

The AI Search Visibility Measurement Workbook

How to design a prompt set, sample answers from ChatGPT search, Google AI features, and Perplexity, log what you see, read referral traffic, and state plainly what cannot be measured.

Contents

01 What you are measuring, and why it is not rank tracking

02 Design the prompt set before you run anything

03 Sampling: how many runs, and how to avoid fooling yourself

04 Logging: one row per run, scored with a fixed rubric

05 Referral traffic and Search Console: useful, incomplete

06 The monthly measurement run, step by step

07 What cannot be measured, and how to say so in a report

Key Takeaways & Sources

01

What you are measuring, and why it is not rank tracking

Classic rank tracking has a stable object: a query returns a list, and your page sits at position seven. AI answers have no such list. ChatGPT search, Google's AI Overviews and AI Mode, and Perplexity each compose a response, and that response may name brands, cite pages, both, or neither. Two people asking the same question can get different answers. So can one person asking twice. Google's documentation says AI Overviews are shown only when its systems judge them additive to classic search, and that they often do not trigger at all. That alone means a prompt that produces an answer today may produce plain results tomorrow.

The practical consequence is that you should measure frequencies rather than ranks. For a fixed set of prompts, run each several times across each engine and record how often your brand is mentioned, how often your domain is cited, and what the answer says about you. The output is a rate with a range, such as appearing in some number of runs out of the total, tracked over months. It is not a position.

It also helps to be clear about what drives eligibility, because measurement is pointless if the page is not eligible in the first place. Google states that a page must be indexed and eligible to show with a snippet to be a supporting link in AI Overviews or AI Mode, and that there are no additional technical requirements or special markup. OpenAI says sites that opt out of OAI-SearchBot will not be shown in ChatGPT search answers, though they can still appear as navigational links. Perplexity recommends allowing PerplexityBot to appear in its results. Check these access conditions before you spend a month logging absences that a robots.txt rule or a firewall explains.

The practical consequence is that you should measure frequencies rather than ranks.

02

Design the prompt set before you run anything

Start from the questions your buyers actually ask. The best raw material is internal: sales-call notes, support tickets, on-site search terms, and the queries already bringing impressions in Search Console. Write each as a person would type it into a chat box, which is usually longer and more specific than a keyword. Include prompts that name no brand, prompts that compare you with named competitors, and prompts that name you directly. The first group measures discovery. The second measures positioning. The third checks whether the engine describes you accurately.

Keep the set small enough to run on a schedule. Thirty to sixty prompts across four or five intent groups is a workable starting size for a single brand, though that figure is a recommendation for manageability, not a statistical threshold. Tag each prompt with an intent group, a funnel stage, and a market or language, then freeze the wording. If you edit a prompt, give it a new ID and start a new series. Add a

short list of control prompts about topics where you have no stake, to remind yourself what a normal answer looks like.

Decide the context for each run in advance and write it down. Logged-in versus logged-out use, location, language, device, and whether the engine's search or browsing feature was active can each change the answer. You cannot hold them all constant forever, but you can record them, which lets you explain a shift instead of guessing at it.

- Discovery prompts: the problem or category with no brand named
- Comparison prompts: you against named alternatives, and alternatives to a named competitor
- Brand prompts: what the engine says about you, your pricing, your reviews, your suitability
- Control prompts: topics with no stake, used to spot changes in the engine rather than in you

Tag each prompt with an intent group, a funnel stage, and a market or language, then freeze the wording.

03

Sampling: how many runs, and how to avoid fooling yourself

Because answers vary, one run per prompt tells you very little. A practical method is to run every prompt several times per engine in one sitting, then repeat the full set on a fixed cadence such as monthly. Five runs per prompt is a sensible floor for a first pass; more is better where you can afford it. Treat that number as a working choice, not a proven requirement. Report the result as a share of runs, and show how many runs sit behind it. A rate of two in five and a rate of forty in a hundred are different amounts of evidence, even though the percentage matches.

