How accurate is AI visibility data?

Short answer: only as accurate as its sampling, its judge and its statistics — and most tools publish none of the three. LLM output is non-deterministic, so any tracker reporting one unqualified number is reporting noise with confidence. Below is the whole method MentionBeat uses, plus a six-point bar you can hold any AI visibility tool to, ours included, in about ten minutes of a demo. Including the limitations most vendors leave out.

Last updated 13 August 2026 · No account required · Every claim here corresponds to shipped, tested code.

PUBLISHED **The Visibility Index, in full**

0.4 × share of voice
+ 0.4 × mention rate
+ 0.2 × recommendation rate

Unbranded prompts only, with a confidence interval from the same bootstrap as its parts. Emitted only when every component exists — never quietly reweighted around a missing one.

SECTION 01

Collection: real engines, real answers

Every response comes from the **real engine's own API**, called with your tracked prompt at a realistic sampling temperature. We do not infer AI answers from Google rankings, and we never silently substitute one model for another.

ENGINE	HOW WE QUERY IT	WEB-GROUNDED?
ChatGPT	OpenAI API	Optional (OpenAI web search)
Claude	Anthropic API	Optional (Anthropic web-search tool)
Gemini	Google Gemini API	Optional (Google Search grounding)
Perplexity	Perplexity Sonar API	Always (natively grounded)
Grok	xAI API	Optional (Live Search over web + X)
DeepSeek	DeepSeek API	**No** — parametric only, and we refuse to label it otherwise
Google AI Overviews	SERP capture (SerpApi)	Always
Google AI Mode	SERP capture (SerpApi)	Always
Microsoft Copilot · Meta AI	**Not yet measured.** No API or SERP surface exists; a browser-capture adapter is on the roadmap. We would rather show a gap than simulated data.	

- **Grounded and parametric answers are reported separately.** A brand can dominate a model's memory yet vanish when the engine searches the live web — collapsing the two hides the difference that matters.
- **Citations come from provider metadata** — Perplexity citation lists, OpenAI URL annotations, Gemini grounding chunks, xAI citations, SERP references — never regex-guessed from answer text.
- **Every call records** cost, latency, tokens and the provider-reported model version, so engine-version drift is visible instead of mysterious.

SECTION 02

Sampling: repetition, and honesty about variance

LLMs are non-deterministic. A single query is an anecdote, not a measurement.

- Each (prompt × engine) cell runs **multiple repetitions** — five by default — at realistic temperature, never one-shot snapshots.
- Prompt sets span an **intent taxonomy** (informational, commercial, comparison, recommendation, problem/solution, verification, fit/limitations, provenance), personas and locales, then get human approval.
- **Geographic sampling states its mechanism.** A locale geo-pins SERP engines natively; chat APIs cannot be geo-pinned without proxies, so they get a recorded location hint instead. Every response stores which mode produced it — a hinted answer is never presented as a natively geo-sampled one.
- Interrupted runs resume idempotently; failed calls are retried and error rates are reported on the run.

Anti-leakage design

Headline visibility uses only **unbranded** prompts — the questions a real buyer asks before they know your name.

Brand-bearing prompts are still tracked, but reported separately and excluded from headline share of voice. Asking an engine “tell me about Acme” and counting the reply as visibility is grading your own exam.

SECTION 03

Scoring: a calibrated judge, not keyword matching

Responses are scored by an LLM judge that extracts brand mentions, position, sentiment, recommendation strength, cited domains and factual claims. Because judges are also LLMs, we treat the judge itself as an instrument to calibrate.

- **Consensus mode** — multiple independent judge passes per response with majority voting, recording the agreement fraction.
- **Human calibration** — the judge is checked against answers a person labelled by hand, drawn by stratum rather than at random, and reported with Cohen's κ, sensitivity and specificity. Those figures then bias-correct the mention rate, so a judge that over- or under-detects cannot silently skew your trend. The whole calibration method is below →
- **A deterministic literal-match safety net** catches misses, with provenance recorded — you can always see *why* something was counted.

Judge calibration: measuring the instrument

The judge answers one question on every response — was your brand named? — and every headline number is built on those verdicts. So we measure how often it gets that call right, against the same responses labelled by hand.

You cannot measure a judge's error rate by asking the judge; that is circular. A person has to look. Those hand-labelled answers are the **gold set**, and everything below is computed against them. A run with no gold labels is reported as **not human-validated**, and its metrics are called directional rather than exact — we would rather say that than let an unchecked judge pass for a measured one.

