

Measured, Not Guessed:

A Field Methodology for AI Visibility (GEO)

AUTHOR
Jun Liu

PRACTICE
JUN & GEO ·
AccessSmith

LOCATION
Cixi, Zhejiang, China

DOCUMENT
WP—01

PUBLISHED
October 20, 2026

VERSION
1.0

LICENSE
CC BY-SA 4.0

LENGTH
4,699 words

■ CONTENTS

§ Executive Summary

1 The Problem: AI Answers Decide Before Websites Do

2 What to Measure: Exposure × Citation, Three Rates, Two Objects

§ Definitions Used in This Paper

3 Sampling Discipline: Why Three Runs

4 The Cognition Audit: What AI Thinks You Sell

5 Fix Operations

6 Reporting Standard & Methodology Note

7 Limitations

R References

A About the Author

L License

■ § Executive Summary

Buyers increasingly ask AI platforms — ChatGPT, Perplexity, Copilot, Gemini, Grok, Claude — for product recommendations before they visit a single website. The answer those platforms give is now the

first shortlist. A brand that is not in the answer is not in the consideration set, and a brand that is in the answer but described wrongly may be worse off than an absent one.

Generative Engine Optimization (GEO) is the practice of measuring and improving a brand's position inside AI-generated answers. This paper sets out the field methodology we use in our audit practice: what to measure, how to sample it, how to audit what AI systems believe about a brand, how fixes are sequenced, and how results must be reported so that anyone can check them.

Five commitments run through the method:

- 1. Measurement splits into two independent halves.** *Exposure* — whether AI systems know your brand, name it, and describe it correctly — and *citation probability* — whether they use your own pages as evidence — fail independently and are fixed by different work. We have measured a brand at 100% mention rate and 0% citation rate in the same audit (our measurement; case detailed in §4).
- 2. Three rates are scored separately, never blended.** Mention rate, citation rate, and recommendation rate are different events with different causes. Published research by Lily Ray (Algorythmic) shows why the distinction matters: in her citation-audit retests, when AI answers cited vendors' own pages, the vendors were still excluded from the answer's recommendations in 69% of cases in June 2026, rising to 83% in September 2026 (published research).
- 3. Single samples are noise.** AI answers are generated fresh each time. Kevin Indig's research (Growth Memo) found that the same prompt, run three times on the same platform, produced consistent citations only 2.2% of the time (published research). Our standard is therefore a fixed question bank, every question sampled three times per platform, reported as appearance rates.
- 4. Cognition precedes visibility.** Before exposure work begins, we audit what AI systems believe the brand *is* — identity, category, business model — claim by claim. Earning more mentions for a misdescribed brand only repeats the error to more buyers.
- 5. Every report carries a methodology note.** Sample sizes, question sets, platforms, dates, and limitations are printed next to the numbers. A score without its method is an advertisement, not a measurement.

One boundary applies throughout: we never promise that an AI platform will recommend a brand. No practitioner can promise that honestly. What can be promised — and checked — is measurement, a prioritized fix list, and re-measurement on a fixed instrument.

■ 1 The Problem: AI Answers Decide Before Websites Do

The classic buying journey ran through a search results page: ten links, a ranking, a click. Measurement followed the page — positions, impressions, click-through rates. AI answers break that model in three ways.

First, the answer is the destination. A buyer who asks an AI platform "which supplier should I consider for X" receives a synthesized recommendation, not a list to work through. Brands are named or not

named; a small number are put forward as choices. The shortlist is written before any website is visited, and most of the brands a search engine would have shown never appear at all.

Second, the answer can be confidently wrong about you. AI systems do not look brands up in a registry. They reconstruct an identity from whatever text about the brand exists where they can read it. Where the footprint is thin, the reconstruction fills gaps with priors — the nearest category, a similar-sounding name, the most common business model. The result is a failure mode that ranking reports cannot see: a brand can be highly "visible" and simultaneously misdescribed (§4 documents a measured case).