Run the engines separately and never average them. ChatGPT search, Google AI features, and Perplexity retrieve and cite differently, and Google itself notes that AI Mode and AI Overviews may use different models and techniques, so the set of responses and links will vary between them. Keep a separate series per engine and per surface. Where a vendor tool does the sampling for you, ask what it samples, how often, from which locations, and whether it uses the consumer product or an API. Those details decide how far the figures resemble what a customer sees. The honest answer from most tools is that they sample a sample.

Guard against your own bias. Run the set in a clean session, not one where you have been researching your own brand. Have a second person score a share of the logged answers independently, and compare the scoring; disagreement tells you which rubric definitions are loose. And resist rerunning a prompt until it returns the answer you hoped for. Log every run, including the ones you dislike.

Where a vendor tool does the sampling for you, ask what it samples, how often, from which locations, and whether it uses the consumer product or an API.

04

Logging: one row per run, scored with a fixed rubric

The log is the actual deliverable, so design it to be boring and complete. Each row is one run: date, engine and surface, prompt ID, run number, context settings, whether an AI answer appeared at all, whether your brand was named, whether your domain was cited as a source, which competitors were named or cited, the cited URL, and a short note on how you were described. Save the full answer text or a screenshot alongside the row so a later reviewer can re-score it. Store the log in a spreadsheet or database the whole team can read; the attached worksheet is a starting layout.

Score three outcomes separately, because they mean different things. A mention means the brand is named in the answer text. A citation means your domain appears among the sources. Accuracy means the claim made about you is correct, which you judge against your own published facts. A brand can be mentioned without a link, linked without being named, or described wrongly. Merging these into one visibility score hides exactly the distinctions you need to act on. If an answer states something false about you, that is a content or entity-clarity problem you can address, and it matters more than a missed mention.

Keep a change log next to the data. Note site launches, content updates, PR coverage, robots.txt edits, and engine changes you have heard about, each with a date. When the rate moves, the first question is what else changed, and the change log is the only reliable way to answer it later.

Field	What to record	Why it matters
Prompt ID and group	Frozen wording, intent group, market, language	Keeps series comparable over time
Engine and surface	ChatGPT search, AI Overviews, AI Mode, Perplexity	Surfaces behave differently; never average them
Run context	Logged in or out, location, device, search feature on	Explains shifts you did not cause
AI answer shown	Yes or no	Many prompts return no AI answer on some runs
Brand mentioned	Yes or no, and how it is described	Separates discovery from accuracy
Domain cited	Yes or no, and which URL	Shows which page is doing the work
Competitors present	Names and citations	Gives a share-of-answer comparison

Field	What to record	Why it matters
Evidence	Saved answer text or screenshot	Allows re-scoring and audit

05

Referral traffic and Search Console: useful, incomplete

Prompt sampling measures what the engines say. Analytics measures what people then do. For Google, there is a real limit: its documentation says sites appearing in AI Overviews and AI Mode are included in overall Search Console traffic, reported in the Performance report under the Web search type, rather than as a separate line. You can see your total Search performance and watch for changes after AI features roll out in your market, but you cannot cleanly isolate the AI portion from that report alone. Google also says that clicks from result pages with AI Overviews have, in its observation, been of higher quality, with more time on site; that is Google's claim, and you should check it against your own conversion data instead of assuming it holds for you.

For other products, look in your analytics for sessions whose source or referrer identifies the AI product, and group them in a custom channel so they stay visible. Check what your own data shows before you build a report around it. Referrer values change, some visits arrive with no referrer at all, and some in-app browsers strip it. Whatever you count this way is a floor on AI-driven visits. Treat a rise in direct or unattributed traffic around the same time as a possible sign of undercounting, not as proof of anything.

Server logs add a different signal. OpenAI, Perplexity, and Google publish crawler user agents, and the logs show whether those bots fetch your pages. Crawling is not the same as citation, and a fetch triggered by a user question can still produce no mention. Perplexity says its Perplexity-User fetcher supports user actions and generally ignores robots.txt rules, which is a reason to read logs rather than rely on robots.txt as a record of who visited. Log data confirms access and timing; it does not confirm that a page was used in an answer.

