Barkhausen AI

WHITEPAPER · BA-W-2026-02

An introduction to AI visibility measurement

When a person asks an AI assistant a question — which schools to consider, which vendor to trust, which clinic to visit — the answer names some organizations and omits others.

“Every leap begins as a crackle.”

DOCUMENT	BA-W-2026-02
SERIES	Whitepaper
DATE	2026
SUBJECTS	Measurement & statistics; Corpora & training data
LICENSE	CC-BY-4.0
AUTHOR	Barkhausen AI

ABSTRACT

When a person asks an AI assistant a question — which schools to consider, which vendor to trust, which clinic to visit — the answer names some organizations and omits others. Being named is AI visibility, and it is not the same as ranking on a search results page. This introduction defines the phenomenon and explains why it must be measured as a probability rather than checked once. It distinguishes the two ways an entity becomes known to an assistant — live retrieval of sources at answer time, and parametric memory formed during training — and summarizes the public evidence that answers are unstable: rewording a question slightly, or asking it again tomorrow, changes the sources and the names returned. It introduces three metrics — Visibility Probability, Share of Voice, and Discovery Depth — and the Barkhausen Ladder, the field's maturity map, pointing to the conventions for formal definitions.

IN BRIEF

Ask an AI assistant a real question and it answers in prose, naming a handful of organizations, products, or places. Whether your organization is among them is your AI visibility. It is different from search: there is usually no ranked list of ten links, the answer is written fresh each time, and the same question asked twice can name different sources and different organizations. That instability is the central problem. Because the answer varies, "did we appear?" is the wrong question; "how often, and how confident are we in that number?" is the right one. Measuring visibility therefore means asking many times and reporting a probability with a margin of error, not taking a screenshot. This document explains what AI visibility is, why it behaves this way, and what it takes to measure it credibly — the starting point for everything else published here.

KEY FINDINGS

1. AI visibility — being named in an assistant's synthesized answer — is a distinct phenomenon from search ranking: answers are generated rather than listed, carry no fixed order, and can differ between two identical or near-identical questions.
2. An entity can become known to an assistant along two different paths: live retrieval, where sources are fetched at answer time (fast, controllable, volatile), and parametric memory, where facts are absorbed during training (slow, durable, largely uncontrollable).
3. Answer variance is documented in public research: in one 2026 study, lightly rewording a question while keeping its intent fixed cut the overlap of the brands an assistant recommended to a Jaccard similarity of about 0.29 — the two recommendation sets sharing fewer than a third of their combined entries — far below the 0.50–0.61 overlap of simply re-running the same prompt [2]; and repeating an identical question on consecutive days left only 34–42% of cited sources and 45–59% of named organizations in common [3].
4. Because a single check can only return "named" or "not named," visibility must be estimated as a probability from repeated sampling and reported with a confidence interval; the field's core metrics are Visibility Probability, Share of Voice, and Discovery Depth (defined in BA-C-2).
5. Common Crawl, a free, roughly monthly web archive, is the shared upstream for most open training corpora — for the one large model with a fully disclosed data composition, it supplied about 82% of the assembled training dataset by token count (and 60% of the tokens actually sampled in training), and 64% of the 47 models surveyed in a 2024 audit used at least one Common Crawl–derived corpus — so what is absent from that pipeline can be absent from many models at once [1, 4, 5].

CONTENTS

Someone opens an AI assistant and types a genuine question: *which universities should I consider for a master's in data science in Germany? Which payroll vendor is best for a fifty-person company? Where can I get a same-day dental crown near me?* A few seconds later a paragraph comes back. It is fluent, it sounds considered, and it names a handful of specific organizations — three schools, two vendors, one clinic — while saying nothing about the dozens of others that could equally have been named.

Whether an organization is among the named few is the subject of this document. This introduction is written for someone who has never had to think rigorously about it: a communications lead who has just been asked "are we showing up in the AI answers?", a researcher entering the field, a journalist trying to describe it accurately. It explains what the phenomenon is, why it does not behave like search, why it has to be measured as a probability rather than checked once, and what a credible measurement looks like. It is the starting point; the conventions and reports referenced throughout carry the formal definitions and the detailed evidence.

