Measuring Brand Visibility Across AI Answer Engines: A Multi-Model Empirical Study

[Authors Redacted for Review]
[Institution Redacted]
[City Redacted], [Country Redacted]

ABSTRACT

As AI answer engines replace traditional search for millions of queries, brands face a measurement gap: no predominant methodology exists to quantify whether or why a brand appears in AI-generated responses. We address this with a large-scale empirical study collecting 110,523 responses from four production AI systems—GPT-4.1, Claude Sonnet 4.5, Gemini 2.5 Flash, and Google AI Overviews—across 50 brands (plus 5 fictitious controls) spanning five verticals and three market tiers (56,803 baseline across two snapshots; 53,720 web-search-enabled across three snapshots). Brands and keywords were drawn from top Google organic rankings, creating a ceiling condition: every brand has already won traditional search, so gaps in AI visibility cannot be attributed to low organic authority.

Our central finding is that keyword-level variation is the strongest marginal predictor of brand mention (marginal $\eta^2 = 32.5\%$), followed by brand identity (11.2%) and intent type (5.2%); these marginal values overlap and should not be summed. Queries vary from category-defining ("best CRM software" for Salesforce) to tangential entity queries ("simparica" for Chewy), and this variation accounts for more than 100 times the variance explained by model choice (0.1%) or market tier (0.3%). A GLMM corroborates this ordering (variational Bayes [VB] convergence warning applies; see Limitations): keyword (ICC = 34.4%) and brand (ICC = 23.0%) random effects dominate, with 42.6% residual. Which AI system answers explains just 0.1%—systems converge on *whether* to mention a brand but diverge markedly in style (hedging: 2.8%–31.7%; word count: 222–675). Market tier explains 0.3%, providing little support for the assumption that larger brands enjoy higher AI visibility. Two distinct patterns of brand absence emerge: *absorption* (healthcare content synthesized without attribution; 94.9% of absent-brand responses in healthcare contain no commercial brand) and *displacement* (competitors fill default rosters in finance, technology, and e-commerce). Citation rates reveal a structural divide: AI Overviews cite brand URLs in 28.7% of responses; standalone LLMs aggregate below 1%. A web-search-enabled condition (n=53,720) confirms this is not a retrieval limitation: enabling web search raises GPT-4.1's overall URL rate to 31.0%—capturing citations to news outlets, review aggregators, and reference sites—but brand-specific citations remain below 3% and brand mention rates fall across all LLMs, indicating a citation preference gap rather than a retrieval gap. Data, scoring code, and prompt sets are released for replication.

KEYWORDS

AI answer engines, brand visibility, generative search, AEO, multi-model evaluation, cross-model agreement

1 INTRODUCTION

1.1 The Measurement Gap

When a consumer asks ChatGPT "what's the best CRM software?" the answer names specific brands—but which ones, and why? This question has no predominant methodology for answering it. The shift from ranked links to generated prose has created a measurement gap: the brands that built their digital presence on SEO now face an environment where their visibility is shaped by opaque model internals rather than indexable page properties alone.

15–25% of informational queries now trigger AI-generated answers [1]. ChatGPT processes 37.5M queries per day with 59% "fan-out" to other sources [2]. Zero-click searches exceed 65% on mobile [3]. Industry data [4] documents 60%+ CTR reduction for queries with AI Overviews in featured position.

Traditional SEO measurement—rank tracking, SERP features, click-through rates [5, 6]—cannot measure brand presence in generative text. A URL either appears in a ranked list or it does not; but a brand can be mentioned, recommended, compared, hedged against, or absorbed into a synthesis without attribution. Commercial tools (e.g., Profound, Semrush AI Overview tracking) have emerged but publish no scoring methodology, provide no confidence intervals, and are not independently replicable.[1] No academic framework exists for measuring, comparing, or explaining brand visibility across AI systems.

1.2 Research Gaps

GEO [7] introduced optimization but used simulated engines, single models, and temperature=0. CC-GSEO-Bench [8] advanced influence measurement but tested one model. Strauss et al. [9] documented the citation crisis but did not measure brand-level visibility. Characterizing Web Search [10] acknowledged non-determinism but called for repeated sampling rather than measuring it.

To our knowledge, no published academic work: (1) measures visibility across multiple production AI systems simultaneously; (2) validates measurement with fictitious brand controls; (3) quantifies how much variance in visibility is explained by each factor; (4) characterizes empirical patterns

[1]We omit Perplexity from our source set; future work should include RAG-based answer engines.

of brand absence—absorption and displacement—and their strategic implications.

1.3 Hypotheses

We formulated five hypotheses from the literature and tested each with a pre-specified methodology. Table 1 summarizes all hypotheses and their verdicts. This structure allows honest reporting of both confirmations and surprises. Measurement integrity is established separately via fictitious brand controls (§4.1).

1.4 Contributions

This paper makes the following contributions:

(1) **Marginal association analysis of visibility predictors**—we quantify how much variance in brand mention each factor explains independently; keyword-level variation dominates (marginal $\eta^2 = 32.5\%$) while model choice explains 0.1% and market tier 0.3%, providing little support for brand-size or model-choice assumptions. (Marginal values overlap due to hierarchical nesting; see §3.5.)

(2) **Two patterns of brand absence: absorption and displacement**—we identify and characterize two distinct mechanisms by which brands fail to appear: in healthcare (a YMYL domain), content is synthesized without attribution (94.9% of absent-brand responses contain no commercial brand); in competitive domains (finance, technology, and e-commerce), absent brands are displaced by named competitors.

(3) **Multi-model empirical measurement at scale**—110,523 responses (56,803 baseline across two snapshots; 53,720 web-search-enabled across three snapshots) collected from four production AI systems (GPT-4.1, Claude Sonnet 4.5, Gemini 2.5 Flash, Google AI Overviews) across 50 brands, 5 verticals, 3 market tiers, and 5 intent types, with scoring code and prompt sets released for replication.

(4) **Cross-model agreement analysis**—models converge on brand absence in 76.2% of prompts but on presence in only 9.4% (Fleiss' $\kappa = 0.647$, substantial agreement), demonstrating that multi-model measurement captures signal that single-model studies miss.

(5) **Construct validation via fictitious brand controls**—5 invented brands with no real-world presence establish a verified 0% measurement floor, confirming that mention rates reflect genuine model behavior rather than scoring artifacts.

(6) **Non-determinism characterization at temperature**=1.0—repeated identical prompts show ∼95% consistency across 4–5 repeats (Claude 94.9%, GPT-4.1 96.2%), though limited repeat counts constrain power to detect low-probability stochastic variation.

2 RELATED WORK

2.1 Generative Engine Optimization

GEO [7] introduced the field with 10K queries across 9 domains using a simulated generative engine. Two visibility metrics were proposed: Position-Adjusted Word Count (citation-centric) and Subjective Impression (LLM-evaluated). Statistics Addition and Quotation Addition improved visibility (+28–41%), while Keyword Stuffing performed *worse* than baseline. Key limitations: simulated engine, single model, no non-determinism treatment.

CC-GSEO-Bench [8] advanced the field with 1,000+ source articles and 5,000+ query-article pairs, measuring three influence dimensions: Exposure, Faithful Credit, and Causal Impact. Strategy effectiveness was shown to be context-dependent. Limitation: single model (gpt-4.1-mini).

E-GEO [15] extended GEO to e-commerce with 7K+ product queries, finding a stable, domain-agnostic optimization pattern. GEO-16 [16] proposed a 16-pillar auditing framework across 70 prompts with 1,702 citations. White Hat SEO [17] and RAID G-SEO [18] explored LLM-aware content optimization strategies, though both focused on single-model settings.

Table 2 summarizes key differences between our study and prior work.

2.2 The Attribution Crisis

Strauss et al. [9] analyzed ∼14,000 LMArena conversation logs, finding 34% of Gemini responses generated without fetching online content and 92% lacking citations. Citation efficiency ranged from 0.19 to 0.45 per relevant page visited. Our LLM citation rate (<1% for Claude, Gemini; 2.3% for GPT-4.1) confirms this for standalone models, while AIO's 28.7% citation rate reveals that search-grounded systems exhibit markedly higher citation rates.

