NOVAXILABS · WHITE PAPER SERIES · NO. 02

The Two-Tier Structure of AI Visibility

Why most AI visibility data is noise — and what a defensible measurement looks like when four credible studies report four different numbers for the same thing.

Author Dr. Adrian Whitmore **Published** 2026-11-09 **Data as of** 2026-10-29

Edition **v1.0** · Novaxilabs Research

Document Information

TITLE	The Two-Tier Structure of AI Visibility: Why Most AI Visibility Data Is Noise, and What to Report Instead
SERIES	Novaxilabs White Paper Series, No. 02
EDITION	v1.0 — first release
PUBLISHED	2026-11-09
DATA AS OF	2026-10-29 (latest source data cited in this edition)
AUTHOR	Dr. Adrian Whitmore — GEO research consultant. PhD in Information Retrieval (University of Cambridge).
CONTACT	info@novaxilabs.com

> **Disclaimer.** This paper is a **secondary research synthesis**. Novaxilabs operates no monitoring sample of its own and presents no original measurement here. Every quantitative figure is taken from a published study and is cited with its sample size, measurement window, and evidence class. Where studies disagree, both numbers are reported and the disagreement is explained rather than averaged away. Figures are directional and dated; engines change behaviour faster than studies replicate.

Revision History

EDITION	DATE	AUTHOR	CHANGE
v1.0	2026-11-09	A. Whitmore	First release. Synthesises seven public measurement studies published between March and October 2026 and argues the two-tier structure of citation stability.

Contents

The executive summary alone answers “so what”. Every number in this paper resolves to a citation in Appendix A.

§ Executive Summary

01 The measurement assumption

02 Scope, sources and method

03 Five findings from the 2026 data

04 What this changes for brands

05 A reporting standard that survives scrutiny

06 Limitations

A Sources & References

B About the Author

C About Novaxilabs

§ Executive Summary

THESIS

AI visibility is not one number. It is **two layers** moving at different speeds: a domain layer that barely changes week to week, and a page layer that is rebuilt almost every day. Most dashboards merge the two, which turns sampling noise into apparent wins and losses.

Six separate studies, published between February and October 2026, measured how stable AI citations actually are. Their headline numbers look irreconcilable. **96.8%** of cited domains show zero change week over week (BrightEdge, week of 1 February 2026) [3]. Yet **69%** of cited domains

change from one day to the next, measured across 530,875 citations (GetMentions AI, June 2026) [6]. And **73.5%** of individual cited URLs appear exactly once and never return, with a median lifespan of zero days (Trakkr, 108,650 URLs over 177 days) [4].

All three are correct. They describe different layers of the same system, at different units — and in the case of [3] and [6], the same layer read over different windows. Once that is understood, the industry's central problem stops being a shortage of data and becomes what it always was: *a unit-of-measurement problem*. A domain can hold its place in the cited set while the page that carries it is swapped out; a weekly average can look flat while every daily reading inside it differs; and a single run can produce a difference that no statistical test would support.

The most consequential finding is not about any one engine. It is that the precision the industry reports is not achievable at the sampling rates the industry uses. To report a citation-share figure with a 95% confidence interval five percentage points wide — roughly ±2.5 points, the resolution at which teams claim "we went up" — one study estimates **40–50 repetitions on Gemini, about 100 on Perplexity, and 150 or more on SearchGPT** [1]. Most dashboards run once. The gap between what is reported and what is measurable is the subject of this paper.

---

01 BACKGROUND

The measurement assumption

Every AI visibility product on the market rests on an assumption inherited from rank tracking: that a number produced today can be compared with the number produced yesterday, and that the difference is a change in your standing. Rank tracking earned that assumption the hard way. Search results were deterministic enough that a movement from position 8 to position 5 meant something, and the tool that reported it could be checked by a person typing the query.

Answer engines break the last link in that chain. Ask the same question twice and the engine may well answer differently, citing a different set of sources, without anything about your pages having changed. This is not a defect to be fixed; it is how a generative system behaves at temperature above zero, with retrieval over an index that updates continuously. The measurement literature barely addressed the question before 2026.

The first rigorous treatment arrived in March 2026. Sielinski's *Quantifying Uncertainty in AI Visibility* reframes the whole exercise: a citation-share figure is not a property of a brand, it is a **sample estimator of an underlying response distribution**. Sampling that distribution once gives a point estimate with an error bar nobody printed. The paper states its central empirical claim plainly: many apparent differences between domains, and many apparent changes across samples, "are not statistically distinguishable from noise" [1].

> “Single-run visibility metrics provide a misleadingly precise picture of domain performance in generative search.”
>
> — Sielinski, arXiv:2603.08924, 39 pages, 13 figures, August 2026 revision [1]

That claim, if taken seriously, invalidates a large share of what the category reports. It does not invalidate the category's purpose. Brands still need to know whether they are in the answer. What changes is the form the answer must take: not "you are at 12.4% today, down from 15.1%", but "your share is 12–16% at this sample size, and the movement you observed is inside that interval".

The practical stakes are asymmetric. A brand that over-reacts to noise burns budget rewriting pages that were never the problem. A brand that under-samples misses a real structural change — and the studies in Section 3 show those do happen, just more slowly and at a different layer than daily tracking suggests.