1 1. What AI visibility is

AI visibility is the tendency of an AI assistant to name a given entity — an organization, product, person, or place — in the answers it generates to relevant questions. When an assistant responds to *which study-abroad agencies are reputable?* by naming three agencies, those three have AI visibility for that question and the rest do not. Being named is the unit of the phenomenon. Everything measured in this field is, at bottom, a refinement of *how often, in what company, and how prominently* an entity gets named.

This matters because the interface is changing where attention lands. A traditional search result page offers a list; the reader scans it and chooses. An assistant's answer offers a recommendation already made — the shortlisting has happened inside the system, out of view, before the reader sees anything. Entities that are named enter the reader's consideration set; entities that are not named are, for that reader in that moment, invisible. The stakes of being named therefore rise as more people begin their decisions by asking an assistant rather than by scanning a results page.

Two clarifications prevent early confusion. First, AI visibility is about *being named in the answer*, not about being cited in a footnote or a source list, though the two are related and both are measured. An assistant can name an organization it does not link to, and can link to a page without naming the organization prominently; a serious measurement distinguishes these. Second, AI visibility is question-specific. An organization can be named reliably for one phrasing of a need and never for another. There is no single "AI visibility score" that is true in general — only visibility for a defined space of questions, which is why measurement begins by specifying the questions.

2 2. Why it is not search ranking

The instinct of anyone with a marketing background is to treat this as search engine optimization (SEO) by another name. That instinct is a good starting point and a misleading destination. Three structural differences separate AI visibility from search ranking.

The output is synthesized, not listed. A search engine returns a ranked list of documents that exist independently of the query; the engine's job is ordering. An assistant *writes* an answer, composing prose that may name entities without linking to any specific document for each one. There is often no list at all, and where there is, it is generated text rather than a retrieved ranking. The question "what position are we in?" frequently has no answer, because there is no ordered set of positions to occupy.

There is usually no stable ranking to hold. In search, a page that ranks third for a query tends to rank third when the same query is repeated a minute later. An assistant's answer is produced by a probabilistic process and can name a different set of entities on each generation. Position, where it exists at all, is not a property you hold; it is an outcome you obtain with some frequency.

The same question, asked twice, can give different answers. This is the difference that reorganizes everything downstream. Search is close to deterministic for a fixed query and index; assistant answers are not. Public research (Section 4) shows that a small change in wording, or simply asking again on another day, can change which sources the system consults and which organizations it names. In search, variability is the exception to be explained; in AI answers, it is the baseline condition to be measured.

The practical consequence is that the familiar SEO artifacts — a rank-tracking dashboard, a "we are number one" screenshot — do not transfer. A screenshot of an assistant naming your organization is a single sample from a distribution, not a status you have achieved. The rest of this document is, in effect, an argument for taking that sentence seriously. Generative engine optimization (GEO) is the emerging name for the practice of improving AI visibility; this publication concerns how to *measure* it, which must come first.

3 3. Two paths to being known

An assistant can name an organization for one of two fundamentally different reasons, and telling them apart is the first genuinely technical idea a newcomer needs.

Path one: retrieval. Many assistants, when answering, fetch documents from a live index at the moment of the query — a search step folded into the answer — and write their response partly from what they just read. This is often called retrieval-augmented generation. If an organization is well represented in the sources the system fetches, it is likely to be named. This path is comparatively **fast** (a new page can begin influencing answers within days), comparatively **controllable** (it responds to the same forces as search visibility — being reachable, being cited by authoritative sources), and comparatively **volatile** (what is retrieved shifts as indexes update and as the system's internal search varies). We call presence obtained this way **retrieval visibility**.