Venkit et al. [14] identified 16 limitations in AI search engines through a 21-participant user study, including hallucination, misattribution, and overconfident language. Citation Alignment [19] showed LLMs are 27% more likely than humans to cite Wikipedia-flagged text. Khalifa et al. [12] demonstrated that attribution behavior is trainable and architecturally dependent.

2.3 Adversarial Manipulation

Pfrommer et al. [11] (EMNLP '24) demonstrated that different LLMs vary significantly in weighting product name vs. document content vs. context position—direct support for model-specificity in cross-model measurement.

2.4 Behavioral Testing of NLP Models

Ribeiro et al. [20] introduced CheckList, a task-agnostic behavioral testing methodology that decomposes model evaluation into capability-specific test types (Minimum Functionality, Invariance, Directional Expectation). Our prompt design adapts this framework: intent-type variations test directional expectations (recommendation prompts

Table 1: Hypotheses with literature basis and empirical verdicts.

ID	Hypothesis	Literature Basis	Verdict
H1	Brand mention rates differ significantly across AI sources	Pfrommer et al. [11]; Khalifa et al. [12]	**Not supported (practically)**—source explains 0.1% of variance; rates converge (16.0%–19.3%) despite divergent behavioral profiles
H2	URL citation rates are uniformly low across all sources	Strauss et al. [9]: 92% of Gemini answers lack citations	**Partial**—LLM aggregate 0.81% (GPT-4.1 2.3%; Claude 0.08%; Gemini 0.10%); AIO 28.7%; web-search LLMs still <3% for brand URLs
H3	Brand tier (enterprise > midmarket > startup) predicts visibility	PageRank logic [13]; domain authority [5]	**Not supported (practically)**—$V = 0.067$ (negligible); continuous Ahrefs DR $\rho = +0.34$ ($p = 0.017$) overall, but within-tier DR is non-significant (enterprise $\rho = +0.28$, ns; midmarket $\rho = +0.06$, ns) due to restricted DR range; startup-tier DR strongly predictive ($\rho = +0.82$, $p < 0.001$, $n = 16$)
H4	Regulated verticals show systematically lower visibility	Venkit et al. [14]; YMYL content sensitivity literature	**Directional support**—Healthcare 12.1% vs. Finance 24.0% (2.0×), $V = 0.120$ (small); the qualitative absorption pattern (§4.5) is the primary finding
H5	Intent type significantly affects brand mention rates	GEO [7]; CC-GSEO-Bench [8] context-dependency	**Confirmed**—4× gap; $V = 0.231$ (largest V)

Table 2: Comparison with prior generative search studies.

Dimension	GEO [7]	CC-GSEO [8]	E-GEO [15]	GEO-16 [16]	This study
Systems	Simulated engine	1 model	1 model	3 (citation only)	4 production models
Primary signal	Citations	Influence score	Visibility rank	Citation quality	Brand mentions
Non-determinism	t=0	N/A	N/A	N/A	Repeated sampling at t=1
Construct validation	None	None	None	None	Fictitious brands
Cross-model comparison	No	No	No	Partial	Full (3 LLMs + AIO)
Scale	10K queries	5K queries	7K queries	70 prompts	110,523 responses‡

‡ 56,803 baseline (2 snapshots) + 53,720 web-search-enabled (3 snapshots).

should surface more brands than informational ones), control repeats test minimum functionality, and paraphrase sets test invariance. Shin et al. [21] showed that prompt phrasing significantly affects knowledge elicitation, motivating our 10-variation-per-keyword design to capture phrasing sensitivity.

2.5 Traditional Search Measurement

Feuerriegel et al. [5] established the SEO measurement gold standard with 67,000 keywords and 6M clicks. Huszár et al. [22] introduced the "performativity gap" concept with bootstrap confidence intervals. Rise of AI Search [23] analyzed 2.8M search results, finding lower response variety and citation concentration in AI search. PageRank [13] and Learning to Rank [24] established SEO foundations; no equivalent exists for LLMs.

3 METHODOLOGY

3.1 Brand Selection and Stratification

We selected 50 brands across five verticals using five criteria: (1) consumer searchability, (2) content presence, (3) tier stratification defensible by market position, (4) within-vertical competition, and (5) cross-vertical diversity. Table 3 shows the full corpus.

SEO-informed selection. Brands were selected from those ranking in the top positions of Google organic search results, using Ahrefs data to identify brands with strong organic visibility across relevant category keywords. This design is intentional: by studying brands that have already "won" in traditional search ranking, we ask a ceiling question—if a brand dominates Google organic results, does that guarantee AI visibility? The answer is a headline finding: it does not.

Table 3: Brand corpus: 50 brands + 5 fictitious controls across 5 verticals and 3 tiers.

Vertical	Enterprise (4)	Midmarket (3)	Startup (3)	Fictitious
Finance	Fidelity, Schwab, Vanguard, PayPal	SoFi, Robinhood, Wealthfront	Mercury, Ramp, Chime	Wynthoral Fin.
Law	LegalZoom, Avvo, LegalShield, Nolo	Rocket Lawyer, ZenBusiness, NW Reg. Agent	Ironclad, Trust & Will, Hello Divorce	Grelvant Law
Healthcare	WebMD, Mayo Clinic, UnitedHealthcare, CVS Health	Zocdoc, GoodRx†	Hims & Hers, K Health, Ro, Cerebral†	Plorantic Health
Technology	Salesforce, HubSpot, Microsoft 365, ServiceNow	Monday.com, Asana, Freshworks	Notion, Linear, Rippling	Jorvelle
E-Commerce	Amazon, Shopify, Walmart, eBay	BigCommerce, Chewy, Wayfair	Bolt, StockX, ThredUp	Velnith Market

†Healthcare has 2 midmarket and 4 startup brands due to the vertical's market structure; all other verticals follow the 4/3/3 split.

Five fictitious brands (one per vertical) were verified non-existent via web search on 2026-02-27. They share keyword sets with real counterparts; any detection constitutes a scoring error.

3.2 Keyword and Prompt Design

Pipeline. Human-curated keywords → LLM-generated prompts (GPT-4o-mini) → human review → approved prompts.

Scale. 750 keywords, 10,508 prompts (7,508 variations + 3,000 controls for real brands; 3,060 total controls including fictitious). Each keyword receives up to 10 prompt variations spanning 5 intent types plus 4 control repeats of an identical seed prompt (14 per keyword). This design follows CheckList's [20] behavioral testing taxonomy: intent variations test directional expectations, controls test minimum functionality, and paraphrase sets test invariance.

Critical constraint. Prompts never mention the target brand, validated automatically via substring matching against brand names and aliases.

Dual-format strategy. Conversational prompts for API models; search-native variants for AIO, reflecting documented query-length differences (Google ∼3.4 words vs. ChatGPT ∼23 words [25]). A format-consistent subset (search-format prompts only, excluding conversational variants) is available in supplementary materials as a robustness check.

Statistical note. The 10 variations per keyword are correlated paraphrases, not independent samples. Effective $N = 750$ keywords for headline statistics.

Keyword composition. Keywords were sourced from Ahrefs organic keyword data for each brand—the same queries where these brands rank highly in Google organic search. Every keyword-brand pair therefore represents a query where the brand has already demonstrated organic search authority, reinforcing the ceiling condition described in §3.1: any gap in AI visibility cannot be attributed to low organic ranking. The 82% average absence rate demonstrates that organic search authority alone does not guarantee AI visibility. Keywords include category queries ("best CRM software"), topic queries ("simparica"), and navigational queries ("pay in 4"). Only 3 of 750 contain the target brand name explicitly. The set is overwhelmingly unbranded.

3.3 Data Collection Architecture

Figure 1 shows the collection pipeline. API models (GPT-4.1, Claude Sonnet 4.5, Gemini 2.5 Flash) are queried via LiteLLM at temperature=1.0 using each provider's native API. Web search and retrieval tools are explicitly disabled for all three LLM sources during the first round of data collection to establish a baseline. These responses reflect only what is encoded in each model's training data, with no live web grounding. All three LLM sources use an identical system prompt ("Answer the user's question directly and helpfully.") to standardize instruction context. Google AI Overviews are collected with geolocation pinned to Chicago, IL, and are search-grounded by design—making AIO the only real-time, retrieval-augmented source in the baseline study. To test whether web access closes the citation gap, a web-search-enabled condition re-queries the same three LLMs with web search tools enabled (n=53,720 responses across three snapshots; §5.1).

