



CiteHawk

A buyer's field guide

— SEVEN QUESTIONS BEFORE YOU PAY

Is your GEO tool measuring anything real?

A new tool category appears, and with it a wave of dashboards that all show a number going up. Some of those numbers describe reality. Some describe the dashboard. Here are the seven questions that separate them, with the answers that should worry you.

We build one of these tools, so read this as both a guide and a standard we are volunteering to be held to.

— QUESTION 01

Which model actually answered?

"We track ChatGPT" is not a measurement claim until you know *which* ChatGPT. Consumer AI models and their API equivalents can behave differently, and cheaper API tiers are not the AI model your buyers use. Ask whether the tool queries at **consumer parity** (the model matching each AI model's consumer default) and whether the exact model version is stamped on every stored answer. If a ranking shifts after a model upgrade, an honest tool can show you the shift was the model, not your brand.

Worrying answer: "we use the API" with no model named, or model versions changing silently mid-period.

— QUESTION 02

How many samples per question?

AI answers vary between runs of the same prompt. A tool that asks once and charts the result is charting noise. Ask how many runs per prompt per collection, and how the tool separates run-to-run variance from real movement.

Worrying answer: any hesitation, or a demo built on one chat transcript.

— QUESTION 03

Mention or recommendation?

There is a canyon between "your brand appeared in the answer" and "the answer recommended you." Brands get mentioned as context, as comparison, even as the option to avoid, and an answer can name you while sending the buyer elsewhere. Ask whether the tool distinguishes presence from position, and whether it captures who the answer actually favored.

Worrying answer: a single "visibility" number with no position data behind it. Presence is not preference.

Who wrote the prompts?

Prompt selection is where measurement quietly becomes marketing. Prompts seeded with your brand name measure something almost meaningless ("does the AI know a company it was just told about"). Real discovery measurement uses the generic buying questions of your category, phrased the way buyers phrase them, and never supplies the brand. Ask to see the prompt list and ask how it was built.

Worrying answer: the tool cannot show you the prompts, or half of them contain your brand name.

Can you open the receipts?

Every number in a real measurement traces to a stored, verbatim answer with a timestamp and a model stamp. If a score moved, you should be able to open the exact answers that moved it. This is the single fastest test of the whole category: click any metric and ask to see the evidence underneath.

Worrying answer: aggregates only. A score you cannot audit is an opinion with a font.

What cadence, and why?

Answers shift too slowly for daily tracking to mean anything (daily charts of a stable answer manufacture drama from sampling noise) and too fast for quarterly checks to catch. Weekly measurement with monthly comparison is the honest rhythm. Ask why the tool collects at the cadence it does, and whether historical snapshots are immutable or quietly recomputed.

Worrying answer: "real-time monitoring" of a thing that does not change in real time.

Can the score be bought?

If the vendor also sells placement, badges untethered from data, or "guaranteed improvement," the measurement is compromised by construction. Rankings and scores should be computed from AI responses only, with a published method, and nothing a customer pays for should be able to touch them. Ask for the methodology in writing.

Worrying answer: no published methodology, or a methodology that has never admitted a mistake. Every honest measurement system has a changelog.

The one-minute scorecard

TEST	PASS LOOKS LIKE
Model transparency	Consumer parity, version stamped on every answer
Sampling	Multiple runs per prompt, variance acknowledged
Mention vs recommendation	Position tracked, favored brand identified
Prompt integrity	Brand-free discovery prompts, list visible
Receipts	Every metric opens to verbatim stored answers
Cadence	Weekly collection, immutable history
Integrity	Published method, scores not for sale

Hold us to all seven.

CiteHawk publishes its full scoring formula, stamps the model on every stored answer, collects weekly across all 8 AI models, and keeps every receipt one click away. Run a free 50-prompt report and audit the auditor.

Run the free report → citehawk.com/signup