AI VISIBILITY MEASUREMENT

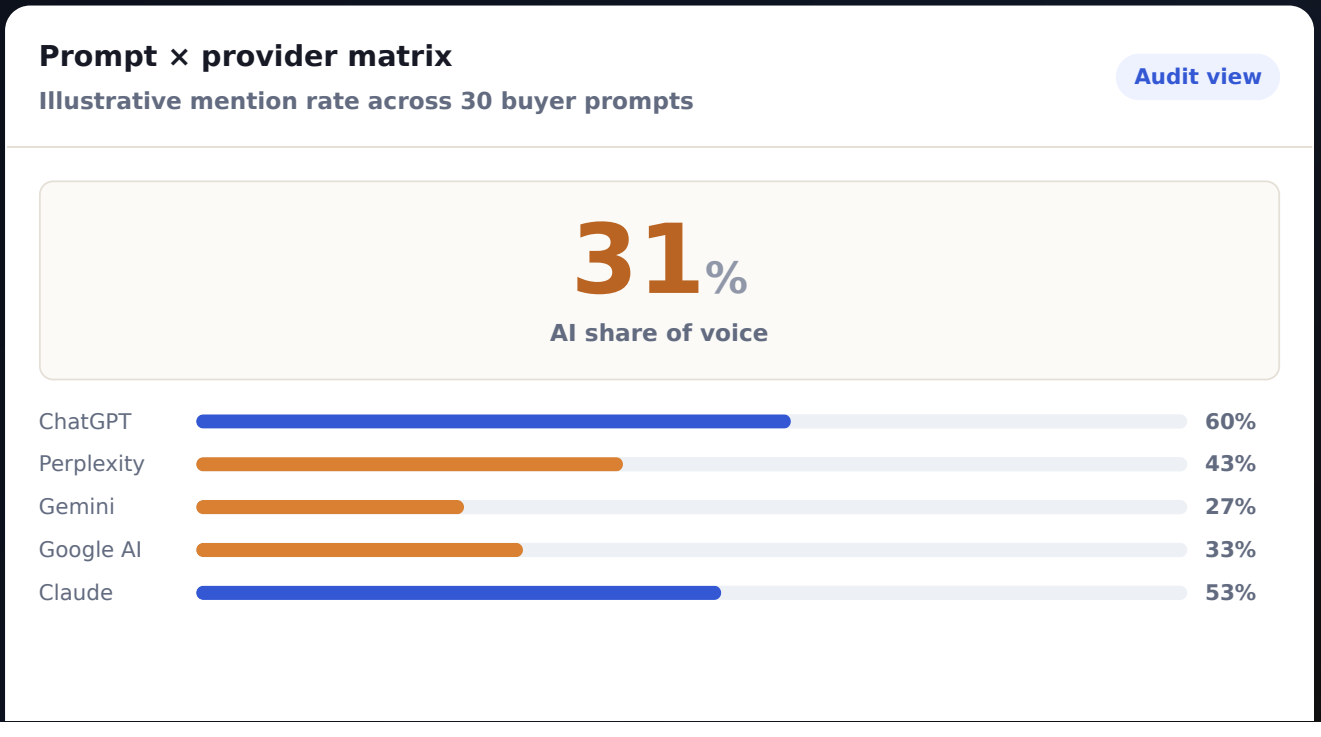
AI visibility audit checklist: how to measure if your brand shows up

A single AI answer is not an audit. It is a screenshot. A real audit is a repeatable benchmark: real buyer prompts, run across the providers your buyers use, scored for mentions, recommendations, citations, competitors, and sentiment — so the fix list comes from evidence, not vibes.

5 prompt types
Brand-direct, category, problem, comparison, and local — in real buyer wording.

9 metrics
From mention rate to repeat-run volatility. Each one answers a different question.

1 matrix
Prompt × provider × competitor × citation source = an actionable diagnosis.



Same brand, same 30 prompts, five providers — and the mention rate more than doubles between the best and worst. A one-provider screenshot would have hidden that entire story.