Third, being used is not the same as being chosen. An answer may cite a vendor's pages for specifications while recommending a competitor in its text. Citation dashboards report this as a win. Lily Ray's repeated citation audits quantify the gap: vendors whose pages were cited were excluded from the answer's recommendations in 69% of cases in June 2026 and 83% in September 2026, and citations to vendors' self-ranking listicles fell 38% over the same research program (all published research, Algorithmic). A measurement practice that counts citations and calls it visibility is reassuring its clients with the weakest of the three signals.

Compounding all three problems, the object being measured is unstable. An AI answer is generated, not retrieved. Kevin Indig's formulation — prompt tracking is polling, not rank tracking (published research, Growth Memo) — is the correct mental model, and his 2.2% three-run citation-consistency figure is its hard evidence. Any methodology that does not design for this instability from the start is measuring noise and reporting it as fact.

The rest of this paper is the methodology we built in response: a two-part measurement frame (§2), a sampling discipline (§3), a cognition audit that runs before visibility work (§4), a fix operation (§5), and a reporting standard that makes every number auditable (§6) — with the method's own limits stated plainly (§7).

■ 2 What to Measure: Exposure × Citation, Three Rates, Two Objects

Definitions Used in This Paper

This box separates what we borrow from what we define. One term belongs to the field: we follow it and claim no ownership of it. Everything else in the box is this methodology's own instrument, defined here so that every number in this paper can be re-derived from these definitions.

FOLLOWED — AN ESTABLISHED TERM.

GEO (Generative Engine Optimization)

— the discipline of measuring and improving a brand's position inside AI-generated answers. The term comes from the research tradition opened by Aggarwal et al., *GEO: Generative Engine Optimization* (2023). This paper uses GEO in that established sense.

DEFINED HERE — TERMS OF THIS METHODOLOGY.

In this methodology, the following terms mean exactly this:

Part One — AI Exposure.

The first half of the audit: whether AI systems know the brand — identify it correctly, name it when buyers ask about the category, and describe it accurately. Read through brand recognition, share of voice, and brand sentiment (§2.1).

Part Two — Citation Probability.

The second half of the audit: whether AI systems use the brand's own pages as evidence when they answer. Read through the citation rate and presence quality (§2.1).

Sampled answer.

One answer returned by one run of one question on one platform. It is the denominator of every rate in this paper: each measurement cell (one object × one platform) holds a fixed bank of five to ten questions, each run three times (§3) — 15 to 30 sampled answers per cell — and every rate is an appearance rate across those runs.

Mention rate.

The share of sampled answers in which the brand is named at all: answers naming the brand ÷ sampled answers.

Citation rate.

The share of sampled answers in which the brand's own pages appear among the answer's sources: answers citing the brand's pages ÷ sampled answers.

Recommendation rate.

The share of sampled answers in which the answer's text puts the brand forward as a choice: answers recommending the brand ÷ sampled answers. The denominator is all sampled answers in the cell, not only the answers that mention the brand. A recommendation is counted when recommendation language (phrases such as "best for," "top choice," "we recommend," or an explicit pick) appears within ±200 characters of a brand mention.

Share of voice (SOV).

An established marketing term; the denominator used here is defined by this methodology. In category-level buyer questions that name no brand, share of voice is this brand's appearances as a share of all brand appearances in the same sampled answers: appearances of this brand ÷ appearances of all brands (this brand plus named competitors).

2.1 Two halves, measured and fixed separately

A GEO audit is split into two parts, because the two parts fail for different reasons and are repaired by different work.

Part one — Exposure. Does the AI system know the brand, name it when buyers ask about the category, and describe it correctly? Exposure is read through three sub-measures:

- **Brand recognition** — whether the system identifies the brand's actual identity: what it makes, where it is, who it sells to.
- **Share of voice** — in category-level buyer questions that name no brand, how often this brand appears relative to competitors.
- **Brand sentiment** — when the brand is mentioned, the tendency of what is said about it.

Part two — Citation probability. When the system answers, does it use the brand's own pages as sources? This is read through the citation rate itself, and through **presence quality** — how cited material is presented in the answer.

