STATE OF THE ART DIGITAL

Intelligence Briefing

Why Brands Disappear From AI Answers

Why an AI visibility score is not a ranking, and what it means for competing in the Chaff Effect

By Steve Coulter, State of the Art Digital

12 July 2026

> *A single AI answer is a snapshot, not a scoreboard. Ask the same question a different way and the picture can change completely. Most businesses are being sold the snapshot as if it were the final score.*

1. Executive summary

New cross-provider research gives a rigorous basis for something that has been true of AI visibility measurement since it began: a single AI visibility score is not a measurement, it is one draw from a system that behaves very differently under slightly different questions. Reword a query and roughly three quarters of the recommended brand set typically changes. Add a location, price point or use case and the change rises further.

The more useful finding sits underneath that instability. When two unrelated AI providers both fail to recommend the same brand, they fail for the same underlying reason in the overwhelming majority of cases. The chaos on the surface sits on top of a small number of stable, diagnosable failure patterns: discoverability, compellingness and positioning.

Those three failure modes only make full sense against two pieces of context this briefing sets out in turn. The first is the Chaff Effect, the residual pool of harder, later-stage enquiries left for most businesses to compete over once AI Summaries have absorbed and resolved the easy, well-established informational queries elsewhere. The second is the distinction between Tier 1 foundation architecture and Tier 2 intent-driven content, which determines which specific failure a brand actually has, and therefore which fix is worth paying for.

2. The point-estimate problem

Controlled testing on AI product recommendations found that a cosmetic rewording of a query, for example changing 'best running shoes for long distance' to 'top long-distance running shoes', produced a brand set with only around a quarter overlap against the original. Adding a constraint such as a price ceiling or location cut the overlap further, to roughly one seventh.

Put plainly: submitting one query once and reading the answer as a ranking is not a measurement. It is a single sample from a distribution whose shape is unknown. Monthly 'AI share of voice' scores built on that basis cannot distinguish a brand that appears reliably from one that appears by chance under one specific phrasing and then vanishes.

The instinctive fix, running the same query repeatedly and averaging the result, does not solve this. It reduces the appearance of volatility without addressing what causes it, and in some cases actively conceals a fixed, repeatable failure behind a reassuring average, a point covered in Section 4.

3. Structure beneath the noise: three failure modes

The more important finding is not how much AI providers disagree with each other on brand recommendations, which is a great deal, but that when they both omit the same brand, the omission can be sorted into matching failure categories in the large majority of cases. That convergence, on independently operated systems, points to shared structure rather than random noise.

Failure mode	What it looks like	What does NOT fix it	What does
Discoverability	Absent from half or more of relevant runs. The model cannot reliably find or identify the brand.	Thought leadership and authority content published before the entity is clearly defined.	Entity clarity first: consistent naming, structured schema, unambiguous category signals.
Compellingness	Retrieved almost every time, but converts to a final recommendation only a fraction of the time.	More content volume. The model already knows the brand exists and is choosing not to recommend it.	Sharper framing and differentiation at the point of selection, not the point of retrieval.
Positioning	Recommendation set swings dramatically depending on who the model imagines is asking.	Generic schema and entity clean-up. The entity is fine; the audience match is not.	Persona-anchored content and clearer audience anchoring across the site.

4. Why averaging does not work

Grounding a model in retrieved facts, or running it several times and checking for consistency, is the standard advice for improving reliability. Recent research on retrieval-augmented systems found that a meaningful share of wrong answers do not vary at all across repeated runs. The system gives the identical wrong answer every time it is asked.

Standard consistency checks are blind to this. A stably wrong answer looks exactly like a stable, correct one. This matters directly for ongoing citation tracking: a score moving from 31% to 34% may represent genuine progress, or it may represent the same fixed exclusion sitting inside ordinary sampling noise. Averaging cannot tell the two apart. Only deliberate variation across phrasing, persona and model exposes it, because a genuinely fixed exclusion shows up as consistent absence across every condition tested.

