Query Intent, Not Google Rank: What Best Predicts AI Citation Behavior

Anthony Lee AI+Automation (aiplusautomation.com) with Claude (Anthropic) — AI research contribution acknowledged

Preprint v2 — February 2026 Not yet peer-reviewed

Abstract

The rapid integration of AI chatbots into consumer search behavior has spawned a cottage industry of Generative Engine Optimization (GEO) advice, much of it built on untested assumptions about how AI platforms select sources for citation. Industry practitioners widely assert that Google ranking determines AI visibility, that community-consensus platforms like Reddit confer citation advantages, and that AI recommendations are too inconsistent to warrant optimization efforts. We tested these claims empirically across four major AI platforms — ChatGPT, Claude, Perplexity, and Gemini — using a multi-study design that combined large-scale query intent classification (*n* = 19,556 queries across 8 verticals), Google rank cross-referencing (120 queries with 360 Top-3 results), server-side fetch verification via Vercel middleware logging, and page-level technical analysis of 479 cited and non-cited pages. Our results challenge all three prevailing claims. First, query intent — not Google rank or domain authority — emerged as the strongest predictor of citation source type at the aggregate level, with intent distributions varying significantly by vertical ($\chi^2(28) = 5{,}195$, $p < .001$, Cramér's $V = 0.258$), though formal predictive modeling showed that at the individual page level, technical page features (AUC = 0.594) outperformed intent (AUC = 0.462) for predicting citation status. Second, Google's Top-3 organic results predicted AI citations poorly: ChatGPT matched only 7.8% of URLs, while Reddit — despite occupying 38.3% of Google Top-3 positions across our sample — received exactly zero AI citations from either platform's API (binomial $p = 3.43 \times 10^{-23}$ for Perplexity). A companion study (Lee, 2026b) subsequently demonstrated that this zero-citation finding is access-channel dependent: the web UIs of the same platforms cite Reddit at rates of 17–44%, suggesting that API-based research may systematically underestimate Reddit's role in AI-generated responses. Third, AI brand recommendations showed substantial within-platform consistency (ChatGPT mean Jaccard = 0.619, 95% CI [0.537, 0.701]), though cross-platform agreement was near-random (all-four-platform Jaccard = 0.036). We further discovered a previously unreported architectural divide: ChatGPT and Claude perform live page fetches during conversations, while Perplexity and Gemini rely exclusively on pre-built search indices — with divergent robots.txt compliance behavior between the fetching platforms. These findings suggest that effective GEO strategy requires intent-aware, platform-specific optimization rather than the one-size-fits-all approach currently advocated by industry practitioners.

Keywords: Generative Engine Optimization, GEO, AI citation behavior, AI search, ChatGPT, Claude, Perplexity, Gemini, query intent, brand recommendations, live page fetch, robots.txt compliance

1. Introduction

When a consumer asks ChatGPT "what's the best running shoe for flat feet?" or turns to Perplexity for "how to start investing," the AI platform must decide which sources to cite in its response. That decision — which pages, brands, and domains appear in an AI-generated answer — now represents a significant and growing channel of consumer influence. As AI chatbots increasingly mediate the relationship between businesses and their potential customers, a new field of optimization has emerged to help businesses earn visibility within these systems.

Generative Engine Optimization (GEO), a term formalized by Aggarwal et al. (2024) in the first academic treatment of the subject, describes strategies for improving content visibility in AI-generated responses. The field has attracted substantial practitioner interest, with industry reports analyzing millions of AI responses (Search Atlas, 2025) and infrastructure providers documenting the surge in AI bot traffic across the web (Cloudflare, 2025). Yet for all this attention, the foundational question remains largely unanswered: what actually determines which sources AI chatbots cite?

The prevailing industry guidance rests on three core claims, each of which has achieved broad acceptance among GEO practitioners despite limited empirical validation. The first claim asserts that Google ranking is the primary predictor of AI citation — that pages ranking well in organic search will naturally appear in AI responses. The second holds that AI platforms preferentially cite community-consensus sources such as Reddit threads and forum discussions, leading practitioners to advocate for user-generated content optimization strategies. The third contends that AI recommendations are essentially random and inconsistent, shifting unpredictably with each query, and that sustained optimization is therefore futile.

We tested all three claims empirically. Our findings contradict each one.

This paper presents a multi-study investigation of AI citation behavior across four major platforms: ChatGPT (OpenAI), Claude (Anthropic), Perplexity, and Gemini (Google). We combined large-scale query analysis, controlled server-side experiments, and page-level technical auditing to construct a more complete picture of how AI chatbots select, retrieve, and present source material. Our work makes six primary contributions. First, we introduce query intent as a previously unexamined variable in GEO research and demonstrate that it is the strongest predictor of citation source type at the aggregate level, while page-level technical features best predict individual citation status — a two-level distinction with direct implications for optimization strategy. Second, we quantify the disconnect between Google organic rankings and AI citation behavior, documenting what we term the *Reddit Paradox*: Reddit's dominance in Google's Top-3 results paired with its complete absence from AI citations. Third, we report the first server-side evidence of a fundamental architectural divide among AI platforms, showing that ChatGPT and Claude perform live page fetches during conversations while Perplexity and Gemini do not. Fourth, we identify which page-level technical factors are statistically associated with AI citation — and which commonly cited factors are not. Fifth, we measure recommendation consistency across platforms, finding it substantially higher than industry discourse suggests. Sixth, we document divergent robots.txt compliance behavior, with practical implications for site owners seeking to control AI access.

Our analysis is guided by four research questions:

RQ1. What determines which sources AI chatbots cite in their responses?

RQ2. How does Google organic ranking relate to AI citation probability?

RQ3. Do different AI platforms exhibit fundamentally different retrieval architectures?

RQ4. How consistent are AI recommendations across repeated queries, and does this consistency vary by platform?

We structure the remainder of this paper as follows. Section 2 reviews the existing academic and industry literature on GEO, AI citation reliability, and LLM recommendation behavior. Section 3 describes our methodology across four core studies and three supporting investigations. Section 4 presents our results. Section 5 discusses implications for practitioners and researchers, challenges prevailing industry narratives, and addresses limitations. Section 6 concludes with a synthesis of our findings and directions for future work.

2. Related Work

2.1 Generative Engine Optimization

The term Generative Engine Optimization was introduced by Aggarwal et al. (2024) at KDD 2024, establishing the first formal framework for understanding how content might be optimized for AI-generated search responses. Using a simulated generative engine (BrightRetrieval paired with an LLM), the authors created GEO-bench — a benchmark of 10,000 queries — and tested nine content optimization heuristics including the addition of citations, statistics, and authoritative language. Their work demonstrated that content modifications could yield up to 40% improvements in visibility, measured through a proposed impression metric. The significance of this contribution lies in its framing: it established that AI-generated search operates by fundamentally different rules than traditional search engine optimization.

Bagga et al. (2025) extended this work into e-commerce with E-GEO, a testbed of over 7,000 product queries paired with 15 rewriting heuristics. Notably, the authors critiqued Aggarwal's impression metric as insufficient for commercial contexts — a page might gain "visibility" in an AI response without generating any commercial value. Their e-commerce focus and attention to practical measurement limitations make E-GEO particularly relevant to our own work, which includes substantial e-commerce query coverage.

Both foundational studies share a critical methodological characteristic: they are *interventional*. They modify content and measure how modifications change visibility. Our work takes a complementary *observational* approach — we measure what currently gets cited in production AI platforms and ask why. This distinction matters because interventional studies tell practitioners what they *can* change, while observational studies reveal what is *most strongly associated with* citation behavior in the wild. The two approaches are not in conflict, but the absence of observational studies has left a significant gap in the literature.