In this article

Most "AI visibility audits" today are one person typing "best [category] tools" into ChatGPT, taking a screenshot, and posting it in Slack. That tells you what one model said, to one phrasing, in one session. Run the same question tomorrow, or in Gemini, or with the words a buyer would actually use, and the answer changes. AI answers vary between identical runs and between providers, which is why one scan is not enough to conclude anything.

An audit worth acting on is a benchmark you can re-run: a fixed set of real buyer prompts, executed across the AI providers your buyers use, scored the same way every time. This article is the checklist for building that — the prompt set, the providers, the nine metrics, the source classification, and how the findings become a fix backlog instead of a slide.

What an AI visibility audit must answer

Before choosing tools or metrics, fix the questions. A complete audit answers six, and each one changes what you do next:

- **Do AI systems mention us at all?** Absence is the cheapest finding to detect and the most expensive to ignore.

- **Do they recommend us** — or just name-drop us in passing while recommending someone else?
- **Do they cite us?** Are our own pages among the sources answers are built from, or is our story being told entirely from third-party pages?
- **Which competitors appear instead?** Who wins the prompts we lose, and how consistently?
- **Which sources are shaping the answer?** The cited URLs are the levers — they tell you where the answer actually comes from.
- **Does visibility change by provider and prompt type?** Strong in ChatGPT and invisible in Gemini is a different problem from weak everywhere.

A screenshot answers, at best, the first question for one provider on one day. The other five are where the strategy lives.

SCREENSHOT "AUDIT"	BENCHMARK AUDIT
1 answer	**1 matrix**
• One prompt, phrased by the marketing team • One provider, one run, one day • No competitor tally, no cited sources • Not reproducible — the next run disagrees • Output: an opinion with a picture	• 20–50 prompts in real buyer wording • Every relevant provider, repeated runs • Mentions, recommendations, citations scored per cell • Competitors and source URLs logged • Output: a fix backlog and a baseline to beat

The unit of an audit is not an answer — it is a prompt × provider matrix you can re-run next month and compare.

Step 1: Build the prompt set

The prompt set is the audit. Get it wrong and every number downstream is precise nonsense. Cover five intent types, because AI systems answer them differently and your visibility differs across them:

- **Brand-direct prompts:** "Is [brand] good?", "Is [brand] legit?", "[brand] reviews". These test what AI believes about you when asked directly.

- **Category prompts:** "best [category] for [buyer or use case]", "top [category] tools for a small team". These are the shortlist-forming prompts where recommendations are won and lost.
- **Problem prompts:** "how do I solve [problem]?" — the buyer doesn't know the category exists yet. Appearing here means AI connects your brand to the underlying pain.
- **Comparison prompts:** "[brand] vs [competitor]", "alternatives to [competitor]". These test whether you exist in the head-to-head frame at all.
- **Local prompts:** "best [service] in [city]", "[service] near me" — essential for any business with a geography.

Source the wording from reality, not from your messaging doc: sales-call recordings, support tickets, review-site complaints, community threads, "people also ask" boxes. Buyers ask "cheap CRM that doesn't suck for a 5-person agency", not "leading customer relationship management solutions". If the prompt set is written in polished internal language, you are auditing a conversation no buyer is having.

Freeze the set

Once the prompt set is built, version it and stop editing it casually. The whole value of an audit is comparability across time — if the prompts drift every month, your trend line is measuring your own edits, not your visibility. Add new prompts as a new, labeled batch.

Step 2: Choose the providers

Run every prompt across the assistants your buyers actually use — as a practical default:

- **ChatGPT** — the largest assistant audience, with web browsing shaping many commercial answers.
- **Perplexity** — citation-forward and retrieval-heavy; a clean window into which sources drive answers.
- **Gemini** — Google's assistant, with its own retrieval behavior and grounding.

- **Google AI Overviews / AI Mode** — AI answers inside the search results your buyers already see. Google's own guidance is that these surfaces build on its core search systems — see its documentation on AI features in Search — so classic technical health and crawlability still gate whether you can appear.
- **Claude** — where relevant to your audience, particularly technical and B2B buyers.

