Barkhausen AI

CONVENTION · BA-C-2

Version 1.0 · Stable

Visibility metrics

This convention defines the metrics a credible measurement of entity visibility in AI assistants must report, and the form each report must take.

“Every leap begins as a crackle.”

DOCUMENT	BA-C-2
VERSION	1.0 (Stable)
EFFECTIVE	2026
SUBJECTS	Measurement & statistics; Conventions & policy
LICENSE	CC-BY-4.0
AUTHOR	Barkhausen AI

ABSTRACT

This convention defines the metrics a credible measurement of entity visibility in AI assistants must report, and the form each report must take. It specifies three primary metrics: Visibility Probability (VP), the probability that an entity is mentioned in an answer to a query drawn from a distribution of real user phrasings, on a given engine, time window, and region; Share of Voice (SoV), the entity's share of all brand mentions in the same answers; and Discovery Depth (DD), the degree of query constraint at which an entity first enters the recommendation set. It adds supporting measures for prominence, citation support, rank-aware list overlap, and sentiment; requires every estimate to carry a confidence interval, sample size, engine version, window, region, and a stated detection method; mandates boundary-valid intervals near zero and one and composition-aware intervals for SoV; and states what claiming conformance requires. Sampling procedure is deferred to BA-C-3.

CONTENTS

A claim that an entity is "visible" in an AI assistant is only as credible as the metric behind it. A screenshot of one favorable answer, or a raw percentage with no sample behind it, does not measure visibility; it records an anecdote. This convention defines the metrics a credible AI-visibility measurement must report, and the form each report must take. It operates at the level of definition and reporting requirement: it fixes what each metric denotes and what a valid report of it must contain. It does not prescribe how the underlying observations are obtained — the sampling design, the coverage of real user phrasings, the number of draws, and the reconciliation of measurement channels are the subject of BA-C-3 and are out of scope here. A well-formatted number obtained from a poorly designed sample is still wrong; this document governs the format and the definitions, and BA-C-3 governs the sample.

Scope. The metrics defined here apply to the measurement of whether, and how, a named entity — a brand, organization, product, or institution — appears in the generated answers of an AI assistant or answer engine, across the queries for which that entity could plausibly be recommended. They are engine-agnostic and category-agnostic. They concern what an engine says in its answers, not whether the entity's content is present in any training corpus; membership in a corpus is a separate question governed elsewhere.

Normative language. The key words MUST, MUST NOT, SHOULD, SHOULD NOT, and MAY in this document carry their established normative meaning. MUST and MUST NOT denote absolute requirements and prohibitions: a report that violates one is not a valid report under this convention. SHOULD and SHOULD NOT denote strong recommendations whose departure requires a stated justification. MAY denotes a genuinely optional element that a report is free to include or omit.

1 1. Estimand: what "visibility" denotes

Before a number can be measured, the quantity it estimates must be defined. This quantity is the *estimand*. Naming it first is what separates a measurement from an impression: it states, in advance and independent of any tool, exactly what a later estimate is an estimate *of*.

For AI visibility the estimand is a probability. Let B be the entity, and let k index an *information need* — a thing a user is trying to find out, such as "which providers should I consider for X." Then the estimand is

$$\mathrm{VP} = \Pr\big(B \text{ mentioned} \mid \text{prompt} \sim Q_k,\ e,\ t,\ g\big).$$

The four conditioning terms are part of the quantity's identity, not decoration. Engine e includes the specific interface and model version, because different deployments of the same underlying model can retrieve and answer differently. Time t fixes the window, because engines are non-stationary. Region g fixes the observation locale, because retrieval is geographically differentiated. A "VP" reported without these is not a well-defined quantity.

The load-bearing term is Q_k. It is the *distribution of real user phrasings* for the information need k: not a single sentence, but a family of semantically equivalent, lexically varied questions that real users actually type to express the same need. "Which study-abroad agencies are good," "recommend some reliable agencies for the United States," and "who should I use to apply to graduate school in the US" are three draws from one such distribution. Visibility is defined over this distribution — as an expectation across the phrasings a real population uses:

$$\mathrm{VP} \ = \ \mathbb{E}_{q \sim Q_k}\big[\,\Pr(B \text{ mentioned} \mid q,\ e,\ t,\ g)\big].$$

Representing an information need by one fixed, canonical sentence implicitly assumes that Q_k is a *degenerate* distribution — a point mass on that single sentence. This assumption is not a harmless simplification. Answer engines are sensitive to phrasing: trivial rewrites of the same question can return

different sources and a different set of recommended brands, and identical prompts issued on different occasions can also diverge. When the phrasing distribution is real and the engine responds to phrasing, a VP estimated from one sentence answers a narrower question — "how visible is B for *this exact wording*" — and is not interchangeable with visibility for the need k. Defining the estimand over Q_k makes the target explicit; how a measurement covers Q_k in practice is a sampling requirement (BA-C-3).