2.2 Citation Reliability in AI Systems

A separate body of work examines the reliability and accuracy of AI citations rather than their optimization. Venkit et al. (2025) conducted a 21-participant user study across You.com, Perplexity, and BingChat, finding that 50–90% of LLM responses were not fully supported by their cited sources. The authors cataloged 16 specific citation, source, and UI issues — establishing empirically what many users have experienced anecdotally: AI citations are frequently unreliable as indicators of actual source content.

Wu et al. (2025) automated this verification process with SourceCheckup, a framework that evaluated LLM citation support across 800 medical questions and 58,000 statement-source pairs. They found that 30–50% of statements were unsupported by the cited sources, with their automated framework achieving 88.7% agreement with medical expert consensus. While their domain (medical information) differs substantially from ours (commercial queries), their finding that citations are structurally unreliable aligns with our discovery that some platforms cite pages they have never actually fetched.

Our server-side fetch verification adds a new dimension to this reliability literature. Where Venkit et al. assessed whether citations *supported* the AI's claims, and Wu et al. assessed whether cited sources *contained* the claimed information, we assessed whether the AI platform *ever contacted* the cited server. The answer, for two of the four platforms we tested, is no — raising questions about the mechanical basis of citation in these systems.

2.3 LLM Recommendation Manipulation

A growing adversarial literature examines how LLM recommendations can be deliberately manipulated. Kumar and Lakkaraju (2024) demonstrated that strategic text sequences — essentially prompt injection at the content level — could push products to the top position in LLM recommendation lists. Filandrianos et al. (2025) tested the influence of cognitive biases on LLM product recommendations, finding that social proof signals consistently boosted brand ranking across Claude, GPT-4o, and Gemini. Their social proof finding is particularly relevant to the "Reddit strategy" we examine: if social proof biases LLM processing, one might expect Reddit's user-generated endorsements to translate into AI citations. Our data demonstrates that this expectation — however theoretically grounded — does not hold in practice.

We position our work in contrast to this adversarial strand. While prior studies have focused on what *can be forced* through manipulation (Kumar & Lakkaraju, 2024; Filandrianos et al., 2025), our study takes a diagnostic approach — measuring what currently predicts citation behavior across production platforms without attempting to alter it. This distinction is methodologically important: understanding natural citation behavior is a prerequisite for assessing whether manipulation studies generalize to real-world conditions.

2.4 Large-Scale Industry Studies

Two large-scale industry analyses provide useful comparison points. Search Atlas (2025) analyzed 5.5 million AI responses from 748,000 queries across Perplexity, Gemini, and OpenAI platforms, representing the largest published dataset on AI citation behavior. Their scale is impressive, but their analysis lacks query intent segmentation, Google rank cross-referencing, and any server-side verification of whether platforms actually fetch the pages they cite.

Cloudflare (2025) has published multiple analyses of AI bot traffic patterns across their content delivery network, documenting the macro-level growth of AI crawler activity. Their data provides infrastructure-level context for our site-level observations — they see the forest of AI bot traffic, while our server-side logging captures the individual trees.

2.5 Research Gap

The existing literature leaves several critical questions unaddressed. No prior work combines server-side fetch verification, query intent taxonomy, Google rank cross-referencing, and multi-platform comparison within a single study. The foundational GEO papers (Aggarwal et al., 2024; Bagga et al., 2025) used simulated engines rather than production platforms. The citation reliability studies (Venkit et al., 2025; Wu et al., 2025) focused on content accuracy rather than retrieval mechanics. The adversarial studies (Kumar & Lakkaraju, 2024; Filandrianos et al., 2025) examined what can be manipulated rather than what naturally occurs. And the large-scale industry analyses (Search Atlas, 2025) provided descriptive statistics without mechanistic explanations.

Research on internal query processing provides additional theoretical grounding for our intent-based approach. Ma et al. (2023) demonstrated that LLMs internally rewrite user queries before retrieval in retrieval-augmented systems — a mechanism that plausibly explains why user query intent maps to specific source types in AI citation behavior.

Our study addresses these gaps by combining observational measurement of natural citation behavior with server-side mechanical verification, intent-based query segmentation, and systematic cross-platform comparison.

---

3. Methodology

3.1 Overview of Experimental Design

We conducted four core studies and three supporting investigations between January and February 2026, testing AI citation behavior across ChatGPT (GPT-4o), Claude (Anthropic), Perplexity, and Gemini (Google). Our experimental infrastructure consisted of API access to each platform where available, server-side request logging via Vercel middleware on a controlled domain (aiplusautomation.com), an SQLite database for structured data collection, and BotSight — a custom-built monitoring tool deployed on the test domain that identifies and logs AI bot traffic by matching incoming request user-agent strings against known AI crawler signatures, recording timestamp, path, user-agent, and country of origin for each hit.

All four core studies share a common design philosophy: we treat AI citation behavior as an observable phenomenon to be measured rather than manipulated. Our approach is diagnostic and observational, complementing the interventional studies that dominate the existing GEO literature. Each study targets a specific research question while contributing evidence to the broader thesis that query intent and platform architecture — not Google rank or domain authority — are the primary drivers of AI citation behavior.

3.2 Notation and Metrics

We adopt the following notation throughout. Let q denote a query with intent type $I_q \in$ *{Informational, Discovery, Validation, Comparison, Review-Seeking}. Let s denote a source page with feature vector* $\boldsymbol{F}_s$. Let $p \in$ {ChatGPT, Claude, Perplexity, Gemini} denote a platform. Our analyses examine the probability of citation $P(\text{cite} \mid I_q, \boldsymbol{F}_s, p)$ through observational measurement rather than parametric modeling, though we supplement this with a formal predictive model in Section 4.6.

We employ several set-based and rank-aware similarity metrics. **Jaccard similarity** between two brand sets A and B is defined as $J(A, B) = |A \cap B| / |A \cup B|$, where $J = 1$ indicates perfect agreement and $J = 0$ indicates no overlap. **Rank-Biased Overlap (RBO)** is a top-weighted rank similarity measure (Webber et al., 2010) that handles lists of different lengths; we use the persistence parameter $p = 0.9$, which places approximately 86% of the evaluation weight on the top 10 items.

For the Reddit Paradox analysis, we compute the binomial probability of observing zero Reddit citations under a random citation model: $P(X = 0) = (1 - \pi)$^n, where $\pi = 138/360 = 0.383$ is Reddit's share of Google Top-3 positions and n is the number of AI citation URL matches. For Perplexity ($n = 107$): $P(X = 0) = (1 - 0.383)$^$107 = 3.43 \times 10^{-23}$.

3.3 Study 1: Query Intent Classification

Data collection. We assembled a corpus of 19,556 queries across 8 commercial verticals: agency services, automation tools, law, marketing, SaaS brands, SaaS products, social media, and supplements. Queries were sourced from keyword research tools, industry forums, and common consumer search patterns, then deduplicated and normalized.

Intent taxonomy. We developed a five-category intent taxonomy based on the functional purpose of the query from the user's perspective: *Informational* (seeking knowledge or explanation), *Discovery* (seeking product or service recommendations), *Validation* (seeking confirmation about a specific entity), *Comparison* (seeking direct comparison between named entities), and *Review-Seeking* (seeking evaluative opinions). Each query was classified using an LLM-based classifier (GPT-4o) operating at temperature 0.1 for classification consistency, with queries processed in batches of 25. The classifier received explicit category definitions with example queries and signal words for each intent type. We validated classification quality through manual review of a stratified random subsample of 200 queries (25 per vertical), achieving 91.5% agreement between the LLM classifications and human judgment. Disagreements were concentrated at the Discovery–Validation boundary (e.g., "is [brand] the best for [use case]" could plausibly be either), and in all cases fell within adjacent categories rather than reflecting gross misclassification. We did not compute formal inter-rater reliability metrics (e.g., Cohen's κ) because the validation was conducted by a single reviewer rather than independent raters — a limitation we acknowledge. We selected this taxonomy because existing search intent frameworks (navigational, informational, transactional) lack the granularity needed to distinguish between the different types of commercial queries that drive AI citation.