Path two: parametric memory. During training, a model reads an enormous body of text and adjusts billions of internal parameters. Facts that appear widely and consistently in that text can become, in effect, part of what the model "knows" without looking anything up. Ask such a model a question with its live search turned off — a *closed-book* question — and it may still name an organization it has never retrieved, because the association was absorbed into its parameters during training. This path is **slow** (it is bounded by training cycles and by a model's knowledge cutoff, so influence is measured in months to a year or more), **durable** (once an association is in the weights it does not fluctuate day to day, and it is hard for a competitor to displace in real time), and **largely uncontrollable** (no one can guarantee that a specific sentence will be memorized). We call presence obtained this way **parametric visibility**, or closed-book visibility. The distinction matters because the two paths have different levers, different timescales, and different evidence. It also carries an honesty boundary that must be stated plainly, because it is easy to overclaim. The mechanics of getting content *into* a training pipeline are increasingly understood; what happens *inside* the model afterward is not guaranteed. **Presence in a training corpus does not imply the model retains or reproduces the content; no causal claim is made.** Corpus entry is a demonstrated upstream mechanism; a model's memory of a specific fact is an unproven downstream step. The correct promise for the parametric path is that effort raises the *probability and expectation* of being remembered correctly — never that a fact is guaranteed to be remembered.

Figure 1 places the two paths side by side: the same public web page reaches an answer either through a live index that refreshes in days, or through a training corpus that is fixed for the life of a model generation.

