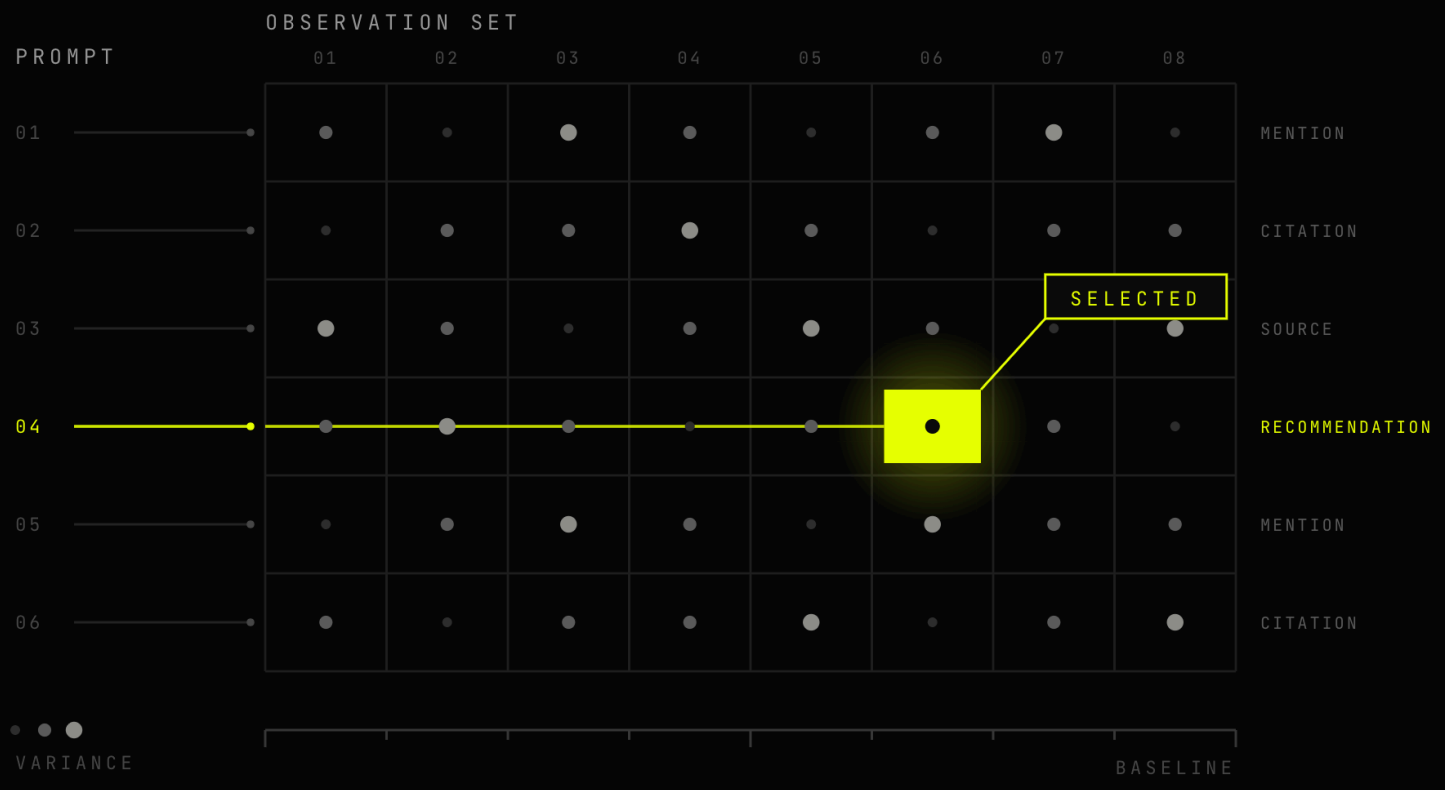
FIELD OF STUDY · AI VISIBILITY MEASUREMENT

38 MIN READ

THE AI SEARCH MEASUREMENT *PLAYBOOK.*

A research-backed white paper on measuring AI visibility across prompts, mentions, citations, competitors, sources, Maps, and answer engines.



AUTHOR
The 6 Signal Research Team
AI Visibility · Measurement

PUBLISHED
June 2026

READ
6signal.co/research/ai-search-measurement-playbook →

WHITE PAPER 38 MIN READ JUNE 1, 2026

The AI Search Measurement Playbook

A research-backed white paper on measuring AI visibility across prompts, mentions, citations, competitors, sources, Maps, and answer engines.