We collected 56,803 baseline responses across two snapshots (snapshot 7: 2026-02-28; snapshot 8: 2026-03-07). Of these, 52,998 have non-null text and form the primary analysis dataset: 18,264 scored AIO responses, 11,817 from Claude Sonnet, 11,101 from GPT-4.1, and 11,816 from Gemini Flash. Results were consistent across both snapshots (overall mention rate: 17.9% vs. 17.3%), providing limited temporal validation of the baseline findings.

3.4 Scoring Approach

We use programmatic scoring across 8 signal categories (23 metrics total; see Appendix B): word-boundary regex for brand mention detection, domain matching for URL citations, list-position parsing for recommendation rank, phrase-list matching for hedging/disclaimer language, VADER for sentiment, and spaCy [26] NER for entity discovery. This captures surface signals reliably but does not capture semantic influence, prominence nuance, or context-aware sentiment.

Known limitations. (1) VADER misses domain-specific sentiment. (2) Phrase lists are hand-curated (Appendix B). (3) Sentiment is confounded with keyword topic valence. (4)

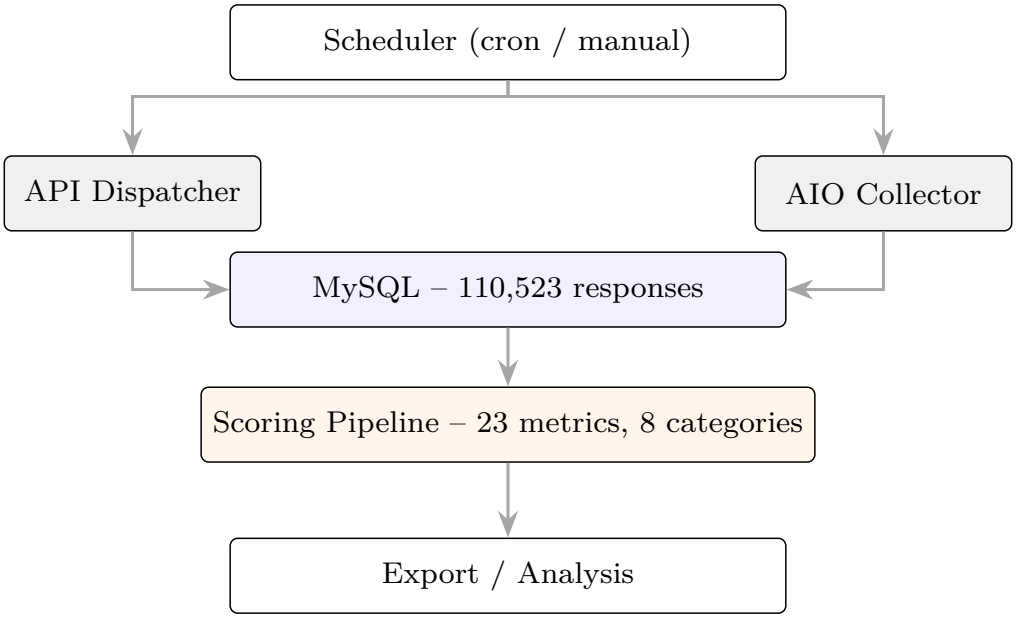

Figure 1: Data collection and scoring architecture. Three LLM APIs and Google AI Overviews feed into a shared response store scored with a 23-metric programmatic pipeline.

Brand mention detection uses canonical names and configured aliases; common shorthand forms (e.g., "Schwab" for Charles Schwab, "CVS" for CVS Health) are detected only where aliases are configured, potentially undercounting mentions for brands with well-known abbreviations. (5) Fictitious brand controls validate the false positive floor (0/356) but do not quantify false negatives or precision for ambiguous brand names (e.g., Linear, Bolt, Mercury, Ro are common English words). A human-labeled validation sample is identified as future work.

3.5 Statistical Approach

Two-level analysis. Level 1: keyword-level aggregation (N=750) with Wilson 95% CIs for proportions (Appendix C). Level 2: phrasing sensitivity (within-keyword SD of mention scores).

Control prompts. 4 repeats of identical seed prompts per keyword (3,060 control prompts) provide a stochastic variance baseline.

Marginal association analysis. Marginal η^2 (one-way ANOVA per factor) quantifies how much variance in `brand_mentioned` each factor explains independently. Because factors are hierarchically nested—keyword within brand, brand within vertical, tier as a property of brand—these are *marginal* contributions that overlap, not an additive partition. We report them for descriptive ranking, not as unique variance claims.

Statistical tests. Chi-square with Cramér's V for categorical comparisons; Fisher's exact where expected cells < 5; Bonferroni correction for multiple comparisons; Cohen's h for pairwise proportion effects.

3.6 Limitations (Upfront)

- **Temporal scope:** All data were collected within an approximately three-week window (2026-02-28 to 2026-03-16). Consistency across snapshots strengthens internal validity, but the study does not constitute long-term longitudinal tracking; changes in model behavior, brand content, or AIO ranking logic over months are not captured.
- **Single sample per prompt variation.** CIs from within-keyword aggregation.
- **Programmatic scoring only.** Primary signal is binary brand mention.
- **English-language only.**
- **Sentiment confound:** VADER + keyword topic valence.
- **Keyword composition:** Includes topic/entity queries with weak brand alignment (e.g., "simparica" for Chewy). This increases the proportion of zero-mention keywords but reflects real-world search behavior (see §4.4).
- **No causal claims.** Observational study throughout.

4 RESULTS

We begin with construct validation (§4.1) to establish measurement integrity, then the control-prompt consistency finding (§4.2) that informs a methodological question. The marginal association analysis (§4.3) provides the organizing framework: keyword (32.5%) > brand identity (11.2%) > intent type (5.2%) > vertical (1.7%) > tier (0.3%) ≈ source (0.1%); these marginal η^2 values overlap due to hierarchical nesting and should not be summed (§3.5). A GLMM corroborates

the qualitative ordering: keyword (ICC = 34.4%) and brand (ICC = 23.0%) random effects dominate, with 42.6% residual (VB fitting produced a convergence warning; estimates are directional and consistent with the marginal η^2 ordering—see §5.6). Subsequent sections unpack each factor.

Across 52,998 scored responses (variations and controls), the average brand mention rate is 17.6%. No mention is the norm: across ∼82% of keyword-brand-source observations—for keywords where these brands rank in top Google organic positions—AI systems do not mention the target brand.

4.1 Construct Validation

Before interpreting real brand results, we must establish that our scoring pipeline does not hallucinate mentions. Five fictitious brands—Wynthoral Financial, Grelvant Law, Plorantic Health, Jorvelle, and Velnith Market—were assigned the same keyword sets as their real vertical counterparts and scored identically. Across 356 scored responses (spanning all sources and intent types), the pipeline detected exactly 0 mentions. No false positives (Wilson 95% CI: [0%, 1.0%]). This validates both the measurement floor and the regex-based detection methodology: all non-zero scores reported below represent genuine brand presence in model outputs.

4.2 Brand Mention Shows High Consistency

A methodological question precedes all results: is brand mention stochastic? GEO [7] assumed yes and set temperature=0 to suppress it. Characterizing Web Search [10] recommended repeated sampling to quantify it. We answer empirically with control prompts—identical text repeated 4–5 times per keyword at temperature=1.0.

Brand mention is highly consistent: Claude (94.9%) and GPT-4.1 (96.2%) produce identical mention outcomes across all control repeats per keyword at temperature=1.0. Across all keyword–source pairs, 78% are always-absent and 11% always-present, with the remainder showing mixed outcomes. The brand-mention decision is consistent with a stable underlying propensity rather than stochastic sampling variation. (With only 4–5 repeats, power to detect a 10% stochastic mention probability is limited; true inconsistency rates may be higher with more repeats.)