Agreement, minus the agreement chance would have produced

Raw agreement flatters a judge. Suppose — purely as an illustration, not a measured figure — your brand turns up in one answer in twenty. A judge that answers “no” every single time agrees with the human 95% of the time and has found nothing.

So Cohen's κ (kappa) is reported beside raw accuracy. Kappa is agreement *after* subtracting the agreement two raters would have reached by chance alone, given how often each of them says yes. A κ of 1 means they agree perfectly; 0 means no better than chance; below 0 means the judge does worse than guessing. That always-says-no judge scores exactly 0 — which is the whole point of reporting it.

One caveat we publish rather than bury: when every labelled answer falls on the same side of the call, chance agreement is already total and kappa has nothing left to correct. It collapses to 1 or 0 and tells you nothing about the judge. Kappa needs a reasonable number of labels, spread across both answers, before it means anything at all.

Correcting a rate for a judge you have measured

The same comparison yields two more numbers. **Sensitivity** is how often the judge catches a mention the human confirmed — a low one means it misses. **Specificity** is how often it correctly passes on an answer that never names you — a low one means it over-detects. Together they describe the judge's bias.

Once both are known, a measured mention rate can be corrected for that bias instead of merely carrying an apology for it. The arithmetic is published, like the Index — this is the **Rogan–Gladen correction**, the standard adjustment for a rate measured with an imperfect classifier:

corrected rate = (observed + specificity − 1) / (sensitivity + specificity − 1)

- **A judge carrying no information gets no correction.** If sensitivity and specificity sum to 1, the judge is doing no better than a coin toss and that denominator collapses to zero. We return no corrected number at all, rather than a figure produced by dividing by almost nothing.
- **A correction that runs off the scale is held at the edge.** The arithmetic can land below 0% or above 100%. It is clamped into the 0–100% range rather than printed as an impossible rate — and a correction pinned to the edge is a sign the sample is too thin to carry it, not a result to act on.

Correcting a number does not make it certain. It moves a number for a bias we measured, and it is only as good as the sensitivity and specificity behind it — both estimated from the same small sample. Calibration does not remove the uncertainty. It states it.

Why the labelled sample is not drawn at random

Human attention is the scarce ingredient, and sampling a run evenly wastes it. In a run where your brand is rarely named, most answers are obvious negatives — fifty clicks of “obviously not mentioned” would score the judge at 100% while testing it nowhere it could fail.

So the review queue is stratified on a signal the judge gets no vote on: does your brand name, an alias, or a locale spelling of it appear literally in the answer, as a whole word? That match is deliberately dumb. It misses implied and transliterated mentions — which is exactly the judgment we want a human to make.

Crossing that signal with the judge's verdict gives four cells, and they are not equally worth a person's time:

Name present, judge said no

A candidate miss: the word was there and the judge passed on it. First into the queue.

Name absent, judge said yes

The judge inferred a mention from something other than the name. Second into the queue, and the cell where a human's read matters most.

Name present, judge said yes

Near-deterministic. Still labelled, but it rarely teaches us anything.

Name absent, judge said no

Near-deterministic, and the cell that would swallow the entire labelling budget if the queue sampled evenly.

The queue serves the two contradiction cells first, then round-robins across all four. Every cell needs its own labels for the re-weighting to work, and alternating between them keeps consecutive expected verdicts different, so a reviewer cannot rubber-stamp a batch. Order within a cell is deterministic, so a labelling session is resumable.

Then the weights undo the sampling bias. Each label counts for the number of answers in its own cell divided by the number labelled there — over-sampled cells weigh less, under-sampled cells weigh more. Accuracy, kappa, sensitivity, specificity and sentiment agreement are all computed on those weights, so the result is an estimate *for the whole run*, not a tally of a deliberately hard sample. Hunting the answers most likely to be wrong is what makes calibration informative; the weighting is what stops it from dragging the reported accuracy below the truth.

A cell nobody labelled carries no weight and is simply unrepresented in the estimate. Answers-per-cell and labels-per-cell are reported beside the result, so the part of the run the estimate does not cover stays visible instead of being averaged over. Sentiment and recommendation agreement are measured only on answers the judge and the human both say mention you — agreement about the sentiment of an absent brand would be agreement about nothing.

Calibration does not make a mention rate certain. It puts a measured bound on the instrument that produced it — and where the instrument has not been checked, it says so rather than implying it has.

Every answer where the judge and the human disagreed is recorded individually, by response. A disagreement is something you can open and read, not a statistic you have to take on faith.

None of this moves your visibility on its own. It tells you how much of the number in front of you is signal — which is the difference between a report you can act on and a chart you have to trust.

