Barkhausen AI

WHITEPAPER · BA-W-2026-04

The measurement gap in AI-visibility tooling

Commercial AI-visibility tools have converged on a recognizable form: a daily score per tracked query, a trend line of those scores, a cited-or-not verdict per topic, and per-engine coverage badges.

“Every leap begins as a crackle.”

DOCUMENT	BA-W-2026-04
SERIES	Whitepaper
DATE	2026
SUBJECTS	Measurement & statistics; Conventions & policy
LICENSE	CC-BY-4.0
AUTHOR	Barkhausen AI

ABSTRACT

Commercial AI-visibility tools have converged on a recognizable form: a daily score per tracked query, a trend line of those scores, a cited-or-not verdict per topic, and per-engine coverage badges. This paper argues the genre, as a class, cannot meet the disclosure floor a measurement requires. Walked against the ten minimum-disclosure items of BA-C-4, the common form omits the load-bearing ones: a daily single run is one Bernoulli draw per query, so the score carries no usable sample size; a line of bare points hides the day-to-day drift that is its own variance; one frozen query per topic measures the wording, not the need; API collection cannot support a claim about what users see; refusals and engine-wide updates go unmarked. The paper states what a conforming tool would show instead, and closes with five questions a buyer can ask. No tool is named or assessed.

IN BRIEF

Software that promises to track whether a brand 'shows up' in AI assistants has settled into a familiar shape: a daily score for each question watched, a line charting that score over time, a green or red 'cited / not cited' mark per question, and badges for each assistant covered. This paper explains why that shape, however polished, cannot be trusted as measurement. Checking a question once a day is a single coin flip per day, so the daily score has no real precision behind it; a plain line invites reading normal day-to-day wobble — which studies show is large — as real change; watching one fixed wording measures that wording rather than the question people actually ask; and collecting answers through a developer interface does not show what ordinary users see. The paper names and grades no product. Instead it describes what an honest tool would show — a margin of error on every point, the number of checks behind each one, clear time windows, a marker when the assistant itself changed — and gives anyone buying such a tool five plain questions to ask before trusting its numbers.

KEY FINDINGS

1. A dashboard that refreshes each tracked query once a day runs one Bernoulli trial per query per day, so its daily 'score' has a sample size of one — a coin flip formatted as a percentage; reaching even a ±10% confidence interval on a mention probability takes about 96 samples per query–engine–window–region cell, and ±5% about 385 (BA-W-2026-01), so a once-daily reading falls roughly two orders of magnitude short of the floor.
2. A trend line of daily point estimates redraws as events the very drift that an interval would have absorbed: an identical, unchanged prompt re-issued on two consecutive days overlapped only 34–42% in its cited sources and 45–59% in its mentioned brands (the brand figure from the study's three qualifying verticals), across four answer engines over 45–46 collection days (2026-01-24 → 2026-03-20) [2], so much of a bare line's day-to-day movement is that variance, not client-driven change.
3. Monitoring one frozen query string per topic measures that wording rather than the information need: rewording a query while holding its intent fixed dropped the overlap of an assistant's recommended-brand set to a Jaccard similarity of about 0.29 (95% CI 0.22–0.36) for a cosmetic rewrite — for sets of comparable size, roughly 55% of each recommendation set replaced — far below the 0.50–0.61 overlap of the same prompt simply re-run, measured across roughly 12,000 runs on two production models (one OpenAI, one Anthropic) [1], and a frozen wording cannot even expose that between-phrasing variance.
4. Collection through an official API cannot, unaided, support a claim about what users see in the consumer interface: the two are different deployments of the same engine whose retrieval, routing, and answer composition can differ, so a brand can be present in one and absent in the other within the same window (BA-C-3), and a genre that collects by API while speaking of user visibility conflates them.
5. A dashboard with no change-point layer cannot separate a client's real movement from a platform-wide engine update — a broad, synchronous score shift across unrelated queries on one date is the signature of an engine change, not of anything the clients did — and a cited-or-not verdict collapses a refusal, which is an availability observation, into 'not cited,' biasing the rate; both are required handling under BA-C-3 and the Barkhausen Criterion.

