The SEO Floor: Measuring Google Rank Distribution of AI-Cited Pages

Anthony Lee aiplusautomation.com anthony@anthonylee991.com

Pre-registration: OSF DOI 10.17605/OSF.IO/FMSRD | Project osf.io/w76y8 | Filed 2026-04-20
Code & data: Zenodo DOI 10.5281/zenodo.19787328 | git tag `study-a-v1.0.2` **Date:** 2026-04-25

Abstract

Whether Generative Engine Optimization (GEO) is a discipline distinct from Search Engine Optimization (SEO), or merely SEO repackaged, has been debated since the rise of consumer-facing AI chat platforms. Empirical resolution requires measuring two things: where on Google's search results AI-cited pages actually rank, and whether content features pre-registered as "GEO levers" predict citation independently of Google rank. We collected 100,411 AI citation events from four production AI platforms (ChatGPT, Perplexity, Claude, Google AI Mode) across 2,000 user queries, and assembled a comparison pool of 165,661 unique URLs from Google top-100 SERPs for the same queries. Combining the citation corpus with the comparison pool yields 114,729 (URL, query) observations for a mixed-effects logistic regression of citation probability on Google rank tier and GEO composite score, with query random intercepts. **Headline finding 1**: While 75.4% of citation events are aggregated to pages outside Google's top 30, *per-page* citation odds span a 34× range across rank tiers — a top-3 page is **7.82× more likely to be cited (OR vs Tier 3, 95% CI 7.28–8.39)** while a rank 31–100 page is **4× less likely (OR=0.23, 95% CI 0.22–0.24)**. The aggregate "75% deep-tier" figure is a denominator artifact: the internet beyond rank 30 has vastly more pages than the top 30, so even at low per-page citation probability the absolute count of deep-tier citations dominates. The two figures must be reported together. **Headline finding 2**: A pre-registered GEO composite of seven content features adds small but real predictive power above Google rank (Z-sum OR=1.06; PCA-1 OR=1.15 per 1 SD), driven primarily by schema markup (OR=1.31). Statistics density and list structure show small positive effects after methodological correction; heading density shows a small negative effect. **Headline finding 3**: The 75% Tier-4+ aggregate is overwhelmingly composed of URLs Google ranks beyond #100 (90% of "Tier 4" events) — not the 31–100 band that the H2 regression speaks to — and these deep-tier citations are 77% one-hit-wonders cited by a single AI platform, with sharp platform divergence on user-generated-content tolerance (Claude 0.6% UGC in deep tier; Perplexity 24%). The Lily-Ray-aligned framing ("AI citation is gated by Google ranking") is empirically supported, with schema markup emerging as the strongest single content-feature predictor inside the gate; whether AI parsers consume schema directly or schema is a proxy for site quality is unresolved by observational data and is the target of a planned follow-up interventional study.

---

1. Introduction

The Generative Engine Optimization (GEO) literature emerged in 2024–2025 as researchers and practitioners began documenting content features associated with AI citation. Aggarwal et al. (2024) [1] published a seminal interventional study using a custom generative engine, identifying statistics density, citations, quotations, and authoritative tone as features that increase visibility in source-

attributed AI responses. Subsequent industry analyses [2][3] reported descriptive associations between content properties and AI citation in production platforms.

Critics from the SEO community, most prominently Lily Ray, argued these findings reflect Berkson's paradox: cited pages are disproportionately top-ranking publisher pages, so any feature common in top-ranking publisher content will appear "associated" with citation regardless of causal direction. Ray's position, articulated in public commentary and direct correspondence with the present author, holds that "indexation is a gate, ranking is the differentiator, GEO is what already-ranking pages do well within tier."

The empirical question this debate hinges on has not been answered at scale with a methodologically sound design. Specifically, prior GEO research suffers from one or both of:

- **Cited-only sampling**: studying features of cited pages without a comparison group of comparable uncited pages from the same query SERPs (collider bias on the citation outcome);
- **No rank measurement**: documenting feature associations without recording where the cited pages actually rank in Google for the eliciting queries.