Prompt *variations* show 38.5% mixed results (Claude) vs. 11.0% for controls—a 3.5× higher rate. Phrasing variation is associated with 3.5× more inconsistency than identical-prompt controls. The 11% inconsistency cases mark the optimization frontier: brands at the mention/no-mention decision boundary.

Healthcare shows the least phrasing sensitivity (23.5% mixed vs. 43.2% for e-commerce), consistent with a visibility ceiling for healthcare brands that is robust to prompt phrasing.

4.3 Keyword Variation Is the Strongest Marginal Predictor

Table 5 presents the central quantitative finding. Which keyword was queried has the largest marginal η^2 (32.5%) among all factors tested. Because keyword has 737 levels (vs. 50 for brand, 5 for vertical, etc.), its high marginal η^2 partly reflects granularity—a categorical variable with many levels will mechanically capture more between-group variance. The finding should be interpreted as: which *specific keyword* was queried is more informative about brand mention than any other single factor. Marginal values overlap due to hierarchical nesting and cannot be summed (see Appendix C).

Keywords with 100% mention rate are category-defining queries: "best CRM software" (Salesforce), "best note-taking app" (Notion), "pay in 4" (PayPal). Keywords with 0% rate (262 of 737 with $n \geq 10$) are topic or entity queries where the brand ranks organically but is not the subject: "simparica" (Chewy), "vitamin d3" (Amazon), "cat noises" (Chewy).

Keyword length. Short keywords (1–2 words): 17.4% mention rate; medium (3–4): 17.3%; long (5+): 11.0%.

Keyword semantic type. Commercial keywords ("best," "compare," "software"): 17.3%. How-to/informational ("how to," "symptoms"): 11.7%. Neither length nor surface type fully explains the distribution; the variation is driven by how specifically each query demands naming a particular brand.

Practical implication. Keyword-level variation has a far larger marginal η^2 than AI system choice (32.5% vs. 0.1%). While these marginal values overlap and cannot be directly compared as unique contributions, the qualitative ordering is robust across specifications: AEO strategy should be keyword-first, not model-first.

4.4 Keyword Variation: Case Studies

The distribution is bimodal: 35% of keywords (262/737) produce zero brand mentions, while 11% (82/737) exceed 50%. This bimodality, combined with keyword identity having the largest ICC among all factors in the supplementary GLMM (Appendix C), supports treating keyword identity as the primary unit of analysis.

ThredUp (49.0% mention rate) illustrates the pattern (Table 6). Seven of its 15 keywords are variations of "online thrifting"—all above 50% mention rate. ThredUp *is* the category for online secondhand clothing; its visibility reflects the nature of its category queries, not brand strength.

Amazon (16.7% overall) is visible only for Amazon-specific products (Kindle, KDP, Prime Video) but invisible for generic product queries. A brand with broad product coverage but no category-defining association achieves low mention rates across queries. This pattern generalizes: Zocdoc (53.4%) owns "find a doctor" while WebMD (0%) is absorbed into general medical knowledge; Vanguard (32.4%) owns "index fund" while Fidelity (15.1%) is a generic brokerage.

Table 4: Behavioral profiles by source.

Metric	AIO	GPT-4.1	Claude	Gemini
Mention Rate	19.3%	17.6%	16.0%	16.5%
Citation Rate	28.7%	2.3%	0.08%	0.10%
Avg Sentiment	0.345	0.235	0.162	0.274
Avg Word Count	251	296	222	675
Hedging Rate	2.8%	7.7%	6.6%	31.7%
Disclaimer Rate	3.2%	5.9%	3.4%	13.5%

Table 5: Marginal association strength (marginal η^2). Values overlap due to hierarchical nesting and should not be summed.

Factor	% Var.	N levels	Definition
Keyword	**32.5%**	737†	Which query was asked; keywords range from category-defining to tangential
Brand identity	11.2%	50	Which brand is being measured
Intent type	5.2%	5	Prompt category (e.g., comparative, informational)
Vertical	1.7%	5	Industry sector (finance, law, healthcare, . . .)
Tier	0.3%	3	Market position (enterprise, midmarket, startup)
Source (AI system)	0.1%	4	Which AI system generated the response

† 737 of 750 keywords with $n \geq 10$ scored responses

Table 6: ThredUp vs. Amazon: keyword-level visibility.

Brand	Keyword	Rate	N
ThredUp	online thrifting	84.8%	33
	used clothes online	82.9%	41
	thrift online	75.6%	41
	consignment store	15.2%	33
Amazon	kindle	69.2%	39
	kdp	64.5%	31
	vitamin d3	0%	38
	magnesium glycinate	0%	38

4.5 The Vertical Gradient and Two Absence Patterns [H4]

H4 directional support. Chi-square: $p < 10^{-127}$; Cramér's $V = 0.120$ (small effect; Table 7). The direction is consistent with H4, but the primary finding is the qualitative distinction between absorption and displacement patterns rather than the magnitude of the vertical gap. *Important caveat applied throughout:* All chi-square p-values are computed at the response level (N=41,388) and are anti-conservative because the 10 prompt variations per keyword are correlated paraphrases (effective $N \approx 750$ keywords). We report p-values for completeness but rely on effect sizes (V, Cohen's h) for interpretation. Keyword-level tests yield qualitatively identical conclusions (see Appendix C).

The pattern differs by vertical. We identify two distinct patterns of brand absence:

Pattern 1: Absorption (Healthcare). When WebMD is absent (100% of its keyword responses), 94.9% of those responses contain *no commercial brand at all*. Outputs contain equivalent medical information—symptoms, treatments, drug interactions—referencing institutions (NIH, PubMed) rather than commercial brands. Healthcare has the most 0%-mention keywords (81, highest of any vertical).

Table 7: Vertical visibility and caution profile.

Vertical	Mention	Hedging	Discl.
Finance	24.0%	12.0%	4.3%
E-Commerce	20.2%	9.6%	1.6%
Technology	20.1%	6.6%	0.6%
Law	12.9%	12.5%	10.3%
Healthcare	12.1%	14.6%	13.8%

Pattern 2: Displacement (Finance, Technology, E-Commerce). When finance, technology, or e-commerce brands are absent, competitors take their place. Co-occurrence analysis reveals stable "default rosters" (Table 8):

Top co-occurring pairs: LegalZoom + Rocket Lawyer (422), Asana + Monday.com (380), HubSpot + Salesforce (295). AIO trigger rates are uniform across verticals (91–95%); the gradient is in answer *content*, not suppression.

Table 8: AI default rosters by vertical.

Vertical	Top Co-Mentioned Brands	Density
Technology	HubSpot (12.3%), Asana (10.5%), Monday.com (9.1%), Salesforce (8.6%)	Dense
E-Commerce	eBay (10.0%), Amazon (9.8%), Shopify (8.2%), ThredUp (4.9%)	Dense
Finance	Vanguard (8.6%), PayPal (8.2%), Schwab (5.2%)	Dense
Law	LegalZoom (11.1%), Rocket Lawyer (9.3%), Nolo (4.9%)	Moderate
Healthcare	Zocdoc (4.9%), GoodRx (4.7%), UnitedHealth (3.7%)	**Sparse**

Table 9: Cross-model agreement on brand mention (5,125 three-way prompts).

Agreement	Count	Prop.	95% CI
All 3: NOT mentioned	4,488	76.2%	[75.1, 77.3]
All 3: mentioned	555	9.4%	[8.7, 10.2]
Partial (1 or 2 of 3)	848	14.4%	[13.5, 15.3]

4.6 Cross-Model Agreement and Behavioral Profiles [H1]

Across 5,125 prompts where all 3 LLMs responded (Table 9):

For context, under independence at the observed ∼18% base mention rate, chance three-way absence agreement would be $(0.82)^3 = 55.1\%$ and chance three-way presence agreement $(0.18)^3 = 0.58\%$. The observed 76.2% absence agreement is above chance, and the 9.4% presence agreement is 16× the chance expectation—models converge on brand inclusion far more than independence would predict. Fleiss' $\kappa = 0.647$ indicates substantial agreement per Landis & Koch benchmarks, though κ is sensitive to marginal distributions; in high-prevalence settings such as this (82% absence), κ underestimates agreement relative to raw observed agreement (85.6%), so $\kappa = 0.647$ is a conservative lower bound on cross-model convergence. The asymmetry is consistent with the marginal association analysis: source explains only 0.1% of *whether* a brand appears, but systems diverge on *which* brands to surface in the 18% of cases where brands appear.