CONTENTS

1 The genre 4

2 The genre against the disclosure floor 4

3 What the genre gets right 7

4 What a conforming tool would show 8

5 Five questions for a buyer 9

6 Limitations 9

1 The genre

A class of software now promises to tell an organization whether it is "visible" in AI assistants. The products differ in detail, but they have converged on a recognizable form, and it is the form — not any one product — that this paper examines. The common shape is a dashboard. For each of a set of tracked queries, it issues the query to one or more answer engines on a regular cadence, usually daily, and records whether the tracked brand appeared. From those daily readings it renders four things: a score, often on a bounded scale, summarizing how present the brand is; a trend line plotting that score over time; a per-query verdict, typically a binary "cited" or "not cited" for the latest reading; and a set of coverage badges showing which engines were checked. Some tools add a competitive view — a share-style figure against named rivals, or a rank. The furniture varies. The genre does not, and no single product need carry every element named here for the common form to be recognizable.

Described this way, the genre is doing something reasonable on its face. It watches over time rather than capturing once; it separates engines; it reduces a messy stream of answers to a number a non-specialist can act on. Those are the instincts of measurement. The argument of this paper is that the genre, as a class, stops short of measurement — not because its builders are careless, but because the dashboard form itself omits the facts that would let anyone judge the numbers it displays. A companion convention in this series, BA-C-4, states the minimum a published AI-visibility claim must disclose to be evaluable at all: ten items, chosen so that any competent measurer can meet them without revealing raw data, query lists, or tooling. The dashboard genre, held against that floor, comes up short on the items that carry the weight.

This paper names no tool and evaluates none. That restriction is not evasion; it is the method. The unit of critique is the practice — the common conventions of a product category — evidenced by what the genre openly shows. The reader is invited to do the evaluation the paper declines to: take whatever tool is in front of you, and hold it against the ten items and the five questions below. Where a tool already discloses them, the critique does not apply to it. Where it does not, the tool is not thereby wrong — it is unevaluable, which for something sold as a measurement is a specific and consequential defect.

2 The genre against the disclosure floor

BA-C-4 lists ten disclosures. The table below walks the dashboard genre against each: what the common form typically shows, and what evaluation requires. The rows where the two columns diverge are the subject of the rest of this section.

#	BA-C-4 DISCLOSURE ITEM	WHAT THE DASHBOARD GENRE TYPICALLY SHOWS	WHAT EVALUATION REQUIRES
1	Entity and information need	The tracked brand and a list of monitored queries	The same — generally the genre's strong point
2	Point estimate with interval and method	A point score and a trend line of point estimates	An interval on every point, with the method named
3	Sample size *n*, per cell	Usually unstated; a daily cadence implies one run per query per day	*n* per query–engine–window–region cell, justified against a target precision
4	Engine, interface, version or date	Per-engine coverage badges; date implied by the daily reading	Engine with interface and version or build, not a brand-name badge
5	Measurement channel	Rarely stated; collection is typically API-based	The channel named, and a consumer-interface basis for user-facing claims
6	Time window	The timestamp of each daily reading	A window over which the estimate is defined, not an instant
7	Region	Sometimes a locale setting; often unstated	The region attached to every number
8	Phrasing representation	One fixed query string per topic	A phrasing distribution per need, with count and coverage disclosed
9	Refusal handling	Not represented; a refusal reads as "not cited"	Refusals recorded and rated as availability observations
10	Detection method	A "cited / not cited" verdict per query	The detection method stated, with its precision and recall

Six of these gaps carry the argument. Each corresponds to an instability that the public research on answer engines has already documented, so the gap is not a hypothetical fussiness but a known way to be wrong.

Sample size: the daily score is one coin flip. A dashboard that issues each tracked query once a day is running one trial per query per day. The outcome of that trial is binary — the brand appeared, or it did not — so a single day's reading for a single query is one Bernoulli draw. BA-W-2026-01 sets out the consequence in full, and it is not re-derived here: a single 0/1 observation of a mention has a standard error as large as 0.5 on the 0–1 scale, and pinning a mention probability to a ±10% interval takes about 96 independent samples per query–engine–window–region cell, ±5% about 385. A once-daily check supplies one. The score the dashboard prints for that query on that day is therefore a coin flip formatted to look like a rate. Aggregating the daily readings into a smoother monthly line does not repair this, for two reasons. The month's sample size climbs only to about thirty — still short of the ninety-odd the loosest useful interval needs — and the aggregation is valid only if the underlying probability held constant across the month, which the drift evidence below says it did not. A larger picture assembled from one-per-day draws inherits the thinness of each draw.