Study A fills this gap. We pre-register a design that:

1. Samples citations from four production AI platforms across 2,000 real user queries spanning 14 verticals (Section 4.1).
2. Constructs a comparison pool by pulling Google's top-100 SERPs for every query (Section 4.3), enabling a true cited-vs-uncited regression.
3. Assigns each cited URL a Google rank tier via SERP join, `site:URL` lookups, and CommonCrawl verification (Section 4.4).
4. Computes seven GEO composite features using extractors imported verbatim from prior research codebases (Section 4.5).
5. Estimates the independent contribution of rank tier and GEO score to citation probability using a mixed-effects logistic regression (Section 4.7).

The pre-registered hypotheses (H1: tier distribution; H2: cited-vs-uncited regression; H3: Tier 5 persistence) are framed as estimation objectives with confidence intervals, not as null-hypothesis tests with rejection regions. The study is publishable regardless of where the point estimates fall. This paper reports those estimates.

2. Prior work

The Princeton GEO benchmark. Aggarwal et al. [1] used a custom generative engine and a small set of content interventions on a sample of pages, reporting that statistics, citations, quotations, and authoritative tone increased visibility weights. The interventional design eliminates the collider-bias concern at the cost of measuring against a synthetic generative engine rather than production platforms (ChatGPT, Perplexity, Claude, Google AI Mode).

Practitioner observational studies. Seer Interactive's 87% Bing-match analysis [2] reports a high overlap between AI-cited URLs and Bing top-20, but conflates URL- and domain-level matching and samples only top-of-SERP. SurferSEO's 36M AI Overviews dataset [3] focuses exclusively on Google AI Overviews and does not stratify by Google rank tier.

Prior internal work. This author's Experiment M (n≈5,000 cited pages, December 2025) [4] reported within-cited associations: cited pages have 2× as many H3 subheadings and 7× the inline statistics

density of uncited peers. Experiment J v2 [5] reported a marginal SEO-vs-GEO contribution analysis on the same corpus (GEO composite added ≈ 4 percentage points of AUC over SEO features alone in a random forest classifier). Both used cited-pages-only sampling; Study A's contribution over Experiment M is the comparison-pool design that explicitly addresses the Berkson critique.

3. Hypotheses

Pre-registered at OSF DOI 10.17605/OSF.IO/FMSRD on 2026-04-20. All three are estimation objectives, not null-hypothesis tests.

H1 (primary, descriptive): Estimate the proportion of AI citation events originating from pages ranking in each of five SEO tiers (1–3, 4–10, 11–30, 31–100, not-indexed). Reported with 95% clustered confidence intervals overall and per platform.

H2 (primary, predictive): Among URLs in the comparison pool (Google top-100 across the 2,000 queries, with confirmed Tier 5 cited URLs appended), estimate the independent effect of (a) Google rank tier and (b) GEO composite score on citation probability using mixed-effects logistic regression with query random intercepts and vertical fixed effects. Tier 5 is excluded from the regression for lack of a bounded uncited denominator (Addendum to v2.1).

H3 (primary, descriptive): Among pages cited by AI but not indexed in Google (true Tier 5, CommonCrawl-verified as on the open web), estimate the proportion that appear in citations within the most recent 30 days of the corpus. 95% CI on the within-window citation rate.

4. Methods

4.1 Citation corpus

Source files were pooled from the GEO_tests research repository covering UI-scraped citations and VPS-captured citation events. The VPS snapshot (2026-04-18) contained 27,769 Vercel-middleware-captured citation events from 2,705 production scrape sessions across the four target platforms. The local UI-scrape corpus contributed 105 JSONL files spanning 14 experiment directories. Study 3 (OpenAI Responses API fan-out data) and Gemini-API grounding-redirect data were excluded a priori as API-only artifacts (Addenda 2 and 3 of the pre-registration log). After post-hoc rescue of three schema-mismatched ecommerce files (Addendum 3), the analytic corpus comprises **100,411 citation events** (94,384 with usable Google-rank tier assignments).