BY MATT VINCENT WALKER 6 SIGNAL

Executive summary

AI search is now a business surface: buyers ask ChatGPT, Perplexity, Gemini, and Google's AI features who to hire, what things cost, and which companies to trust — and the answers name some businesses and omit others. Yet most companies "measure" their AI visibility with a screenshot: one prompt, one engine, one day, one conclusion. That is not measurement. It is anecdote collection, and it routinely produces both false confidence and false panic.

This playbook lays out a disciplined alternative: a fixed prompt set written in real buyer language; systematic runs across every engine that matters; a strict vocabulary separating mentions, citations, recommendations, and sources; consistent logging; variance-aware interpretation; and — the part most measurement programs skip — a hard link between what the numbers show and what work gets done next.

Three design principles govern everything here:

01 **Repeatability beats cleverness.** A mediocre prompt set run identically for six months outperforms a brilliant one that changes every week.

02 **Absence is data.** "The engine did not name us" and "no AI answer appeared at all" are results to log, not failures to retry.

03 **Measurement that doesn't change the work plan is decoration.** The output of this system is a prioritized fix list, not a dashboard.

Nothing in this playbook guarantees citations or recommendations — no honest methodology can. What it guarantees is that you will know where you stand, whether you're moving, and what to do next.

Short answer

To measure AI visibility credibly: build a frozen set of 15–25 prompts in the language buyers actually use (mostly non-branded); run them on a fixed cadence across ChatGPT, Perplexity, Gemini, Google's AI surfaces, and Maps; log every run with the same fields — mentioned, position, sentiment, competitors named, sources cited; score results at the prompt×engine level and roll them up into mention rate, share of voice, and per-engine rates; treat run-to-run variance as expected and read trendlines, not single runs; and convert every persistent gap into a specific fix. One screenshot is a mood. A logged, repeated, multi-engine protocol is a measurement.

Part 1: Why AI visibility measurement is hard

Traditional rank tracking had convenient properties: a deterministic-ish list, a stable position number, mature tooling. AI answers have none of them.

Answers are generated, not retrieved from a fixed list. The same question can yield different shortlists on different runs, because generation samples from a distribution. Any measurement design that assumes stability will misread noise as signal — in both directions.

Phrasing changes retrieval. "Best plumber in Red Oak" and "who should I call about a slab leak in Red Oak" are different retrieval events with potentially different winners. A buyer population asks a distribution of questions; a measurement program has to sample that distribution, not a single point in it.

Engines diverge — legitimately. Each engine has its own retrieval pipeline, grounding sources, and answer style. In our client testing it is routine to see the same business named in most runs on one engine and almost never on another during the same week. That divergence is not measurement error. It is the finding.