This definition also settles a common confusion. A single answer to a single prompt is one *Bernoulli realization* of the estimand — a coin flip whose bias is the probability being estimated — not the probability itself. Observing that the assistant mentioned the entity once establishes only that the outcome is possible, not how probable it is. The metric is a parameter; an observation is a draw from it. Everything downstream — the confidence interval, the sample size, the interval method near the boundaries — follows from taking that distinction seriously.

2 2. Visibility Probability (VP)

VP is the primary visibility metric. For an entity B and information need k, it is the probability, defined in §1, that B is mentioned in an answer to a prompt drawn from Q_k on a specified engine, window, and region. It is a Bernoulli parameter $p \in [0, 1]$.

Estimation is the estimation of a proportion. From n answers, of which B is mentioned in x, the point estimate is $\hat{p} = x/n$, with standard error

$$\mathrm{SE}(\hat{p}) \;=\; \sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}}.$$

Under the normal approximation, the 95% confidence interval has half-width $E \approx 1.96\,\mathrm{SE}(\hat{p})$. Inverting this relationship gives the sample size required to reach a target precision E:

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

The worst case is $p = 0.5$, where the variance is largest; sizing for that value is conservative and phrasing-independent. The two precision points a report will most often quote are:

TARGET 95% INTERVAL HALF-WIDTH E (AT $P = 0.5$)	DRAWS N REQUIRED
$\pm 10\%$	≈ 96
$\pm 5\%$	≈ 385

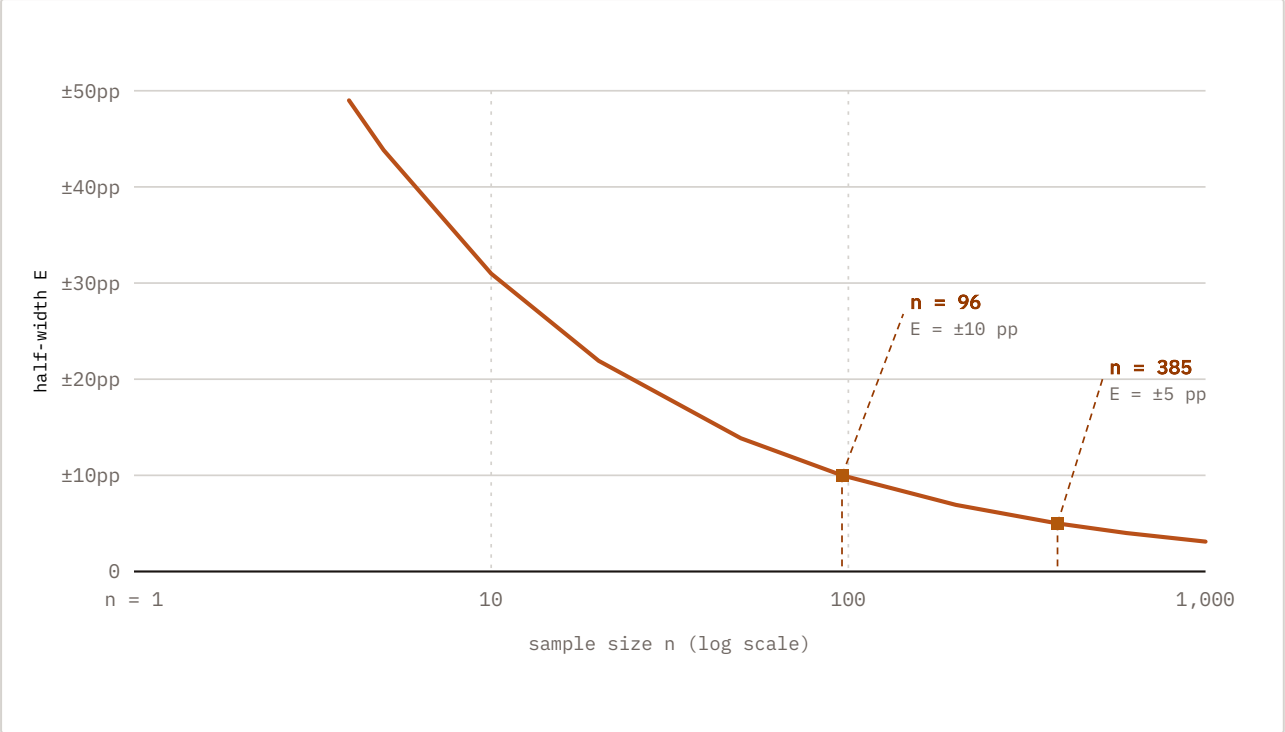

Figure 1. The cost of precision. The half-width of a 95% confidence interval narrows only with the square root of the sample size, so each halving of the margin quadruples the draws required: ±10 percentage points needs about 96 draws, ±5 needs about 385. The curve is drawn from n = 4 (±49 points); at n = 1 a single draw's interval spans the entire scale and is omitted.