---

02 METHOD

Scope, sources and method

This is secondary research, and this paper says so plainly. Novaxilabs operates no monitoring sample of its own. Every figure below is a published finding from a named study, reported with its sample size, its measurement window, and its evidence class. What this paper adds is not new data; it is a structural reading of studies that appear to contradict each other, and the translation of that reading into a reporting standard.

Evidence classes

Claims are graded on a five-level scale used across this series: **Official** (vendor documentation), **Academic preprint** (arXiv posting with disclosed methods and data, not yet formally peer-reviewed), **Independent measurement** (third-party sample with stated size and window; no instance in this edition), **Vendor-reported** (a tracking company publishing its own platform data), and **Analysis** (reasoned inference, explicitly not a claim of access to engine internals).

One uncomfortable fact deserves to be stated before the findings, because it shapes how they should be read: **five of the six studies cited here are vendor-reported.** BrightEdge, Trakkr, Ahrefs, GetMentions AI and SISTRIX all measured their own tracking data and published the result; only Sielinski's paper is an academic source with no commercial interest in the answer. References [7] and [8] are supplementary pages from the same Trakkr dataset as [4]: they are cited for specific figures, not counted as separate studies. There is no third-party auditor for AI citation data. That is the state of the field, not a limitation of this paper — and it means the honest way to read any figure in this category, including this one, is to check its stated sample and window first, and its headline second.

Why absolute numbers are never compared across studies

The studies do not share a methodology and this paper does not pretend they do. SISTRIX tracked 82,619 prompts weekly for 17 weeks across six countries [2]. BrightEdge compared consecutive weekly snapshots on its AI Catalyst platform [3]. Trakkr followed 108,650 URLs for 177 days across eight models [4]. Ahrefs followed 37,500 citations daily for 20 days across five tools [5]. GetMentions ran 2,398 queries once a day for 7 days across four engines [6]. A percentage from one of these cannot be subtracted from a percentage in another, and where this paper places them side by side it is comparing *direction and structure*, never magnitude.

What this paper does not claim

It does not claim to know how any engine selects sources internally. It does not claim these findings hold for every market: the only multi-country dataset here covers six Western markets and is not evidence about others. It does not claim the structure is permanent — every study is dated, and the paper's own evidence shows engine behaviour moving faster than any study replicates.

A note on two widely circulated figures. Two numbers appear in most industry write-ups of this topic and are deliberately excluded here. The first is a "23 million sources" study attributed to Writesonic; its primary report could not be located, and only secondary summaries were available at the time of writing. The second is

a claim that 73.4% of cited URLs appear once; the primary study reports **73.5%**, so the primary figure is used throughout [4]. Where a number could not be traced to a primary page, it was dropped rather than repeated.

---

03 FINDINGS

Five findings from the 2026 data

Each finding below states one claim, names the study that supports it, and reports the sample it rests on. Where two studies disagree, both numbers appear — because in this category the disagreement is usually the finding, not an error to be resolved.

FINDING 01

Citation stability is a property of the layer you measure

VENDOR-REPORTED

Three studies, three units, three answers. At the **domain** layer, **96.8%** of cited domains showed zero change week over week, and of the 3.2% that moved, 87% were losses rather than gains (BrightEdge, week of 1 February 2026) [3]. At the **page** layer, **73.5%** of individual cited URLs appeared exactly once and never returned, with a median lifespan of **zero days** (Trakkr, 108,650 URLs, 177-day window) [4]. And inside each answer there is a third unit: SISTRIX found that **86.5%** of prompts have a stable core of one to five domains while the remaining domains rotate at **89% per week** [2].

FIGURE 1 — THE SAME CITED SET, READ AT TWO UNITS

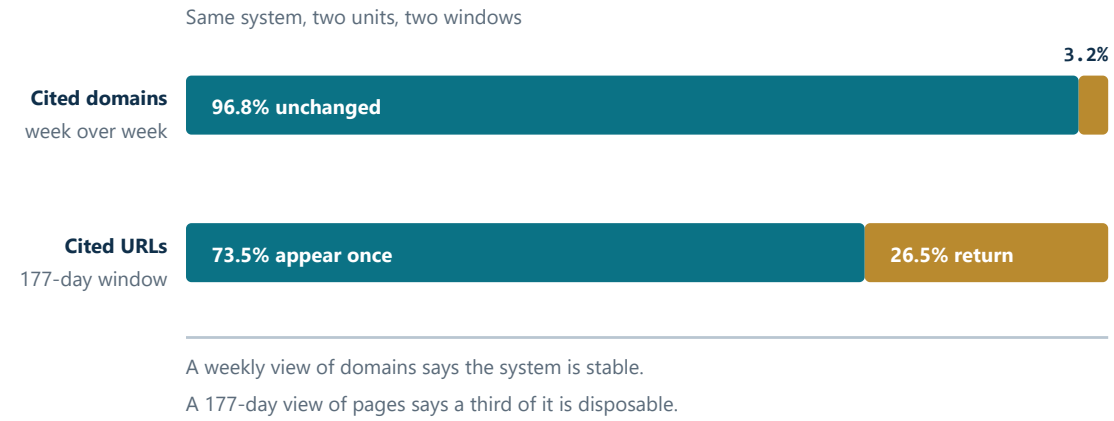


Figure 1 — Neither bar is wrong. The domain layer holds while the page carrying it is replaced, which is why a brand can be "always cited" and "never cited on the same page" at once.