Intervals: the trend line plots its own noise. The genre's signature object is the trend line: daily point estimates connected into a curve. A line invites the eye to read its every rise and fall as an event — a gain to celebrate, a dip to explain. But the day-to-day movement of an answer engine is already characterized, and most of it is not events. Re-issuing an identical, unchanged prompt on two consecutive days overlapped only 34–42% in the sources it cited and 45–59% in the brands it mentioned — the brand figure computed on the three verticals that met the study's brand-detection threshold — measured across four

answer engines over 45–46 collection days (2026-01-24 → 2026-03-20) [2]. The brand figure is the load-bearing one for a visibility dashboard: even for an unchanged question on an unchanged engine, roughly two-fifths to a half of the mentioned brands turn over from one day to the next. That turnover is exactly the quantity a confidence interval is meant to express. A dashboard that plots bare points and joins them with a line has taken that variance and redrawn it as motion along the line — the wobble is the interval, wearing the costume of a trend. BA-C-2 requires an interval on every reported estimate, computed by a boundary-valid method (Wilson or Clopper–Pearson) where the estimate sits near zero or one, as it often is for a challenger brand or a niche entity. A line without a band cannot tell its reader whether Tuesday differs from Monday by more than the engine's own daily noise.

Phrasing: one frozen query measures the wording. To produce a stable daily series, the genre holds the query text fixed: the same sentence, issued each day, for each topic. Freezing the wording makes the series look steadier, but it does so by hiding a large source of variation rather than removing it. Rewording a query while holding its underlying intent fixed — the difference between two natural ways of asking the same thing — dropped the overlap of an assistant's recommended-brand set to a Jaccard similarity of about 0.29 (95% CI 0.22–0.36) for a cosmetic rewrite — on comparably sized sets, roughly 55% of each recommendation set replaced — and to about 0.14 when the rewrite added a qualifying constraint; both sit far below the 0.50–0.61 overlap of the identical prompt simply re-run, measured across roughly 12,000 runs on two production models (one OpenAI, one Anthropic) [1]. Two consequences follow for a dashboard. First, a tool tracking one wording is measuring that wording, not the information need behind it; the brand may be present for the sentence the tool happens to use and absent for the other sentences real users type, or the reverse. Second, because the wording never varies, the tool is structurally blind to this variance — its numbers are quiet not because the quantity is stable but because the instrument only ever pokes it in one place. BA-C-3 requires each information need to be represented by a distribution of real phrasings, with the number of phrasings and the coverage strategy disclosed, precisely so that the estimate is of the need and not of a sentence.

Channel: an API reading is not what users see. A dashboard collects at scale, and scale means automation, which in practice means an official API. The claim the dashboard then renders — that a brand does or does not "show up" when people ask — is a claim about the consumer interface, the surface real users see. BA-C-3 treats these as different deployments of the same engine, not two windows onto one: retrieval configuration, model routing, tool invocation, and answer composition can all differ between the API and the consumer product, so a brand can be absent from an engine's consumer answers while present in its API answers within the same window, or the reverse. A claim about what users see must therefore be measured on the consumer interface, or be accompanied by a stated calibration of the collection channel against consumer-interface samples for the same cells and window. A genre that collects by API and speaks of user visibility, with no channel stated and no calibration shown, has quietly substituted one deployment for another and reported the result as though the substitution were free.

Refusals: a decline is not a zero. The genre's per-query verdict is binary: cited, or not cited. That vocabulary has no cell for a third outcome that occurs in practice — the engine declining to answer at all, because the query trips a filter or touches a restricted topic. Under a two-valued scheme a refusal is recorded as "not cited," indistinguishable from a fully answered response in which the brand simply did not appear. But the two are different observations. A refusal is a fact about the query's availability, not about the brand's standing within an answer, and folding it into "not cited" biases the reported rate — down, when an unavailable query is counted as a visibility failure, and up elsewhere, when refusals are silently dropped rather than counted. BA-C-3 requires refusals to be recorded as availability observations

and the refusal rate reported per cell alongside the visibility figure. A dashboard with only two colors cannot carry that third fact.