Barkhausen AI · BA-C-2 §2 (E = 1.96·√(p(1−p)/n) at p = 0.5)

Figure 1 plots this relationship across the full range of n on a logarithmic scale, marking the two anchor points. These figures are the reason a single observation cannot be a VP. At $n = 1$ the standard error at $p = 0.5$ is 0.5: the estimate is pure noise, and no confidence interval narrower than the entire $[0, 1]$ range can be justified. Reporting a probability requires enough independent draws to constrain it; the table states how many, and BA-C-3 governs how those draws are obtained so that they are genuinely informative about Q_k rather than about one phrasing.

The normal approximation fails near the boundaries. When $\hat{p}$ is close to 0 or 1 — as it often is for a challenger brand or a niche entity — the symmetric interval $\hat{p} \pm E$ can extend below 0 or above 1, an impossible claim (for example, $\hat{p} = 0.05$ with a wide margin yields a negative lower bound). In this regime a report MUST NOT use the normal interval. It MUST instead use the **Wilson score interval** [2] or the **Clopper–Pearson exact interval** [3], both of which stay within $[0, 1]$ and are standard in proportion estimation. With $z = 1.96$ for 95% confidence, the Wilson interval is

$$\frac{\hat{p} + \frac{z^2}{2n} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z^2}{4n^2}}}{1 + \frac{z^2}{n}}$$

The Clopper–Pearson interval is more conservative (guaranteed coverage at or above the nominal level) and is appropriate when strict coverage matters more than interval width. A report SHOULD name which interval method it used; near the boundaries it MUST be Wilson or Clopper–Pearson rather than the normal approximation. Figure 2 contrasts the two constructions at $\hat{p} = 0.05$, $n = 40$, where the normal lower bound falls below zero.