Sources: BrightEdge, consecutive weekly snapshots, 2026-02-01 week [3]; Trakkr Research Study 009, 108,650 URLs across 200 brands, 2025-10-04 to 2026-03-30 [4]. The two bars use different units and windows and are not on a common scale.

The two-tier reading resolves what otherwise looks like a contradiction between the most careful studies in the field. It also changes what "our AI visibility dropped" can mean. If a page fell out of an answer while the domain stayed in the cited set, the domain layer will report no change, the page layer will report a loss, and both are accurate descriptions of the same event. A report that mixes the two produces a number nobody can act on.

SISTRIX's framing is the most operationally useful one published so far: every response has a **fixed core and a carousel**. For 86.5% of prompts there is a small stable set of domains that persists across weeks and months; everything else is reallocated weekly. The question a brand should ask is therefore not "am I in the answer?" but "am I in the core, or in the carousel?" [2] ANALYSIS

That distinction has a time dimension the daily dashboards miss. Trakkr's brand-level series — 857,138 daily reports covering 10,991 brands over ten months — puts the **median time from peak to half** at 31 days, and finds only **173 of 1,314 brands** tracked for 90 days or more holding a steady citation presence [4]. Losing the core is a slow event measured in weeks. Losing a page is a fast event measured in days. Reporting both at the same cadence guarantees that the fast event drowns out the slow one.

FINDING 02

The platforms are three different architectures, not three settings of one machine

ACADEMIC + VENDOR

The engines do not cite at the same volume, and they do not rotate at the same rate. Median citations per answer: **Gemini 36–43**, **Perplexity 19–22**, **SearchGPT 5–7** (Sielinski, three consumer topics, repeated sampling) [1]. Weekly domain churn: **AI Mode 56%**, **ChatGPT Search 74%**, with AI Overviews completely stable for just over half of queries (SISTRIX, 82,619 prompts, six countries, 17 weeks) [2].

FIGURE 2 — MEDIAN CITATIONS PER ANSWER, BY PLATFORM

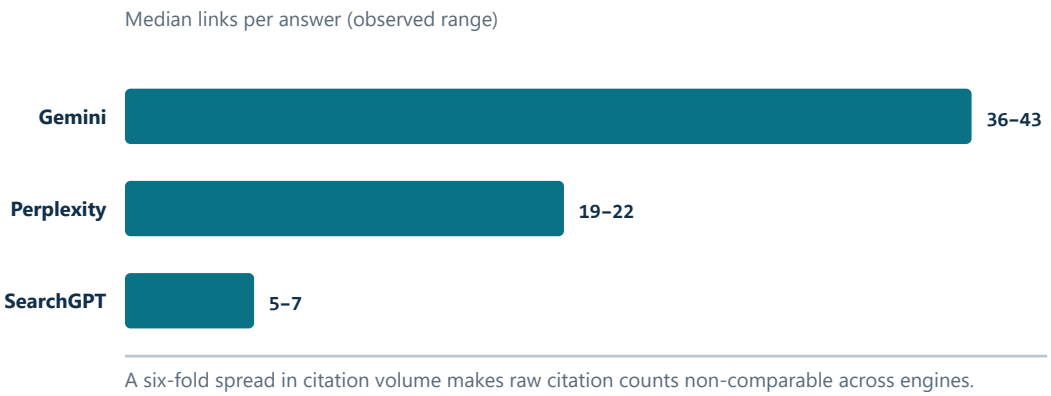


Figure 2 — Gemini distributes the answer across dozens of links; SearchGPT concentrates it into a handful. Counting how many times a brand appears therefore measures the engine's citation style before it measures the brand.

Source: Sielinski, arXiv:2603.08924, three consumer topics (bird feeders, multivitamins, running gear), nine daily collections plus ten-minute high-frequency sampling [1]. Percentages of share, not raw counts, are the comparable metric.

This is the mechanical reason raw citation counts cannot be compared across engines, and the reason "share of citations" is the only metric with a defensible cross-platform meaning [1]. A brand cited twice inside a SearchGPT answer that carries six links has roughly a third of that answer; the same two mentions inside a Gemini answer carrying forty links is a rounding error. Both may be shown to a client as "two citations".