What calibration cannot do

The correction bounds the judge's error. It does not erase it, and it will not:

- Rescue a judge that is simply wrong — one no better than chance produces no correction at all
- Make κ meaningful on a handful of labels
- Cover a cell nobody labelled
- Make a corrected rate exact — sensitivity and specificity are estimates too
- Stand in for the raw answers, which stay stored and checkable

SECTION 04

Statistics: confidence intervals, or it didn't happen

- Headline metrics carry **95% confidence intervals from a cluster bootstrap (2,000 resamples) that resamples prompts, not individual responses** — repetitions of one prompt are correlated, and treating them as independent understates uncertainty. That is the standard mistake in this category.
- Small slices use Wilson intervals; every rate is shown with its interval.
- **Trend deltas are flagged "likely real" only when intervals don't overlap.** We would rather tell you a wiggle is noise than sell you a dashboard of it.
- Share of voice comes in two variants: simple mention share, and position-weighted, where a #1 recommendation counts for more than a footnote. Both are de-duplicated per response.
- Demand-weighted and unweighted metrics use the same bootstrap; modelled weights are labelled as estimates.
- A power calculator recommends repetition counts before you spend budget.

The Visibility Index is published arithmetic

Our composite score is not a proprietary mystery. It is **0.4 × share of voice + 0.4 × mention rate + 0.2 × recommendation rate**, computed on unbranded prompts only, carrying a confidence interval from the same cluster bootstrap as its parts. It is emitted only when every component exists — never silently

reweighted around a missing one.

Those weights are chosen for explainability, and they carry that provenance in the product. As measured before/after experiments accumulate alongside AI-referred traffic, we fit outcome-calibrated weights against them — and any weight change would be announced and versioned, never slipped into your trend.

Demand weighting

Not every question is asked equally often, so prompts can carry a demand weight and the headline metrics can be reported weighted as well as flat. Weighted and unweighted figures come from the same bootstrap, so the two are directly comparable.

Where a weight is a modelled estimate it is labelled as one, in the product and here. An estimated weight is **relative**, on a 0–100 scale within your own prompt set — never a claim about absolute search volume. It is built from a log-scaled demand proxy (keyword and Trends volume, optionally calibrated against your own Search Console impressions), multiplied by mild intent-class priors: chat usage skews toward informational and comparison questions and away from purely transactional ones, so those classes are nudged up and down by no more than 25%. The priors are priors, not measurements, which is exactly why they are kept small.

A prompt we have no demand signal for keeps a uniform weight and is flagged as unweighted, rather than being quietly assigned a number.

Comparative lenses, and the floors they respect

Some questions only the whole measured field can answer, so these lenses aggregate across projects — under hard floors, stated with every result.

Engine weather

Per-engine volatility: within-run answer flips, run-to-run rate swings, and cited-source churn. Each score names exactly which components it stands on, and a component we cannot compute stays absent rather than counting as zero.

Peer percentiles

Where your number sits among projects measured the same way. Withheld, with the reason shown, until a cohort holds at least five peer projects — a percentile against two peers is both meaningless and a privacy leak.

Source authority

Which third-party domains shape AI answers across the field. A domain appears only when at least three unrelated projects' answers cite it, and only ever as shares, so no row is traceable to any one customer's data.

The lift database

What a kind of content change — an FAQ block added, schema markup fixed, an entity listing claimed — has measurably moved, pooled across comparable experiments. Confounded experiments are excluded, and no number is shown below three measured experiments of that kind.

SECTION 05

Accuracy: grounded in your approved facts

We don't only count mentions — we fact-check what engines *say* about you against your product brief: a provenance-tracked store of approved facts, each carrying its source and human-review status. This section is about grounding *our* outputs in facts you approved; for the separate job of correcting what an engine says about you, see accuracy monitoring and corrections (mentionbeat.com/product/accuracy).

- Claims that **contradict** an approved fact are reported as hallucinations, bound to the specific fact they contradict, so the finding is directly actionable.
- Claims we **cannot verify** are reported separately. Unverifiable is not the same as false.
- Cited domains are classified own, competitor or third-party, mapping your off-site corroboration footprint.

SECTION 06

Experiments: measured lift, not vibes

Content changes are evaluated as **baseline → treatment experiments**: per-metric lift with confidence intervals, significant only when those intervals separate.

A delta is only as trustworthy as the comparison behind it. If the questions changed, the answer changed for a reason that has nothing to do with your content — so we check that first and say so, rather than letting a setup change take the credit.