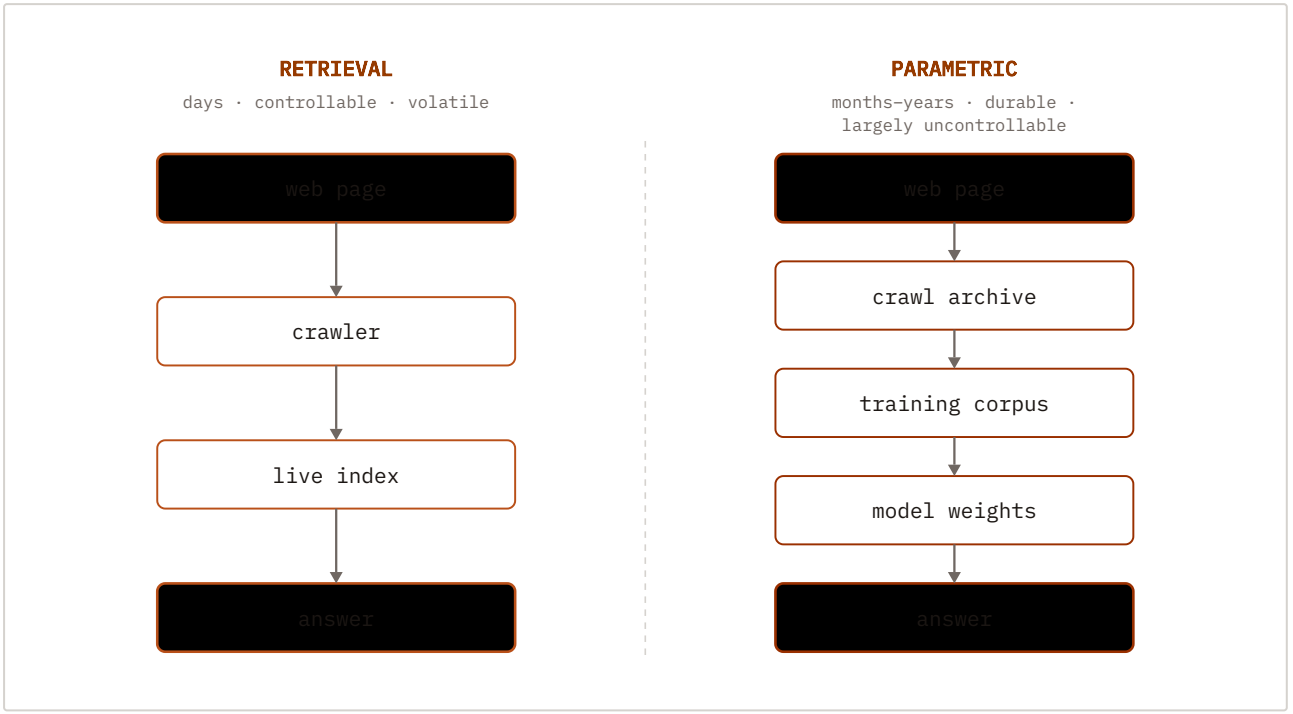

Figure 1. The two paths by which an entity becomes known to an assistant. The retrieval path (left) runs through a live index and turns over in days; the parametric path (right) runs through a training corpus into model weights and is fixed for the life of a model generation. Both begin at a public web page and end at an answer.

Barkhausen AI · BA-W-2026-02

One public fact grounds why the parametric path is worth attention even though it is slow and indirect. Most open training corpora are not independent; they descend from a common upstream. **Common Crawl**, a nonprofit that has archived the public web on a regular cadence since 2008 and releases the result for free, is the raw material behind most of them [4]. For the one large model with a fully disclosed data composition, Common Crawl supplied about 82% of the assembled training dataset by token count — and about 60% of the tokens actually sampled during training [1] — and a 2024 audit found that 64% of the 47 models it examined had used at least one Common Crawl–derived corpus in pretraining [4]. The most modern openly documented corpus is drawn *entirely* from Common Crawl, built from 96 snapshots spanning 2013 to 2024 and totaling roughly 15 trillion tokens of text [5]. The practical reading is not that any one model is knowable, but that this shared pipeline is a single front door: what is absent from it can be absent from many models at once, and what is present in it is at least eligible to reach many models over time.

4 4. Why it must be measured as a probability

Return to the sentence that separates this field from search: the same question, asked twice, can give different answers. If that is true, then “did we appear?” is the wrong question, because its answer is not stable enough to act on. A single check tells you what happened on one draw from a distribution — nothing more. The right question is *how often do we appear, and how confident are we in that number?*

The public evidence that a single check is untrustworthy comes from several independent directions. In a 2026 study of roughly 12,000 runs across two production models — one from OpenAI, one from Anthropic — rewording a question while keeping what it asked for fixed sharply changed which brands the system

recommended: a light rewrite cut the overlap of the recommended set to a Jaccard similarity of about 0.29 (the two sets sharing fewer than a third of their combined entries), and a rewrite that added a constraint cut it to about 0.14 — both far below the 0.50–0.61 overlap seen when the identical prompt was simply re-run, which shows the change came from the wording rather than ordinary randomness [2]. Separately, repeating the *identical* question on two consecutive days left only 34–42% of the cited sources in common across the two days, with the set of named organizations somewhat more stable at 45–59% overlap [3]. The wording changed only cosmetically; the day did not change the world; the answers changed anyway. This is why visibility is estimated, not checked. Ask a defined question many times under controlled conditions, count how often the entity is named, and the fraction is an estimate of an underlying probability. Formally, if an entity is named in k of n independent samples, the point estimate is $\hat{p} = k/n$, and it is reported not as a bare number but with a **confidence interval** — a range that expresses how much the estimate could move if the sampling were repeated. The interval narrows as n grows. A useful intuition without the algebra: pinning a probability near one-half to within a few percentage points takes on the order of hundreds of repeated questions, not one and not ten. A screenshot is a sample of size one; its confidence interval spans nearly the whole range from zero to one (Figure 2). The sampling requirements — how many samples, over what window, disclosed how — are the subject of BA-C-3.