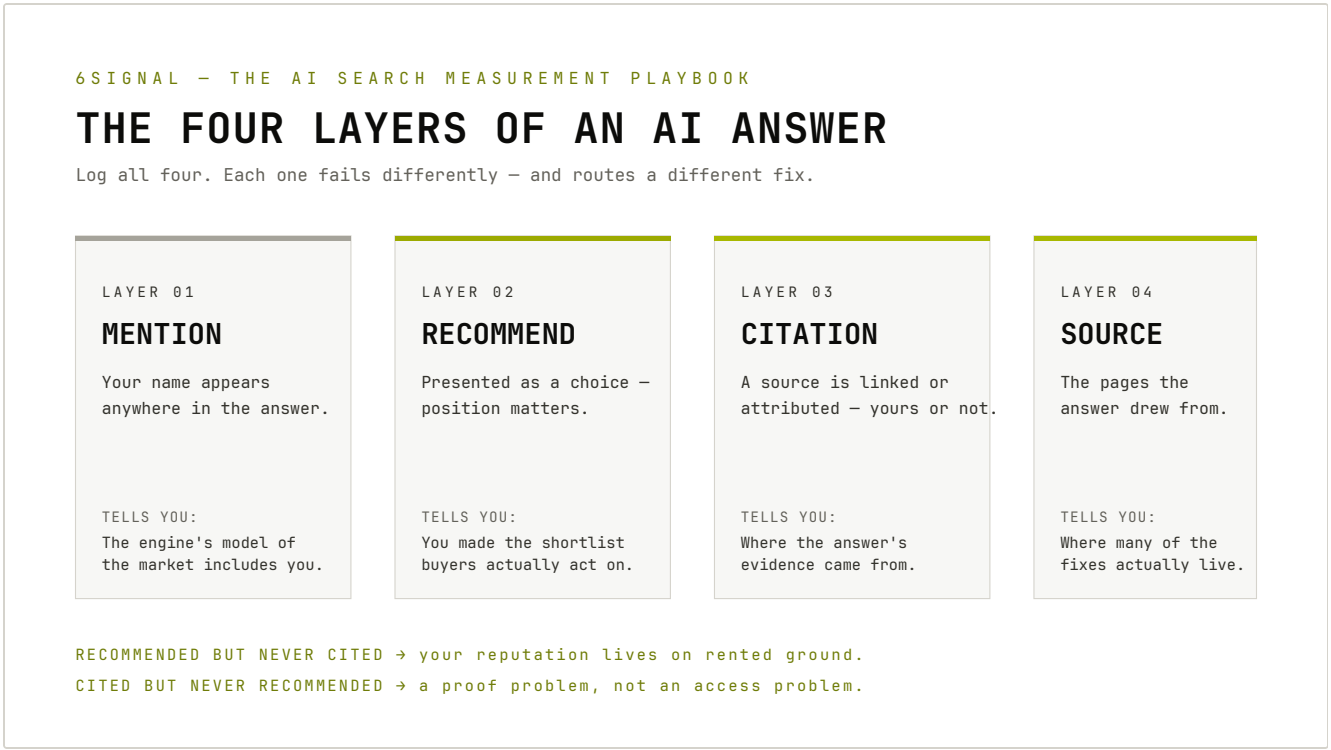
The surfaces are heterogeneous. Google AI Overviews may simply not appear for a query. Maps is a ranked list, not a generated paragraph. Chat engines produce prose with optional citations. A useful protocol has to normalize results across surfaces without pretending they're the same thing.

And the stakes are asymmetric. In a ten-link world, position four still got seen. In a generated answer naming two or three companies, everyone else is invisible for that ask. This is the compression we described in Ranking Is Not the Same as Being Recommended — and it's why guessing is no longer an acceptable substitute for measuring.

Part 2: The vocabulary — mention, citation, recommendation, source

Sloppy vocabulary produces sloppy strategy. Four terms, kept strictly separate:

Term	Definition	What it tells you
Mention	Your business name appears anywhere in the answer	The engine's model of the market includes you — minimum viable visibility
Recommendation	The answer presents you as a choice, often ordered	You're on the shortlist; position among named options matters
Citation	The engine links/attributes a source page (yours or not)	Where the answer's evidence came from; your site can be cited without you being recommended, and vice versa
Source	The underlying pages the answer drew on — directories, reviews, forums, your site	The raw material of the answer; the layer where many fixes actually live



Two diagnostic patterns fall straight out of this vocabulary:

- **Recommended but never cited:** your reputation is carried entirely by third-party pages. That position is real but fragile — you don't control the evidence. The fix lane is owned content that can carry the load.
- **Cited but never recommended:** engines read your pages for facts but don't put you on the shortlist. That's usually a proof/prominence problem — reviews, corroboration, entity strength — not a content-access problem.

The third-party layer is important enough that we wrote a separate analysis of it: The Source Ecosystem.

Part 3: Prompt set design

The prompt set is your instrument. Design it once, carefully, and then leave it alone.

Size: 15–25 prompts for a single-market local business. Fewer under-samples; many more becomes a chore nobody sustains.

Composition — three intent bands:

- **Decision prompts (roughly half the set).** "Best [trade] in [city]." "Who should I call for [emergency] in [city]?" "[Trade] near me with good reviews." These are the money prompts: someone asking is close to hiring.
- **Problem prompts (a quarter).** "Why is my water heater making a popping sound?" "How much does drain cleaning cost in [city]?" These measure whether you exist at the research stage that precedes the decision stage.

- Verification prompts (a quarter, including brand). "[Business name] reviews." "Is [business name] legit?" Branded prompts measure something different from non-branded ones — see below.

Branded vs. non-branded — keep the distinction sharp. A branded prompt ("tell me about X-Act Plumbing") tests *what engines say when asked about you.* Almost any established business "wins" its branded prompts, which is exactly why they flatter. Non-branded buyer prompts ("best plumber in Red Oak") test *whether engines think of you at all* — the contested, valuable question. Measure both; never blend them into one number. A report that quietly pads mention rate with branded prompts is lying politely.