06

The monthly measurement run, step by step

Once the prompt set and log exist, the routine should take a few hours, not a week. The table below is the sequence, with a gate at each step so a skipped check shows up before it contaminates the month's numbers.

Step	Owner	Gate to pass before moving on
Confirm crawler access for Googlebot, OAI-SearchBot, and PerplexityBot in robots.txt and at the firewall	Technical SEO	Access rules reviewed this month; any blocks are deliberate and documented

Step	Owner	Gate to pass before moving on
Freeze the prompt set and note any new prompts as new IDs	SEO lead	No existing prompt wording has changed since the last run
Run every prompt the agreed number of times per engine, in a clean session	Analyst	Run count per prompt and engine matches the plan; context settings logged
Save answer text or a screenshot for every run	Analyst	Every row has an evidence link
Score mention, citation, and accuracy with the written rubric	Analyst, with a second reviewer on a sample	Reviewer disagreements resolved and the rubric clarified where needed
Pull Search Console and analytics for the same period	Analytics lead	AI referral sessions grouped in one channel and labelled as a minimum
Update the change log and write the findings with method, run counts, and caveats	SEO lead	Every finding names its denominator and a next action, or is dropped

07

What cannot be measured, and how to say so in a report

Nobody outside the engines can see the true population of prompts users type, or how often each is asked. A tool's prompt volume figures are estimates, often built from panels, partner data, or its own sampling, and the method behind them may not be disclosed in a way you can verify. Personalised and conversational answers also depend on history you cannot see. And sampling can only show that a result is possible and how often it appeared in your runs; it cannot tell you what the median customer saw.

A defensible report therefore leads with the method and the denominator: which prompts, which engines, how many runs, in what period, from what context. It gives rates with the run count, not bare percentages. It separates mention, citation, and accuracy. It pairs the sampled data with analytics and Search Console, and it says what that pairing can and cannot prove. And it avoids attributing revenue to AI answers on the strength of a handful of referral sessions.

Finally, remember that measurement does not change the underlying work. Google says the foundations of good SEO still apply to its AI features and that no special optimisation is required. The measurable levers are mostly the ordinary ones: crawlable, indexable pages, clear and correct statements of what you do, and independent mentions elsewhere. If the log shows an engine describing you wrongly, fix the source it is drawing on. If it shows you absent from comparison prompts, the issue is likely content or authority rather than a missing tag.

- State the method, the number of runs, and the date range before any result
- Report mention, citation, and accuracy separately

- Call referral figures a minimum and say why
- Never present modelled prompt volume as observed demand
- Name the next action tied to each finding, or leave the finding out

Key Takeaways

- AI answers vary from run to run. Any single screenshot is an anecdote; a rate across repeated runs of a fixed prompt set is the closest thing to a measurement.
- Build the prompt set from real customer questions and sales-call language, grouped by intent, and freeze it. Changing prompts mid-series destroys the trend.
- Log mention, citation, and sentiment separately. Being named in an answer, being linked as a source, and being described accurately are three different outcomes.
- Referral traffic from AI products is a floor, not a count. Some visits arrive without a usable referrer, and Google reports its AI features inside ordinary Search Console totals.
- Report ranges and caveats. No third-party tool can see the full population of prompts people actually type, and an honest report says so.

Sources

1. AI features and your website (Google Search Central) — <https://developers.google.com/search/docs/appearance/ai-features>
2. Overview of OpenAI crawlers (OpenAI documentation) — <https://platform.openai.com/docs/bots>
3. Perplexity crawlers (Perplexity documentation) — <https://docs.perplexity.ai/guides/bots>
4. ChatGPT search (OpenAI Help Center) — <https://help.openai.com/en/articles/9237897-chatgpt-search>
5. Performance report for Search results (Search Console Help) — <https://support.google.com/webmasters/answer/7576553>