Market competition is not a third part but a cross-cutting appendix: every figure above is reported next to the same figures for named competitors, on the same questions, in the same runs. A rate without a competitor baseline is a number without a scale.

The two halves are empirically independent. In the anonymized demo measurement detailed in §4, a brand scored a mention rate of 100% and a citation rate of 0% across the same answer set (our measurement). If the audit had reported one blended "visibility score," that structure — known but never used as evidence — would have been invisible, and with it the entire fix strategy.

2.2 Three rates, scored independently

Within both halves, three events are recorded separately for every answer:

- **Mention rate** — the share of sampled answers in which the brand is named at all.
- **Citation rate** — the share of sampled answers in which the brand's own pages appear among the answer's sources.
- **Recommendation rate** — the share of sampled answers in which the answer's text actually puts the brand forward as a choice. Operationally, we score a recommendation when recommendation language (phrases such as "best for," "top choice," "we recommend," or an explicit pick) appears within a ± 200 -character window around the brand mention.

The three rates fail differently, which is why they are never averaged into one score. A low mention rate is an awareness problem: the platform does not associate the name with the category. A low citation rate is a content problem: the brand's pages do not state facts in a form the platform can lift. A low recommendation rate is a trust problem: third parties have not vouched for the brand where platforms read. Three problems, three fix lists (§5).

2.3 Two objects: the brand and each product

Every rate is measured twice — once for the **brand** and once for **each product or SKU** separately. AI systems can know a brand well and still never surface the specific product a buyer would purchase; conversely a product can circulate while its maker's identity is garbled. The fixes differ by object: brand gaps are repaired mainly with third-party signals, product gaps mainly with page-level citability. Collapsing the two objects into one measurement hides which lever to pull.

2.4 Six platforms, reported per platform

Measurement covers the six platforms buyers actually ask: ChatGPT, Perplexity, Copilot, Gemini, Grok, and Claude. We measure the platforms as products — the surface a buyer meets, with its browsing, source selection, and presentation layers — not as raw model endpoints. Results are reported per platform and never blended into a cross-platform average: a brand can be a consistent answer on one platform and absent on another, and the average of those two facts helps no one decide anything.

2.5 Answer-quality scoring inside each cell

Rates say whether a brand appears; a secondary rubric says how good the appearance is. Within each measurement cell (one object × one platform), sampled answers are also scored 0–100 on three sub-dimensions — **relevance** (is the answer actually answering the buyer's question), **authority** (how strong are the sources it relies on, on a tier ladder running from government/academic and major professional sources at the top, through ordinary sites and community content, to anonymous or unknown sources at the bottom), and **freshness** (is the information current). Cells in the exposure half weight relevance most heavily; cells in the citation half weight authority most heavily. The rubric is diagnostic — it tells the fix team whether a weak cell is a content problem, a source problem, or a staleness problem — and it never substitutes for the three rates, which remain the headline numbers.

■ 3 Sampling Discipline: Why Three Runs

A measurement is only as honest as its sampling. Because AI answers are generated fresh on every run, a single sample of a prompt is closer to a coin flip than to a reading. The evidence is Kevin Indig's: the same prompt run three times on the same platform produced consistent citations only 2.2% of the time (published research, Growth Memo). A "visibility score" built on single runs records which way the coin landed.

Our standard has five elements.

A fixed question bank. Each measurement cell is tested with a fixed set of five to ten real buyer questions. Questions are drawn from prompt research — the phrasings buyers actually use, gathered from search and community data — and validated in client interviews before they enter the bank. Quality of questions beats quantity of questions: eight real questions sampled properly tell you more than a hundred invented prompts sampled once. The bank is then frozen, because a fixed instrument is what makes month-over-month comparison honest — the questions do not change between readings.

Three samples per question per platform. Every question in the bank is run three times on each of the six platforms. Rates are reported as *appearance rates* across the runs — a brand's position on a question is 0 of 3, 1 of 3, 2 of 3, or 3 of 3 — rather than as a single yes or no.