Where a prompt suite did change between two runs, we also recompute both numbers over **only the prompts they share**, giving you a like-for-like delta beside the raw one.

What trips the confound guard

Any of these between baseline and treatment marks the result confounded instead of reporting it as a win:

- The engine set changed
- The repetition count changed
- The brand profile version changed
- The prompt suite changed
- An engine silently resolved to a new model version

SECTION 07

Live and modelled data are labelled, end to end

The authoring studio can preview engine behaviour through a persona simulation when no API key is configured. Every such run carries its label from the moment it is created to the moment it is read.

Modelled data never appears in measurement reporting as though it were live. If a vendor key is missing you see a gap or a label, never an imitation of the real thing.

The same rule governs this website. A number we publish either came from a real run and carries its interval, or it is cited and dated to someone else.

The three labels

Live

Answers from a real engine API call. The only kind that reaches measurement reporting.

Modelled

A persona simulation used for authoring preview. Never counted, never charted as visibility.

Mixed

A view combining both. Labelled as such, so the blend is never mistaken for a measurement.

SECTION 08

The bar — hold any tool to it, including us

LLM output is non-deterministic, so a tracker reporting a single unqualified number is reporting noise with confidence. These six requirements follow from that one fact. They are not our preferences — they are what honest measurement of a non-deterministic system requires, and each takes minutes to verify in a demo.

REQUIREMENT	WHY IT MATTERS	MENTIONBEAT
Repetition and confidence intervals	One sample per prompt makes every week-over-week “change” indistinguishable from noise.	Multiple repetitions per (prompt × engine); 95% CI on every headline metric; deltas flagged only when intervals separate.
Uncertainty computed at the prompt level	Repetitions of one prompt are correlated; treating them as independent samples understates uncertainty — the standard mistake.	Cluster bootstrap of 2,000 resamples over prompts, not responses; Wilson intervals on small slices.
A calibrated judge	If an LLM scores the answers, that judge has its own error rate. Uncalibrated, it can shift a headline number by more than the change you are trying to detect.	Consensus voting, validation against human gold labels (Cohen's κ, sensitivity and specificity), Rogan–Gladen bias correction on mention rates.
Grounded and parametric reported separately	A web-grounded answer and a from-memory answer measure different channels; averaging them hides the difference that matters.	Every response is classified, and the two are never averaged together.

REQUIREMENT	WHY IT MATTERS	MENTIONBEAT
No brand-name leakage in headline metrics	Asking “tell me about [Brand]” and counting the reply measures nothing.	Headline metrics come only from unbranded category prompts; brand-bearing prompts are a separately reported diagnostic slice.
Immutable raw answers	If only scores are stored, the methodology can never be audited or improved retroactively — the dashboard becomes the only evidence of itself.	Every raw response is stored, and every metric is recomputable from the raw log without re-spending a single API call.

Evaluating any AI-visibility tracker, this one included? Ask how it handles these six. The answers separate measurement from screenshots faster than any feature list.

“Which tool has the most accurate data?” is the wrong shape of question — no vendor can answer it about itself credibly, and we are not going to pretend otherwise. The answerable version is **“which tool can show me its error bars, its judge’s validation scores, and its raw answers?”** Those are checkable in a demo. A tool that cannot produce all three is not more or less accurate than its rivals — it is simply unmeasured, which is worse.

SECTION 09

Known limitations — the part most vendors skip

A methodology that lists no limitations is marketing. These are ours, stated plainly.

API answers, not consumer UI answers

We query engine APIs; consumer web interfaces can route to different configurations. A browser-capture adapter with a published API-versus-UI reconciliation report is planned. Until then, our results measure the engines' API-served answers.

No consumer panel

We have no clickstream data of real user prompts. Prompt sets are generated and human-curated, and demand weights are estimates unless you supply real volumes.

Geographic sampling is locale-of-prompt

Not IP-of-origin. Region-parameterised querying is in development; SERP engines already accept country and language parameters, chat APIs do not.

SERP-captured engines depend on the capture

AI Overviews and AI Mode are read through a SERP provider's rendering. The absence of an AI answer for a query is itself recorded as a data point.

Calibration is only as good as the labels behind it

The judge's error rate is measured against answers a person labelled by hand, and that attention is scarce. A run with no gold labels is reported as not human-validated; a run with few carries its label counts beside the figures. A thin sample bounds the judge loosely — see the calibration method for what the correction does and does not fix.

Non-determinism is managed, not eliminated

Repetition and intervals bound it; they do not make an LLM deterministic. Treat every point estimate as the centre of an interval.