Analysis. We computed the distribution of intent types across the full corpus with Wilson score confidence intervals for each proportion. We tested whether the distribution was uniform using a χ^2 goodness-of-fit test and assessed the dependence of intent distribution on vertical using a χ^2 test of independence with Cramér's V as the effect size measure.

3.4 Study 2: Google Rank vs. AI Citation

Data collection. We selected 120 queries across 10 commercial verticals (Technology, Health, Finance, Travel, Home Improvement, Education, Food, Fitness, Legal, and E-commerce) balanced between informational and commercial intent. For each query, we retrieved the Top-3 organic Google results using the Google Custom Search API, then issued the same queries to ChatGPT (via API) and Perplexity (via API), collecting all cited URLs from each platform's response. We limited this study to ChatGPT and Perplexity because both platforms consistently provide explicit URL citations in their responses, enabling systematic URL-level matching. Claude and Gemini were excluded from this particular analysis: Claude frequently declines commercial recommendation queries (approximately 45% refusal rate in our consistency experiment) and provides citations less consistently, while Gemini's citation format during the study period made automated URL extraction unreliable. Both platforms are included in our other studies where their response characteristics permitted systematic comparison.

Matching methodology. We performed URL-level and domain-level matching between Google's Top-3 results and each platform's citations. A URL match required the exact URL (after normalization of trailing slashes, protocol, and www prefix) to appear in the AI platform's cited sources. A domain match required only the root domain to appear. We computed match rates for each platform, each Google position (1st, 2nd, 3rd), and each vertical.

Reddit analysis. During initial data exploration, we observed that Reddit URLs appeared frequently in Google's Top-3 results but were absent from AI citations. We formalized this observation into a specific test: for each platform, we computed the number of AI citations matching Reddit domains and tested whether the observed count of zero could have arisen by chance under random citation, using a binomial test with the base rate set to Reddit's proportion of Google Top-3 results (138/360 = 0.383).

Query type comparison. We further segmented queries into commercial and informational categories and compared AI citation match rates across these categories using χ^2 tests for each platform.

3.5 Study 3: Platform Architecture (Live Fetch Behavior)

Design. This study used server-side request logging to determine whether each AI platform performs real-time HTTP requests to target web pages during conversations — a question that prior literature has not addressed with direct evidence. We deployed Vercel middleware on aiplusautomation.com that logged every incoming request with timestamp, path, user-agent string, HTTP status, and IP address.

Phase 1: Initial test (January 6–7, 2026). We conducted an initial experiment using an llms.txt file deployment as a trigger, asking each of the four AI platforms to retrieve information from our site. This test initially concluded that only ChatGPT performed live page fetches. However, this conclusion was based on an incomplete understanding of Claude's retrieval architecture.

Phase 2: Correction and formal verification (February 14, 2026). Subsequent investigation revealed that Claude has a two-tool architecture: *web_search* queries a search index without contacting target servers, while *web_fetch* sends a real HTTP GET request with the user-agent string "Claude-User/1.0." During the January test, Claude had sufficient cached context from its search index and memory system to answer without triggering web_fetch, leading to our initial misclassification. We corrected this through five formal sub-tests.

Test 1 reproduced Claude's live-fetch behavior across two sessions and seven page fetches, with Vercel log correlation confirming each request. **Test 2** deployed a never-before-seen page (/test-fetch-experiment) containing unique verification facts (a passphrase, a fictional product name, and a verification number) that could not exist in any training data or search index. Claude retrieved all four verification facts correctly, with the corresponding server hit confirmed in Vercel logs. **Test 3** formally compared web_search and web_fetch: a web_search query for the test page produced zero server hits (confirming index-only behavior), while web_fetch produced a confirmed server hit. **Test 4** tested robots.txt compliance using fresh pages in both allowed and disallowed paths. After discovering memory contamination in the conversational interface (Claude presented memorized content as live fetch results without disclosure), we resolved this via an API-based retry using the Anthropic Messages API with the web_fetch_20250910 beta tool — eliminating the memory system entirely. The API test confirmed that Claude checks robots.txt before each fetch call and refuses requests matching Disallow directives: the control page (/test-allowed-v2/fresh-page) was fetched successfully, while the blocked page (/test-blocked/fresh-page) was refused with a "no permission" error and zero server hits. **Test 5** replicated the same protocol with ChatGPT, finding that ChatGPT fetched all requested pages — including the blocked page — with no robots.txt check observed in server logs.

For Perplexity and Gemini, we found no conversational user-agent hits in BotSight monitoring data or Vercel logs across either the January or February test periods. A Perplexity background crawler (PerplexityBot) was observed during the February tests, but this reflected background indexing activity rather than conversational fetching. Gemini's responses were consistent with retrieval from Google's internal search index, with no direct server contact observed.

3.6 Study 4: Page-Level Technical Factors

Data collection. We crawled 479 unique pages using Playwright — a headless browser automation tool — comprising 241 pages that had been cited by at least one AI platform and 238 pages that appeared in the same query contexts but were not cited. Pages were drawn from the URL datasets collected in Studies 1 and 2.

Feature extraction. We measured 26 features across each page, organized into binary features (presence/absence of schema markup, self-referencing canonical tag, mobile viewport meta tag, proper heading hierarchy, about page link, contact page link, author attribution, cookie banner, paywall signals, login wall, and HTTPS) and continuous features (word count, content-to-HTML ratio, H1/H2/H3 heading counts, internal link count, external link count, total link count, page size in bytes, load time in milliseconds, popup/modal element count, affiliate link count, ad network signal count, and schema count).

Analysis. For binary features, we used χ^2 tests with Yates' continuity correction (or Fisher's exact test when expected cell counts were below 5) comparing cited and non-cited pages, computing odds ratios with 95% confidence intervals. For continuous features, we used Mann-Whitney *U* tests with rank-biserial correlation as the effect size measure. All 26 *p*-values were corrected for multiple comparisons using the Benjamini-Hochberg false discovery rate (FDR) procedure at $\alpha = .05$.

3.7 Supporting Studies

Three additional investigations provided corroborating evidence for our primary findings.

Brand recommendation consistency. We issued 50 product recommendation queries (25 entity-anchored and 25 generic) to each of four platforms (ChatGPT, Claude, Perplexity, and Gemini), repeating each query three times for a total of 600 API calls. We computed pairwise Jaccard similarity and Rank-Biased Overlap (RBO) between response triplicates for each query, as well as cross-platform overlap. Entity-anchored queries named a specific brand (e.g., "should I buy the AirPods Pro"), while generic queries did not (e.g., "best budget gaming headset").

Content freshness. We measured the publication date of sources cited by Perplexity across 110 queries, categorized into three topic velocity tiers (high, medium, and low), and compared source age distributions against Google's organic results for the same queries.

Temporal stability. We re-issued 40 queries (20 to ChatGPT, 20 to Perplexity) approximately five weeks after the original experiment to measure cross-temporal stability of brand recommendations, computing Jaccard similarity between original and retest brand sets.

4. Results

4.1 Query Intent Predicts Citation Source Type

The distribution of query intent across our 19,556-query corpus was heavily skewed toward two dominant categories. Informational queries constituted 61.3% of the corpus (*n* = 11,986, 95% CI [60.6%, 62.0%]), followed by Discovery at 31.2% (*n* = 6,106, 95% CI [30.6%, 31.9%]). The remaining three categories collectively accounted for just 7.5%: Validation at 3.2% (*n* = 617), Comparison at 2.3% (*n* = 456), and Review-Seeking at 2.0% (*n* = 391). This distribution was significantly non-uniform ($\chi^2(4)$ = 26,898, *p* < .001).