5. How this differs by brand size

A companion audit of 37,000 recommendation runs found the instability is not evenly spread. Category leaders are retrieved almost every time but convert that retrieval into an actual recommendation in only 25 to 41% of cases, a pure compellingness problem. Specialist and regional players are absent entirely from roughly half of all

runs, a discoverability problem. Mid-market brands are the most exposed to persona effects, with their recommendation set changing by up to 75% depending on the imagined buyer, while category leaders barely move under the same test.

This has a direct read-across for any sector built on local or regional trust, automotive retail and estate agency among them. A national dealer group and a small independent agency are not fighting the same battle, even when both operate in the same category. The larger brand's problem is more likely to be compellingness. The smaller one's is more likely to be discoverability. A single headline score across both would obscure exactly the distinction that determines which fix is worth paying for.

6. The Chaff Effect: competing for what is left behind

None of the above happens in a vacuum. AI Summaries and Overviews already resolve most straightforward, well-established informational queries directly, citing a small and fairly consistent set of sources to do so. The searcher gets an answer and never clicks through. A handful of cited sites capture that mention. Everyone else was never in contention for it.

This is the subject of our earlier report, The Chaff Effect (formerly published as The Shrinking, Souring Pool), which found a zero-click rate of 74% on informational queries against 31% on transactional ones. The gap between those two figures is the point: the easy, generic, top-of-funnel questions are the ones AI answers directly and completely, while harder, later-stage, more specific enquiries are far more likely to still require a genuine visit to a website.

What is left for most businesses to compete over, then, is not the full spread of search demand. In practice this is a residual pool of enquiries too specific, too niche, or too embedded in a particular buyer's circumstances for a generic AI answer to resolve outright, though no methodology currently exists that can reliably sort individual searches into 'already resolved elsewhere' and 'still contestable'. What can be said with more confidence is the general shape of the shift: the demand still worth winning is, on the whole, later-stage and harder-won than it used to be. That is a reason for realism about what any AI visibility audit is actually measuring, and a caution against any claim that the residual pool itself can be precisely defined or targeted.

7. Tier 1 and Tier 2: locating exactly where a brand fails

The Chaff Effect explains why the competitive field has narrowed. It does not explain why a specific brand is failing to be cited within it. That is a separate, structural question, addressed by a second framework: the distinction between Tier 1 AI-optimised website foundation architecture and Tier 2 intent-driven, citation-optimised sitemap and content.

Framework layer	What it covers	Corresponding failure mode
Tier 1: AI-optimised website foundation architecture	Crawlability, structured data, entity clarity and machine-readable technical signals. The base layer a model needs to find and correctly identify a brand at all.	Discoverability
Tier 2: intent-driven, citation-optimised sitemap and content layer	Built on top of Tier 1. Aligns site structure and content to actual query intent, so a brand that can already be found is understood, differentiated and chosen.	Compellingness and positioning

This distinction gives the three failure modes from Section 3 a physical home. Discoverability sits at Tier 1: a brand the model cannot reliably find or parse has a foundation problem, and no amount of Tier 2 content work will fix it. Compellingness and positioning sit at Tier 2: a brand with sound foundations that still is not chosen, or is chosen inconsistently, has a framing and intent-matching problem sitting above an already-adequate base.

Our earlier report, Search Doomsday Is Q3 2027, set out the market-level version of this same logic: as Tier 1 technical proficiency closes across enough of the competitor population, the competitive battleground relocates

almost entirely to Tier 2. Section 5's finding, that category leaders are overwhelmingly retrieved but not chosen, while smaller and regional players are simply absent, is an early, visible instance of exactly that shift. The leaders have already closed Tier 1 and are now fighting purely on Tier 2 terms. Much of the rest of the field has not yet closed Tier 1 at all.

8. Evidence strength and how to hold it

The underlying research is largely the work of one team, published as unreviewed preprints through May and June 2026, and has not yet been independently replicated. The precise instability figures are sensitive to which similarity metric is used and should be treated as directional rather than exact. A vendor-sponsored counter-study claiming prompt wording barely matters relies on averaged data across tens of thousands of responses, which is the same point-estimate approach this briefing argues against, and its own secondary findings, including keyword-style prompts lifting visibility by up to 25% and ranking-format prompts surfacing around 20% more brands, support the brittleness case more than they undermine it.