Healthcare and law show highest absence consensus (∼85%) and lowest presence consensus (5–7%). Comparative queries produce the most presence agreement (20.2%); informational queries the most absence agreement (92.3%).

H1: Mention rates (not supported). Mention rates cluster narrowly (16.0%–19.3%), with source explaining just 0.1% of variance. However, behavioral profiles diverge markedly: hedging rates span 2.8%–31.7% and word counts span 222–675 across sources (Table 4).

AIO is most commercially oriented (highest sentiment, least hedging). Claude is most neutral (lowest sentiment). Gemini is most cautious (highest hedging, 3× Claude's word count). Within-Google divergence (Gemini vs. AIO) is consistent with deliberate product differentiation.

The Citation Divide [H2]. URL citation rates vary dramatically by source type: AIO cites brand URLs in 28.7% of responses (leveraging search-index grounding), GPT-4.1 in 2.3%, while Claude (0.08%) and Gemini (0.10%) produce near-zero citations. **H2 partially confirmed for LLMs, rejected for AIO.** The LLM-only citation rate is 0.81%—citation-based measurement fails for standalone LLMs but works for search-grounded systems. A web-search-enabled condition (n=53,720; §5.1) confirms this is not merely a retrieval limitation: with web search enabled, GPT-4.1's overall URL rate rises to 31.0% but brand-specific citations remain at 2.4%, aggregate brand URL citation stays below 3%, and brand mention rates fall across all three LLMs (16.6% → 14.2%).

4.7 Market Tier Does Not Predict Visibility [H3]

H3 not supported. While statistically significant (χ^2: $p < 10^{-41}$), the effect size is negligible (Cramér's $V = 0.067$); tier explains 0.3% of variance (Table 5).

Only technology follows the expected hierarchy. Stratifying by keyword type explains the pattern: category-defining keywords (>25% mention rate) yield ∼50% mention regardless of tier (enterprise 52.0%, midmarket 47.4%, startup 49.1%), while tangential keywords cluster at 4–6% across all tiers. The apparent tier effect is largely explained by differences in keyword type density across tiers—startups with category-defining queries (ThredUp = online thrifting) outperform enterprise brands with diffuse query portfolios (Amazon = everything).

Domain authority as a continuous signal. To probe whether the null tier result masks a genuine authority effect, we supplemented the categorical analysis with continuous Ahrefs Domain Rating (DR) scores for all 50 brand domains (retrieved 2026-03-01, range DR 57–96). Across all brands and LLM sources, Spearman $\rho = +0.34$ ($p = 0.017$)—weak but significant; AI Overviews shows negligible association ($\rho = +0.17$, ns), consistent with retrieval-based rather than parametric selection.

Stratifying by tier reveals a ceiling effect that explains the aggregate. Enterprise brands occupy a compressed DR band (73–96, $\sigma = 5.8$), leaving too little variance for the signal to surface ($\rho = +0.28$, ns; $n = 20$). Midmarket shows near-zero correlation ($\rho = +0.06$, ns; $n = 14$) for the same reason. Within startups, however, where DR ranges from 57 to 92 ($\sigma = 8.3$), domain authority is strongly predictive of mention rate ($\rho = +0.82$, $p < 0.001$, $n = 16$; leave-one-out range 0.80–0.90). The startup correlation partly reflects this tier's wider DR range. The aggregate weak result is a restricted-range artifact: by selecting brands that have already "won"

Table 10: Mention rate by tier × vertical.

Vertical	Enterprise	Midmarket	Startup
E-Commerce	18.9%	9.1%	**32.5%**
Finance	**30.8%**	8.4%	19.7%
Healthcare	12.7%	**34.5%**	7.4%
Law	**17.3%**	16.8%	5.9%
Technology	**25.2%**	13.4%	13.6%

Table 11: Mention rate by intent type. Cohen's h is computed relative to the informational baseline.

Intent Type	Rate	Cohen's h
Comparative	30.1%	0.607
Constrained	26.3%	0.524
Recommendation	23.1%	0.449
Problem-solving	11.0%	0.126
Informational	7.5%	(baseline)

traditional search, our corpus suppresses the variation needed to detect a monotonic DR effect.

Counter-examples suggest a threshold rather than monotonic interpretation. Using LLM-only mention rates (excluding AIO, which is retrieval-augmented): WebMD (DR = 92, 0%), Mayo Clinic (DR = 93, 0.6%), and Nolo (DR = 73, 5.9%)—all from low-visibility YMYL verticals (healthcare and law; confounding with H4)—demonstrate that high domain authority does not ensure LLM visibility when queries are tangential to the brand's core category. Conversely, Notion (DR = 92, 75% mention) shows that a startup-tier domain can dominate when its content is semantically entangled with query categories in training data. The data suggest a sufficiency threshold in the DR 70–80 range (mention rates rise from ∼5% below to ∼15–20% above), though within-bin variance is large and formal changepoint analysis on 50 brands is underpowered; above this range, query specificity becomes the decisive factor.

Competitor comparison. The alignment effect is further visible in competitor analysis: target brands (those ranking in Google's top 5 for a keyword) are mentioned at 17.4% by LLMs, 6× higher than same-vertical competitors (2.92%) and far above other-vertical brands (0.19%; n=18,039 LLM responses with non-null text). Within the top 5, however, exact rank position shows no linear association with LLM mention rate ($\rho = +0.04$, ns; coded 1=best), though AIO exhibits modest rank-sensitivity ($\rho = -0.103$, $p = 0.007$).

4.8 Intent Type and Position Dynamics [H5]

H5 confirmed. χ^2: $p < 10^{-300}$; Cramér's $V = 0.231$—the largest Cramér's V in the study (Table 11).

The ranking is model-independent: comparative > constrained > recommendation > problem-solving > informational across all 4 sources, with a 4× gap between the highest and lowest intent types (30.1% comparative vs. 7.5% informational).

Position dynamics. When brands appear in ranked lists (LLM sources only; see caveat below), position is cross-model stable. ThredUp holds #1 across all 3 LLMs in e-commerce; Ro holds #1 in healthcare; LegalZoom #1 in law.

Position analysis carries an important caveat: brands mentioned in low-visibility verticals are a highly selected sample. Ro's 47.3% rank-#1 share (fraction of healthcare ranked-list responses in which Ro holds the top position) reflects the bimodal nature of healthcare brand presence under the absorption pattern—almost always absent, but prominent when present.

Selection bias in position data. Position rank varies by intent (LLM sources only; AIO's citation-style output does not yield list-position data, so its higher mention rates are excluded from this analysis): constrained queries yield the best average rank (5.12) despite lower LLM mention rates (22.6%), while comparative queries produce more LLM mentions (26.9%) at worse average positions (7.42). By tier, startups achieve rank #1 most often (32.6% of ranked mentions vs. 20.8% enterprise, 17.4% midmarket)—likely because startup mentions occur in narrower, more definitive contexts. These position statistics condition on brand appearance, a highly selected subset; they should not be interpreted as unconditional rankings.

5 DISCUSSION

We discuss five implications of the findings above.

5.1 The Citation Divide

Citation behavior reveals a large gap between system types. AIO cites brand URLs in 28.7% of responses; standalone LLMs cite at far lower rates (GPT-4.1: 2.3%, Claude: 0.08%, Gemini: 0.10%). This renders citation-based measurement—including GEO's Position-Adjusted Word Count—ineffective for LLMs but highly informative for search-grounded systems. The implication for practitioners: AIO is not just the most brand-friendly surface for mentions, but also the only one that provides verifiable attribution. Khalifa et al. [12] show citation is technically fixable in LLMs but commercially unmotivated.

Web-search extension. We collected 53,720 responses from the same three LLMs with web search tools enabled across

Table 12: Baseline vs. web-search-enabled LLMs: brand citation and mention rates (n=53,720).