FIGURE 3 — WEEKLY TURNOVER OF CITED DOMAINS, BY PLATFORM

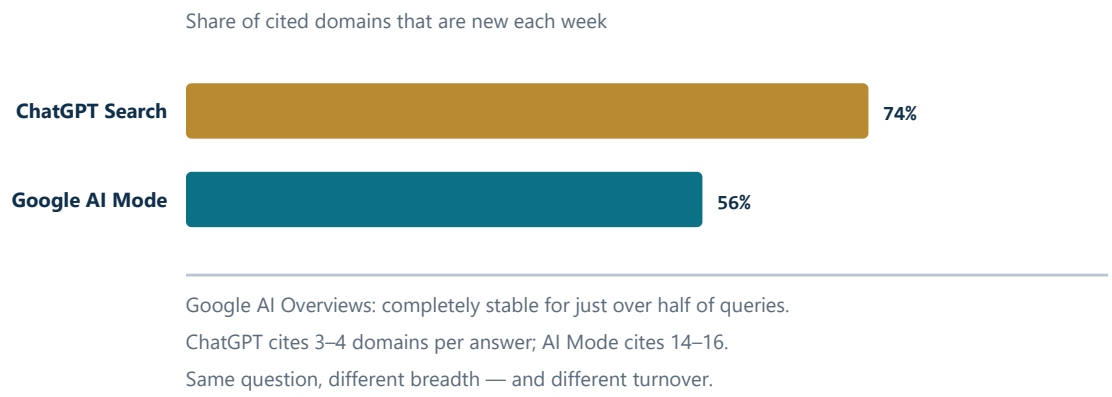


Figure 3 — Turning over 74% of a three-domain answer every week is a structurally different problem from turning over 56% of a fifteen-domain answer. Averaging them produces a number that describes no engine.

Source: SISTRIX AI Research Index, 82,619 qualified prompts, 1,548,213 snapshots, six countries, 17 weeks (2025-12-17 to 2026-04-08) [2]. AI Overviews is reported qualitatively by the study ("completely stable for just over half of queries") and is therefore not plotted on the same scale.

SISTRIX adds a result that undercuts the assumption of a universal content strategy. In Google AI Mode, **80% of core domains are in German** for German queries and 85% of core URLs are evergreen, with product and shop pages dominating the core while guides and news rotate through. In ChatGPT Search the profile inverts: core domains score **16/20 on E-E-A-T** against 14/20 for peripheral ones, institutional and documentation sources form the core, and **68% of core content is in English even for German queries** [2]. The two platforms reward near-opposite content profiles, and no single page can occupy both cores comfortably.

One more number from the same study deserves attention because it is the strongest argument against planning citation strategy around press coverage: only **1.4% of news articles** cited as sources remained permanently in the citation set [2]. Editorial news is a one-way ticket. Whatever survives is evergreen by construction.

FINDING 03

At the sampling rates the industry uses, most reported movement is indistinguishable from noise ARXIV PREPRINT

To produce a citation-share figure whose 95% confidence interval is five percentage points wide — about ±2.5 points, the resolution at which a team says "we gained two points" — the published estimate is **40–50 repetitions on Gemini, about 100 on Perplexity, and 150 or more on SearchGPT** (Sielinski, arXiv:2603.08924) [1]. Dashboard sampling is usually one run per query per day.

FIGURE 4 — REPETITIONS NEEDED FOR A ±2.5-POINT CITATION-SHARE INTERVAL

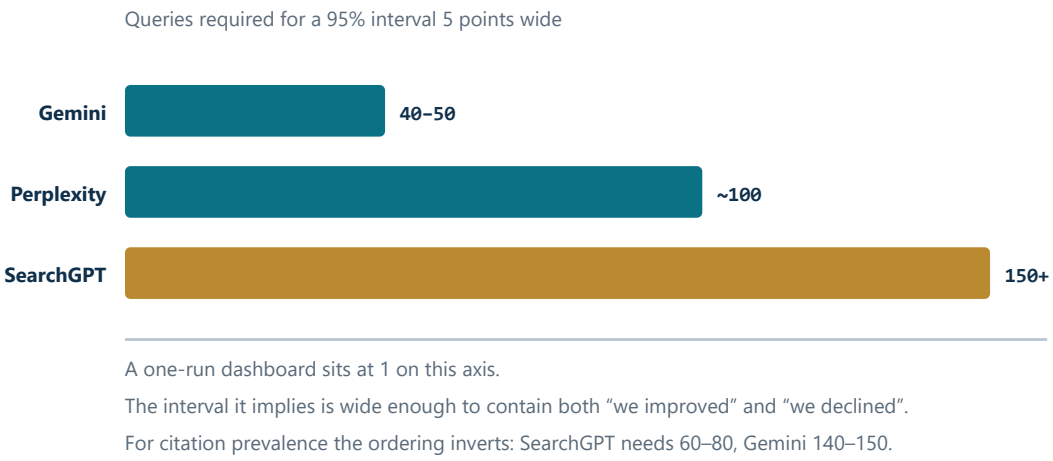


Figure 4 — The precision brands assume is a property of the sample size, not of the tool. The required size differs by platform and by metric, which is the point the paper makes: one universal dashboard setting cannot produce comparable precision across engines.

Source: Sielinski, arXiv:2603.08924, three consumer topics, nine daily collections plus ten-minute sampling [1]. Figures are the study's empirical convergence points, not a universal law. The same paper reports non-monotonic convergence, so the practical advice is to fix the sample size in advance rather than stop when the interval happens to look narrow.