The reason you cannot sample one provider and extrapolate: they retrieve from different indexes, weight different sources, and routinely disagree about the same business — one recommends you, another has never heard of you. That disagreement is itself a finding, and it is common enough that we wrote a separate breakdown of how AI Overviews, ChatGPT, Perplexity, and Gemini see the same business differently. Audit the providers your buyers use, not the one you personally like.

Step 3: Track the right metrics

Score every prompt × provider cell the same way, every audit. Nine metrics cover the picture, and each answers a distinct question:

- **Mention rate** — the share of answers where your brand appears at all. The floor of visibility.
- **Recommendation rate** — the share where the answer actively suggests you for the buyer's need. Mentioned-but-not-recommended is a positioning problem, not an awareness problem.
- **Citation rate** — how often your own domain appears among the cited sources. You can be recommended entirely from pages you don't control; this metric tells you whether you own any part of your own story. (Closely related: AI citation share, your slice of all citations in your category.)
- **Citation source domains** — the actual URLs behind the answers. This list is your lever inventory: every fix in step 5 traces back to it.
- **AI share of voice** — your mentions as a share of all brand mentions across the prompt set. The single number executives will remember.
- **Competitor frequency** — which competitors appear, how often, and on which prompt types. The competitor who wins your category prompts may not be the

one on your battlecards.

- **Sentiment and context** — how you are described when you do appear: recommended with caveats? framed as the budget option? described with outdated facts?
- **Provider disagreement** — the spread between your best and worst provider. A wide spread localizes the problem to specific retrieval ecosystems.
- **Repeat-run volatility** — how much answers change across identical runs. High volatility means you are on the model's bubble: sometimes retrieved, sometimes not. It also sets the error bars on every other metric above.

Concretely, this is what a filled-in slice of the matrix looks like — one example prompt per intent type, scored per provider, for a fictional small-business accounting firm. Every cell gets the same three marks: Mentioned, Recommended, Cited. The pattern, not any single cell, is the finding: recommended on ChatGPT's shortlist, cited only by Perplexity, invisible everywhere on the problem prompt, and rescued on local only by a Google profile.

PROMPT	CHATGPT	PERPLEXITY	GEMINI	GOOGLE AI
Brand-direct "Is Ledgerline Accounting legit?"	M R C	M R C	M R C	M R C
Category "best accounting firm for ecommerce startups"	M R C	M R C	M R C	M R C
Problem "two years behind on bookkeeping — what do I do?"	M R C	M R C	M R C	M R C
Comparison "Ledgerline vs Countwell for online sellers"	M R C	M R C	M R C	M R C
Local "best small-business accountant in Austin"	M R C	M R C	M R C	M R C

M Mentioned R Recommended C Cited (own domain) M grey = not in this run

Illustrative scores for five of the 20–50 prompts, one per intent type. This grid — not a screenshot — is the artifact the audit should produce: freeze the prompts, score every cell the same way, and re-run it monthly so

the numbers are comparable.

Report ranges, not single runs

Because answers vary run to run, every metric should come from repeated runs per prompt — and be reported as a rate with its volatility, not as "we appear" or "we don't". A recommendation rate that swings between 20% and 60% across runs is a different (and more fixable) situation than a stable 40%.

Step 4: Classify the cited sources

The cited URLs are the most actionable data the audit produces, but only after you classify them. Bucket every citation — for prompts you win *and* prompts competitors win — by source type:

Owned website

Your pages. If these never show up, AI is telling your story entirely from sources you don't control.

Local profiles

Google Business Profile and equivalents — the backbone of local prompts and "near me" answers.

Directories & review sites

G2, Capterra, Yelp, industry directories. Aggregated, structured, and heavily retrieved for "best X" prompts.

Forums & Reddit

Community threads where real users compare options. Increasingly cited for trust-shaped questions.

YouTube & news

Video walkthroughs, editorial coverage, press. Third-party corroboration that models treat as independent.

Comparison & listicles

"Top 10 X" roundups. Whoever is in them gets retrieved; whoever isn't is invisible on shortlist prompts.

Six citation buckets. The distribution across them — for you and for the competitors who beat you — is the diagnosis.

