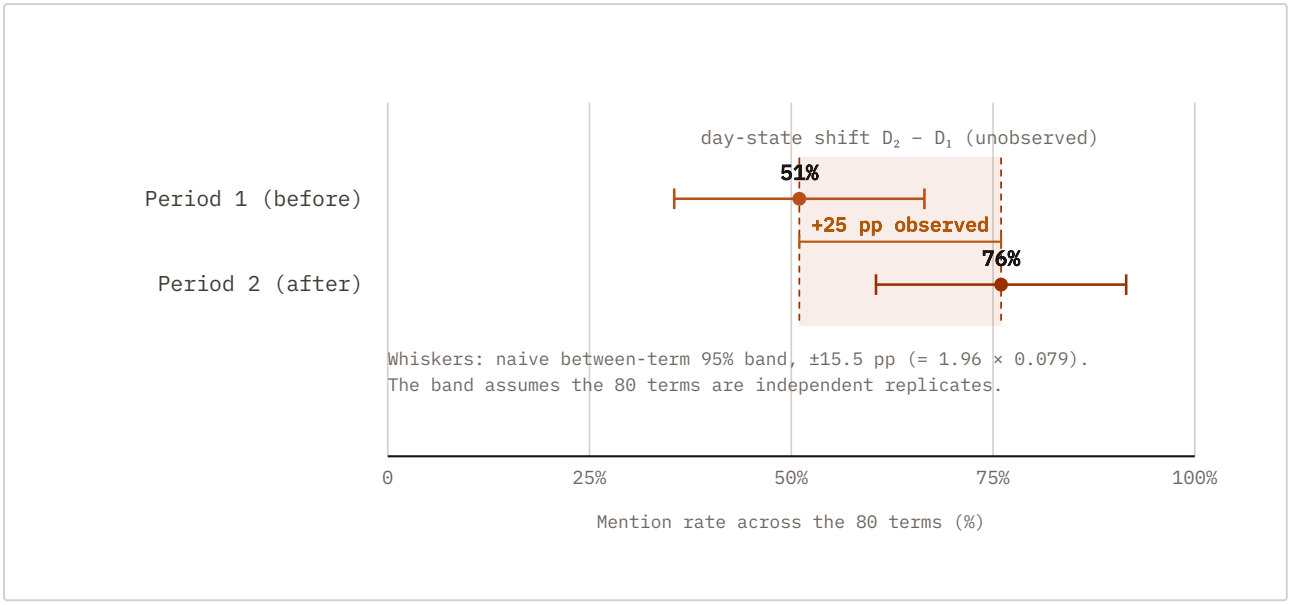

Figure 3. The worked example, read honestly. A brand's mention rate across 80 terms rises from 51% to 76% between two single-day measurements — a 25-point jump that is nominally 3σ under a naive two-proportion test. But the two periods are one measurement day each, so the engine's unobserved day-state shift $D_2 - D_1$ is perfectly confounded with the intervention: at 45–59% same-prompt brand overlap [2], drift alone can move roughly 20 of the 80 terms, a swing the size of the whole jump. The whiskers are the naive between-term 95% band (±1.96 × 0.079 ≈ ±15.5 pp) and assume the 80 terms are independent replicates — the assumption the design cannot support.

Constructed illustration; drift envelope from Don't Measure Once (arXiv:2604.07585) [2]

The taxonomy and the example converge on one point. Single-shot verification does not merely produce imprecise numbers; it produces numbers whose imprecision is invisible, because the procedure never generates the interval that would reveal it. A screenshot cannot be under-powered, because it never claimed a power. That is precisely the problem.

4 What valid measurement requires

The remedy is not a larger portfolio of one-shot checks. It is repeated sampling within each cell, to a stated precision, over an explicit window and version, across a distribution of phrasings, with the results pooled and monitored correctly. The requirements below are the minimum a competent statistician would demand before treating a visibility number as an estimate. They are specified normatively in BA-C-2 (metric definitions and reporting requirements) and BA-C-3 (sampling design, sample size, and interval estimation, together with windows, versioning, change-point monitoring, and refusal handling); this section states them and their public-evidence rationale.

Sample to a declared interval width. For a mention probability estimated as $\hat{p} = k/n$ from n independent samples, the 95% interval half-width is $E \approx 1.96\sqrt{\hat{p}(1-\hat{p})/n}$. Inverting gives the sample size needed for a target precision:

$$n \approx \frac{1.96^2\, p(1-p)}{E^2}$$

The variance $p(1-p)$ is largest at $p = 0.5$, so planning for that worst case yields the table below [3]. The declared precision, not intuition, sets the sample size.