Engine change: a synchronous drop is a platform event. The most consequential gap is the one a single client never sees. Answer engines are updated silently — weights, retrieval, and system prompts change without announcement — and such an update shows up as a large, same-direction shift across many unrelated queries on the same date. To a dashboard watching one client, that shift looks like the client's own movement: every tracked query dips together, and the tool renders a plunging line that invites a story about what the client did or failed to do. The signature of an engine update is breadth and synchrony — a coordinated jump across a large fraction of unrelated queries on one date — and it is legible only to an observer looking across many queries and clients at once, which is exactly the vantage a per-client dashboard discards. BA-C-3 requires online change-point monitoring to flag such shifts as likely engine changes rather than genuine movement, and the Barkhausen Criterion it defines makes the distinction a precondition for any change claim: a visibility change counts only when it is significant against proper intervals, sustained across at least two consecutive windows, clear of any flagged engine-change point, and controlled for the multiplicity of the many queries tested at once. A dashboard with no change-point layer cannot meet the Criterion even in principle. Worse, it will bill the platform's reset to the client — as a triumph if the reset happened to lift the brand, as a failure if it lowered it — and neither attribution has any basis.

3 What the genre gets right

A critique of a practice owes the practice its due, and the dashboard genre has real merit that the gaps above should not obscure.

Its founding instinct is correct. The genre watches visibility over time instead of capturing it once, and that is the harder and better thing to build. The alternative the genre has already rejected — a single screenshot filed as proof — is worse on every axis; BA-W-2026-01 is largely an argument against exactly that single-shot verification, and a daily series, whatever its per-day thinness, at least refuses the screenshot. The ambition to produce a longitudinal record is the right ambition, and it is the precondition for everything the conventions ask on top of it.

Where a tool keeps engines separate rather than pooling them into one blended score, it is honoring a real requirement, not adding a nicety. Instability differs sharply between engines — the same public study that puts consecutive-day source overlap at 34–42% across four engines reports, for the same prompts within a single 24-hour span, per-engine source overlap ranging from about 0.23 on one engine to about 0.51 on another [2] — so a figure pooled across engines describes none of them, and BA-C-3 requires estimates to be reported per engine for that reason. A per-engine dashboard has that part of the structure right even where the per-cell arithmetic is thin.

And as a way to surface individual answers, a daily crawl is genuinely useful. Reading the actual sentences an engine produces about a brand — noticing a confident factual error, a stale claim, a competitor recommended in the brand's place — is valuable work, and a tool that collects those answers daily is a good instrument for it. That is qualitative monitoring, and it is valuable precisely as a stream of anecdotes to be read, not as a rate to be trusted. BA-C-4 is explicit that a statement framed as anecdote and presented as nothing more — "here is one answer we received" — is outside its scope; the failure begins only when a count and a trend line are laid over the anecdotes and the composite is presented as a measurement. The genre's data-collection layer is often sound. It is the statistical layer above it that the disclosure floor indicts.

The honest summary is that the dashboard genre is a capable qualitative monitoring surface wearing the furniture of a measurement instrument. The surface is worth having. The furniture is what misleads.

4 What a conforming tool would show

Nothing in the conventions demands that a tool be less useful, less automated, or less legible. A conforming tool would look much like the genre, with a different set of objects on the screen — bands where there are now lines, windows where there are now instants, flags where there are now silent resets. The sketch below is descriptive, not a recommendation of any product, and each element follows directly from BA-C-2, BA-C-3, and BA-C-4.

Every plotted point would carry an interval, not sit bare on a line. The band would be wide when the day's sample is small and would narrow as the sample grows, so that a reader sees the precision of each reading directly and cannot mistake a one-draw wobble for a move. Near zero or one — where estimates for challenger and niche brands often live — the band would be computed by a boundary-valid method rather than a normal approximation that can run off the end of the scale.

Each point, and each cell, would show its sample size. A reader could see that today's figure rests on one draw or on a hundred, and the design would make the difference visible rather than rounding both to the same clean percentage.

Every figure would name the window it summarizes, not merely the moment it was taken. "This week," or a labeled window such as 2026-W27, describes a quantity; a bare timestamp describes an instant that cannot be compared to anything.

Each number would carry a channel badge — consumer interface or API — and a claim about what users see would be sourced from the consumer interface or explicitly calibrated against it. The tool would not let an API reading silently stand in for the user's experience.

A detected engine change would appear as a marked, dated line across the series, so that a platform-wide reset is visibly separated from client movement, and change claims would be gated on the Barkhausen Criterion — significant, sustained across at least two windows, clear of any flagged engine-change date, and adjusted for the many queries tested at once — rather than fired on a single day's jump.