Table 1. Query Intent Distribution (*N* = 19,556)

Intent Type	Count	Proportion	95% CI
Informational	11,986	61.3%	[60.6%, 62.0%]
Discovery	6,106	31.2%	[30.6%, 31.9%]
Validation	617	3.2%	[2.9%, 3.4%]
Comparison	456	2.3%	[2.1%, 2.6%]
Review-Seeking	391	2.0%	[1.8%, 2.2%]

More importantly, the intent distribution varied significantly by vertical ($\chi^2(28)$ = 5,195, *p* < .001, Cramér's *V* = 0.258, a medium effect size). This vertical dependence is not merely statistically significant — it is structurally meaningful. Agency services and law verticals were overwhelmingly Discovery-skewed (67.6% and 64.6% Discovery, respectively), reflecting consumer behavior patterns in which users seek providers rather than information. SaaS brands were overwhelmingly Informational (87.3%), consistent with users seeking to understand products they have already identified. SaaS products showed an even split between Informational and Discovery (43.8% and 44.8%, respectively), while supplements showed elevated Validation and Review-Seeking rates relative to other verticals — consistent with the health-adjacent nature of supplement purchases.

These intent profiles carry direct implications for citation behavior. In terms of the citation probability framework introduced in Section 3.2, the results indicate that *P*(cite | *Iq*, ***Fs***, *p*) varies substantially as a function of *I_q*: informational queries tend to cite authoritative reference sources such as institutional websites, government pages, and established publications, while discovery queries draw from review aggregators, comparison sites, and direct brand pages. A strategy optimized for one intent type may be counterproductive for another. We return to this point in the Discussion and formalize this comparison in the predictive model (Section 4.6).

4.2 Google Rank Does Not Predict AI Citation

The relationship between Google organic ranking and AI citation was weak overall and varied dramatically by platform. Across 120 queries and 360 Google Top-3 results, Perplexity matched 29.7% of Google's exact URLs in its citations, while ChatGPT matched only 7.8% — a significant difference (χ^2 = 55.47, *p* < .001, Cramér's *V* = 0.278). At the domain level, the gap narrowed but remained substantial: Perplexity matched 33.6% of domains versus ChatGPT's 12.2% (χ^2 = 45.41, *p* < .001, *V* = 0.251).

Table 2. Google Rank vs. AI Citation Match Rates (*N* = 360 Google Top-3 Results)

Metric	Perplexity	ChatGPT	Test Statistic	*p*
Exact URL match	29.7% (107/360)	7.8% (28/360)	χ^2 = 55.47	< .001
Domain match	33.6% (121/360)	12.2% (44/360)	χ^2 = 45.41	< .001
Reddit citations	0	0	—	—

These figures mean that 92.2% of ChatGPT's citations came from pages outside Google's Top-3, and 70.3% of Perplexity's did the same. Position within Google's Top-3 showed no significant effect for ChatGPT ($\chi^2(2) = 0.31$, $p = .856$) and a non-linear pattern for Perplexity, where Position 2 showed the highest match rate (41.7%) compared to Position 1 (22.5%) and Position 3 (25.0%) ($\chi^2(2) = 12.47$, $p = .002$). Neither platform showed significant variation across verticals (Perplexity: $\chi^2(9) = 7.73$, $p = .562$; ChatGPT: $\chi^2(9) = 15.34$, $p = .082$).

Query type mattered considerably. Both platforms showed higher match rates for informational queries than for commercial queries: Perplexity's informational match rate was 41.7% versus 17.8% for commercial queries ($\chi^2 = 23.46$, $p < .001$), and ChatGPT's was 11.1% versus 4.4% ($\chi^2 = 4.69$, $p = .030$). This finding reinforces the role of query intent as a mediating variable — even the modest connection between Google rank and AI citation is itself intent-dependent.

The Reddit Paradox. The most striking result in this study was the complete absence of Reddit citations from both AI platforms' APIs despite Reddit's extraordinary dominance in Google's organic results. Reddit URLs occupied 138 of 360 Google Top-3 positions (38.3%) across our query sample. Yet neither ChatGPT nor Perplexity cited Reddit a single time across their API responses. Under a random citation model — if AI platforms selected from Google's Top-3 with equal probability regardless of domain — the probability of observing zero Reddit citations across Perplexity's 107 URL matches was $p = 3.43 \times 10^{-23}$. For ChatGPT's 28 matches, the probability was $p = 1.32 \times 10^{-6}$. Both platforms systematically exclude Reddit from their API citations, a finding that directly contradicts the widely promoted "Reddit strategy" in GEO practitioner advice.

Important qualification. A companion study (Lee, 2026b) subsequently tested the same platforms through their consumer-facing web UIs rather than APIs, using browser automation to collect citation data across 100 queries. The web UI results diverge dramatically: Google AI Mode cited Reddit in 44% of queries, Perplexity in 20%, and ChatGPT in 17%. Only Claude maintained zero Reddit citations across both access channels. This access-channel divergence means that the Reddit Paradox as documented here is specifically an *API* phenomenon. The practical implications differ depending on which access channel mediates the user's interaction — a distinction we did not anticipate when designing this study. Readers should interpret the zero-citation finding throughout this paper as applying to API-based retrieval; the broader picture of Reddit's role in AI responses is more nuanced than this study alone suggests.

4.3 Platform Architecture: Two Fetch Paradigms

Our server-side logging revealed a fundamental architectural divide among the four AI platforms — a distinction that, to our knowledge, has not been previously reported with direct mechanical evidence.

Table 3. Platform Live-Fetch Behavior

Characteristic	ChatGPT	Claude	Perplexity	Gemini
Live page fetch	Yes	Yes	No	No
User-agent	ChatGPT-User/1.0	Claude-User/1.0	N/A	N/A
Fetch behavior	Proactive	Demand-driven	Index only	Index only
robots.txt check	Not observed	Once per session	N/A	N/A
robots.txt compliance	Not compliant	Compliant	N/A	N/A
Search index source	Bing	Unknown	Proprietary	Google Search

ChatGPT and Claude both send real HTTP GET requests to target servers during conversations, but their behaviors differ meaningfully. ChatGPT fetches proactively: in our tests, it retrieved the requested page and then fetched the site's homepage two seconds later without being asked — a pattern consistent with speculative pre-fetching. ChatGPT's requests used the user-agent string "ChatGPT-User/1.0" and were not preceded by any robots.txt check in our server logs. When directed to fetch a page explicitly blocked by robots.txt for ChatGPT-User, ChatGPT retrieved the page successfully and returned its full content, including all embedded verification facts.

Claude’s architecture is more layered. Its web_search tool queries a search index without contacting target servers — confirmed by zero server hits during dedicated search-only tests. Only the web_fetch tool triggers a real HTTP GET request, and this tool is invoked selectively: when Claude’s existing context (training data, memory, and search snippets) is sufficient to answer a query, web_fetch is never called. This demand-driven architecture explains our initial January misclassification, in which we concluded that Claude did not perform live fetches. Claude had sufficient cached context to answer our test queries without ever needing to contact the server.

Claude’s robots.txt behavior followed a session-level caching model. In our first session, the robots.txt request preceded the first page fetch by 3.0 seconds; in the second session, the gap was 1.1 seconds. Subsequent fetches within the same session did not trigger additional robots.txt checks. When we added a Disallow directive mid-session, Claude continued to fetch the newly blocked path — the cached robots.txt did not reflect the change. However, in a fresh session (tested via the Anthropic API to eliminate memory contamination), Claude checked the updated robots.txt and correctly refused to fetch the blocked path, returning a “no permission” error with zero corresponding server hits.

Perplexity and Gemini showed no evidence of conversational page fetching. No conversational user-agent strings appeared in our server logs or BotSight monitoring data across either the January or February testing periods. Both platforms appear to rely entirely on pre-built search indices — Perplexity on its proprietary index, and Gemini on Google’s internal search infrastructure.

This architectural divide has profound practical implications. For ChatGPT and Claude, the actual HTML content of a page matters — these platforms read the page in real time. For Perplexity and Gemini, only the information captured in search snippets, metadata, and cached index entries matters, because no server contact occurs during conversations. A single optimization strategy cannot serve both paradigms effectively.