The temporal scope as captured is 2026-02-04 through 2026-04-17 (~10 weeks), narrower than the "Q3 2025 through April 2026" framing in the filed protocol, which was an author error inherited from program-level documentation (Addendum 1). The 90-day primary window pre-registered for sensitivity captures 99.96% of the corpus, making the primary-vs-robustness window distinction effectively cosmetic in this dataset.

4.2 Query reformulation (Method A)

User queries in the corpus are conversational and situation-first (median 18 words, often containing first-person pronouns and personal context). Searching Google with such queries literally measures Google's weak conversational handling rather than how cited pages rank for the underlying intent. We therefore reformulate every user query with Gemini 3.1 Flash Lite into a keyword-intent search version

(Method A), with deterministic temperature=0 for reproducibility. The original user query is retained on every record.

4.3 Comparison pool

For each of the 2,000 keyword-reformulated queries, we pulled Google's top-100 and Bing's top-100 SERPs via Thordata's SERP API. Google produced 213,814 SERP rows; Bing's actual SERP depth caps at ~20 results regardless of pagination (verified on both Thordata and SerpApi), yielding a smaller Bing pool. Every URL in the Google top-100 becomes an observation in the comparison pool — regardless of whether it was cited.

URLs cited by AI but not present in our Google top-100 SERP pull are appended to the pool with their tier assignment from `site:URL` Thordata lookups (~1,000 lookups) and CommonCrawl verification.

4.4 SEO tier classification

Each (URL, query) observation is assigned a tier:

Tier	Definition	Source
1	Google rank 1–3	SERP join
2	Google rank 4–10	SERP join
3	Google rank 11–30	SERP join
4	Google rank 31–100	SERP join
5	Not indexed	`site:URL` returns zero AND CommonCrawl confirms URL exists

URLs in the citation corpus that are not in our Google top-100 but for which `site:URL` returns positive results are documented as a "ranked but >100" subgroup. Per the pre-reg, this subgroup is included in descriptive statistics for H1 (counted as Tier 4 in the H1 distribution) but excluded from H2 regression because there is no bounded uncited denominator for the >100 region (Addendum to v2.1).

4.5 GEO composite features

Seven query-independent features per protocol §5.2 plus an answer-first coverage feature computed per (URL, query) pair:

Feature	Source extractor	Operationalization
Statistics density	Princeton `count_statistics`	Stats-pattern matches per 1k words
Heading density	Princeton `structured_formatting`	(H2 + H3) per 1k words
Answer-first coverage	Experiment M `extract_semantic_features`	Query content-words found in first 200 words / query content-word count
Comparison signals	Experiment M `extract_content_classification`	"vs / versus / compared to" pattern density per 1k words
Primary-source score	Experiment M `extract_content_role_features`	`stats + technical − aggregator_score` (Princeton-derived)
Schema presence	Crawler boolean flags	Sum of has-Article / has-FAQ / has-Product / has-Review / has-Organization
List structure	Princeton `structured_formatting`	(Bullets + numbered + dashes) per 1k words

All extractors are imported directly from the source codebases (`princeton_replication/step1_extract_princeton_features.py`, `experiment_M/extract_features.py`) — no feature is reimplemented in this study's code. (Note: the first H2 run did silently reimplement these, producing flawed coefficients on `stats_density` and `list_structure`. The error was caught on review and corrected; the side-by-side comparison and rationale are documented in Addendum 9 of the pre-registration log.)

Feature crawl: every URL in the Google top-100 SERP pool plus every cited URL outside top-100 was crawled via Playwright (n=165,661 unique URLs queued). 115,248 URLs yielded sufficient text after cleaning for feature extraction.

Composite construction (per protocol §5.2):

- **Z-sum**: each feature Z-scored across the analytic cohort, then summed equal-weighted (primary composite, comparable to Experiment J v2).
- **PCA-1**: first principal component of the seven Z-scored features.
- **Cronbach's α**: reported as internal-consistency check.