	Brand URL Citation		Brand Mention	
Source	**Baseline**	**Web Search**	**Baseline**	**Web Search**
GPT-4.1	2.3%	2.4%	17.6%	14.6%
Claude Sonnet	0.08%	0.09%	16.0%	14.7%
Gemini Flash	0.10%	0.07%	16.5%	13.3%
Aggregate	0.81%	0.85%	16.6%	14.2%

three snapshots (Table 12). This condition extends the baseline analysis in two ways.

Citation behavior. GPT-4.1's overall URL citation rate rises sharply (2.3% → 31.0%), but brand-specific URL citations remain flat. Claude and Gemini brand citations stay near zero. LLMs with web search cite news outlets, review aggregators, and reference sites, but rarely attribute to brand domains (aggregate brand URL citation: 0.85

Mention rates and behavioral profiles. Web search consistently suppresses brand mention rates across all three LLMs (aggregate LLM mention: 16.6% → 14.2%). Models with retrieval access appear to synthesize content more generically, naming specific brands less even when those brands are relevant. Behavioral profiles shift as well: Claude responses lengthen substantially (222 → 333 words) while Gemini responses shorten (675 → 557 words); Claude hedging increases (6.6% → 9.0%) while Gemini hedging decreases (31.7% → 18.8%). Retrieval augmentation interacts differently with each model's response style, though the brand-mention suppression effect is consistent across all three.

5.2 Why Keyword Variation Dominates

The marginal association analysis is our central empirical result. Keyword-level variation has marginal $\eta^2 = 32.5\%$, source 0.1%, tier 0.3%; a supplementary GLMM corroborates the qualitative ordering (VB convergence warning applies; Appendix C). AI visibility is most strongly associated with which specific query is asked, not which AI system answers it or how large the brand is. The observed spread—from near-100% mention rates on category-defining queries to near-zero on tangential ones—suggests that the driving factor is how specifically a query demands naming a particular brand.

The ThredUp/Amazon case (§4.4) illustrates the pattern: ThredUp *is* the category for "online thrifting"; Amazon lacks a single defining category. The Zocdoc/WebMD case shows a similar pattern: specific transactional content ("book a doctor") is associated with higher mention rates than encyclopedic content ("symptoms of diabetes"), plausibly because the former requires naming a service while the latter can be synthesized from parametric knowledge.

This thesis is partially supported by a preliminary embedding-based operationalization of query-brand semantic specificity: cosine distance between query embeddings and brand-page embeddings yields Spearman $\rho = 0.42$ ($p < 10^{-32}$) and explains $\Delta R^2 = 0.065$ of mention-rate variance beyond the fixed effects—a promising but preliminary signal. A planned content attribute analysis will examine whether brand page language predicts visibility independent of domain authority, providing a more direct test of whether category-defining content drives the keyword-dominance finding.

Table 13: Organic ceiling test: LLM visibility for brands already in Google's top 5. All 737 keywords have the target brand in Google's top 5 organic results.

Metric	**Value**
Keywords with 0% LLM mention	355 (48.2%)
Keywords mentioned by all 3 LLMs	241 (32.7%)
Keywords mentioned by ≥1 LLM	382 (51.8%)
Overall LLM mention rate	17.4%
Competitor comparison (n=18,039):	
Target brand (top-5 ranked)	17.4%
Same-vertical competitor	2.92% (6× lower)
Other-vertical brand	0.19% (92× lower)
When AIO mentions the brand (>25% rate):	
LLMs also mention	152/168 (90.5%)
LLMs do not mention	16/168 (9.5%)

Organic dominance does not guarantee AI visibility. Every brand in our corpus holds a top-5 Google organic position for its study keywords—the ceiling of traditional search success. Yet 82% of AI responses omit the target brand, and nearly half of all keywords (48.2%) receive zero LLM mentions (Table 13). Winning organic search is not sufficient for AI visibility.

Position *within* the top 5 does not differentiate further: rank #1 and ranks #2–5 yield statistically indistinguishable LLM mention rates ($\rho = +0.04$, ns). AIO shows a modest rank-#1 advantage ($\rho = -0.103$, $p = 0.007$), plausibly reflecting shared Google infrastructure. Only one-third of keywords (32.7%) achieve mention from all three LLMs. Because our corpus is restricted to top-ranking brands, we cannot make claims about whether organic authority is irrelevant to AI visibility in general—only that among brands that have already won traditional search, additional organic authority yields diminishing returns.

5.3 Absorption vs. Displacement

Brand absence exhibits different patterns depending on vertical:

Absorption is the pattern where AI outputs contain equivalent information to authoritative brand content but without attribution. This may reflect an unintended consequence of E-E-A-T optimization [27]: authoritative brand content is absorbed into parametric model knowledge and reproduced in AI outputs without source attribution. WebMD's medical content, Mayo Clinic's guidelines—the information appears

but the brand names do not. When these brands are absent, 94.9% of responses contain *no commercial brand at all*, with outputs referencing institutional sources (NIH, PubMed, CDC) as detected via entity discovery [26].

Displacement occurs when competitors appear instead. In finance, technology, and e-commerce, the default roster redistributes attention—a zero-sum dynamic.

These patterns have different strategic implications. Brands exhibiting the absorption pattern would benefit from content so specific and transactional that the brand name becomes part of the answer. Brands exhibiting displacement would benefit from creating content that addresses comparative and constrained queries—the two intent types associated with the highest mention rates (comparative 30.1%, constrained 26.3%; Table 11)—to position themselves within the keyword categories where AI default rosters form.

5.4 AIO as Most Commercial Surface

AIO is the most commercially oriented surface across every metric: highest mention rate (19.3%), highest sentiment (0.345 compound), lowest hedging (2.8%), and lowest disclaimer rate (3.2%). One possible explanation is that AIO inherits ranking signals from its underlying search index, which is optimized for commercial relevance; alternative explanations include differences in retrieval-augmented generation or fine-tuning objectives.

The within-Google divergence is notable. Gemini Flash, sharing the same corporate parent, produces 11× more hedging, 4× more disclaimers, and 2.7× longer responses (675 vs. 250 avg words). Despite shared corporate parentage, the two systems produce markedly different behavioral profiles (whether they share training data is unknown). This is consistent with different design objectives: AIO may be tuned for commercial utility while Gemini may be tuned for conversational caution, though the exact tuning objectives are not publicly documented.

Yet *which* AI system answers explains just 0.1% of variance in brand mention—a warning against overindexing on any single surface. The brand-mention decision is overwhelmingly predicted by the query (keyword) and the target (brand), not which AI system is queried.

5.5 The Vertical Gradient and Fairness

Although the effect size is small ($V = 0.120$), the 2.0× gap (Healthcare 12.1% vs. Finance 24.0%) raises a fairness question that Venkit et al. [14] anticipated: AI systems may exhibit systematically different competitive landscapes across verticals. Healthcare brands face a structural ceiling that is cross-model (86.9% unanimous absence), phrasing-robust (23.5% sensitivity), and brand-free by default. A healthcare startup investing in AEO faces fundamentally different odds than a fintech startup—not solely because of market dynamics, but plausibly because of structural factors in how AI systems handle medical content—though the specific mechanism (alignment training, safety tuning, or other factors) is not identifiable from observational data.

The absorption pattern is consistent with much of this. AIO trigger rates are uniform across verticals (91–95%), so the gap is in answer *content*, not in whether AI answers appear. Models are equally willing to generate AI answers for healthcare queries—they are simply less willing to name commercial brands in those answers. This is arguably good for consumers (fewer conflicts of interest in medical advice) but creates an uneven competitive landscape that brands cannot overcome through content strategy alone.

5.6 Limitations

Temporal scope: all data fall within an approximately three-week window (2026-02-28 to 2026-03-16); snapshot consistency strengthens internal validity but the study does not track changes over months as models update and brand content evolves. *Programmatic scoring only:* binary brand mention is the primary signal; LLM-as-judge scoring would add nuance. *Keyword composition:* topic/entity queries increase the proportion of zero-mention keywords and drive the keyword-dominance finding; a different keyword set would shift variance ratios. *Marginal association analysis:* marginal η^2 values overlap due to hierarchical nesting and differences in factor granularity (750 keyword levels vs. 5 for vertical). A BinomialBayesMixedGLM corroborates the qualitative ordering: keyword ICC = 34.4%, brand ICC = 23.0%, residual = 42.6%, with intent type as the strongest fixed predictor (marginal $\eta^2 = 5.2\%$; Table 5). Note: VB fitting produced a convergence warning; the qualitative ordering is consistent with marginal η^2 results. *Restricted-range DR:* brands were selected from top Google organic rankings (DR 57–96); the weak DR correlation cannot generalize to brands with low domain authority. *Mechanism ambiguity:* all findings are correlational, not causal.