The paper's second practical warning is subtler and easier to get wrong. Confidence intervals do not always narrow smoothly as more queries are added; on SearchGPT they can narrow and then widen again, because the underlying distribution shifts during sampling. A practitioner who stopped at a conveniently narrow point would report an interval that *understates* the true uncertainty [1]. The safe rule is to commit to a sample size in advance — derived from the platform and the topic — rather than to stop when the number looks acceptable.

Put those two findings together and the implication for reporting is unavoidable. A dashboard that samples once and prints a two-decimal percentage is not measuring a small difference; it is

printing an artefact. The number is real arithmetic on real data, and it carries no inferential content beyond the one sample it came from. The gap of a few points that a client asks about each month is, at that sample size, the measurement's own uncertainty expressed as apparent performance.

FINDING 04

Stability is a property of the question category, not of the engine

VENDOR-REPORTED

Across five AI tools tracked daily for 20 days, the same first cited source survived far more often on health topics (**56–79%** of the time) than on fashion topics (**13–23%**) (Ahrefs, 37,500 citations) [5]. Brand-name queries were the most stable segment overall — a brand's own site stayed cited about **75%** of the time day to day, and about **60%** on Gemini — while "best X" questions were among the least stable, at roughly **75% of cited pages changing daily**.

FIGURE 5 — KEEPING THE SAME FIRST CITED SOURCE, BY TOPIC

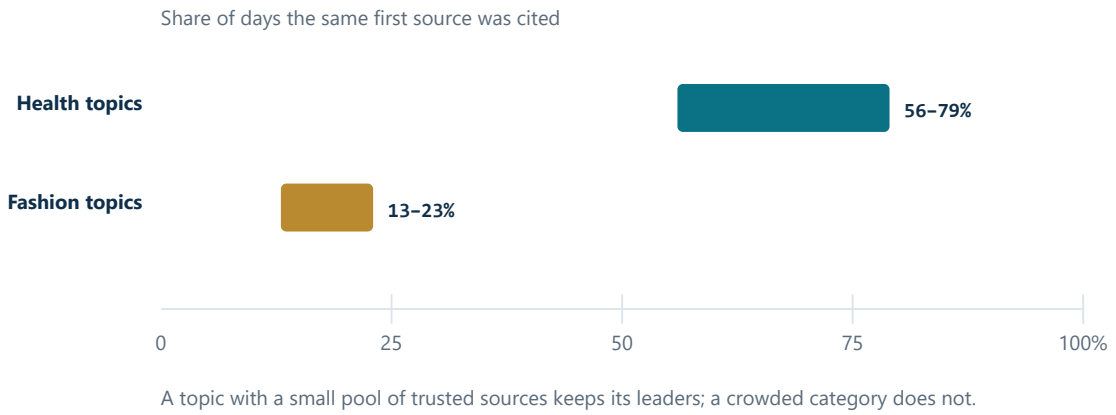


Figure 5 — The same engine, the same 20 days, the same tracking method — and a four-to-five-fold difference in source retention driven entirely by the topic. Aggregate "AI visibility stability" numbers average this away.

Source: Ahrefs, 37,500 citations across five AI tools, tracked daily for 20 days, published 2026-10-29 [5]. Ranges are the study's reported spans across question categories; bars show the span, not a point estimate.

SISTRIX reaches the same conclusion from a different direction, and adds the mechanism. Brand queries drift about **20% less** than the average (45% versus 56% churn-in), which sounds reassuring until the co-citations are examined: the twelve to fifteen other domains cited alongside a brand **change at 70% per week** [2]. The brand's own domain is anchored; the space around it is reallocated weekly. That is where competitive displacement actually happens, and it is invisible to a metric that only asks whether the brand appeared.

The planning consequence is specific. A brand that tracks only its own name will see a stable line and conclude that nothing is moving, while the set of sources cited beside it turns over every week. A brand that tracks only category terms will see constant movement and conclude the category is unmeasurable. Both are looking at half the structure.

FINDING 05

Short-window stability routinely hides long-window drift

ARXIV PREPRINT

On Gemini, consecutive-day ranking correlations averaged **ρ = 0.866** (bird feeders), **0.806** (multivitamins) and **0.913** (running gear). Comparing the first day with the last day of the same nine-day window, the correlations fell to **0.689**, **0.726** and **0.900** (Sielinski) [1]. Every adjacent pair looked acceptable; the window as a whole had drifted.