Why three, and not thirty. Three samples separate the three states a fix list actually acts on — *never*, *sometimes*, *always* — at a cost that lets a small or mid-size brand afford monthly measurement. The method trades decimal precision for cadence and trend, deliberately, and says so in every report (§6). A

practitioner who needs finer resolution can raise the run count; what is not acceptable is reporting a single run as if it were a rate.

Spot checks are labeled spot checks. Single-run lookups have legitimate uses — a quick check of how one question is answered today. In our workspace they exist, and they are labeled as spot checks wherever they appear, never as measurements and never in a rate.

Repetition over time beats size at a point. The same frozen bank is re-run on schedule (\$5). Any single reading contains generation noise; the trend across readings, on an unchanged instrument, is the signal a business can act on. This is the polling discipline applied literally: fixed questionnaire, stated sample, repeated waves, movement reported between waves.

■ 4 The Cognition Audit: What AI Thinks You Sell

Visibility work assumes the platforms know what the brand is. That assumption fails often enough that we audit it first, in a dedicated step that runs before any exposure or citation work: the **brand cognition audit**.

4.1 A measured case

The following case is from our own audit work, anonymized; the figures are our measurement. The subject is a home-appliance OEM in Cixi, Zhejiang — an established manufacturer selling to overseas buyers under its own brand, with a website, certifications, and a product line it has produced for years. By traditional measures it is perfectly findable: search its name and you find it.

We asked AI systems, in buyer language, what this brand is and what it makes. Three systems gave three different answers, and all three were wrong:

- **The wrong industry.** One system described the brand as a maker of industrial laundry equipment — commercial machines for laundromats and hotels. The factory makes household appliances for homes. Its category sits near laundry; the system reached for the most documented neighboring category and stopped there.
- **The wrong name.** A second system rendered the brand's name as a different, similar-sounding name, attributing the company to another business entirely. Every fact it then produced belonged to the wrong entity — worse than invisibility, because a buyer's inquiry is now aimed at someone else.
- **The wrong business model.** A third system, asked in Chinese, guessed the brand was an online shop selling small appliances — a retailer, not a manufacturer. The single fact its buyers care about most was not missing; it was replaced.

The same measurement produced the brand's demo scores: **mention rate 100%, citation rate 0%, share of voice 54.5%** (our measurement). Read together, the pattern is exact: every system knew the name, none used the company's own pages as a source (citation rate 0% across all twelve answers collected), and so each reconstructed the company from fragments and priors — and improvised

wrongly. Had we reported only the first two numbers, the client would have celebrated a visibility position that was, in substance, three circulating errors.

4.2 The method

The cognition audit has three steps.

1. **Ask directly.** Put buyer-style questions about the brand itself to each platform — what does this company make, where is it based, who does it sell to, is it a maker or a seller — in the languages its buyers use.
2. **Verify claim by claim.** Break each answer into individual factual claims and grade every claim against ground truth: *correct*, *wrong*, or *invented*. Averages hide errors; a claim ledger does not.
3. **Trace each error to its source.** For every wrong claim, establish the likely origin — a namesake company, a stale directory entry, an outdated page, a pure category assumption. The fix differs by origin: a namesake collision needs disambiguation signals; a stale listing needs correction at the listing; an invented fact needs the true fact published, in quotable form, where systems actually read.

4.3 The order is not negotiable

Cognition repair comes first, exposure second, recommendation last. There is no value in earning more mentions for a brand the answers describe wrongly — each new mention repeats the error to another buyer. Identity first is not a stylistic preference; §4.1 is what the reverse order produces.

■ 5 Fix Operations

Measurement without repair is a weather report. Our engagements run in a fixed order — **Audit** → **Fix** → **Monitor** — and the fix work splits along the same line as the measurement: exposure fixes and citation fixes are different trades.