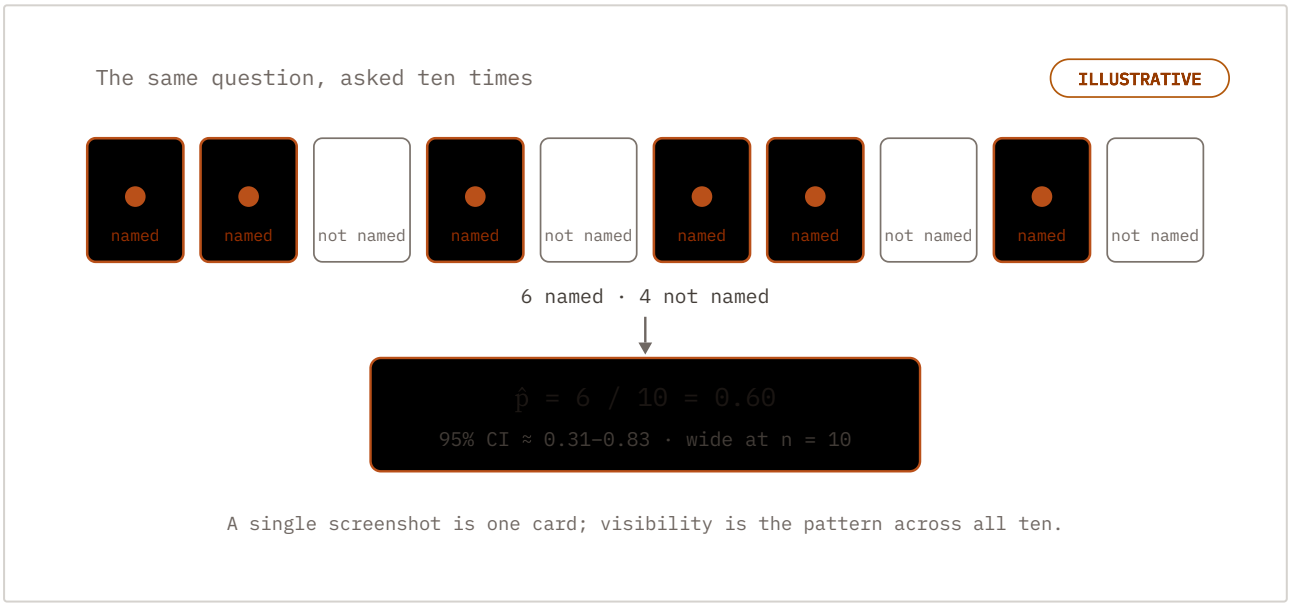

Figure 2. Illustrative. A single answer is one draw: it either names the entity or it does not. Asking the same question ten times yields ten outcomes — here the entity is named in six and not in four — which together estimate a probability with a wide confidence interval. The six-of-ten split and the interval shown are illustrative, not a measured result.

Barkhausen AI · BA-W-2026-02

This reframing gives the field its three core metrics, defined formally in BA-C-2 and introduced here only at the level of intuition:

- **Visibility Probability** (**VP**) — the probability that an entity is named in an answer to a defined question, estimated from repeated sampling and always reported with its confidence interval. VP is the base measurement; the other two build on it.
- **Share of Voice** (**SoV**) — of all the naming that happens in an answer, the fraction that is yours: your mentions relative to the mentions of competitors named alongside you. Where VP asks "how often am I named at all," SoV asks "when the category is discussed, how much of the conversation is about me?" It tends to be steadier than VP when an engine's behavior shifts across the board, because relative standing survives changes that move every absolute number together.

- **Discovery Depth** (**DD**) — how specific a question has to become before an entity enters the recommended set. A broad question (*study-abroad agencies*) may name only the largest few; add constraints (*affordable study-abroad agencies in Munich specializing in engineering PhDs*) and a different, longer set appears. An entity with shallow discovery depth is named even for broad questions; one with deep discovery depth surfaces only once the question narrows to its niche. DD measures reach across the space of questions, not just frequency within one.

There is a companion idea for deciding when a measured change is real. Because the numbers move on their own, a claim that visibility has genuinely improved should be held to a threshold: the change must be statistically significant and sustained across consecutive windows — and it must survive two further checks, that it does not coincide with a platform-wide engine change and that it is not one lucky find among many comparisons — not a one-period blip that variance could have produced. This publication adopts **the Barkhausen Criterion** for that test; the reasoning connects to the physics the name borrows, discussed in Section 7.