FIGURE 6 — ADJACENT-DAY STABILITY VERSUS FIRST-TO-LAST DRIFT

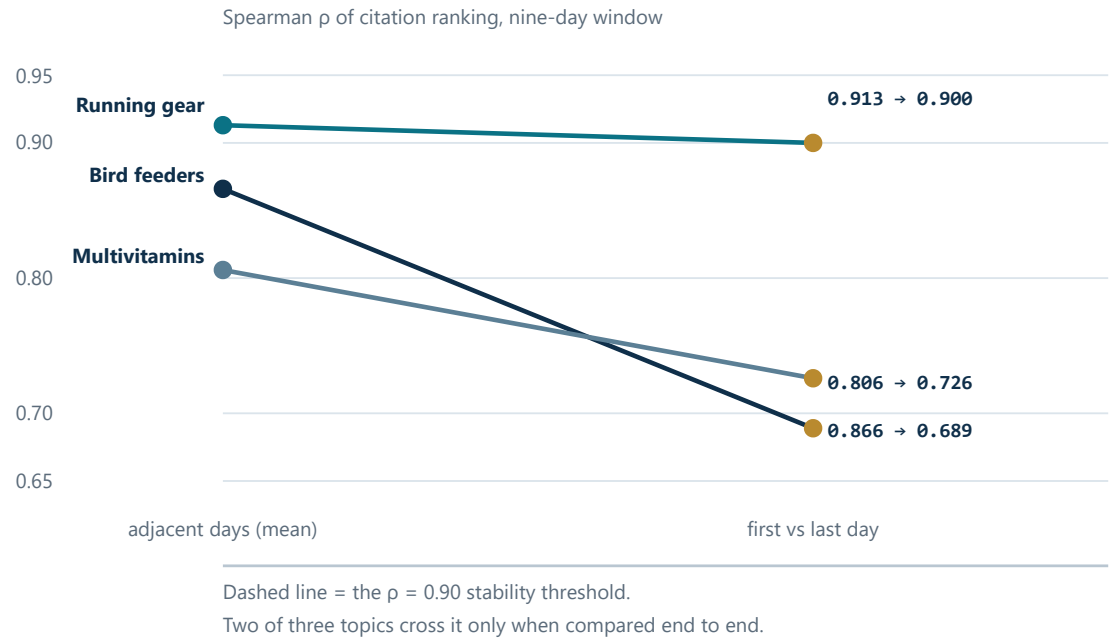


Figure 6 — Consecutive-pair comparison is the default in every daily dashboard, and it is structurally blind to cumulative drift. Two of the three topics look stable pair by pair and are not stable across the window.

Source: Sielinski, arXiv:2603.08924, Table 6, Gemini, nine daily collections per topic [1]. Adjacent values are the mean of eight consecutive-pair correlations; the endpoint values compare the first and last collection.

The same paper records a case at the opposite extreme, which is useful because it shows the failure is not universal. SearchGPT on bird feeders had five consecutive pairs with intervals narrow

enough to trust, and **none of them reached the 0.90 stability threshold** — the mean interval width was 0.225, close to the limit at which the comparison stops being informative at all [1]. On the other two topics the pairs were not even sufficient to compute. A daily dashboard would have reported a number for each of those days.

There is a clean operational reading of this. Pairwise comparison answers "did it change since yesterday", which is a question about noise. Endpoint comparison answers "did it change over the period we care about", which is the question a quarterly review actually asks. A measurement system that only does the first will keep confirming that nothing is wrong while the thing it was hired to watch quietly moves.

---

04 IMPLICATIONS

What this changes for brands

None of this argues that AI visibility should stop being measured. It argues that the unit of measurement has to change, and that three habits in current practice are actively misleading.

Two supplementary figures from Trakkr's public data pages make the same point on different samples. The live citation index reports that **74%** of cited links appear in one answer and never again — a second sample behind the 73.5% longitudinal figure, not an independent replication [7]. The same group's 1.3-million-citation analysis finds **Wikipedia alone accounting for roughly 17% of all AI citations**, which is what a core layer looks like when a single domain occupies it [8]. A carousel of thousands of pages rotates underneath a handful of structurally unchanging sources.

1. Stop treating "cited" as a binary state

The data says every answer has a core and a carousel [2]. A brand is either in the core — persistent across weeks, and the thing a competitor has to displace — or it is in the rotation, where presence today predicts very little about presence next week. Reporting a single "you are cited in 34% of answers" figure collapses the two states into one number, and the number is dominated by the volatile half. The useful question is what share of tracked prompts a brand holds a core position on, and how that share moves month to month.

2. Stop reporting daily deltas as performance

Daily movement is real, large, and almost entirely uninformative at one sample per query [1][6]. A daily series is worth keeping — it is how an outage, a model update, or a sudden de-indexing announces itself — but it is not a performance metric. The number a stakeholder should see weekly is a share with an interval; the number worth escalating is a break in the series, not a two-point wobble.

3. Stop averaging engines into one score

Gemini cites 36–43 links per answer, SearchGPT 5–7, and the two rotate sources at different rates on different content profiles [1][2]. A blended "AI visibility score" across engines is an average of quantities with different units, and it moves when the mix of prompts changes rather than when the brand does. Per-engine numbers, reported separately, are less satisfying to look at and far more useful to act on.

There is a budget consequence too. If the core layer turns over on a scale of weeks to months — Trakkr's median brand takes 31 days to fall from peak to half [4] — then the intervention cycle should be monthly, not daily. Content published to chase a daily swing is optimising against noise. Content published to enter a core set is competing on the timescale the data actually supports.

---

05 STANDARDS

A reporting standard that survives scrutiny

The six rules below are not a house style. Each one follows directly from a measured property of the data, and each one exists to prevent a specific mistake that the studies above document. A report that omits them is not wrong in its arithmetic; it is wrong in what it claims the arithmetic means.