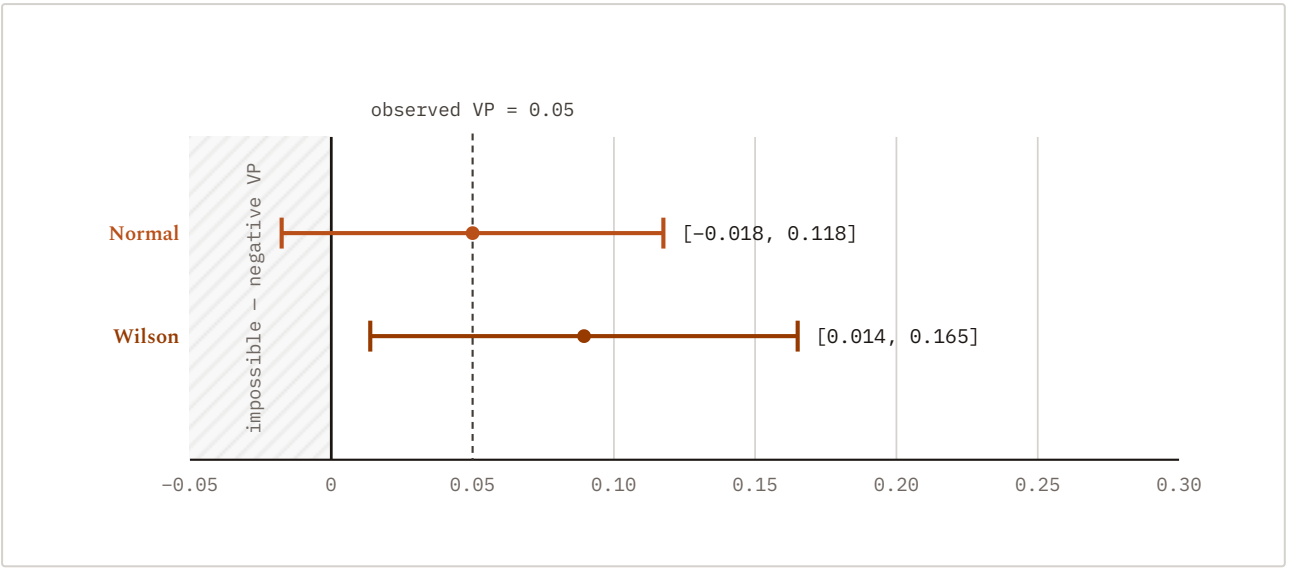

Figure 2. Near the boundary the normal interval breaks. For an observed visibility of 0.05 from n = 40 draws, the normal (Wald) interval [−0.018, 0.118] places its lower bound below zero — an impossible negative visibility (hatched). The Wilson score interval [0.014, 0.165], centered at 0.089, stays within the 0-to-1 range. This is why a report must use Wilson or Clopper–Pearson wherever an estimate lies near zero or one.

Barkhausen AI · BA-C-2 §2 (p̂ = 0.05, n = 40, 95% level)

A valid VP report MUST carry: the point estimate $\hat{p}$; a confidence interval with its confidence level and method; the sample size n; the engine, including interface and version or the observation date; the time window; and the region. The information need k MUST be identifiable at the level of the phrasing distribution it represents. These are consolidated in §6.

3 3. Share of Voice (SoV)

VP measures an entity in isolation. Share of Voice places it against its competitors in the same answers. Within the set of answers sampled for an information need, SoV is the entity's share of all brand mentions:

$$\mathrm{SoV}_B = \frac{m_B}{m_B + \sum_{c \in C} m_c},$$

where m_B is the count (or weight) of mentions of B, C is the competitor set, and m_c is the count (or weight) of mentions of competitor c. Mentions may be counted plainly, or weighted by prominence using the position-adjusted scheme of §5.1; a report states which.

SoV is more robust to engine-wide retrieval shifts than absolute VP. When an engine changes its retrieval behavior and every brand's absolute mention rate moves together — up in a more generous regime, down in a stricter one — the *relative* standing of the entity is more stable than its absolute probability. Because it divides out shifts that affect the whole answer set, SoV is well suited as a long-run metric, tracking competitive position across time windows in which absolute VP may drift for reasons unrelated to the entity. The concept adapts the long-standing "share of voice" measure from advertising and media analysis to generated answers; the position-weighted variant connects to the position-adjusted mention measurement of the founding generative-engine-optimization literature [1].

SoV is still an estimate from finite sampling, so it inherits the reporting requirements of §6: it MUST carry n, a confidence interval, the engine and version, the window, and the region. Its interval demands care beyond the binomial formula, because SoV is a share of a composition rather than a simple proportion, and the mentions within one answer are not independent trials. An SoV interval MUST come from a construction that respects that structure — a nonparametric bootstrap that resamples whole answers, or a

Dirichlet–multinomial model over the per-answer mention counts, are both acceptable; applying the binomial standard-error formula to the share is not. A report SHOULD name the construction it used. Two further requirements are specific to SoV. First, the value depends entirely on the competitor set: an SoV computed against a different set of competitors is a different number and is not comparable. A valid SoV report MUST publish the competitor set it used. Second, a report MUST state the weighting scheme — plain mention count versus position-adjusted weight — because the two can rank an entity differently.

4 4. Discovery Depth (DD)

VP and SoV both presuppose an information need and ask how the entity fares within it. Discovery Depth asks a prior question: how much must a user narrow their request before the entity appears at all?

Real users rarely stop at a bare query. They add qualifying attributes, and the answer's recommendation set changes as they do. A broad need ("MBA programs") admits many candidates; each added constraint ("MBA programs in Germany," then "English-taught MBA programs in Germany under a set budget") shrinks the set and shifts who is recommended. Discovery Depth is the degree of constraint at which the entity *first enters* the recommendation set — for example, the number of qualifying attributes that must be specified before the entity appears.

DD complements the probability metrics. An entity that appears only under heavy constraint is *discoverable* — a sufficiently specific user will find it — but not broadly *visible*; an entity that appears on broad, lightly constrained needs has low Discovery Depth and high reach. Two entities can share a similar VP on a narrow need while differing sharply in DD, and that difference is meaningful: it is the difference between being recommended to everyone asking about a category and being recommended only to those who already know what they want.