5.1 Exposure fixes: third-party signals

Platforms learn what a brand is mostly from places other than the brand's own site. Exposure repair therefore works on the third-party footprint:

- **Consistent entity facts everywhere they are listed** — the same name, category, location, and description across directories, registries, marketplaces, and public entity records, so the reconstruction (§4) has consistent material to work from.
- **Earned mentions in the places platforms read** — trade press, industry publications, comparison and review content, and genuine community discussion. These are also the sources that feed the recommendation sentence (§1), which is why exposure work and recommendation work are the same investment seen from two sides.
- **Disambiguation and correction at the source** — where a namesake collision or a stale listing is feeding an error, the correction is made at that listing, not argued with on the brand's own pages.

We do not buy placements, spin up sock-puppet discussion, or manufacture reviews. Apart from the ethics, platforms discount exactly these signals — the 38% decline in citations to self-ranking vendor listicles (published research, Lily Ray, Algorythmic) is the visible edge of that discounting.

5.2 Citation fixes: pages worth quoting

Citation repair works on the brand's own site, on making its facts liftable:

- **Answer-first rewriting.** Facts are restated in self-contained, quotable passages — a paragraph a system can lift whole, without reassembling the page. Most corporate sites fail here: the fact exists, but smeared across context, modals, images, or PDFs, so the system reads it and still reaches elsewhere for a cleaner statement of the same fact.
- **Correct facts, published once, kept current.** The canonical statement of each core fact — what the company makes, where, since when, under what certifications — lives on the site in plain text, and wrong versions circulating elsewhere are corrected where they stand.
- **Structure machines can use.** Valid structured data (schema), clean crawl access for AI readers, and an llms.txt manifest — filed, in our practice, under hygiene rather than strategy. Our own before-and-after measurements, on sites where llms.txt was the only change, show near-zero direct effect on citation rates (our measurement). We ship it anyway: it is cheap, it breaks nothing, and a site that labels itself carefully for machine readers is usually a site whose pages are worth reading. What we refuse to do is sell it as a visibility lever.
- **Depth where the questions are.** Pages are built or deepened around the actual buyer questions in the measurement bank, so the citable statement exists for the questions buyers really ask.

5.3 Fix-list discipline

Every fix item on the list names three things: the evidence behind it (the answers and claims it responds to), the rate it is meant to move, and its priority. Items that cannot be tied to a measured gap do not make the list. The list is finite and priced before work begins; there are no open-ended engagements without numbers attached.

5.4 Monitor: the same instrument, on schedule

After fixes ship, the frozen question bank is re-run monthly on the same standard — same questions, three runs per platform, per-platform reporting. Each cycle reports what moved, what did not, and what happens next. AI answers drift as platforms update models and sources; a one-time fix decays quietly, and only scheduled re-measurement catches the decay while it is still cheap to correct.

■ 6 Reporting Standard & Methodology Note

A measurement practice is defined by what it is willing to show. Our reporting standard has four rules.

Rule 1 — Rates, not scores. The headline of every report is the three rates (mention, citation, recommendation), per platform, per object (brand and each SKU), with competitor figures alongside. No

blended visibility score is reported, because a blend conceals exactly the structure the client needs to act on (§2.2).

Rule 2 — Evidence is attached. The full text of the sampled answers is preserved and delivered with the report, including the sources each answer cited. Any number in the report can be traced to the answers it came from, and any classification can be re-checked by a human reader.

Rule 3 — The fix list is part of the report. Findings arrive with the prioritized fix list (§5.3), each item tied to the rate it targets. A report that ends at diagnosis has done the easy half.

Rule 4 — Every report carries a methodology note. The note is printed with the numbers, in the report itself, not in a separate document. Our standard template:

Methodology note. Questions: a fixed bank of [5–10] real buyer questions per measurement cell, drawn from prompt research and validated in client interviews; the bank is unchanged between reporting periods. Sampling: each question run 3 times per platform; all rates are appearance rates across runs. Platforms: ChatGPT, Perplexity, Copilot, Gemini, Grok, Claude — measured as the products buyers use, reported per platform, never averaged. Period: [start date] – [end date]. Scoring: mention, citation, and recommendation scored independently; a recommendation is scored when recommendation language appears within ± 200 characters of a brand mention. Evidence: full answer texts retained and available with this report. Limitations: samples are not censuses of platform behavior; platform models and source selection change without notice; results describe the tested questions, languages, and period, and are not a guarantee of future answers. This report does not promise, and no GEO practice can promise, that any platform will recommend a brand.