Element	Common practice	Defensible practice	Evidence
Unit	One blended “visibility percentage” per brand	Domain-level share with an interval; page-level movement kept as an internal diagnostic, never in a client report	Domain layer 96.8% stable week to week [3]; 73.5% of URLs appear once [4]
Sampling	One run per query per day	A fixed, pre-committed repetition count per platform: 40–50 on Gemini, about 100 on Perplexity, 150+ on SearchGPT	±2.5-point interval requirements [1]
Cadence	Daily movement presented as performance	Weekly trend on the domain and core layers; daily series used only for anomaly detection	56–74% weekly turnover [2]; roughly three quarters of cited pages changing daily [5]
Change test	A two-point difference is reported as up or down	A change is reported only when it falls outside the confidence interval for that sample	Noise floor renders many comparisons untestable [1]
Cross-engine	Engine scores summed into one index	Per-engine numbers, compared on citation share only, never on raw citation counts	Six-fold citation volume spread [1]; 56% vs 74% turnover [2]
Question mix	All prompts averaged together	Brand prompts, category prompts and “best X” prompts reported separately	56–79% vs 13–23% first-source retention [5]

Two of these rules will feel like a downgrade, because they remove numbers a client has been reading for a year. Replacing a daily percentage with a weekly interval makes the report shorter and, at first glance, less informative. It is more informative in the only sense that matters: it can be defended. A figure with no stated sample size cannot be challenged, and a figure that cannot be challenged is not evidence.

> “If a visibility report cannot answer ‘what was your sample size and window’, the number it contains is not a measurement — it is an impression with two decimal places.”
>
> — Inference from the sampling thresholds in [1], stated here as a reporting rule rather than a finding

The standard also has a competitive use. In a category where every vendor publishes a different number for the same brand, the vendor that states its sample size, its window and its interval is the one whose number can be reconciled with the others. That is not a marketing position; it is the only position from which a measurement can be compared at all.

06 LIMITATIONS

Limitations

This paper's weaknesses are specific and worth stating in full, because a synthesis inherits every weakness of its sources and adds a few of its own.

No original measurement

Novaxilabs operates no monitoring sample. Nothing here was measured by the publisher, and no figure is presented as independent confirmation of another. The paper's contribution is structural: it reads six studies as one system and shows that their apparent contradictions are differences of unit, window and question type.

Most of the evidence is vendor-reported

Five of the six studies are published by companies that sell citation tracking. Their methodology sections are unusually detailed, and where they state limitations this paper repeats them — but none has been audited by a third party, and each has an obvious interest in the phenomenon being real. Only the arXiv paper [1] is academic and disinterested. A reader who discounts all vendor-reported numbers is left with the methodological core of this paper intact: the two-tier structure and the sampling thresholds both rest primarily on [1] and on the internal consistency of [2]–[6].

Narrow coverage, in every dimension

The multi-country data covers six Western markets and nothing else [2]. The sampling-threshold work covers three consumer product topics [1]. The longest window is ten months [4]; the shortest is seven days [6]. None of these can be generalised to a category, a language or a market without checking, and this paper does not attempt to.

The mechanism is not explained

This paper describes behaviour, not causes. It can say that domain-level citation is stable while page-level citation is not; it cannot say why, because none of the studies isolates the variable. Retrieval re-ranking, index refresh cycles, answer-length policy and query rewriting all plausibly

contribute, and separating them would require controlled perturbation that no published study has attempted. Any causal claim attached to these numbers — including the flattering ones — is inference, not measurement.

One study would close most of the gaps

The most conspicuous absence in the field is a single study that combines four properties at once: multiple languages, multiple engines, a window of at least six months, and enough repetitions per query to attach an interval. Every existing dataset holds three of the four. That study does not exist in the public record as of this paper's data cutoff, which is why this paper's claims are structural rather than predictive.

Why eight references, not twenty

The first paper in this series carried twenty sources because its question — which page structures survive citation — has a large literature behind it. This question does not. Every public dataset that measures citation stability at scale belongs to one of five tracking companies, and only one academic paper treats the statistics directly. Padding the list with secondary commentary would have added length and removed traceability, so the list stops at the eight pages that actually carry the figures used here. That is a judgement about what can be checked, not a claim that the literature is thin.

How to use this paper

As a reading key, not a verdict. Applied to a vendor report, it asks three questions in order: which layer is this number measuring, what is its sample size, and what is its window. Most reports fail at the first question. That failure is usually not dishonest — it is the default output of a tool built to sample once and display a percentage — but it is the reason two credible vendors can report incompatible numbers for the same brand in the same month without either of them lying.

Readers who want the primary sources should go to them directly. This paper is a map of where the evidence agrees and where it conflicts; the conflicts are more useful than any summary of them.

A

REFERENCES

Sources & References

Every source below was opened and read at its primary page before any figure was quoted from it. No number in this paper is taken from a secondary compilation. Where a widely circulated figure could not be traced to a published original — the 23-million-citation claim discussed in §02 — it was omitted rather than attributed loosely.