4.4 What Makes a Page Citable

Our analysis of 479 pages (241 cited, 238 not cited) across 26 technical features revealed seven features that remained statistically significant after Benjamini-Hochberg FDR correction at $\alpha = .05$.

Table 4. Significant Page-Level Technical Factors (After FDR Correction, *n* = 479)

Feature	Cited	Not Cited	Test	*p*_adj	Effect [95% CI]
Word count (median)	2,582	1,859	U = 34,237	.006	r = –0.194 [–0.279, –0.106]
Schema count (median)	1.0	1.0	U = 33,742	.006	r = –0.177 [–0.262, –0.089]
Canonical is self	84.2%	73.5%	χ^2 (Yates)	.037	OR = 1.92 [1.23, 3.02]
Has schema markup	73.9%	62.6%	χ^2 (Yates)	.047	OR = 1.69 [1.14, 2.49]
Internal links (median)	123	96	U = 32,752	.037	r = –0.142 [–0.229, –0.053]
Total links (median)	164	134	U = 32,771	.037	r = –0.143 [–0.230, –0.054]
Content-to-HTML ratio (median)	0.086	0.065	U = 32,453	.047	r = –0.132 [–0.219, –0.043]

Word count showed the strongest effect among continuous features. Cited pages had a median word count of 2,582 compared to 1,859 for non-cited pages (Mann-Whitney U = 34,237, p_adj = .006, rank-biserial r = –0.194). This represents a meaningful but not overwhelming difference — cited pages were approximately 39% longer at the median. Schema markup was present on 73.9% of cited pages versus 62.6% of non-cited pages (OR = 1.69, 95% CI [1.14, 2.49], p_adj = .047), and pages with self-referencing canonical tags were nearly twice as likely to be cited as those without (OR = 1.92, 95% CI [1.23, 3.02], p_adj = .037).

A note of calibration for practitioners: while these seven features reached statistical significance, their effect sizes are modest. The rank-biserial correlations for continuous features range from r = –0.132 to r = –0.194, and the odds ratios for binary features fall below 2.0. These are associational signals in an observational design — not silver bullets. A page with 2,582 words, schema markup, and a self-referencing canonical tag is somewhat more likely to be cited than one without these features, but many cited pages lacked them and many non-cited pages had them. We caution against treating any single feature as a reliable citation predictor in isolation.

Equally notable were the features that showed *no* significant association with AI citation after multiple comparison correction. Popup and modal elements — frequently cited by practitioners as detrimental to AI visibility — showed no difference between cited and non-cited pages (U = 27,698, p_adj = .606, r = 0.034). Author attribution, commonly recommended as a trust signal, was nonsignificant (χ^2, p_adj = .522). Paywall signals, load time, page size, and affiliate link counts were all nonsignificant. These null results are practically important because they challenge specific pieces of widely circulated GEO advice.

4.5 Supporting Evidence

Recommendation consistency. AI brand recommendations were considerably more consistent within platforms than industry discourse suggests. Across four platforms and 50 queries (each repeated three times), ChatGPT showed the highest within-platform consistency with a mean Jaccard similarity of 0.619 (95% CI [0.537, 0.701]) and a perfect match rate — all three runs returning the identical brand set — of 30%. Claude, Perplexity, and Gemini showed lower but still non-trivial consistency: mean Jaccard scores of 0.316 (95% CI [0.259, 0.380]), 0.331 (95% CI [0.276, 0.391]), and 0.255 (95% CI [0.199, 0.314]), respectively. The difference across platforms was significant (Kruskal-Wallis H = 47.11, p < .001, ε^2 = 0.237). Post-hoc pairwise comparisons with Bonferroni correction confirmed that ChatGPT's consistency was significantly higher than all three other platforms, while the differences among Claude, Perplexity, and Gemini were not significant.

Table 5. Recommendation Consistency Across Platforms (*n* = 50 Queries Per Platform)

Platform	Mean Jaccard	95% CI	Perfect Match	Top-1 Consistent
ChatGPT	0.619	[0.537, 0.701]	30%	70%
Perplexity	0.331	[0.276, 0.391]	2%	40%
Claude	0.316	[0.259, 0.380]	4%	38%
Gemini	0.255	[0.199, 0.314]	2%	48%

The top-1 consistency metric — whether the same brand appeared first across all three runs — was 70% for ChatGPT, 48% for Gemini, 40% for Perplexity, and 38% for Claude. This indicates that even on the less-consistent platforms, the leading recommendation was preserved in roughly two out of five queries.

Cross-platform agreement told a strikingly different story. When we compared the brand sets returned by all four platforms for the same queries, the mean Jaccard similarity was 0.036 (median = 0.026). Pairwise cross-platform comparisons ranged from 0.120 (ChatGPT vs. Perplexity) to 0.182 (Claude vs. Gemini). In practical terms, asking the same question to different AI platforms yields almost entirely different brand recommendations.

Temporal stability. Re-testing 40 queries after five weeks produced a mean cross-temporal Jaccard of 0.307 (median = 0.225, n = 40). A Wilcoxon signed-rank test confirmed that temporal Jaccard values were significantly greater than zero (W = 666, p = 8.32×10^{-8}), indicating that cross-temporal overlap was not attributable to chance. The top-1 brand from the original test was still present in the retest recommendations 65% of the time. Google rank match rates remained largely stable at 82.5% across original and retest observations. These results indicate that while the specific set of recommended brands shifts over time, the leading recommendation and the general source selection pattern show moderate, statistically significant stability.

Content freshness. Source age varied significantly by topic velocity for Perplexity (Kruskal-Wallis H = 21.54, p < .001), with high-velocity topics yielding a median source age of 8.5 days, medium-velocity topics 196 days, and low-velocity topics 530 days. Google showed the same velocity-dependent pattern (H = 15.85, p < .001). Importantly, we found no significant difference between Perplexity and Google source ages within any velocity tier — both engines cited similarly fresh sources for rapidly changing topics and similarly older sources for stable topics.

4.6 Predictive Modeling

To directly address whether intent (*Iq), page features (**Fs***), or their combination best predicts individual page citation, we fit a series of nested logistic regression models estimating P(cite | Iq, $\boldsymbol{Fs}$, p) using the page-level dataset (n = 479 pages, 241 cited, 238 not cited). The binary outcome was citation status (1 = cited by any AI platform, 0 = not cited). We compared four models of increasing complexity:

- **M0** (Intent only): Query type (commercial vs. informational) as the sole predictor.
- **M1** (Page features only): Seven significant page-level features from Table 4 — word count, schema count, internal link count, total link count, content-to-HTML ratio, canonical-is-self, and has-schema-markup.
- **M2** (Page features + Intent): All M1 features plus query type.
- **M3** (Page features + Intent + Domain): All M2 features plus query domain (10 verticals).

All continuous features were standardized prior to fitting. Models used L2 regularization for cross-validation and no regularization for full-sample coefficient estimation. We evaluated predictive performance using 5-fold stratified cross-validation with AUC-ROC as the primary metric.

Table 6. Nested Model Comparison for Page-Level Citation Prediction (*n* = 479)

Model	Predictors	AUC (5-fold CV)	SD	McFadden R^2	AIC	BIC
M0	Intent only	0.462	0.075	0.001	667.6	675.9
M1	Page features	0.594	0.048	0.033	657.8	691.2
M2	Features + Intent	0.578	0.064	0.034	659.8	697.3
M3	Features + Intent + Domain	0.539	0.053	0.048	668.1	743.1

The results reveal a clear pattern. Query intent alone (M0) performed below chance (AUC = 0.462), indicating that the commercial/informational distinction has virtually no predictive power for individual page citation. Page-level technical features (M1) were the strongest predictors, achieving AUC = 0.594 — modest but significantly above chance and significantly above M0 (paired t-test on fold AUCs: t = 7.66, p = .002). Adding intent to page features (M2) did not improve prediction: the likelihood ratio test was nonsignificant (LR = 0.08, df = 1, p = .78), and the cross-validated AUC was numerically lower (0.578 vs. 0.594). Adding query domain (M3) further degraded cross-validated performance (AUC = 0.539), suggesting overfitting with the additional 9 domain parameters.