TARGET 95% INTERVAL HALF-WIDTH E (AT $P = 0.5$)	REQUIRED N PER TERM–ENGINE–WINDOW–REGION CELL
±10%	≈ 96
±5%	≈ 385

The gap between this and single-shot practice is not a matter of degree but of kind. Giving "62% mentioned" the scientific meaning of "±5%" requires on the order of a hundred to a few hundred samples per term, per engine, per window. One screenshot supplies one. The reason to sample in the hundreds is not thoroughness for its own sake; it is that nothing narrower than that produces an interval a decision can rest on.

Use bounded intervals near the extremes. The normal approximation behind the formula above breaks down as $\hat{p}$ approaches 0 or 1: at a 5% mention rate with modest n it can place the lower bound below zero, an impossible value for a probability. Near the bounds — which is exactly where many real terms sit, since most brands are mentioned rarely for most questions — the **Wilson score interval** [4] or the **Clopper–Pearson exact interval** [5] must be used instead. Both are standard and available in any statistics library; reporting a naive interval that runs off the end of the probability scale is a marker of the same innumeracy as reporting no interval at all.

Sample a phrasing distribution, not a sentence. The estimand conditions on Q_k, the distribution of ways a real user expresses information need k, and the rewrite-sensitivity evidence [1] shows why a single sentence cannot stand in for it. Valid measurement draws prompts across a set of phrasings that reflects real user language — including, where the audience warrants, more than one language and region — so that VP estimates the mention probability over how the question is actually asked, not over one arbitrary wording. The construction and validation of that phrasing set is specified in BA-C-3.

Pool phrasings with partial pooling. Once each term carries several phrasings, each sampled some number of times, estimating every phrasing independently wastes information and lets small-sample phrasings swing wildly. The correct treatment is a hierarchical model in which phrasings are random effects under their term, so each phrasing's estimate is shrunk toward the term's overall mean by an amount set by its own sample size: phrasings with ample data keep their signal, thin ones borrow strength from the term. The reported quantity for each term–engine–window–region cell is then a posterior mean with a 95% credible interval, not a bare ratio.

Record the window and the engine version on every number. Because the distribution drifts, a visibility figure without a date describes nothing durable, and because engines differ and update silently, a figure without an engine-and-version label is not portable. Every VP should be reportable in a form that carries its evidence with it, for example:

> *VP = 0.62 (95% CI 0.53–0.70, Wilson; n = 120, engine X vYYYY-MM, region DE, window 2026-W27)*

A number in this form can be checked, compared, and reproduced. A screenshot cannot; a bare "62% visible" cannot. The difference between the two is the difference between a measurement and a claim.

Monitor for change points. Drift and discontinuity are not only noise to be averaged away; the discontinuities are informative. When a large number of unrelated terms move in the same direction on the same day, the cause is almost never a coincidence of optimization work — it is an engine update. Running an online change-point procedure such as a CUSUM control chart or Bayesian online change-point detection on each term's VP time series converts that pattern from an unexplained shock ("the numbers dropped and we don't know why") into a detected, dated event. This is also what enforces the

"sustained" half of the Barkhausen Criterion: a jump that a change-point monitor shows reverting at the next engine shift was never a durable gain.

Record refusals as data. When an engine declines to answer — because a topic trips a safety filter, or for any other reason — that refusal is an observation about the term's availability, not a missing value to be silently dropped. The refusal rate is itself a metric; discarding refusals inflates the apparent mention rate among the answers that remain. BA-C-3 requires refusals to be stored and reported alongside mentions.

Two companion metrics complete the picture and inherit the same discipline. **Share of Voice** (**SoV**) — a brand's share of all brand mentions in an answer — is more robust to engine-wide drift than absolute VP, because when the whole retrieval layer shifts, relative standings move less than absolute probabilities; it is the better basis for a long-run key metric. **Discovery Depth** (**DD**) is the degree of query constraint at which a brand first enters an engine's recommendation set — how far a user must narrow a broad need, by adding qualifying attributes, before the brand appears — so a brand recommended on a bare category query has lower DD, and broader reach, than one that surfaces only under heavy constraint. Both are defined in BA-C-2, and both are estimates with intervals, subject to every requirement above. Where within an answer a brand appears — first line versus seventh item — is a separate concern, the mention-prominence measure of BA-C-2 §5.1.