6 CONCLUSION

We measured brand visibility across 4 production AI answer engines—110,523 responses (56,803 baseline; 53,720 web-search-enabled) from 50 brands across 5 verticals. Of five hypotheses (Table 1), one was confirmed (H5: intent $V = 0.231$), one received directional support (H4: healthcare 2.0× lower than finance (12.1% vs. 24.0%), $V = 0.120$; the absorption/displacement distinction is the primary contribution), one partially confirmed (H2: LLM aggregate citation rate 0.81% vs. AIO 28.7%; even with web search enabled, GPT-4.1 overall citation rises to 31% but brand-specific citation remains <3%—a citation preference gap, not a retrieval limitation), and two were not supported in practice (H1: source explains 0.1% of variance; H3: tier $V = 0.067$, though continuous DR predicts mention within startups at $\rho = +0.82$). Measurement integrity was established separately: fictitious brand controls yielded 0/356 mentions with no false positives (§4.1).

The marginal association hierarchy: keyword (32.5%) > brand identity (11.2%) > intent type (5.2%) > vertical (1.7%) > tier (0.3%) ≈ source (0.1%); these marginal η^2 values overlap and should not be summed. AEO strategy

should be keyword-first, not model-first. Two patterns of brand absence—absorption (healthcare) and displacement (finance, technology, and e-commerce)—suggest different content strategies. Brand mention is ∼95% consistent at temperature=1.0 in limited repeats. Among brands that already dominate Google rankings, organic search authority adds little incremental LLM visibility—the average mention rate is only 17.6% despite uniformly high domain authority. Cross-model measurement is essential: 76.2% absence consensus vs. 9.4% presence consensus (Fleiss' $\kappa = 0.647$).

Despite limitations in temporal scope, scoring depth, and causal inference, the methodology is fully open and the findings are replicable. For practitioners: optimize for category-defining keywords, not model-specific tactics; measure across multiple AI systems, not just one; and recognize that healthcare brands (a YMYL domain) face structurally lower AI visibility ceilings than brands in finance and technology. Priority extensions include temporal snapshots for stability testing, LLM-as-judge scoring to quantify absorption depth, human-labeled validation of scoring precision for ambiguous brand names, and a content attribute analysis formalizing the semantic specificity signal described in §5.2. By providing open data, validated methodology, and falsifiable claims, this study establishes an initial empirical foundation for AI answer engine measurement.

REFERENCES

[1] Advanced Web Ranking. AI Overview Trigger Rates and CTR Impact. Industry Report, 2025.
[2] Nectiv. ChatGPT Query Volume and Fan-Out Analysis. Industry Report, 2025.
[3] SparkToro/Datos. Zero-Click Search Trends. Industry Report, 2025.
[4] Ahrefs. AI Overview CTR Impact Study. Industry Report, Feb 2026.
[5] S. Feuerriegel et al. Understanding the Impact of SERP Features on Search Behavior. In *Proc. SIGIR*, 2023.
[6] D. Lewandowski and S. Schultheiß. Public Awareness and Attitudes Towards Search Engine Optimization. arXiv preprint arXiv:2204.10078, 2022.
[7] P. Aggarwal, V. Murahari, T. Rajpurohit, A. Kalyan, K. Narasimhan, and A. Deshpande. GEO: Generative Engine Optimization. In *Proc. KDD*, 2024.
[8] Y. Chen et al. CC-GSEO-Bench: A Benchmark for Generative Search Engine Optimization. Preprint, 2025.
[9] P. Strauss et al. Do AI Search Engines Give Credit Where It's Due? Investigating the Attribution Crisis. Preprint, 2025.
[10] E. Kirsten, J. Grosse Perdekamp, M. Upadhyay, K. P. Gummadi, and M. B. Zafar. Characterizing Web Search in the Age of Generative AI. arXiv preprint arXiv:2510.11560, 2025.
[11] D. Pfrommer et al. Ranking Manipulation for Conversational Search Engines. In *Proc. EMNLP*, 2024.
[12] M. Khalifa et al. Source-Aware Training Enables Knowledge Attribution in Language Models. Preprint, 2024.
[13] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank Citation Ranking: Bringing Order to the Web. Technical Report, Stanford, 1999.
[14] P. N. Venkit et al. Search Engines in an AI Era: Understanding Challenges and Opportunities. In *Proc. FAccT*, 2025.
[15] A. Bagga et al. E-GEO: Generative Engine Optimization for E-Commerce. Preprint, 2025.
[16] A. Kumar and L. Palkhouski. AI Answer Engine Citation Behavior: An Empirical Analysis of the GEO-16 Framework. arXiv preprint arXiv:2509.10762, 2025.
[17] N. Bardas, T. Mordo, O. Kurland, M. Tennenholtz, and G. Zur. White Hat Search Engine Optimization Using Large Language Models. arXiv preprint arXiv:2502.07315, 2025.
[18] X. Chen, H. Wu, J. Bao, Z. Chen, Y. Liao, and H. Huang. RAID G-SEO: Role-Augmented Intent-Driven Generative Search Engine Optimization. arXiv preprint arXiv:2508.11158, 2025.
[19] K. Ando and T. Harada. Aligning Large Language Model Behavior with Human Citation Preferences. arXiv preprint arXiv:2602.05205, 2026.
[20] M. T. Ribeiro, T. Wu, C. Guestrin, and S. Singh. Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. In *Proc. ACL*, 2020. (Best Paper Award)
[21] T. Shin, Y. Razeghi, R. L. Logan IV, E. Wallace, and S. Singh. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proc. EMNLP*, 2020.
[22] F. Huszár et al. Measuring the Performative Power of Search Engines. Preprint, 2024.
[23] S. Aral, H. Li, and R. Zuo. The Rise of AI Search: Implications for Information Markets and Human Judgement at Scale. arXiv preprint arXiv:2602.13415, 2026.
[24] T.-Y. Liu. Learning to Rank for Information Retrieval. *Foundations and Trends in IR*, 2009.
[25] Semrush. Query Length Statistics Across Search and AI Platforms. Industry Report, 2025.
[26] M. Honnibal, I. Montani, S. Van Landeghem, and A. Boyd. spaCy: Industrial-Strength Natural Language Processing in Python. Zenodo, 2020. https://doi.org/10.5281/zenodo.1212303
[27] Google. Search Quality Rater Guidelines: E-E-A-T Update. 2024.

DATA AND CODE AVAILABILITY

All data, prompts, scoring code, and per-brand results are available at https://github.com/keygrip/aeo-research. The repository includes the full brand corpus with aliases, prompt generation pipeline, scoring codebook (regex patterns, phrase lists, VADER/spaCy configuration), per-brand results across all sources, control prompt analysis, and per-keyword visibility breakdowns.

A FULL BRAND CORPUS

Table 14 lists all 50 brands with aliases used for mention detection, primary domains for URL citation matching, and tier classification. Five fictitious brands (one per vertical) were verified non-existent via Google search on 2026-02-27: Wynthoral Financial, Grelvant Law, Plorantic Health, Jorvelle, and Velnith Market (356 scored responses, 0 mentions).

B SCORING CODEBOOK

The programmatic scoring pipeline produces 23 metrics across 8 signal categories. All scoring is deterministic (no API calls) and re-runnable. Full source code is available in the supplementary repository.

AIO preprocessing. Google AI Overview responses contain inline citation markers (e.g., "K Health +3") that are stripped via regex before scoring to prevent sentiment contamination.

C STATISTICAL DETAILS

Confidence intervals. All proportion estimates use Wilson score intervals at 95% confidence:

$$\frac{\hat{p} + \frac{z^2}{2n} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z^2}{4n^2}}}{1 + \frac{z^2}{n}}$$

where $z = 1.96$. Wilson intervals are preferred over Wald intervals for proportions near 0 or 1, which is common in our data ($\sim$82% of keyword-brand pairs show 0% mention rate).