This convention defines DD as a concept and a reporting requirement, not as a probing protocol. How constraints are enumerated and traversed is a measurement operation deferred to BA-C-3. As a reporting requirement, a DD report MUST: state the constraint dimension or dimensions along which "depth" is defined (the attributes whose specification constitutes narrowing), because there is no single canonical way to order how constrained a need is; identify the engine with version, the window, and the region; and treat the first-appearance depth as itself an estimate. That last point matters because the answers on which DD is read are the same stochastic answers that underlie VP: the depth at which an entity first appears varies across draws, so a DD figure SHOULD be reported with the sampling uncertainty around it, not as a single exact integer.

5 5. Supporting measures

A binary "mentioned or not" is the floor. The measures below add resolution and MAY be reported alongside the primary metrics. Each carries its own reporting requirement; none replaces VP, SoV, or DD.

5.1 Mention position and prominence

Not every mention is equal. A brand named in the opening sentence with a clause of description is worth more than the seventh item in a list. Prominence captures this. The established construction is the **Position-Adjusted Word Count** (**PAWC**) from the founding generative-engine-optimization literature [1], which weights each mention by its position in the answer and by the number of words allocated to it, rather than counting mentions as identical units. A report that includes a prominence measure MUST state the weighting function it applies — how position and length map to weight — because "prominence" is otherwise undefined, and because a prominence-weighted SoV (§3) depends on exactly this choice.

5.2 Citation-backed versus unsupported mention

Engines produce some mentions from retrieved sources and others without retrieval support. A mention that sits within a passage anchored to a cited source is *citation-backed*; one produced with no such support is *unsupported*. The distinction matters for stability: citation-backed mentions are more traceable and tend to be more reproducible than mentions the model emits from parametric memory alone, and the two carry different implications for how a visibility position might be influenced. A report that distinguishes them MUST define how "citation-backed" is determined — for instance, that the mention falls within a span the engine attributes to a citation — so that the category is auditable rather than asserted.

5.3 Rank-aware overlap for list comparisons

Visibility work constantly compares ranked lists — the recommendation sets an engine returns across repeated runs, phrasings, engines, or windows. Plain set overlap, such as the Jaccard similarity, discards order: it cannot distinguish "the same entities in a different order" from "different entities altogether," and the two describe very different kinds of instability. **Rank-biased overlap** (**RBO**) [4] is an established similarity measure for indefinite rankings that weights agreement at the top of the lists more heavily, and it MAY be reported alongside set-overlap measures wherever ranked recommendation sets are compared. A report that uses RBO MUST state the persistence parameter it applied, because that parameter sets how top-weighted the comparison is and two RBO values computed under different parameters are not comparable.

5.4 Sentiment and risk

Being mentioned is not the same as being mentioned well. Sentiment records the polarity of a mention (positive, neutral, negative); risk flags record specific hazards, such as a factual error about the entity, outdated information, a negative reputational framing, or confusion with a different entity of similar name. These measures depend on a classifier, and classifiers have systematic error. Therefore, if a report includes sentiment or risk, it MUST publish the classifier's accuracy on a labeled evaluation set alongside the metric, and it MUST retain the underlying answer text for audit. Reporting a sentiment breakdown without the classifier's measured accuracy treats the classifier's output as ground truth, which it is not; such a report is not valid under this convention. Sentiment and risk are indicative signals to be read with their error rate in view, never exact counts.

6 6. Reporting requirements

The governing rule is simple: an estimate never appears without the facts needed to judge it. A bare percentage — "62% visible" — is not a measurement under this convention; it is an anecdote wearing a number. Every reported VP, SoV, or DD estimate MUST be accompanied by the following, whether inline or in an adjacent, unambiguously linked note:

REQUIRED WITH EVERY ESTIMATE	WHAT IT FIXES
Point estimate	The value being claimed
Confidence interval	Its precision — with confidence level and interval method named
Sample size n	How many independent draws support it
Engine + interface + version/date	Which system, in which deployment, when observed
Time window	The period the estimate describes
Region	The observation locale
Information need	The need k, at the level of its phrasing distribution
Detection method	How a "mention" was identified in answer text

A percentage reported without its sample size n is prohibited; it is the single most common defect in visibility reporting and the clearest marker of an unmeasured claim. A report SHOULD state the interval method explicitly, and near the boundaries it MUST use Wilson or Clopper–Pearson (§2). For SoV, the competitor set, weighting scheme, and interval construction are additionally required (§3); for DD, the constraint dimensions (§4); for sentiment or risk, the classifier's accuracy on a labeled set (§5.4).