1 Sielinski, R. (2026). *Quantifying Uncertainty in AI Visibility: A Statistical Framework for Generative Search Measurement.* arXiv:2603.08924v3 [stat.AP], 39 pages, 13 figures. First version 9 March 2026; revised 26 August 2026. arxiv.org/abs/2603.08924 **ARXIV PREPRINT**

2 SISTRIX. (2026, 1 May). *AI Citation Drift: How stable are sources in AI search results?* 82,619 qualified prompts, 1,548,213 snapshots, 2,556 classified URLs, six countries (DE, US, UK, IT, ES, FR), three platforms, 17 weeks (17 December 2025 – 8 April 2026). sistrix.com/blog/ai-citation-drift **VENDOR-REPORTED**

3 BrightEdge. (2026, 6 February). *AI Search Citations: How Much Do They Really Change Week to Week?* Consecutive weekly snapshots, sample week of 1 February 2026. help.brightedge.com — weekly AI search insights **VENDOR-REPORTED**

4 Trakkr Research. (2026, 30 March). *The Half-Life of AI Citations* (Study 009). 108,650 URLs across 200 brands sampled from 960, eight models, 177-day window; plus a brand-level series of 857,138 daily reports covering 10,991 brands. trakkr.ai/trakkr-research/citation-decay **VENDOR-REPORTED**

5 Ahrefs. (2026, 29 October). *How Stable Are AI Citations? 10 Things We Learned Tracking 37,500 Citations.* Five AI tools, 20 days of daily tracking. ahrefs.com/blog/how-stable-are-ai-citations **VENDOR-REPORTED**

6 GetMentions AI. (2026, 9 July). *AI Citation Volatility: A 530,875-Citation Study.* 2,398 real queries from 56 brand accounts, four engines, seven consecutive days, June 2026; 67,144 answers, 181,225 URLs, 55,387 domains, 46,259 day-pair comparisons. Measured at domain level. getmentions.ai/blog/ai-citation-volatility-study **VENDOR-REPORTED**

7 Trakkr. (2026). *Trakkr Data* — continuously updated citation datasets, current as of 28 October 2026. Used here only for the one-answer-only share of cited links and the daily page-view series. trakkr.ai/data **VENDOR-REPORTED**

8 Trakkr. (2026). *The Complete, Data-Backed Guide to Earning AI Citations.* 1.3 million citations across 60,209 domains, including domain-concentration figures. trakkr.ai/guides/how-to-get-cited-by-ai **VENDOR-REPORTED**

Verification note. All eight links returned HTTP 200 when checked on 9 November 2026, and each was read at its original publication rather than through a press summary. Where a secondary compilation restated a primary figure differently, the primary figure was used: Trakkr's citation half-life discussion is widely restated as 73.4%, and the published study says 73.5%. Where two studies measure the same concept with different units — for example 96.8% domain stability against 73.5% URL singularity — both are reported side by side and neither is averaged into the other.

Evidence classes used in this paper. Official — a published standard or regulator document. **Academic preprint** — an arXiv posting with disclosed methods and data, not yet formally peer-reviewed, here only [1]. **Independent measurement** — a third-party measurement house publishing its own data. **Vendor-reported** — a company selling citation tracking publishing results about the phenomenon it sells, which is five of the six studies here. **Analysis** — this paper's own inference, labelled where it appears. No study in this field has been audited by a disinterested third party, and this paper does not claim otherwise.

B

AUTHOR

About the Author



Dr. Adrian Whitmore

Research Lead, Novaxilabs

Adrian leads measurement research at Novaxilabs, where he works on the statistical foundations of AI citation tracking: how many samples a visibility claim needs, what a stable metric looks like when the underlying system is stochastic, and where the field's published numbers disagree with each other.

This is the second paper in the series. The first, *Beyond Rankings — What Actually Gets a Page Cited in AI Answers*, examined which structural properties of a page survive selection by an answer engine. This one steps back and asks a prior question: whether the numbers used to report that selection are precise enough to mean anything. Both papers work exclusively from published third-party measurement; Novaxilabs operates no monitoring sample of its own, and the company's own platform data is not used as evidence here. Correspondence on methodology, including disagreements with the readings above, is welcome at info@novaxilabs.com.

C

PUBLISHER

About Novaxilabs

Novaxilabs measures how brands appear in AI-generated answers — across engines, across languages, over time. The platform tracks citation presence at the domain and page layers separately, reports intervals rather than point estimates, and keeps brand prompts and category prompts in separate series, for the reasons set out in §03 and §05 of this paper.

Layer-aware. Domain stability and page turnover are tracked as two different metrics, because they move on two different clocks.

Sample-honest. Every figure carries its repetition count and window. Where a claim cannot be supported at the sampling rate in use, the platform says so instead of printing a percentage.

Cross-market. The same tracked prompt set is run in eight languages across eight regional sites, so a shift in one market is visible against the others rather than averaged into them.

See your baseline AI visibility.

Run a tracked set of prompts against the engines that matter in your market, at a sampling rate that produces an interval you can defend in a review.

Start with a free scan | **How measurement works**

One email, the PDF, no newsletter. Privacy

Novaxilabs — AI visibility measurement, reported at a resolution that survives scrutiny.

Home · Resources · Pricing · Services · Paper No. 01

© 2026 Novaxilabs. Citation permitted with attribution. All third-party figures remain the property of their publishers; this paper's contribution is the reading of them as one system.