Geographic coverage: include every town in the service area, not just the headquarters city. Answers change with the place name — a business can be strong in its home town's answers and absent one town over. (Our town-level testing exists precisely because this gap is common.)

The freeze rule: once the set is built, changes are versioned events. Add prompts if the business adds services or towns — and record the date — but never silently swap prompts. A drifting instrument produces fictional trendlines.

Part 4: Engine coverage

Minimum viable coverage for a local business, and what each surface is actually measuring:

- ChatGPT — the largest consumer chat surface; with browsing, answers ground in live web sources. Log the sources it cites.

- Perplexity — citation-forward by design; the clearest window into *which* pages carry local authority. If you're absent here, the source layer usually explains it.
- Gemini — grounded in Google's index and local data; in our testing it often responds earliest to Business Profile and entity fixes.
- Google AI Overviews / AI Mode — Google's own generative surfaces. Google's guidance is explicit that these have "no additional requirements… nor other special optimizations necessary" beyond standard Search eligibility — which means your fundamentals are what's being measured. Critically: an overview not appearing for a query is a loggable state, not a failed test. Nobody wins a surface that doesn't render.
- Maps — not generative, but it feeds everything and is itself the highest-intent local surface. Log actual rank position for the core category queries. Note that a single-point Maps check understates reality: rankings vary block by block, which is why grid-based scanning exists.
- Claude / Copilot — secondary for most local trades today; worth quarterly spot checks rather than biweekly instrumentation. If your buyers skew commercial/B2B, promote Copilot into the core rotation.

Cadence: every two weeks. Weekly adds noise and burnout; monthly is too slow to connect fixes to effects. The correct cadence is the one that is still running in month six.

Part 5: Logging and scoring

The log schema. Every run writes the same row:

```
date | prompt_id | prompt_text | engine | appeared (did an answer render at all)
mentioned (y/n) | position (int or null) | sentiment (pos/neutral/neg)
competitors_named (list) | sources_cited (list of domains) | notes
```

A spreadsheet is fine. A database is better. Screenshots-in-a-folder is neither.

Judging rules — decide them before you start:

- "Mentioned" means the business is explicitly named. Not implied, not "local licensed plumbers," not a sister brand.
- Position = order among *named businesses*, not paragraph placement.
- Sentiment is coarse by design (positive / neutral / negative). Fine-grained sentiment on generated text is false precision.
- Same judge, same rules, every cycle — whether the judge is a person or an automated classifier. Consistency outranks sophistication.

The rollups — the 6Signal AI Visibility Scorecard:

| Metric | Definition | Why it matters | |---|---|---| | Mention rate | % of prompt×engine runs where you're named | Headline visibility number | | Per-engine mention rate | Same, split by engine | Different engines fail for different reasons — this routes the fix | | Recommendation position | Avg. position when named | Being fifth of five is not being first of three | | Share of voice | Your mentions ÷ (yours + competitors') across the set | Catches the case where you improve but the market improves faster | | Citation share | % of cited domains that are yours vs. directories vs. competitors | Reveals whether you or third parties carry your reputation | | Coverage breadth | % of towns/services with at least one strong result | Finds the dark spots averages hide | | Win/loss events | Prompt×engine flips ✗→✓ and ✓→✗ this cycle | The honest progress metric |

Score each dimension, weight decision prompts highest, and resist inventing a single blended "AI score" for external bragging — composite scores hide exactly the routing information that makes measurement useful.

Part 6: Interpreting variance

Variance is the part that breaks naive programs. Rules that keep you honest:

01 **Never conclude from one run.** A prompt×engine cell earns a verdict only across multiple cycles. Flickering cells (✓ one cycle, ✗ the next) are real — they usually mean you're borderline in that retrieval, which is itself actionable information.