Mention detection is itself a measurement component, not plumbing. Every metric above is computed from detected mentions, so detector error propagates into all of them. A report MUST state how mentions were detected — for example, lexicon matching against a competitor set of disclosed size, or model-based entity resolution — and SHOULD publish the detector's precision and recall on a labeled sample of answers. Where a conclusion rests on a small difference between estimates, an undocumented detector is grounds to treat the difference as unestablished.

A conforming inline form makes every required field visible at once, for example:

> *VP = 0.62 (95% CI 0.53–0.70, Wilson; n = 120; Engine X, official API, sampled 2026-07-08; region: DE)*

Because engines are non-stationary, a report SHOULD carry a perishability statement making explicit that the figures describe the engine only as sampled in the stated window (see §8). How the n draws are collected so that they estimate the intended estimand — coverage of Q_k, independence, and reconciliation across measurement channels — is not specified here; those are the sampling requirements of BA-C-3, and a report SHOULD cite the sampling convention under which its draws were obtained.

7 7. Conformance

A report claims conformance with this convention only when every MUST requirement applicable to the metrics it publishes is satisfied and every disclosure of §6 is present. The citation form is: *reported in conformance with BA-C-2 v1.0*. Partial conformance MUST be stated as named exceptions — "conformant with BA-C-2 v1.0 except the sentiment-classifier accuracy disclosure (§5.4)" — never implied by silence.

Conformance with this convention concerns metric definition and reporting form only. The validity of the underlying sample — its size, phrasing coverage, channel, window discipline, and change-claim handling — is governed by BA-C-3, and a report SHOULD state its conformance with both conventions or with neither; a BA-C-2-conformant report built on a non-conforming sample is well-labeled noise.

8 8. Limitations

These metrics define *what* to report and in *what form*, not *how* to sample. The validity of any VP, SoV, or DD number rests on the sampling design meeting BA-C-3. A confidence interval quantifies sampling variability only; it says nothing about systematic bias. A narrow interval computed over a sample that does not represent Q_k, or that is drawn through a channel whose behavior differs from the interface users actually see, is precisely and confidently wrong. Precision is not accuracy.

VP is defined relative to a specified phrasing distribution Q_k. The choice of that distribution is a modeling decision, and two measurements that adopt different distributions for the "same" need are not comparable, even on the same engine in the same window. The distribution itself is therefore a legitimate axis of disagreement between measurers and should be treated as part of the estimand rather than an implementation detail.

Mention detection is itself a classification problem — entity resolution, coreference, and disambiguation of similarly named entities — and its error propagates into every metric above even when the detection method is disclosed as §6 requires. Published precision and recall describe the labeled sample they were computed on, not every future answer. Sentiment and risk classification carry systematic error as a matter of course and are to be read as indicative, never as truth (§5.4).

SoV is only as comparable as its competitor set; across different sets it is a different quantity (§3). DD depends on the constraint ordering chosen, and there is no single canonical ordering of how constrained a need is, so DD figures compare cleanly only within a fixed constraint scheme (§4).

Finally, results are perishable. These results describe the engines as sampled during the stated window; engines change without notice, and results should be assumed perishable. A metric that was accurate on the day it was measured may misdescribe the same engine a month later, not because the measurement was wrong but because the object measured has moved. Reporting the window is what keeps a perishable result honest.

REFERENCES

1. Aggarwal, Murahari, Rajpurohit, Kalyan, Narasimhan, Deshpande; Proc. 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2024); arXiv:2311.09735. *GEO: Generative Engine Optimization* (2024). https://arxiv.org/abs/2311.09735 Accessed 2026-07-08. [archived]
2. Wilson, E. B.; Journal of the American Statistical Association, 22(158), 209–212. *Probable Inference, the Law of Succession, and Statistical Inference* (1927). https://doi.org/10.1080/01621459.1927.10502953 Accessed 2026-07-08.
3. Clopper, C. J.; Pearson, E. S.; Biometrika, 26(4), 404–413. *The use of confidence or fiducial limits illustrated in the case of the binomial* (1934). https://doi.org/10.1093/biomet/26.4.404 Accessed 2026-07-08.
4. Webber, W.; Moffat, A.; Zobel, J.; ACM Transactions on Information Systems, 28(4), Article 20. *A similarity measure for indefinite rankings* (2010). https://doi.org/10.1145/1852102.1852106 Accessed 2026-07-09.

HOW TO CITE

Barkhausen AI (2026). Visibility metrics. https://barkhausen.ai/conventions/visibility-metrics/