The last sentences are deliberate. The red line of this practice — stated in every report and every engagement — is that we promise measurement, fixes, and re-measurement. We do not promise recommendations, rankings in AI answers, or any specific platform behavior, because none of those is ours to promise.

■ 7 Limitations

A methodology paper that does not state its own limits is a sales document. Ours are these.

Samples are not populations. A fixed bank of five to ten questions per cell, sampled three times, is a deliberate trade of precision for cadence (§3). The figures support direction, prioritization, and trend — the decisions fix work actually needs. They do not support decimal-place claims about a platform's total behavior, and we do not make such claims.

The platforms are moving targets. Models are retrained, retrieval sources change, and answer formats are redesigned without notice. A methodology survives this only by freezing its own instrument and re-baselining honestly when a platform's behavior shifts discontinuously; results across such a shift are not strictly comparable, and reports say so when it happens.

Collection surfaces are disclosed. In our current implementation, platform answers are collected through API access, and for some platforms that access routes through intermediary (proxy) model endpoints rather than the platform's own native first-party API. A proxy endpoint may differ from the consumer product surface in browsing behavior and source selection. We regard closing this gap — native first-party access for all six platforms — as standing work, and we disclose the current state here rather than let readers assume otherwise. Readers comparing our figures with measurements taken on consumer surfaces should keep this difference in mind.

Language and market scope. Buyer questions are asked in the languages and market framings of the brand's actual buyers. Results measured in one language do not transfer automatically to another; the cognition case in §4 includes an error that appeared only when the question was asked in Chinese. Each language a brand sells in needs its own bank.

Classification involves judgment. Recommendation detection (recommendation language within a fixed window around a mention) and sentiment reading are heuristics; edge cases can be misclassified. This is why full answer texts are retained with every report (§6, Rule 2): any figure can be re-derived and any classification disputed, by a human, from the same evidence.

The case study is one case. The figures in §4 come from a single anonymized demo measurement. They illustrate what the method detects; they are not market statistics and are not presented as such.

■ R References

- Indig, K. — Growth Memo. Research on prompt-tracking consistency: the same prompt run three times on one platform produced consistent citations in 2.2% of runs; "prompt tracking is polling, not rank tracking." (Published research.)
- Ray, L. — Algorhythmic. Citation-audit retests, June 2026 and September 2026: vendors whose pages were cited were excluded from answer recommendations in 69% (June) and 83% (September) of cases; citations to vendors' self-ranking listicles fell 38%. (Published research.)
- Liu, J. — JUN & GEO measurement practice. Anonymized cognition-audit case (mention rate 100%, citation rate 0%, share of voice 54.5%); before-and-after llms.txt measurements showing near-zero direct effect on citation rates. (Our measurement, 2026.)

■ A About the Author

Jun Liu is a GEO consultant based in Cixi, Zhejiang, China, and the founder of AccessSmith, a GEO services practice. He spent 22 years in product management and 13 years in e-commerce in China, leading product categories that scaled to hundreds of millions in revenue. He works under the name JUN & GEO — "Jun" is his name; GEO, Generative Engine Optimization, is the discipline he practices. He publishes GEO research at jungeo.com and GEO notes in Chinese on his WeChat official account. Engagements and questions: hi@accesssmith.com.

■ L License

© 2026 Jun Liu. This work — *Measured, Not Guessed: A Field Methodology for AI Visibility (GEO)*, v1.0 — is licensed under the Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0). You are free to share and adapt the material for any purpose, including commercially, provided you give appropriate credit and distribute your contributions under the same license. Full license text: <https://creativecommons.org/licenses/by-sa/4.0/>