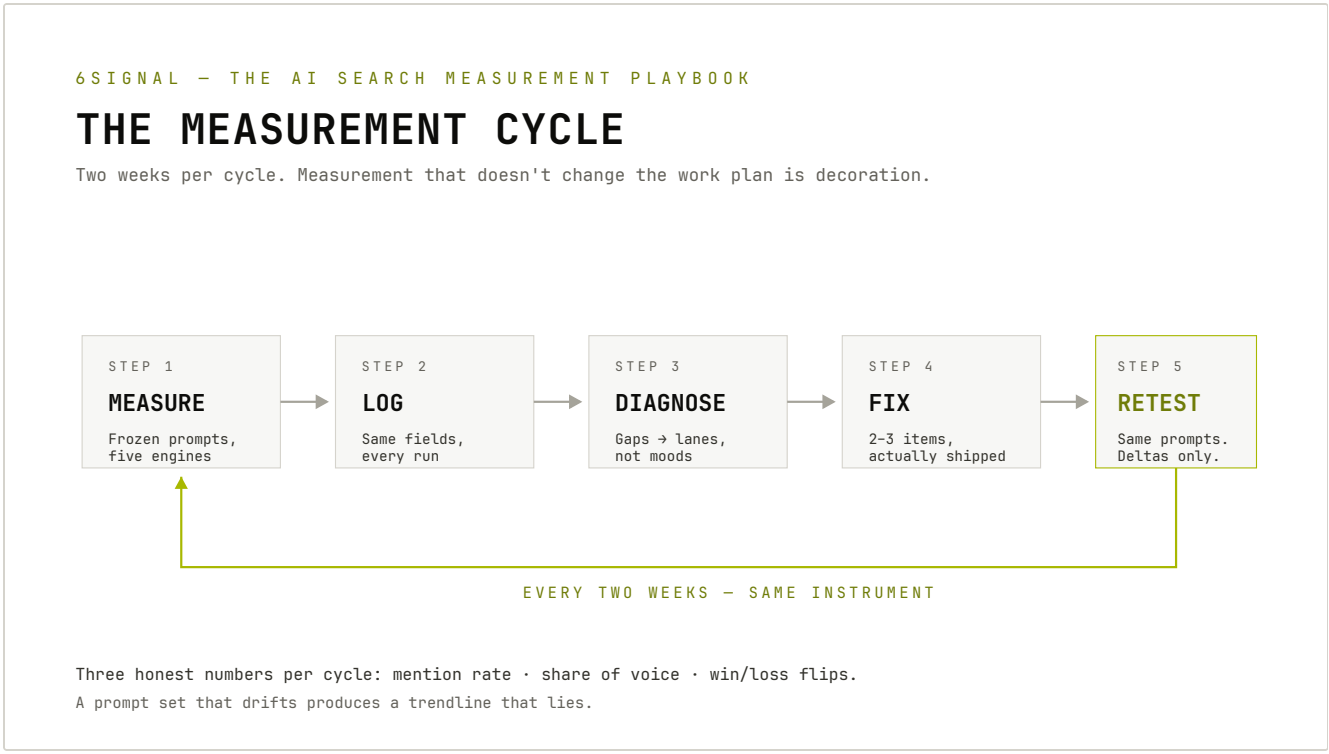
02 **Distinguish three kinds of zero.** *Never named* (structural absence — the big fix lane), *named then lost* (a competitor or source shift worth investigating), and *surface absent* (no overview rendered — not your loss).

03 **Expect uneven engine response to fixes.** Profile and entity work tends to show up in Google-adjacent surfaces first; content and citation work often lands in Perplexity/ChatGPT later. Divergence after a fix cycle is normal; total stasis everywhere is diagnostic.

04 **Do not re-roll for a better answer.** Running a prompt five times and logging the best result is the methodology equivalent of weighing yourself until the scale flatters you. If you use multi-run sampling, log all runs and score the rate.

05 **Cross-cycle trendlines are the product.** The question is never "what did Gemini say Tuesday"; it's "what direction has Gemini moved over three cycles, and did that follow our fix?"

Part 7: From measurement to fixes



Every persistent gap maps to a lane. This table is the entire point of the program:

| Persistent observation | Likely diagnosis | Fix lane | |---|---|---| | Low mentions across all engines | Entity unclear; content doesn't answer buyer prompts | Entity cleanup; answer-ready service pages | | Strong chat engines, weak Maps | Local signals lagging | GBP completeness, review velocity, citations | | Strong Maps, weak chat engines | Web content/source footprint thin | Question content, directory presence, PR/mentions | | Competitors cited from directories you're absent from | Source gap | Get listed and consistent where the engines already look — directories keep winning AI answers for a reason | | Cited but not recommended | Proof gap | Review specificity, third-party corroboration | | Visible in home city, dark in satellite towns | Location gap | Town pages, GBP service area, town-specific proof | | Named but with stale/wrong facts | Record inconsistency | Align site, profile, directories — same facts everywhere |

Cycle discipline: each measurement cycle ends with the fix list re-ranked, two or three items promoted into actual work, and everything else explicitly deferred. Measurement systems die from producing observations faster than anyone acts on them.