5 A cautionary case: llms.txt

The discipline this paper argues for applies to techniques as much as to metrics: a claim that a tactic *works* is a claim about a measurable effect, and it stands or falls on evidence of that effect. llms.txt is a useful case because it is adopted at scale and promoted, in some quarters, as a lever on AI visibility — while the public evidence for the promised effect is absent.

The proposal is a plain-text file at a site's root that offers language models a curated guide to the site's content. Adoption is real and substantial: the file has been deployed on well over a hundred thousand sites [9], and one vendor survey reports that two AI companies have stated support for the convention [7]. That is the strongest thing that can be said for it, and it is a statement about deployment, not about effect.

Against the effect, the public record is consistent. A published analysis of roughly 300,000 domains found no statistically measurable association between the presence of llms.txt and the frequency of AI citation; removing the variable from the analysts' predictive model *improved* its accuracy, the signature of a variable that contributes noise rather than signal [6]. Independent server-log analysis of about 137,000 sites has found that AI crawlers rarely request the file, and that a request in any case does not establish that its contents are used at answer time [9]. No major model provider has publicly confirmed consuming llms.txt at answer time. And the posture of at least one major search provider has been openly skeptical: a Google search advocate compared llms.txt in mid-2025 to the long-deprecated keywords meta tag and stated that Google Search does not support it [8] — even as an llms.txt file appeared on Google's own Search Central developer documentation in December 2025 and was removed within hours [10]. Adoption of the file across the web also remains in the single-to-low-double-digit percentages by most samples, so it is neither universal nor, on the evidence, consequential.

The honest reading is not that llms.txt is harmful. It costs a few minutes to deploy and carries no evident downside, and within specific ecosystems whose operators have said they support it there may be marginal value. The reading is that it belongs in the category of cheap hygiene, deployed without a promise, and never in the category of a demonstrated lever. Presence in a training corpus does not imply the model retains or reproduces the content; no causal claim is made — and by the same standard, presence of a file a model has not been shown to read does not imply the model reads it. A publication or a

vendor that presents llms.txt as a driver of AI visibility is making exactly the move this paper began by criticizing: converting a plausible-sounding intervention into a claimed effect without the measurement that would distinguish the two. The correct response to "does it work?" is not a screenshot of a site that deployed it and later got cited. It is a controlled estimate — and where one exists, at 300,000 domains, it points the other way.

6 Conclusion

The through-line is a single distinction. Visibility in an answer engine is a distribution, and the number people want — how visible is the brand — is a property of that distribution's center. A screenshot, and a once-a-day query at one fixed phrasing, sample that distribution once. One sample estimates a distribution's center to a precision of essentially nothing, and it cannot separate a real change from the variation that the public evidence shows is large at every time scale: roughly half the recommended brands turning over under a light reword (Jaccard ≈ 0.29 against a 0.50–0.61 plain re-run) [1], 45–59% brand overlap for an identical prompt a day apart [2], and continuous monthly churn of the underlying index.

What replaces it is not exotic. It is the ordinary statistics of estimating a proportion, applied with the field's specifics respected: sample each term to a declared interval width — on the order of a hundred to a few hundred draws, from $n \approx 1.96^2 p(1-p)/E^2$ [3]; use Wilson or Clopper–Pearson intervals near the bounds [4][5]; sample a distribution of phrasings rather than a sentence; pool phrasings with shrinkage; stamp every number with its window and engine version; monitor for change points; and record refusals as the observations they are. A claim that clears the Barkhausen Criterion — significant against a properly estimated interval, sustained across the engine's drift, clear of a coincident engine update, and controlled for testing many terms at once — is a measurement. A claim that does not is an anecdote, however many decimal places it wears. The normative form of these requirements is BA-C-2 and BA-C-3.

> ***About the name.*** *The Barkhausen effect is the observation that a ferromagnet, placed under a smoothly increasing field, does not magnetize smoothly: its domains flip in sudden, discrete jumps, and a coil around the sample picks the jumps up as crackling noise. The analogy is exact enough to be useful. Under the steady drive of optimization work, an entity's visibility does not climb smoothly either; it advances in discrete, noisy jumps against a background of drift, and the whole problem is telling a genuine jump from the Barkhausen noise around it. That is a statistical task, not a photographic one.*

7 Limitations

The instability figures cited here come from specific studies of specific engines over specific windows in 2026. These results describe the engines as sampled during those windows; engines change without notice, and results should be assumed perishable. The exact percentages should not be read as constants of the engines, only as evidence that the underlying distributions are wide and moving.

