

MEASUREMENT · SHARE OF VOICE · AI VISIBILITY · BRAND MESSAGING

Why AI share of voice is the wrong metric, and what to measure instead

Share of voice counts presence but can't distinguish narrow dominance from broad visibility. In AI search, that distinction is everything. A better metric measures citation consistency across a wide query set.

AIVIARA RESEARCH · MAY 2026

Share of voice has a straightforward origin. In advertising, it measured the proportion of category media spend controlled by a brand. In social media, it counted the proportion of brand-relevant conversations. In SEO, it tracked the share of organic keyword impressions. The formula is the same across all contexts: brand metric divided by total market metric, multiplied by 100. Sprout Social describes it as "the measure of the market conversation your brand owns compared to your competitors." HubSpot calls it the metric that "allows you to compare brand awareness on different marketing channels against your competitors."

For B2B SaaS vendors, where buyers increasingly turn to ChatGPT or Perplexity before visiting any vendor website, AI visibility has become a live commercial concern, and share of voice has become the default way to measure it. When commercial AI visibility tools emerged, they applied this same logic to AI-generated responses. Semrush defines its AI share of voice as "the percentage of mentions your brand receives in AI-generated answers compared to competitors in your market." Profound defines it as "Your citations / Total citations across competitors." Tools including Peec and Otterly use functionally similar approaches. As Discovered Labs noted in a review of these platforms, none of the three explicitly detail their specific share of voice methodologies, but the underlying logic is consistent: count presence, divide by the total, express as a percentage.

This is not an unreasonable starting point. The problem is that it is only a starting point, and the AI search environment exposes its limits in ways the advertising and SEO contexts did not. Share of voice counts presence. In AI search, presence is not the same thing as consistent, broad-based representation, and the metric does not register the gap.

A cybersecurity vendor might appear in nearly every AI response to "what is zero trust architecture" while being absent from the broader enterprise security queries its buyers actually run during shortlisting. A share of voice score covering both scenarios treats them identically. The metric does not know the difference.

What share of voice measures, and what it misses

The foundational critique of share of voice is not new. Sprout Social states it plainly: "While this formula provides a basic understanding of SOV, it doesn't capture everything. Qualitative factors, such as the impact and relevance of mentions, also play a crucial role." In advertising this limitation was manageable. Presence in a media buy is relatively uniform. A brand appearing in a 30-second television slot is, broadly speaking, present in a comparable way across placements. The counting shortcut was defensible because the underlying phenomenon was relatively stable.

AI-generated responses are not like television slots. The Rank Masters identifies several compounding problems specific to AI search: non-deterministic outputs (the same prompt produces different responses across runs), prompt universe bias (the score depends entirely on which prompts are being tracked), nuance loss (a prominent recommendation in the opening sentence and a passing mention buried in a caveat are treated identically under a binary presence count), and attribution difficulty (the inability to trace which AI interaction influenced a given buyer decision). Each of these would individually degrade the metric. Together, they produce a measurement that is unstable, prompt-dependent, and blind to the character of the mentions it is counting.

There is also a structural volatility problem that binary presence counts do not surface. The non-deterministic nature of AI output means citation presence is inherently unstable as a measurement baseline. The Rank Masters notes that the same prompt can produce different outputs across runs. A brand's share of voice score can oscillate substantially from one measurement period to the next in ways that reflect platform behaviour rather than anything the brand has done. A metric this sensitive to conditions outside the brand's control is telling you something about the AI search environment; it is not telling you much about your brand's actual standing within it.

Quattr, writing about SOV in SEO contexts, identifies data fragmentation, missing contextual factors, and misalignment with actual traffic as structural limitations of the approach. These concerns apply with at least as much force to AI search, where the gap between raw mention count and anything resembling outcome relevance is wider than in the keyword-based world SOV was originally adapted from.

This is not an argument against measurement

The critique above is sometimes read as a case for abandoning quantitative tracking altogether. That is not the argument here. The problem is not that teams are trying to measure AI visibility. The problem is that the metric they are using rewards a kind of performance that does not map onto what consistent AI visibility actually looks like.

A high share of voice score can be produced by a brand that dominates on a narrow cluster of prompts while being entirely absent from a broad range of relevant queries. It can be produced by a brand whose mentions are uniformly peripheral (present in the response, technically, but not as a recommendation). It can spike and collapse with platform-level citation pattern changes that have nothing to do with the brand's actual footprint. What all of these scenarios have in common is that they satisfy the metric while failing to describe a brand that is well-represented across the AI search environment.

The measurement goal is not zero. It is the right question. And the right question for AI visibility is not "what percentage of tracked citations does my brand own?" It is something closer to: "Across a broad and diverse set of relevant queries, how consistently is my brand cited by AI engines?" That is a question about breadth and consistency, not about the ratio of one brand's presence to another's. Share of voice does not distinguish between the two.

What concentration looks like in practice

The extent to which citation performance concentrates on a narrow slice of prompts rather than distributing across a broad query set is visible in the citation data that has been published about the AI search environment.

