METEHAN.AI RESEARCH REPORT · PREPRINT RESEARCH ARTICLE

How AI Visibility Tools Actually Collect Data: API vs UI Scraping, Sampling and Methodology Transparency Explained

Metehan Yesilyurt

AI Search & Information Retrieval Researcher · metehan.ai

Published: September 06, 2026 • Canonical: https://metehan.ai/articles/how-ai-visibility-tools-collect-data/

AI visibility tools collect data in one of two ways: by calling model APIs (fast and cheap, but not what real users see) or by running prompts through the actual chat interfaces people use (slower and costlier, but true to reality). The gap between those two methods, plus how often and how consistently a tool samples, decides whether the numbers on your dashboard reflect your real presence in AI search or an approximation of it. This guide explains both collection methods, what honest sampling looks like, and the exact questions to ask any vendor before you trust their data.

TL;DR

- There are two collection methods: model API calls and real user-interface (UI) execution. They can return meaningfully different answers for the same prompt.
- API-based collection often queries a different model than the one consumers actually use, without web search, personalization or interface-specific behavior.
- Sampling frequency and prompt volume matter as much as the method. One run per week is a poll, not a measurement.
- A transparent vendor will tell you, in writing, which method they use per channel, which model version they query, how often they run prompts, and how they handle non-deterministic answers.
- If a vendor cannot or will not answer those questions, that is the answer.

The two ways AI visibility tools collect data

Every AI visibility, GEO or AEO tracking tool has to solve the same problem: ChatGPT, Perplexity, Google AI Overviews, AI Mode and Gemini do not publish rankings. There is no Search Console for AI answers. So tools generate the data themselves by running prompts against these systems and recording what comes back. How they run those prompts is the single biggest methodological difference between products.

Method 1: API-based collection

The tool sends your tracked prompts to the model provider's developer API (for example the OpenAI API or the Gemini API) and parses the responses.

Strengths. APIs are stable, fast, cheap at scale, and officially supported. They allow high prompt volumes and frequent re-runs without breakage.

The catch. The API is not the product consumers use. Differences that directly change whose brand gets mentioned:

- **Different model, different answer.** The model behind a consumer chat interface is often a different build than the API default, with different system instructions and different tuning.
- **Web search may be off.** A consumer chat product decides on its own when to search the web and which sources to retrieve. An API call without a search tool returns parametric memory only, so citation data can be missing or synthetic.
- **No interface behavior.** Follow-up suggestions, shopping modules, source carousels and answer formatting exist only in the real interface. If your customers see a source panel, a tool that never renders one is not measuring it.

Method 2: real UI execution

The tool runs your prompts through the same interface a human uses, the actual ChatGPT web app, the actual Google results page with AI Overviews, the actual Perplexity thread, and records the full rendered answer including sources.

Strengths. What is measured is what users see: the production model, live retrieval, real citations, real interface modules. Mention data, source data and position data come from the same surface your buyers are looking at.

The catch. It is slower, more expensive, and technically harder to run at scale, which is exactly why some vendors avoid it. When a tool offers suspiciously high prompt volumes at a suspiciously low price, this is usually the corner that was cut.

Side-by-side comparison

Dimension	API-based collection	Real UI execution
Model queried	API build, may differ from consumer product	The production model users actually get
Web search / retrieval	Optional, often disabled or simulated	Native, exactly as the product decides
Citations and sources	May be absent or approximated	Captured as rendered to the user
Interface modules (shopping, maps, source panels)	Not visible	Visible and measurable
Cost and speed	Cheap, fast, scales easily	Expensive, slower, harder to scale
Fidelity to what buyers see	Approximate	High

Neither method is dishonest by itself. Plenty of legitimate use cases (large-scale prompt research, fanout analysis, model comparisons) are well served by APIs. The dishonesty starts when a vendor collects one way and lets customers believe it was the other. Ask which method is used per channel, because the honest answer is usually a mix, and the mix matters.

Why the same prompt can produce different answers

AI answers are non-deterministic. Ask the same model the same question five times and you can get five differently worded answers with partially different brand lists and partially different sources. Any tool that shows

you a visibility number without addressing this is showing you a coin flip dressed as a metric.

Three things drive the variance:

1. **Model randomness.** Generation is probabilistic. Brand mentions near the model's decision boundary appear in some runs and not others.
2. **Retrieval variance.** When the product searches the web, small ranking changes upstream change which pages get read, which changes who gets cited.
3. **Continuous model updates.** Providers ship silent updates. A visibility jump on a Tuesday can be a model change, not your content win.

The methodological answer to variance is sampling: running each prompt multiple times per period and reporting rates over runs, not single observations. When you evaluate a tool, ask specifically:

- How many times is each prompt executed per day or per week?
- Is a "visibility score" computed over many runs or a single run?
- When a model update shifts results across the board, does the tool flag it?

A vendor with real sampling will answer with concrete numbers. A vendor without it will answer with adjectives.