5 5. A short tour of the Barkhausen Ladder

Measurement answers "how visible are we?" It does not by itself answer "what would it take to become more visible?" That is the work of a maturity model. The **Barkhausen Ladder** (BA-C-1) arranges the requirements of AI visibility into nine levels, from technical hygiene up to the emerging frontier of agent-readiness. It is a map, not a to-do list, and its purpose here is orientation: to show a newcomer the whole terrain at a glance and where measurement sits within it.

LEVEL	NAME	WHAT IT COVERS
BL-0	Search hygiene	The fundamentals: pages that can be reached, fetched, and parsed; clean canonical URLs; valid markup; nothing accidentally blocked. If a page cannot be read, nothing above it matters.
BL-1	Crawler access	Knowingly deciding which automated crawlers — those that feed retrieval indexes and those that feed training corpora — are allowed, and writing access rules that say what their owners intend.
BL-2	Structured data and entities	Machine-readable identity: standard markup and an entity record precise enough that an assistant can be sure which organization of a given name you are.
BL-3	Content form	Writing in the shape assistants can lift cleanly: self-contained, well-formed, fact-dense passages rather than diffuse prose.
BL-4	Distribution	Being present and cited where the engines look — authoritative third-party sources, directories, and reference works, not only your own site.
BL-5	Measurement	Estimating visibility as a probability with confidence intervals: the subject of this document and of the conventions. Nothing above this level can be managed without it.
BL-6	Algorithmic optimization	Systematically improving how often and how favorably an entity is surfaced, evaluated against measured baselines. Treated here at the level of what and why, not how.
BL-7	Corpus presence	Appearing durably in the training-data pipeline so a model can name you without retrieving anything — the parametric path of Section 3.
BL-8	Agent-readiness	Being usable by autonomous agents that act on answers, through structured actions and machine-consumable interfaces — the emerging edge of the field.