If Z-sum and PCA-1 H2 coefficients differ by >0.5 SE, the discrepancy is reported in main results.

4.6 Method B validation

For records with captured AI fan-out sub-queries, we run a parallel Method B analysis: assign tier based on the best Google rank across the platform's actual fan-out queries rather than the Method A reformulated keyword. Cohen's κ between Method A and Method B tier assignments serves as construct validity.

Method B was performed on ChatGPT (via OpenAI Responses API, `gpt-5.4-mini`, N=200) and Gemini (via Google GenAI API with search grounding, `gemini-3.1-flash-lite-preview`, N=200, serving as proxy for Google AI Mode). Perplexity was dropped from κ validation (Addendum 4) due to a mid-study Perplexity Pro rate-limit change and a confirmed scraper channel gap. Claude was excluded a priori (uncappable web_search cache).

Pre-registered decision rule: $\kappa \geq 0.85 \rightarrow$ Method A validated; $0.70 \leq \kappa < 0.85 \rightarrow$ partial validation; $\kappa < 0.70 \rightarrow$ Method A invalidated, restrict to fan-out subset. We deviate from the strict rule for ChatGPT

(Addendum 5) — see Section 5.3.

4.7 Statistical model

H2 primary specification:

```
cited_ij ~ Bernoulli(logit⁻¹(
    β₀ + β₁ · SEO_tier_i + β₂ · GEO_score_i + β₃ · vertical + u_query_j
))
```

where `i` indexes the (URL, query) observation and `j` indexes the query. SEO tier enters as 4-level categorical (Tier 5 excluded) with Tier 3 (rank 11–30) as reference. `u_query` is a query-level random intercept. Estimation: `lme4::glmer` (R 4.5.3) with Laplace ML and bobyqa optimizer; `optimx::nlminb` fallback on convergence failures.

Time Epoch was pre-registered as a fixed-effect covariate and was dropped from the primary specification (Addendum 7) for two reasons: the corpus's 10-week span provides insufficient range for the step-function model-release signal the covariate was designed to capture, and uncited comparison-pool observations have no citation timestamp (assigning one would fabricate an empirical referent). Confidence intervals are reported using the Wald approximation rather than profile-likelihood as pre-registered (Addendum 8); at n=114k Wald and profile CIs are numerically indistinguishable for coefficients well separated from zero, while profile-likelihood computation across the 15-model sensitivity battery was not tractable within the analysis timeline.

4.8 Pre-registration and addenda

Pre-registration filed 2026-04-20 at OSF (DOI 10.17605/OSF.IO/FMSRD). Nine addenda are public on the project wiki at osf.io/w76y8 documenting (1) the date-range correction, (2) Gemini-API exclusion, (3) ecommerce schema rescue, (4) Perplexity κ scope, (5) ChatGPT κ deviation, (6) H3 reframing, (7) Time Epoch drop, (8) Wald CIs, and (9) the feature-pipeline correction. Readers checking compliance with the original filing should read the addendum log alongside the protocol.

---

5. Results

5.1 H1: Citation distribution across SEO tiers (descriptive)

Across 94,384 analytic citation events:

Tier	Google rank	Count	% of all citations
1	1–3	6,980	7.4%
2	4–10	8,000	8.5%
3	11–30	8,259	8.8%
4	31–100 (and indexed-but->100)	70,204	74.4%
5	Not indexed (CommonCrawl-confirmed open-web)	941	1.0%
(Tier-5-ambiguous)	Not in Google, not in CC	6,027	(separate bucket)

Cumulative Tier-4+: 75.4% (95% clustered CI 74.7–76.1%) — i.e., three-quarters of AI citations are to pages outside Google's top 30 for the keyword-intent version of the eliciting query. Only 7.4% are to top-3 pages; only 15.9% to the top 10.