The rewrite-sensitivity study [1] measures a single commercial-recommendation setting on two production models (one OpenAI, one Anthropic) over a single collection period, so its exact magnitudes should not be assumed to transfer to other engines, domains, or languages. Its design does include a same-prompt rerun baseline, which is what lets it attribute most of the observed movement to the rewording rather than to run-to-run randomness; but it is a single-day study and does not establish how paraphrase sensitivity compounds with day-to-day drift over longer horizons.

The day-to-day drift figures [2] come from a single study of four answer engines across four German-language verticals, collected from servers in one region across 45–46 collection days (2026-01-24 → 2026-

03-20); they describe overlap between consecutive days and should not be extrapolated to arbitrary time gaps, other regions or languages, or engines outside that study. Its brand detection relies on lexicon matching, which can miss brands named by synonym or paraphrase. The monthly-turnover range is single-source industry monitoring, is not independently corroborated, and is reported only as directional context, never as an established value.

The worked example in this paper is a constructed, neutral illustration chosen to make the arithmetic transparent; it is not a measurement of any actual campaign, engine, or product, and its specific numbers carry no empirical weight beyond the point they demonstrate. The sample-size formula assumes independent draws and the normal approximation, both of which real monitoring must handle carefully — independence fails when phrasings within a term are correlated, which is the reason partial pooling is required rather than optional, and the normal approximation fails near the bounds, which is the reason exact intervals are required there.

This paper specifies requirements and does not reproduce any particular operator's implementation of them; the normative specifications live in BA-C-2 and BA-C-3, and reasonable implementations may differ in detail while meeting the same requirements. Finally, the paper concerns retrieval-time visibility as sampled from answer engines; it makes no claim about whether or how a model reproduces content from its training corpus, which is a separate question addressed elsewhere.

REFERENCES

1. Jack, Lehman, Maloney, Xu; arXiv:2605.27440. *Paraphrase brittleness in production retrieval-augmented commercial recommendation: reproducibility below the rerun-stability baseline* (2026). https://arxiv.org/abs/2605.27440 Accessed 2026-07-08. [archived]
2. Schulte, Bleeker, Kaufmann; arXiv:2604.07585. *Don't Measure Once: Measuring Visibility in AI Search (GEO)* (2026). https://arxiv.org/abs/2604.07585 Accessed 2026-07-08. [archived]
3. William G. Cochran (Wiley). *Sampling Techniques (3rd ed.)* (1977).
4. Edwin B. Wilson (Journal of the American Statistical Association 22(158):209–212). *Probable inference, the law of succession, and statistical inference* (1927).
5. C. J. Clopper & E. S. Pearson (Biometrika 26(4):404–413). *The use of confidence or fiducial limits illustrated in the case of the binomial* (1934).
6. SE Ranking. *Analysis of llms.txt and AI-citation frequency across roughly 300,000 domains* (2026). https://seranking.com/blog/llms-txt/ Accessed 2026-07-08. [archived]
7. Presenc.ai. *State of llms.txt 2026* (2026). https://presenc.ai/research/state-of-llms-txt-2026 Accessed 2026-07-08. [archived]
8. Google's John Mueller, reported by Search Engine Journal. *Google says llms.txt is comparable to the keywords meta tag; Google Search does not support llms.txt* (2025). https://www.searchenginejournal.com/google-says-llms-txt-comparable-to-keywords-meta-tag/544804/ Accessed 2026-07-08. [archived]
9. Ahrefs. *Server-log analysis of llms.txt fetch rates across roughly 137,000 sites* (2026). https://ahrefs.com/blog/llmstxt-study/ Accessed 2026-07-08. [archived]
10. Matt G. Southern, Search Engine Journal. *Google's llms.txt Guidance Depends On Which Product You Ask* (2026). https://www.searchenginejournal.com/googles-llms-txt-guidance-depends-on-which-product-you-ask/575431/ Accessed 2026-07-09. [archived]

HOW TO CITE

Barkhausen AI (2026). Measuring AI visibility: statistical requirements and common failures. https://barkhausen.ai/research/measurement-statistics-whitepaper/

Barkhausen AI · Est. MMXXVI · *Every leap begins as a crackle.*