Table 14: Complete brand corpus with detection aliases and domains. Aliases marked with ∗ were configured in the scoring pipeline; unmarked aliases are known alternate forms not used in scoring (see Section 3.4, limitation 4).

Vertical	Brand	Tier	Aliases	Domain
Finance	Fidelity Investments	Enterprise	Fidelity	fidelity.com
	Charles Schwab	Enterprise	Schwab	schwab.com
	Vanguard	Enterprise	—	vanguard.com
	PayPal	Enterprise	—	paypal.com
	SoFi	Midmarket	—	sofi.com
	Robinhood	Midmarket	—	robinhood.com
	Wealthfront	Midmarket	—	wealthfront.com
	Mercury	Startup	—	mercury.com
	Ramp	Startup	—	ramp.com
	Chime	Startup	—	chime.com
Law	LegalZoom	Enterprise	Legal Zoom	legalzoom.com
	Avvo	Enterprise	Martindale-Avvo	avvo.com
	LegalShield	Enterprise	Legal Shield	legalshield.com
	Nolo	Enterprise	—	nolo.com
	Rocket Lawyer	Midmarket	RocketLawyer	rocketlawyer.com
	ZenBusiness	Midmarket	Zen Business	zenbusiness.com
	NW Registered Agent	Midmarket	Northwest Registered Agent	northwestregisteredagent.com
	Ironclad	Startup	—	ironcladapp.com
	Trust & Will	Startup	TrustAndWill	trustandwill.com
	Hello Divorce	Startup	HelloDivorce	hellodivorce.com
Healthcare	WebMD	Enterprise	—	webmd.com
	Mayo Clinic	Enterprise	MayoClinic	mayoclinic.org
	UnitedHealthcare	Enterprise	UHC, United Healthcare	uhc.com
	CVS Health	Enterprise	CVS, Aetna	cvs.com
	Zocdoc	Midmarket	—	zocdoc.com
	GoodRx	Midmarket	Good Rx	goodrx.com
	Hims & Hers	Startup	Hims, ForHims	forhims.com
	K Health	Startup	—	khealth.com
	Ro	Startup	—	ro.co
	Cerebral	Startup	—	cerebral.com
Technology	Salesforce	Enterprise	SFDC, Salesforce.com∗	salesforce.com
	HubSpot	Enterprise	—	hubspot.com
	Microsoft 365	Enterprise	M365, Office 365	microsoft.com
	ServiceNow	Enterprise	—	servicenow.com
	Monday.com	Midmarket	Monday∗	monday.com
	Asana	Midmarket	—	asana.com
	Freshworks	Midmarket	—	freshworks.com
	Notion	Startup	Notion.so∗	notion.so
	Linear	Startup	—	linear.app
	Rippling	Startup	—	rippling.com
E-Commerce	Amazon	Enterprise	—	amazon.com
	Shopify	Enterprise	—	shopify.com
	Walmart	Enterprise	—	walmart.com
	eBay	Enterprise	—	ebay.com
	BigCommerce	Midmarket	Big Commerce	bigcommerce.com
	Chewy	Midmarket	—	chewy.com
	Wayfair	Midmarket	—	wayfair.com
	Bolt	Startup	—	bolt.com
	StockX	Startup	—	stockx.com
	ThredUp	Startup	—	thredup.com

Table 15: Complete scoring pipeline: 23 metrics across 8 categories.

Category	Metric	Type	Algorithm
Brand Mention	brand_mentioned	bool	Word-boundary regex over canonical name + aliases (case-insensitive, possessive-aware)
	brand_mention_count	int	Total match count across all variants
	brand_first_mention_position	int	Character offset of earliest match
	brand_variants_found	list	Which name variants triggered matches
URL Citation	url_cited	bool	Extract URLs via regex, compare registered domain to brand domain
	cited_urls	list	All matching URLs with position classification
Rec. Rank	recommendation_rank	int	Parse numbered and bullet lists, match brand variants against items
	total_recommendations	int	Total items in detected list
Structure	response_format	str	Cascade: table → mixed → numbered → bullets → prose
	response_word_count	int	Whitespace splitting
	brand_word_count	int	Words in brand-mentioning sentences
	has_headers	bool	Markdown headers or bold lines
	flesch_kincaid_grade	float	Textstat library
Language	hedging_language	bool	13-phrase list ("it depends," "your mileage may vary," etc.)
	confidence_language	bool	10-phrase list ("the best," "hands down," etc.)
	disclaimer_present	bool	11-phrase list ("consult a professional," "not financial advice," etc.)
	refusal	bool	8-phrase list ("I can't provide," "I cannot recommend," etc.)
	clarification_request	bool	Regex pattern matching
Factual Claims	factual_claim_count	int	Count sentences with %, $, years, or user counts
Entity Disc.	discovered_entities	list	spaCy NER: ORG and PRODUCT entities (deduplicated)
	all_entities_mentioned	list	All known-brand mentions (competitive landscape)
Sentiment	brand_sentiment_score	float	VADER compound, mean over brand-mentioning sentences
	brand_sentiment_label	str	Positive ($\geq$0.05), negative ($\leq$–0.05), neutral

Marginal association analysis. Marginal η^2 is computed as one-way ANOVA per factor: $\eta^2_{\text{marginal}} = SS_{\text{between}}/SS_{\text{total}}$ for each factor independently. Because the outcome is binary, η^2 from ANOVA is equivalent to R^2 from a linear probability model; we use this approximation for descriptive comparison. Because factors are hierarchically nested—keyword within brand, brand within vertical, tier as a property of brand—these marginal values *overlap* and do not sum to an additive partition. Additionally, keyword identity has 737 levels, giving it more degrees of freedom than other factors (50 for brand, 5 for vertical); this mechanically increases its explanatory capacity and should be considered when comparing marginal η^2 across factors. The ∼48% unexplained variance includes phrasing variation, factor interactions (keyword × intent, brand × source), and residual stochasticity. A BinomialBayesMixedGLM with keyword and brand as random effects corroborates the qualitative ordering: keyword ICC = 34.4%, brand ICC = 23.0%, residual = 42.6% (logit-scale ICCs using the conventional $\pi^2/3$ residual variance; VB fitting convergence warning noted). Intent type is the strongest fixed predictor (marginal $\eta^2 = 5.2\%$; Table 5); Cramér's $V = 0.231$ is from the chi-square analysis (§4.8), not the GLMM.

Effect sizes. Chi-square tests use Cramér's $V = \sqrt{\chi^2/(n \cdot \min(r-1, c-1))}$. Key results: vertical × mention ($V = 0.120$, small), tier × mention ($V = 0.067$, negligible), intent × mention ($V = 0.231$, small-to-medium for df^*=4). Cohen's h for pairwise proportions: comparative (30.1%) vs. informational (7.5%) yields $h = 0.607$ (medium; computed from exact unrounded proportions).

Multiple comparisons. Bonferroni correction is applied within each family of pairwise comparisons (vertical: 10 pairs; intent: 10 pairs; source: 6 pairs; tier: 3 pairs; $\alpha_{\text{adjusted}} = 0.005$ for 10-pair families). All five hypotheses are evaluated at $\alpha = 0.05$ without cross-hypothesis correction, as they address substantively distinct questions. All reported results survive even a global Bonferroni threshold of $\alpha = 0.01$.

Power. With $N = 27{,}164$ scored responses and baseline mention rate ∼18%, the study detects 2pp differences at power > 0.99 and 1pp differences at power > 0.90. Underpowered for per-keyword comparisons where $n < 20$.

Non-independence. The 10 prompt variations per keyword are correlated paraphrases. Effective sample size is $N = 750$ keywords for headline statistics, not 20,408 responses. Within-keyword aggregation is used for all primary analyses. Chi-square p-values are reported at the response level; given the non-independence, these are anti-conservative (inflated significance). Cramér's V, which normalizes by N, is the primary effect-size measure; however, V remains influenced by non-independence in the underlying χ^2 statistic. Keyword-level chi-square tests yield qualitatively identical conclusions (see main text).