Among page features, internal link count showed the largest standardized coefficient (β = 0.73, OR = 2.07), followed by content-to-HTML ratio (β = 0.25, OR = 1.29), schema count (β = 0.19, OR = 1.21), and canonical-is-self (β = 0.19, OR = 1.21). Total link count had a negative coefficient (β = –0.76, OR = 0.47), likely reflecting collinearity with internal links — pages with many links but few internal ones (i.e., heavy external linking) were less likely to be cited.

Interpreting the tension with our broader findings. These results may appear to contradict our earlier finding that intent is the strongest predictor of citation behavior. The apparent tension resolves when we distinguish between two levels of analysis. The associational evidence from Study 1 — where intent predicted *citation source type distributions* with Cramér's $V = 0.258$ across five intent categories and eight verticals — operates at the aggregate query level. It shows that informational queries lead to qualitatively different citation patterns (e.g., more authoritative institutional sources) than commercial queries (e.g., more review and comparison sites). The predictive model here operates at the *individual page level*, asking whether a specific page was cited given its technical features and the query type. At this level, intent does not help distinguish cited from non-cited pages — but page features do, modestly. Both findings are valid and complementary: intent shapes *what kinds of sources* AI platforms prefer, while page-level features influence *which specific pages* within those categories are selected. This two-level distinction has important practical implications. Content strategists should first consider whether their content matches the intent profile that AI platforms seek for a given query category, and then optimize page-level technical features to improve selection probability within that intent-appropriate pool.

---

5. Discussion

5.1 Challenging Industry Narratives

Our results directly challenge four widely circulated claims in the GEO practitioner community.

"Google rank equals AI visibility." This claim assumes that optimizing for traditional search engine rankings automatically yields AI citation. Our data demonstrates otherwise. ChatGPT cited Google Top-3 URLs only 7.8% of the time, meaning that over 92% of its citations came from pages outside Google's highest-ranked results. Even Perplexity — which showed the strongest connection to Google rankings among the platforms we tested — matched only 29.7% of Top-3 URLs. Furthermore, this already modest connection was itself intent-dependent: informational queries showed higher match rates than commercial queries for both platforms. Google rank is not irrelevant to AI citation, but it is far from the primary determinant that practitioners assume it to be.

"AI prefers community consensus." The recommendation to "post on Reddit because AI prefers user-generated content" is perhaps the most widely promoted piece of GEO advice. Our Reddit Paradox finding appeared to demolish this claim: Reddit occupied 38.3% of Google's Top-3 positions across our sample yet received exactly zero citations from either ChatGPT or Perplexity's APIs (binomial $p = 3.43 \times 10^{-23}$). However, a companion study (Lee, 2026b) has since demonstrated that this exclusion is access-channel specific. When the same platforms are queried through their consumer-facing web UIs, Reddit citations appear at rates of 17–44% — with validation-intent queries ("best X for Y") producing Reddit citation rates as high as 71% on Google AI Mode and 46% on Perplexity. The community consensus claim is therefore not demolished but *bifurcated*: AI APIs systematically exclude Reddit, while AI web UIs actively cite it — particularly for the query types where community consensus is most relevant.

This bifurcation narrows the scope of the mechanistic explanations we can offer. The domain-blocklist hypothesis and content-quality filter hypothesis remain viable for API pipelines specifically, but cannot explain the overall pattern — the same platforms that exclude Reddit via API include it via web UI. The most parsimonious explanation is that API and web UI retrieval systems use different source filtering configurations, possibly reflecting different licensing arrangements, different retrieval-augmented generation (RAG) pipelines, or different product decisions about source diversity. The training-time hypothesis remains relevant for a distinct channel of influence: a companion correlation analysis (Lee, 2026b) found that

Reddit's brand consensus rankings correlate strongly with AI brand recommendations (ρ = .554), suggesting that Reddit shapes model weights even when it is not cited.

This finding is consistent with Filandrianos et al.'s (2025) observation that social proof biases LLM *processing* — and the web UI data suggests that these processing biases *do* translate into citation behavior, but only through certain access channels. The relationship between Reddit and AI citation is more complex than either the "Reddit strategy" advocates or our original API-only data suggested.

"AI is random — don't bother optimizing." Our consistency data contradicts the claim that AI recommendations are chaotically unpredictable. ChatGPT's mean Jaccard similarity of 0.619 and 70% top-1 consistency rate indicate that the same brand is recommended first in seven out of ten queries. Even on the less-consistent platforms, top-1 retention hovered around 40–48%. The pattern is consistent enough to reward systematic optimization — particularly for securing the top recommendation position, which showed the highest stability across all platforms. That said, the near-zero cross-platform overlap (Jaccard = 0.036) means that a brand recommended by ChatGPT is unlikely to also be recommended by Claude or Gemini for the same query. Consistency exists *within* platforms but not *across* them.

"All AI platforms work the same." The 2-vs-2 architectural split we documented — ChatGPT and Claude performing live page fetches while Perplexity and Gemini rely on pre-built indices — represents a fundamental structural difference with direct optimization implications. For the fetching platforms, actual page content matters: word count, content structure, schema markup, and server-side rendering all directly influence what the platform reads and can cite. For the index-only platforms, the page itself is never read during conversations — only its representation in the search index matters. A one-size-fits-all optimization strategy that ignores this divide will necessarily underperform a platform-aware approach.

Connecting architecture to citation patterns. The architectural divide provides a mechanistic foundation for several of our other findings, though the causal chains remain partially hypothesized. First, the relevance of page-level technical features (Section 4.4, 4.6) may be architecture-dependent: live-fetch platforms like ChatGPT can directly evaluate page content — word count, schema markup, content structure — and incorporate these signals into citation decisions, whereas index-only platforms like Perplexity and Gemini rely on whatever representation their search index provider (Bing for Perplexity, Google's internal index for Gemini) has extracted and stored. The page-level associations we observed in Table 4 likely reflect the disproportionate influence of the fetching platforms in our sample. Second, the Reddit Paradox may have distinct explanations across the architectural divide: index-only platforms may exclude Reddit because their index providers filter it at the indexing stage (perhaps due to licensing considerations), while live-fetch platforms could theoretically retrieve Reddit pages but may apply content-quality filters during the LLM's internal query rewriting step — what Ma et al. (2023) describe as the reformulation phase where the model decides which sources merit citation. Third, the intent-dependence of Google rank match rates (Section 4.2) may reflect a deeper interaction: for informational queries, both Google and AI platforms may converge on the same authoritative institutional sources through independent quality signals, whereas for commercial queries, AI platforms may apply internal preference heuristics — favoring comparison sites, official product pages, or expert reviews — that diverge from Google's ranking algorithm. These hypothesized causal chains are testable: future work could compare the page features of AI-cited pages separately for each platform to determine whether the associations in Table 4 are driven primarily by the fetching platforms, and could probe whether index source (Bing vs. Google vs. proprietary) correlates with the types of sources cited.

5.2 Implications for Practitioners

These findings suggest a fundamentally different approach to GEO than what is currently advocated. We propose an intent-aware, platform-specific framework.

Intent-first strategy. Rather than optimizing uniformly for "AI visibility," practitioners should begin by identifying the dominant intent types in their vertical. An agency or law firm — where Discovery queries dominate — should optimize for appearing in recommendation contexts, with detailed service descriptions, comparison-ready content, and clear differentiation. A SaaS brand — where Informational queries dominate — should prioritize authoritative educational content, technical documentation, and expertise signals.