Refusals would form their own series, not a silent fold into "not cited," so that a query which is frequently declined reads as unavailable rather than as a visibility failure. And each topic would rest on a disclosed number of real phrasings, its estimate pooled across them, so that the figure describes the information need rather than one frozen sentence.

Reduced to a single number, the result would carry its own evidence, in the conforming form BA-C-4 sets out:

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

A number in that form can be checked, compared, and defended when the engine moves. A bare score of 62, plotted on a line, cannot. The distance between the two is not a matter of polish; it is the distance between a measurement and a claim.

5 Five questions for a buyer

The critique above reduces to five questions any buyer of an AI-visibility tool can ask before trusting its numbers. None asks the vendor to reveal a query list, a model, or any trade secret; each asks only for one item of the BA-C-4 disclosure floor, which any measurer can meet without surrendering confidential material. They are written to be used.

1. **What is the sample size behind each number — how many times is each query asked, per engine, per day, and per plotted point?** If the answer is "once a day," the daily score is a single coin flip, and the smoother long-run line is built from coin flips (BA-C-4, item 3).
2. **Does every number carry a confidence interval, and by what method?** A line of bare points cannot show whether a rise or a dip exceeds the engine's own day-to-day noise, and near zero or one the interval must be a boundary-valid one (BA-C-4, item 2).
3. **How many distinct phrasings stand behind each tracked topic, and how were they chosen?** One frozen query string measures that wording, not the information need, and cannot even see the between-phrasing variance the evidence shows to be large (BA-C-4, item 8).
4. **Through which channel are answers collected, and does it match the surface where you claim users see us?** Collection through an API cannot, unaided, support a claim about the consumer interface; the tool should name the channel or show a calibration between them (BA-C-4, item 5).
5. **How are engine-wide changes and refusals handled?** A synchronous score drop across many clients on one date is a platform update, not any one client's decline, and a refusal is an availability outcome, not a zero mention; a tool that flags neither will misattribute both (BA-C-4, item 9 and the change-claim requirement of §3).

A tool that can answer all five is disclosing enough to be judged, and may well be excellent. A tool that cannot is not thereby proven wrong — it is unevaluable, and for a product sold as measurement, being unevaluable is the defect that matters, because it removes the buyer's ability to tell an excellent tool from a poor one at all.

6 Limitations

This paper critiques a genre, and a genre is a moving target. The common form described here is the dashboard practice as its conventions stand in mid-2026; the product category is young and changing quickly, and this critique will date as the genre does. These observations describe the practice as it is now, and should be assumed perishable in the same way the engines themselves are.

The unit of analysis is the practice, not any product, and no product is assessed, ranked, or endorsed here in either direction. Individual tools may already do several of the things the conforming-tool sketch describes; where a tool discloses the items the genre typically omits, the audit simply does not apply to it, whoever makes it. Nothing here should be read as a claim about any specific tool's quality, and none was examined for the purpose of making one.

The evidence base is deliberately narrow, and its narrowness is worth stating plainly. This paper rests on the public instability literature [1][2] and on the conventions of this series (BA-C-2, BA-C-3, BA-C-4, and the argument of BA-W-2026-01); it does not rest on a survey of tools. No tool inventory, feature census, or hands-on audit of any product was conducted. The genre is characterized from its publicly observable conventions — the objects these tools show on screen and describe in their own material — and the reader

is invited to test that characterization against whatever tool is in front of them. A characterization built this way can be wrong about a particular product precisely because it looked at none; that is the cost of the no-vendor rule, accepted openly.

The instability figures carry their own limits. The rewrite-sensitivity magnitudes [1] come from a single commercial-recommendation setting on two production models over one collection period, and the day-to-day drift figures [2] from four engines across four verticals over 45–46 collection days in early 2026; neither set of numbers should be treated as a constant of answer engines in general. They are cited as evidence that the underlying distributions are wide and moving — which is all the argument needs — not as fixed properties. These results describe the engines as sampled during their stated windows; engines change without notice, and results should be assumed perishable.

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 2026-07-10. [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 2026-07-10. [archived]

HOW TO CITE

Barkhausen AI (2026). The measurement gap in AI-visibility tooling. https://barkhausen.ai/research/measurement-gap-tooling/