Semrush conducted a study (Luke Harsel, November 2025) analysing 230,000 prompts over 13 weeks across ChatGPT, Google AI Mode, and Perplexity, covering more than 100 million total AI citations. The headline finding on concentration is striking: in early August 2025, ChatGPT cited Reddit in close to 60 percent of prompt responses, and Wikipedia in roughly 55 percent of AI prompt responses. Two domains, capturing the majority of citations across a 230,000-prompt corpus.

60%

of ChatGPT prompt responses cited Reddit in August 2025. Wikipedia appeared in roughly 55%. Two domains capturing the majority of citations before brand-level competition begins.

SEMRUSH · NOVEMBER 2025 · 230,000 PROMPTS · 100M+ CITATIONS

A brand tracking its own share of voice within a category is operating against a backdrop where citation mass is concentrated at this scale in a small number of platforms before brand-level competition begins.

The cross-platform picture is worse. The Digital Bloom's 2025 AI citation report found that only 11 percent of domains are cited by both ChatGPT and Perplexity. A brand achieving a high share of voice within one platform's citation pool may have essentially no presence in another. Share of voice scores computed within a single platform present this as a ceiling; they do not surface the cross-platform coverage gap.

Only **11%** of domains are cited by both ChatGPT and Perplexity.

THE DIGITAL BLOOM, 2025 AI CITATION REPORT

What these findings illustrate is that the distribution of AI citations is not a relatively flat market with many participants dividing the space, where percentage share is a useful coordinate. It is a highly concentrated environment where breadth of coverage (being cited across many relevant prompts and multiple platforms) is a qualitatively different kind of achievement from dominating a narrow slice. A metric that divides one brand's count by the total count within a tracked set does not capture that distinction.

A different framing

In bibliometrics, researchers faced an analogous problem decades before AI search existed. Raw citation counts were the standard measure of academic impact, and they had a well-understood flaw: a single highly-cited paper could inflate a researcher's total indefinitely, making it impossible to distinguish between narrow dominance on one paper and sustained, broad contribution across a body of work.

J.E. Hirsch proposed the h-index in 2005 (arXiv:physics/0508025, published in PNAS 102(46):16569–72) specifically to solve this. The h-index is defined as "the number of papers with citation number higher or equal to h." A researcher has an h-index of h if exactly h of their papers have each been cited at least h times. The metric's key property is that

narrow dominance cannot inflate it. A researcher with one paper cited 10,000 times and nine papers cited zero times has an h-index of 1. Sustained, broad-based contribution is what drives the number up.

Measurement theory had this problem first. A brand cited at high frequency on a narrow set of prompts, with minimal presence elsewhere, is to AI visibility what a single viral paper is to academic impact: impressive on the headline number, misleading as a characterisation of breadth.

The Generative Brand Mention Framework (GBMF), published by Aiviara Research, approaches this problem by making breadth intrinsic to the score. GBMF's core visibility measure, MV (brand-mention visibility), is computed as the mean mention rate across a declared prompt set of 30 to 50 queries, spanning categorical, comparative, use-case, and problem-solution queries. MV is computed per engine first, then averaged equally across engines to a headline figure. A brand that dominates a narrow cluster while absent from most of the prompt set cannot inflate MV by concentrating on a few high-performance prompts. The denominator is the full declared set, not just the prompts where the brand appears. Whether breadth-based visibility measures prove more predictive of commercial outcomes than share-of-voice approaches is an open question, but they address a structural property the current generation of tools mostly ignores.

What this means for how you track AI visibility

For B2B marketing teams tracking whether their brand appears during vendor evaluation queries, four specific risks follow from relying on share of voice as your primary AI visibility metric.

The first is concentration masking. A brand's share of voice score can be driven almost entirely by strong performance on a small number of prompts while its broader presence across the relevant query set is negligible. The percentage looks healthy; the distribution is not.

The second is platform inflation. If you are computing share of voice within a single platform, you are measuring a ratio that does not reflect cross-platform visibility. Given that only 11 percent of domains appear across both ChatGPT and Perplexity, a strong share of voice score on one platform may coexist with near-total absence on another. The metric cannot surface this.

The third is volatility conflation. AI citation patterns shift across runs, platforms, and time, driven by platform model updates rather than anything a brand has done. Share of voice

scores can swing substantially between measurement periods in ways that carry no strategic signal. Tracking a volatile ratio over time can create an impression of meaningful movement when what is actually happening is platform-level instability.

The fourth is methodological opacity. The commercial tools that publish AI share of voice scores do not disclose their formulas. Scores are not reproducible or comparable across platforms, which makes cross-tool validation impossible and any time series vulnerable to quiet methodological changes on the vendor's end.

None of these risks are hypothetical edge cases. They are predictable consequences of applying a ratio metric to an environment characterised by concentration, cross-platform fragmentation, high volatility, and proprietary score construction.

The better measurement question for AI visibility is whether a brand is consistently cited across a broad and diverse set of relevant queries, rather than what percentage of a tracked citation pool it controls. Share of voice will tell you something. It will not tell you that.

Aiviara is building infrastructure for monitoring AI brand citations and factual accuracy across LLM platforms. Early access information is available at aiviara.com.