Platform-specific optimization. For ChatGPT and Claude, content depth and structure matter directly because these platforms read the actual HTML. Longer, well-structured pages with schema markup and proper content-to-HTML ratios showed statistically significant associations with citation (Table 4). Server-side rendering is essential for Claude, which cannot execute JavaScript — our tests showed that JS-rendered pages returned only the HTML shell. For Perplexity and Gemini, optimization should focus on meta descriptions, title tags, and the information captured in search index snippets, since no direct page contact occurs.

The Reddit strategy — it depends on the channel. Our API data showed zero Reddit citations across all 120 queries, 10 verticals, and both platforms tested — which initially suggested that Reddit presence was irrelevant for GEO. However, companion research (Lee, 2026b) using browser automation to test the same platforms through their web UIs found Reddit citation rates of 17–44%, with validation queries producing rates as high as 71% on Google AI Mode. This means Reddit's value as a GEO strategy is access-channel dependent. For practitioners building applications on AI APIs (chatbots, automated workflows, programmatic content), Reddit investment yields no citation return. For practitioners seeking visibility in consumer-facing AI interfaces — where the majority of end users interact — Reddit presence can be a significant source of AI citations, particularly for recommendation and comparison queries. The practical advice is therefore nuanced: Reddit is not universally misallocated, but its value is confined to the web UI channel and to specific query intent types.

5.3 Implications for Researchers

Query intent as a required variable. Our finding that intent distribution varies significantly by vertical (Cramér's $V = 0.258$) and mediates the relationship between Google rank and AI citation suggests that future GEO research should control for query intent. Our predictive modeling (Section 4.6) further revealed that intent operates at the aggregate level — shaping source type distributions — rather than at the page level, where technical features dominate. This two-level distinction means that studies aggregating across intent types risk producing results that are artifacts of their query distribution rather than genuine platform behaviors.

Server-side verification as a methodological standard. Our discovery that some platforms cite pages they never fetched — and that one platform's memory system can fabricate fetch results — underscores the need for server-side verification in AI retrieval research. Observing what an AI platform *claims* to have retrieved is insufficient; researchers must independently verify whether server contact occurred. We encountered this limitation directly when Claude's memory system presented cached content from a prior conversation as live fetch results in a fresh session, producing zero server hits despite claiming successful retrieval. Only server-side logs revealed the discrepancy.

Cross-platform comparison design. The near-zero cross-platform overlap (Jaccard = 0.036) we observed means that findings from single-platform studies may not generalize. A result demonstrated on ChatGPT cannot be assumed to hold for Claude, Perplexity, or Gemini. Multi-platform designs should be the norm in GEO research, not the exception.

5.4 Limitations

We acknowledge several limitations that qualify our findings and suggest directions for future work.

Non-unified study timing. Our experiments were conducted across a two-month window (January–February 2026) as a series of related investigations rather than a single pre-registered study. While this allowed us to iterate and correct methodological errors — such as the initial misclassification of Claude's fetch behavior — it means that platform behavior may have changed between earlier and later experiments. A pre-registered replication with unified timing would strengthen these conclusions.

Google rank analysis platform coverage. The finding that Google rank does not predict AI citation (Section 4.2) is directly demonstrated only for ChatGPT and Perplexity. Claude and Gemini were excluded from this analysis due to Claude's high commercial query refusal rate and Gemini's inconsistent citation format during the study period. We infer that the finding likely extends to these platforms based on their architectural differences — Claude’s demand-driven fetch and Gemini’s reliance on Google’s internal search index suggest different retrieval pathways than Google’s organic ranking algorithm — but this inference has not been directly measured. A replication with manual citation extraction from Claude and Gemini would close this gap.

E-commerce weighting. Our query intent corpus of 19,556 queries was drawn primarily from commercial verticals, which may not represent the full distribution of queries posed to AI chatbots. Users asking about science, history, current events, or other non-commercial topics may encounter different citation patterns. Our cross-domain validation across SaaS, healthcare, local services, and finance provides some evidence of generalizability, but broader domain coverage would increase confidence.

Platform behavior changes. AI platforms update their models and retrieval systems frequently. Our findings represent a snapshot of platform behavior as of January–February 2026. ChatGPT’s robots.txt non-compliance, Claude’s demand-driven fetch architecture, Perplexity’s index-only approach, and Gemini’s reliance on Google’s internal search may all change with future updates. Longitudinal monitoring studies would help establish which of our findings reflect stable architectural choices versus transient implementation details.

Observational design and intent confounding. Our study is observational, not experimental. We can report that query intent, content depth, and schema markup are *associated* with AI citation, but we cannot establish causal relationships. It is possible that these features correlate with unmeasured confounders — for example, that longer pages are also better written, and writing quality rather than length drives citation. More specifically, the page-level analysis (Study 4) compares cited versus non-cited pages without stratifying by query intent or vertical. If certain intent categories both favor longer pages and generate more citations, the observed word count association may partly reflect intent-driven selection effects rather than a direct page-level signal. We attempted to address this through the predictive model in Section 4.6, which includes both intent and page features simultaneously, but a fully stratified analysis with a larger, intent-annotated page sample remains a priority for future work. Interventional studies that manipulate individual features while controlling others would be needed to establish causation.

Live-fetch sample size. Our server-side verification of platform fetch behavior (Study 3) rests on 13 confirmed page fetches across two platforms. This sample is sufficient to establish the *existence* of live-fetch behavior — a single confirmed fetch with server-side log correlation refutes the null hypothesis that a platform never contacts target servers. However, the robots.txt compliance conclusions are more narrowly grounded: Claude’s compliance was confirmed through one API-based test with a control and a blocked path, and ChatGPT’s non-compliance through one analogous test. While the results were unambiguous (Claude refused the blocked path; ChatGPT fetched it), replication across a broader set of robots.txt configurations — including partial Disallow rules, crawl-delay directives, and varying user-agent specificity — would strengthen these conclusions considerably.

Other sample size constraints. Beyond the live-fetch study, individual sub-analyses have smaller samples than the aggregate numbers suggest. The Google rank study used 120 queries; the temporal retest included 40 queries. Additionally, Claude declined to provide product recommendations for approximately 45% of queries in our consistency experiment, reducing the effective sample size for that platform. Larger-scale replications would increase statistical power for detecting smaller effects.

Claude’s commercial query refusal rate. Claude declined to provide product recommendations for approximately 45% of commercial queries in our consistency experiment — a rate far exceeding that of any other platform tested. This is not merely a sample size issue; it represents a fundamentally different platform behavior that shapes the observable citation landscape. Claude’s refusal likely reflects a deliberate design choice around commercial safety or neutrality, but it creates a selection effect: the queries Claude *does* answer may not be representative of the full commercial query space. For practitioners, this suggests that Claude’s citation behavior may be less relevant for commercial GEO than the other three platforms, though potentially more relevant for informational queries where it responds freely. For researchers, Claude’s refusal behavior is itself a variable worth studying — the conditions under which it engages versus refuses may reveal latent intent categories or risk thresholds that influence citation selection when it does respond.

Memory contamination. Our discovery that Claude's memory system can fabricate fetch results — presenting memorized content from prior conversations as live retrievals without disclosure — represents a challenge for any research using Claude's conversational interface. We addressed this by using the stateless API for our critical robots.txt compliance test, but researchers should be aware of this behavior and design accordingly.