Part 8: Sample spreadsheet schema

For teams starting in a spreadsheet, three tabs:

Tab 1 — `prompts : prompt_id | text | intent_band (decision/problem/verification) | branded (y/n) | town | service | date_added`

Tab 2 — `runs` : the log schema from Part 5, one row per prompt×engine×cycle.

Tab 3 — `cycle_summary : cycle_date | mention_rate | share_of_voice | per-engine rates (5 cols) | wins | losses | top_gap_1/2/3 | fixes_shipped_last_cycle`

The third tab is the one you show the owner. The second is the one that keeps everyone honest.

Common mistakes

01 Branded-prompt padding — blending "tell me about us" prompts into the headline rate. Always segment.

02 Prompt drift — quietly rewriting prompts until the numbers improve.

03 Single-engine generalization — treating ChatGPT as "AI" and skipping the surfaces where local buyers actually are (Maps, AI Overviews).

04 Screenshot governance — deciding strategy from whichever screenshot most recently alarmed someone senior.

05 Dashboard terminus — a beautiful tracking sheet feeding zero work orders. If Q2's measurement didn't change Q3's plan, the program failed regardless of how pretty it was.

06 Ignoring the "no overview" state — scoring AI Overviews absence as a loss when the surface simply didn't render for that query.

07 False precision — decimal-point sentiment scores and blended indices on top of a generative process with real variance. Report ranges and directions, not fake exactness.

What to fix first

If you have no measurement program: build the prompt set this week (Part 3), run one full baseline across five engines, and log it properly. That's one afternoon of work and it converts every future visibility argument from opinion to observation.

If you have a casual program: freeze the prompt set, add the mention/citation/recommendation/source distinction to your logging, and start recording win/loss events per cycle.

If you have a real program: audit the last ninety days for the only failure that matters — observations that never became work orders.

What this means for operators

You don't need to become a data analyst. You need three numbers on a schedule — mention rate, share of voice, wins/losses — and the discipline to let them pick your next fix instead of your instincts. The full DIY version of the basic test lives in How to Test Whether AI Recommends Your Business; the honest framing of the whole discipline in How to Measure AI Visibility Without Lying to Yourself.

Limitations and caveats

Be suspicious of anyone in this space who doesn't state these plainly:

- **Engines change without notice.** Retrieval pipelines, grounding sources, and answer formats shift. A methodology survives this; any single number may not.
- **Your probes aren't your buyers.** Logged-out, clean-context testing approximates a population of real users with histories, locations, and account personalization. Treat results as directional, not census-grade.
- **Sample sizes are small.** Twenty prompts × five engines is 100 cells — enough for trendlines and gap-finding, not for statistical significance claims. Don't dress the output in confidence it doesn't have.
- **Nobody can guarantee inclusion.** Google explicitly provides no mechanism to pay or petition for AI feature inclusion, and chat engines offer nothing of the kind either. Measurement tells you where you stand and what to try. The verbs remain "improve" and "verify," never "guarantee."

Want this run for you instead of by you?

This playbook is the system we operate for clients — fixed prompt sets, six engines, logged verdicts, biweekly cycles, and a fix list that changes every cycle based on what the engines actually said.

Book the Visibility Audit

Sources and further reading

- Google Search Central: AI features and your website — developers.google.com/search/docs/appearance/ai-features
- Google Search Essentials — developers.google.com/search/docs/essentials
- Google Business Profile: How to improve your local ranking — support.google.com/business/answer/7091
- Google Search Console: Performance report documentation (AI feature traffic under "Web") — support.google.com/webmasters
- Pew Research Center (2025): analysis of user behavior on Google search results pages with AI Overviews — pewresearch.org
- 6Signal: The AEO Field Manual — /research/aeo-field-manual-answer-engine-optimization

NEXT STEP

See where you *actually stand.*

The AI Visibility Audit runs your company through all six layers — GEO, AEO, LEO, VEO, PEO, IEO — and delivers instant results. $27. Specific to your business, your trade, and your market.

GET THE AUDIT

6signal.co/visibility-check