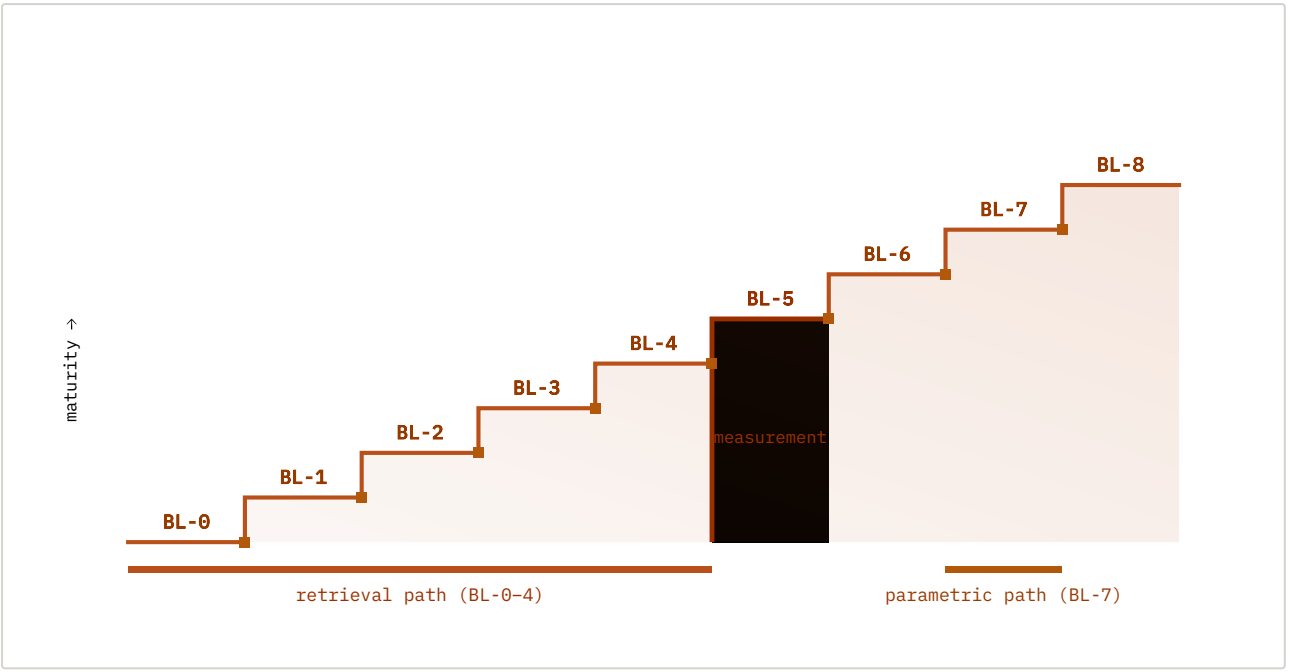

Figure 3. The Barkhausen Ladder in brief: nine cumulative levels, BL-0 through BL-8, with measurement (BL-5) highlighted as the subject of this document. The two paths of Section 3 map onto different rungs — the retrieval path is served mostly by BL-0 through BL-4, the parametric path by BL-7. Full definitions are in BA-C-1.

Barkhausen AI · BA-W-2026-02

Three things are worth noticing about the shape of the ladder. Measurement (BL-5) sits in the middle by necessity: the levels beneath it are prerequisites you can verify by inspection, while the levels above it cannot be pursued responsibly without the ability to tell whether they are working. The two paths of Section 3 map onto different rungs (Figure 3) — retrieval visibility is served mostly by BL-0 through BL-4, parametric visibility by BL-7. And the ladder is ordered by dependency, not by difficulty or value: a well-known organization can be strong at distribution while failing at hygiene, and the failure at the bottom will cap everything above it. The full level definitions, observable criteria, and an assessment checklist are in BA-C-1.

6 6. What good evidence looks like

If the central claim of this field is that visibility is a probability, then the central discipline is reporting that probability honestly. A credible AI-visibility number is not a percentage on its own; it is a percentage wearing its evidence. Four things travel with every number worth trusting.

A sample size. A visibility figure without an n is uninterpretable, because the same figure can be near-certain or near-meaningless depending on how many samples produced it. "Named 62% of the time" means one thing at $n = 400$ and almost nothing at $n = 3$.

A time window. Because engines change and indexes turn over, a number describes a period, not a permanent state. "62% during the last week of June 2026" is a claim that can be checked and re-run; "62%" with no window is a claim about nothing in particular.

An engine and version. Different assistants behave differently, and the same assistant behaves differently across versions and updates. A number that does not say which system, sampled when, cannot be compared to anything or reproduced by anyone.

Sources you can check. The underlying study should name its primary sources — the papers, the official documentation, the released datasets — so a reader can follow the claim to its origin rather than taking it on trust.

A worked example of the house format makes the convention concrete: *named in 62% of answers (95% confidence interval 53–70%, n = 120, [engine and version], sampled 2026-06-22 → 2026-06-28)*. Every element is load-bearing. Strip the interval and the reader cannot tell a real difference from noise; strip the window and the number cannot be reproduced; strip the *n* and the interval could be anything. A claim missing these elements is not necessarily wrong, but it is unfalsifiable, and unfalsifiable claims are the ones this field must learn to discount. The minimum disclosure that any visibility claim should carry is set out in BA-C-4, and the statistical failure modes — single-shot screenshots presented as status, comparisons across undisclosed windows, percentages without intervals — are examined in the companion whitepaper BA-W-2026-01.

One caution belongs with every reported result, and this publication states it verbatim wherever engine measurements appear. **These results describe the engines as sampled during the stated window; engines change without notice, and results should be assumed perishable.** A visibility number is a photograph of a moving subject. It is worth taking, and worth dating.