The shared-failure finding, that independent providers converge on the same failure category when they both omit a brand, does not depend on the disputed similarity metrics and is the more defensible half of the argument. Treat the exact percentages as illustrative and the three-mode framework as the durable takeaway.

9. What this means for brand owners

Most of the AI visibility industry is still selling a number. A dashboard, a monthly percentage, a line that moves up or down. The research above says plainly that the number was never the thing worth buying. A brand absent from AI answers for a Tier 1 reason needs a completely different response to a brand absent for a Tier 2 reason, and no single score can tell the two apart.

This is the case for treating AI visibility work as a diagnosis rather than a scoreboard, and for holding realistic expectations about the shrinking, harder-won pool described in Section 6, even though that pool cannot currently be isolated or measured directly. It is also worth being honest about the state of the discipline itself: citation analysis is new, the systems being measured change faster than any methodology can be finalised, and nobody, this firm included, would claim to have it fully solved. What follows is the current, tested toolkit for engaging with that problem properly, not a claim that the problem is closed.

The OPTIMUM 7-Dimension AI Summary and GEO Audit is the diagnostic starting point. Rather than reporting a single visibility figure, it is built to separate discoverability, compellingness and positioning across both Tier 1 and Tier 2, so the output is a specific, actionable reason for a gap rather than a percentage with no explanation behind it.

The OPTIMUM AI Citation and Competitor Gap Analysis examines who is actually being cited for a brand's core query set across AI answers, and identifies the specific gap between that brand and the sources currently capturing those citations.

The OPTIMUM 2026 Minimum AI SEO / GEO Website Technical Standard is the published yardstick for Tier 1. It sets out, in concrete and testable terms, what foundation architecture actually needs to include for a brand to be reliably discoverable at all, ahead of any Tier 2 work being worthwhile.

OPTIMUM AI-Optimised, Intent-Driven, AI Summary Citation-Optimised Sitemap Development is the Tier 2 counterpart to that standard. Where the Technical Standard defines the foundation a brand needs to be found, this maps and restructures the sitemap and content layer above it so a brand that can already be found is actually understood, differentiated and cited. It has been delivered for a four-branch estate agency group, where a premium brand and a distinct rental operation needed to be woven into a single, coherent sitemap architecture rather than treated as separate sites competing for the same citations.

The OPTIMUM Citation Gap Methodology is the analytical method underpinning the above. It applies distribution-based testing across phrasing, persona and model, rather than a single-query point estimate, directly addressing the measurement flaw this briefing has argued against throughout.

Taken together, these five give a brand a way to find out which specific failure it has, at which layer, and where the realistic opportunity for citation actually lies. That is a materially different, and more useful, starting point than another monthly percentage.

10. Recommended action sequence

Measure the distribution first, diagnose the failure mode second, apply the matching fix third. The current market convention runs this in reverse: apply a fix, watch whether the score moves, and assume causation. That approach cannot separate a fix that worked from one that simply coincided with normal run-to-run variance.

A workable minimum: one core query intent, run across five to ten phrasings, two to three buyer personas, and at least two AI models with live web search enabled. That is enough spread to see whether a brand appears consistently or sporadically, whether persona conditioning reshuffles the result, and whether the pattern holds across models. The shape of that spread is the diagnosis.

Related reading

Your Business Is Becoming Invisible: the founding report in this series, setting out the general scale of the zero-click and AI-answer threat to owned website traffic.

The Chaff Effect (formerly The Shrinking, Souring Pool): the source of the 74% and 31% zero-click figures referenced in Section 6, and the fuller case for why AI answers are absorbing a growing share of straightforward search demand.

Search Doomsday Is Q3 2027: the source of the market-level Tier 1 and Tier 2 thesis referenced in Section 7. All three are available free at stevecoulter.co.uk.

Sources

Cheung, D. (2026). Why Does My Brand Disappear From AI Answers? danielkcheung.com, 12 July 2026.

Underlying research: Jack, Lehman, Maloney and Xu (2026), multiple papers, arXiv, May to June 2026; Julka, S. (2026), arXiv; Sielinski, R. (2026), arXiv. Full citation list available in the source article.