The distribution is diagnostic. Heavy directory citations with zero owned citations means your profile pages are working and your website isn't. A competitor winning problem prompts via forum threads means their community footprint — not their content team — is what you're actually competing with. Same visibility gap, opposite fixes.

Step 5: Turn findings into fixes

This is where a matrix audit pays for itself: each finding pattern maps to a specific action, tied to the source bucket that produced it.

- **Missing owned citation** on a prompt you care about → create or rework an answer-first page for exactly that question: the question in the heading, the direct answer in the first sentences, structured data and honest dates behind it. Google's AI optimization guidance says the same thing search quality always has: unique, useful content for people — there is no special trick, and no one can guarantee you a slot.
- **Missing third-party trust** — you're absent from the directories, review sites, and communities the answers cite → pursue real listings, real reviews from real customers, and genuine community participation. Never fabricated reviews or planted mentions: engines cross-check sources, and manufactured trust is both a policy violation and a discoverable one.
- **Bad sentiment or stale facts** — you appear, described wrongly or grudgingly → fix the underlying record: correct outdated pricing and claims on your own pages, address the recurring complaint that reviews keep surfacing, and make your facts consistent everywhere AI looks.

- **Competitor dominates one source type** → take source-specific action. If they win via a comparison listicle, pitch inclusion in those roundups; if via Reddit, earn a legitimate presence in those communities; if via their own site, study which of their pages get cited and build a better answer to the same question.

Prioritize by revenue weight, not by metric: a losing category prompt in your core segment outranks a wobbly brand-direct prompt every time.

Simple scoring rubric

Give each prompt/provider result a structured grade instead of a vague pass/fail. This makes the audit repeatable and makes monthly movement easier to defend.

- **0:** brand absent, competitor wins, or answer contains harmful/wrong facts.
- **1:** brand mentioned but not recommended, weak citation, or uncertain sentiment.
- **2:** brand recommended with accurate facts and a relevant owned or trusted third-party citation.

Export prompt, provider, run number, answer text, mentioned brands, cited URLs, source type, sentiment, factual errors, and recommended fix. A screenshot alone is not an audit artifact.

Step 6: Write it up — the reporting template

A repeatable audit deserves a repeatable report. Seven sections, in this order:

1. Executive summary — share of voice, the headline gap, the top three fixes. One page.

2. Prompt set — the versioned list, by intent type, so the benchmark is auditable.

3. Visibility by provider — mention / recommendation / citation rates per provider, with volatility.

4. Competitor winners — who wins which prompt types, and how often.

5. Cited sources — the classified URL inventory behind the answers.

6. Fix backlog — each fix tied to the prompt, provider, and source gap that justifies it.

7. 30/60/90-day plan — 30: owned answer-first pages and fact corrections. 60: third-party listings, reviews, community presence. 90: re-run the identical audit and report movement.

The 90-day re-run is the point of the whole exercise. Because the prompt set is frozen and the scoring is fixed, the second audit is directly comparable to the first — and "mention rate on category prompts in Gemini went from 20% to 45%" is a sentence a screenshot can never produce.

The takeaway

An AI visibility audit is not a generic score and not a screenshot. It is a prompt × provider × competitor × citation-source matrix: prompts buyers actually ask, the engines they actually use, whether you are mentioned, recommended, and cited in each cell, who appears instead, and which URLs shaped the answer. The fix backlog comes from the cited-source gaps — not from a generic "write more content" recommendation — and the same matrix, re-run on a cadence, tells you whether the fixes worked.

You can run this in a spreadsheet. It works. It is also exactly the kind of repetitive, consistency-critical work machines do better — which is why Plastorium turns this workflow into a scan and a report instead of a spreadsheet: your prompt set, run across providers with repeated runs, scored for mentions, recommendations, citations, competitors, and sentiment, every time the same way.

FAQ

▸ **How many prompts should an AI visibility audit include?**

▸ **What is the difference between a mention, a recommendation, and a citation?**

▶ **Which AI providers should an audit cover?**

▶ **How often should I repeat an AI visibility audit?**