7 7. About the name

The name is borrowed from physics, and the metaphor is exact enough to be worth a paragraph. In 1919 Heinrich Barkhausen was studying how iron becomes magnetized. Increase the magnetic field around a piece of iron smoothly and continuously, and one might expect its magnetization to rise just as smoothly. It does not. It advances in a rapid series of tiny discrete jumps — internal magnetic domains snapping past obstacles one after another — which Barkhausen made audible as a crackle in a loudspeaker. The drive is continuous; the response is a staircase of sudden steps. This is the Barkhausen effect. AI visibility behaves the same way. Effort is continuous — pages are published, sources accumulate citations, an entity record is corrected and enriched — while for stretches nothing visible changes. Then an index refreshes, a snapshot is crawled, or a new model generation ships, and the entity appears across many questions at once. Progress is real but arrives in jumps, which is precisely why it must be measured rather than watched: the eye sees a flat line and then a surprise, while the instrument sees the underlying probability rising toward the threshold at which the jump becomes visible. A **visibility jump** that clears the Barkhausen Criterion — statistically significant and sustained rather than a momentary flicker — is what a genuine improvement looks like in the data. The etymology is offered because it is true and because it names the thing well; nothing about the work depends on the physics beyond the shape of the picture.

8 8. Where this leads

A newcomer who has followed this far now has the field's frame. AI visibility is being named in a generated answer; it is not search ranking; an entity becomes known through retrieval or through parametric memory, on very different timescales; the answers vary, so visibility is a probability to be estimated with confidence intervals, not a status to be checked; the metrics for that estimate are Visibility Probability, Share of Voice, and Discovery Depth; the full terrain is the Barkhausen Ladder; and credible evidence always carries its *n*, its window, its engine and version, and its sources.

The natural next steps are the conventions that make each of these precise. BA-C-1 defines the Ladder and its levels. BA-C-2 defines the three metrics formally. BA-C-3 sets the sampling requirements — how many samples, over what window, disclosed how — and BA-C-4 states the minimum any published visibility

claim must disclose. The whitepaper BA-W-2026-01 works through the statistics of measurement and the ways it commonly fails. This document is the door; those are the rooms.

9 Limitations

This is an introductory synthesis, and its limits should be as visible as its claims.

It is conceptual, not empirical. It reports no new measurements of its own; the numbers it cites come from external public research, and its purpose is orientation rather than evidence. Readers who need current figures should consult the reports and censuses, where measurements are made under a disclosed protocol.

The public studies it summarizes are themselves bounded. The rewrite-sensitivity finding [2] comes from a single commercial-recommendation setting on two production models, measured over one collection period; it does include a same-prompt rerun baseline — which is what lets it attribute most of the change to the wording rather than to run-to-run randomness — but its exact magnitudes should not be assumed to hold for other engines, domains, or languages. The day-to-day drift figures [3] come from a single study of four answer engines across four German-language verticals in one region, and describe the systems studied over that window, not assistants in general.

Findings about specific engines are perishable. The systems described here change without notice, and any numerical result should be assumed to describe only the window in which it was collected. Where this document states probabilities and overlaps, they illustrate a general condition — that answers vary — rather than a fixed property of any current system.

Finally, the honesty boundary of Section 3 bears repeating as a limit on interpretation. The evidence that content can be engineered into a training pipeline is not evidence that a given model remembers a given fact. Presence in a corpus is an upstream mechanism; parametric memory of a specific fact is a downstream step that public evidence does not establish. Claims about the parametric path should be read as statements about probability and expectation, never as guarantees.

REFERENCES

1. Brown et al. (NeurIPS). *Language Models are Few-Shot Learners* (2020). https://arxiv.org/abs/2005.14165 Accessed 2026-07-08. [archived]
2. 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]
3. 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]
4. Baack, Mozilla Foundation. *Training Data for the Price of a Sandwich: Common Crawl's Impact on Generative AI* (2024). https://www.mozillafoundation.org/en/research/library/generative-ai-training-data/common-crawl/ Accessed 2026-07-08. [archived]
5. Penedo et al. (NeurIPS Datasets & Benchmarks). *The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale* (2024). https://arxiv.org/abs/2406.17557 Accessed 2026-07-08. [archived]

HOW TO CITE

Barkhausen AI (2026). An introduction to AI visibility measurement. https://barkhausen.ai/research/primer-ai-visibility/

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