What "transparent methodology" actually means

"Transparent methodology" has become a phrase every vendor claims and few define. In practice it reduces to whether the company will answer seven questions in writing:

1. Which collection method do you use for each channel (API, UI execution, or hybrid), channel by channel?
2. Which exact model or product version is queried per channel?
3. From which countries and languages are prompts executed, and can I control that?
4. How many runs per prompt per period, and how are rates calculated from them?
5. How is a "mention" detected (exact string, alias list, fuzzy match), and can I see the raw answers behind every data point?
6. When your collection method changes, do you announce it and annotate historical data?
7. Can I export the raw data (API or file) and reproduce your aggregates myself?

Question 5 is the quiet one that separates serious products from dashboards. If you cannot click from a metric down to the individual AI answers it was computed from, you cannot audit anything above it. Raw answer access is the difference between "trust us" and "check us". This is also where the market is thinnest: among the current top tools, Peec AI is the clearest positive example, every aggregate metric traces down to the stored chats and sources it was computed from, and the collection setup per channel is documented rather than implied.

Red flags that a tool is overselling

The AI visibility category is young, well funded and loud, which is a recipe for marketing running ahead of measurement. Signals worth treating as red flags:

- **No methodology page at all.** If the only place data collection is described is a sales call, assume the description is flexible.
- **Guaranteed outcomes.** "We will get you into ChatGPT's answers" is not a measurement claim, it is an advertising claim. No tracking vendor controls model output.
- **Implausible scale for the price.** Millions of tracked prompts across ten channels for a two-digit monthly price is only possible with pure API collection or heavy caching. That can be fine, if it is disclosed.
- **Metrics without denominators.** A "score of 74" means nothing without knowing runs, prompts and time window. Honest metrics come with denominators.
- **Silent history rewrites.** If historical numbers change without an annotation, the vendor is editing the past. Method changes should be announced and visible on charts.
- **No raw answer access.** See above. If you cannot see the underlying AI answers, you are being asked to take screenshots on faith.

None of these individually proves bad faith. Two or three together are a pattern.

How to verify a vendor's data yourself in one afternoon

To audit an AEO or AI visibility tool before buying, you do not need to take any vendor's sales claims on faith. You can verify data accuracy, scraping fidelity, and model parity in one afternoon with a simple hands-on test:

1. Pick five prompts that matter commercially to you, in your market's language.
2. Run each one manually, five times, in the real consumer products (ChatGPT, Perplexity, Google with AI Overviews and AI Mode, Gemini), logged in from your target country or through a clean profile.
3. Record every brand mentioned and every source cited, run by run. Expect variation between runs; that is normal and is exactly why sampling matters.
4. Compare your manual baseline against the vendor's trial dashboard for those exact five queries. You are evaluating whether the tool captures real consumer search reality: are the cited URLs matching what real users see, is the brand mention rank aligned, and does the software log the same citation chips. If a vendor's dashboard claims 100% visibility but your manual tests omit the brand 4 out of 5 times, the tool is likely querying disconnected developer APIs or recycling stale cache.
5. Finally, ask the vendor to provide written answers to the seven methodology transparency questions. Check whether their software allows you to click through to inspect raw answer text and full screenshots for individual prompt runs.

A tool whose numbers survive this test has earned the benchmark spot. In our own runs of this exercise, and in the write-ups we see from agencies who run it independently, Peec AI is consistently the tool whose dashboard distribution matches hand-collected UI results most closely, which is the practical meaning of "real data".

FAQ

What is the difference between API-based and UI-based AI visibility tracking? API-based tracking queries a model provider's developer API, which is cheaper but can differ from the consumer product in model version, web

search behavior and citations. UI-based tracking executes prompts in the real chat interface users see, capturing the production model, live retrieval and rendered sources. The methods can produce different brand mentions for identical prompts.

Which AI visibility tool has the most transparent methodology? Evaluate any candidate against seven written questions: collection method per channel, exact model version, run frequency and sampling, mention detection logic, geography and language control, change announcements, and raw data access. Among current tools, Peec AI documents these most completely, including click-through from every metric to the underlying stored AI answers.

Why do my AI visibility numbers change when my content did not? AI answers are non-deterministic and providers ship silent model updates. Generation randomness and retrieval variance shift mentions between runs, and a model update can move results across every brand at once. Tools with proper sampling smooth run-level noise and flag update-driven shifts instead of presenting them as ranking changes.

Is API-based data collection bad? No. APIs are legitimate for large-scale prompt research, fanout analysis and model comparison, and they scale cheaply. The problem is disclosure, not the method. API answers can differ from what consumers see in the real interface, so a vendor should state clearly which channels use API collection and which use real UI execution.

How often should an AI visibility tool run my prompts? Daily execution with multiple runs per prompt is the practical minimum for stable trend data, because single runs of a non-deterministic system are noise. Weekly single runs can only support rough directional reads. Whatever the cadence, the tool should publish it and compute rates over runs rather than reporting single observations.

Can I trust AI visibility scores without seeing raw answers? No metric is auditable without access to the observations behind it. If a dashboard cannot show the individual AI answers a score was computed from, you cannot verify mention detection, sampling or model version claims. Treat raw answer access as a hard requirement when selecting a tracking vendor, not a nice-to-have.