Reproducibility and platform versioning. Our experiments were conducted during January–February 2026 using the following platform access methods and model versions: ChatGPT via the OpenAI Chat Completions API (model: `gpt-4o`), Claude via the Anthropic Messages API (model: `claude-3-5-sonnet-20241022`), Perplexity via the Perplexity API (model: `sonar`), and Gemini via the Google Generative AI API (model: `gemini-2.0-flash`). The intent classifier used GPT-4o via the OpenAI API. All API interactions used default parameters except where noted (e.g., temperature 0.1 for classification consistency). Google rank data was collected via the Google Custom Search JSON API. Server-side fetch verification used Vercel middleware logging on a Next.js application and BotSight analytics. Because all four AI platforms update their models and retrieval systems without advance notice — often multiple times per month — our findings represent a snapshot of behavior at the time of data collection. We recommend that future researchers record the specific model version strings, API endpoints, and dates of data collection to enable meaningful comparison across studies. The temporal stability test (Section 4.5) provides some evidence that citation patterns persist across at least a five-week window (W = 666, p < .001), but longer-term monitoring is needed to distinguish findings that reflect stable architectural design choices from those that may shift with model updates.

Intent classification validation. The LLM-based intent classifier is a central methodological component of this study, and its quality directly affects the validity of intent-related findings. Our validation was limited to a single reviewer's manual assessment of 200 queries rather than formal inter-rater reliability testing with multiple independent annotators. While the 91.5% agreement rate and the pattern of disagreements (concentrated at adjacent category boundaries) suggest reasonable classification quality, we cannot rule out systematic biases — for example, if the LLM and the human reviewer share similar biases about intent categorization. Additionally, the five-category taxonomy was developed for this study and has not been independently validated against existing intent classification benchmarks. Future work should employ multiple independent raters to compute Cohen's κ or Fleiss' κ, and should test the taxonomy's discriminant validity by comparing it to established search intent frameworks.

6. Conclusion

We set out to test three foundational claims that dominate current GEO practitioner advice: that Google rank determines AI visibility, that community consensus platforms confer citation advantages, and that AI recommendations are too inconsistent to optimize for. Our data contradicts all three.

Google ranking showed weak and inconsistent predictive power for AI citation. ChatGPT cited Google Top-3 URLs only 7.8% of the time. Reddit — despite occupying 38.3% of Google's Top-3 positions — received exactly zero AI citations via API, a finding so statistically extreme (binomial p = 3.43 × 10^{-23}) that it cannot be attributed to sampling variation. However, companion research (Lee, 2026b) has since shown that this exclusion is access-channel specific: web UIs of the same platforms cite Reddit at rates of 17–44%, meaning that the Reddit Paradox is an API phenomenon rather than a universal property of AI citation behavior. AI recommendations were substantially more consistent than claimed, particularly on ChatGPT (mean Jaccard = 0.619, top-1 consistency = 70%), though cross-platform agreement was near-random (all-four Jaccard = 0.036).

AI citation behavior operates at two distinct levels, and the strongest predictors differ between them. At the aggregate level, query intent and platform architecture are the dominant forces shaping *what kinds of sources* are cited — intent distributions vary dramatically by vertical, explaining why optimization strategies that work in one industry may fail in another. At the individual page level, technical features — content depth, schema markup, internal linking structure, and content-to-HTML ratio — are the strongest predictors of *which specific pages* are selected (AUC = 0.594), while intent adds no additional predictive power (likelihood ratio p = .78). The 2-vs-2 architectural split between fetching platforms (ChatGPT, Claude) and index-only platforms (Perplexity, Gemini) means that the same page presents entirely different information to different AI platforms. Commonly cited optimization factors like popup elements, author attribution, and page load time showed no significant association with citation.

We contend that the GEO field needs to move from a one-size-fits-all paradigm — "optimize the same way for all AI platforms" — to an intent-aware, platform-specific framework. The advice to "rank on Google and AI will follow" is not merely incomplete; for 92% of ChatGPT's citations and 70% of Perplexity's, it is empirically wrong. The advice to "post on Reddit for AI visibility" is contradicted by API data but supported by web UI data (Lee, 2026b), making it access-channel dependent rather than categorically wrong. And the claim that "AI is too random to optimize for" underestimates the consistency that exists — particularly for the top recommendation position.

Future work should pursue three priorities. First, longitudinal studies tracking how AI citation behavior evolves with platform updates would help distinguish our findings that reflect stable architectural design from those that may shift with model changes. Second, interventional experiments that manipulate individual page features while controlling others would establish the causal relationships that our observational design can only suggest. Third, extending query intent analysis beyond commercial verticals to medical, scientific, legal, and educational domains would test whether the intent-driven citation patterns we observed generalize across the full spectrum of AI-mediated search.

The businesses and practitioners seeking to influence AI citation behavior face a complex and rapidly evolving landscape. We offer this work not as a definitive map, but as an empirically grounded starting point — one that replaces assumption with measurement and intuition with data.

Author Note

This paper was co-developed by Anthony Lee and Claude (Anthropic, Claude Opus 4.6 and Claude Sonnet 4.5). The human author provided the research hypothesis, experimental design, data collection infrastructure (including API integrations, Vercel middleware logging, BotSight monitoring, and SQLite database architecture), all manual platform tests (Experiment A), and editorial direction. The AI contributed literature synthesis, statistical analysis, experimental automation scripts, data processing pipelines, and prose drafting. Both authors iteratively refined the arguments through collaborative dialogue across multiple sessions, with the human author retaining final editorial authority over all claims and interpretations.

The collaborative nature of this research is not incidental to its subject matter. We studied AI citation behavior using an AI system as a research tool — a circumstance that required particular attention to methodological independence. The potential conflict of interest is most acute for Study 3, where Claude's own fetch and robots.txt compliance behavior was under examination. We addressed this through two design choices. First, the critical robots.txt compliance test was conducted via the stateless Anthropic Messages API rather than the conversational interface, eliminating the possibility that Claude's memory system could influence results. Second, all server-side observations were independently verifiable through Vercel function logs — the AI contributor could not influence what appeared in those logs. We report both favorable findings (Claude's robots.txt compliance) and unfavorable ones (Claude's memory contamination behavior, Claude's lower consistency scores relative to ChatGPT) without selective emphasis, and we invite replication of all findings by independent researchers.

References

Aggarwal, P., Murahari, V., Rajpurohit, T., Kalyan, A., Narasimhan, K., & Deshpande, A. (2024). GEO: Generative Engine Optimization. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (KDD 2024). https://doi.org/10.1145/3637528.3671882

Bagga, A., Stein, D., Subramanian, S., & Hsu, D. (2025). E-GEO: A testbed for Generative Engine Optimization in e-commerce. *arXiv preprint*. arXiv:2511.20867

Cloudflare. (2025). AI bot traffic analysis reports. Retrieved February 14, 2026, from https://blog.cloudflare.com

Filandrianos, G., Konstantopoulos, I., Pantazis, O., Tsimpoukelli, E., & Panagiotakos, D. (2025). Bias beware: Cognitive biases on LLM-driven product recommendations. *arXiv preprint*. arXiv:2502.01349

Kumar, A., & Lakkaraju, H. (2024). Manipulating large language models to increase product visibility. *arXiv preprint*. arXiv:2404.07981

Lee, A. (2026b). Reddit doesn’t get cited (through the API): Training data influence, access-channel divergence, and the shadow corpus in AI brand recommendations. *Preprint*.

Ma, X., Gong, Y., He, P., Zhao, H., & Duan, N. (2023). Query rewriting in retrieval-augmented large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (EMNLP 2023).

Search Atlas. (2025). Comparative analysis of LLM citation behavior. Retrieved February 14, 2026, from https://searchatlas.com

Webber, W., Moffat, A., & Zobel, J. (2010). A similarity measure for indefinite rankings. *ACM Transactions on Information Systems*, 28(4), Article 20. https://doi.org/10.1145/1852102.1852106

Venkit, P. N., Eswaran, K., Gupta, U., Eslami, M., & Kumaraguru, P. (2025). Search engines in an AI era: The false promise of factual and verifiable source-cited responses. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency* (FAccT 2025).

Wu, J., Luo, Y., Xu, Z., Pan, L., & Peng, N. (2025). SourceCheckup: An automated framework for evaluating LLM citation support. *Nature Communications*, 16, Article 3615. https://doi.org/10.1038/s41467-025-58551-6