Per-platform (Method A):

Platform	Tier 4+ rate	Method B κ_binary (top-30 vs deeper)
Google AI Mode	77.0%	**0.800** (95% CI 0.71–0.89) — strong validation
ChatGPT	82.3%	0.419 (95% CI 0.28–0.56) — partial validation
Perplexity	67.1%	Method-A-only (Addendum 4)
Claude	67.8%	Method-A-only (no fan-outs)

The directional claim — most AI citations are outside Google's top 30 — holds across all four platforms (range 67–82%). Strongest Method B support is Google AI Mode (κ=0.80, near-almost-perfect on the binary H1 claim). ChatGPT's partial κ (0.42) is driven primarily by within-top-30 sub-tier disagreement; binary top-30-vs-deeper agreement remains 70.4%.

5.2 What is behind the 75% Tier-4+ headline?

The aggregate Tier-4+ statistic conceals four exploratory findings about the structure of deep-tier citations.

5.2.1 The Tier 4 bucket is overwhelmingly "rank > 100", not "rank 31–100"

The 70,204 Tier-4 citation events resolve into:

Provenance	Events	Share of Tier 4
`site_lookup` (Google indexes the URL but rank > 100)	63,489	90.4%
`serp_pool_match` (URL in our Google top-100 at rank 31–100)	6,096	8.7%
`site_lookup_indexed` (auxiliary verification path)	619	0.9%

The Tier-4 H2 regression coefficient (OR=0.23 vs Tier 3, see Section 5.4) only describes the 8.7% slice of citations to URLs ranked 31–100 in our SERP pool. The other 90% of "Tier 4" citations go to URLs Google indexes but ranks beyond #100 — a region for which we have no bounded uncited denominator. Practical implication: the H2 regression cannot speak to citation probability for rank-unknown-but->100 URLs; H1's descriptive statistic is the only quantity available for that subgroup.

5.2.2 Deep-tier citations are sprawling one-hits, not consistent picks

Across the 100,411 citation events covering 53,925 unique cited URLs:

Tier	Unique URLs	Total events	% of cited URLs cited only once	% cited by only 1 platform
1	1,793	6,797	33%	45%
2	2,633	7,887	48%	61%
3	3,306	8,401	59%	74%
4	42,250	70,361	**77%**	**92%**
5	652	942	83%	96%

Top-3 pages that are cited tend to be cited *repeatedly* and across *multiple platforms* — they function as stable answer sources. Tier-4+ pages that are cited are 77% one-hit-wonders, 92% single-platform: there is no consistent "deep-tier signal" to retro-engineer at the page level. The deep-tier citation pattern is closer to AI doing fuzzy semantic retrieval into a long tail of 43,000+ pages than to AI building consistent trust in a narrow set of deep-tier sources.

5.2.3 Platforms diverge sharply on UGC tolerance in deep tier

Domain-category breakdown of each platform's Tier-4+ citations:

Platform	UGC / Social	Academic / Gov	Publisher / Other
ChatGPT	16.3%	2.8%	80.9%
Google AI Mode	21.5%	1.6%	76.9%
Claude	**0.6%**	0.2%	99.2%
Perplexity	**24.3%**	1.2%	74.5%

Claude essentially refuses to cite UGC sources in the deep tier (0.6% — rounding-error level). Perplexity is heavily UGC-reliant in deep tier (~1 in 4 deep citations is Reddit / YouTube / social). ChatGPT and Google AI Mode are intermediate. This is a sharp architecture-level difference not visible in aggregate tier statistics, with practical implications for publishers tracking platform-specific citation patterns.

5.2.4 Deep-cited pages have different content profile than top-30 cited pages

For the subset of cited pages that we crawled and feature-extracted (top-30 cited n=19,348; deep cited n=54,634):

Feature	Top-30 cited (mean)	Deep cited (mean)	Direction
stats_density (per 1k)	8.95	9.88	Deep is *more* stats-dense
heading_density (per 1k)	11.76	12.81	Deep has *more* H2/H3 per word
list_structure (per 1k)	1.20	1.55	Deep has *more* list items per word
word_count (mean)	3,195	2,538	Deep is shorter
has_canonical	90.8%	82.8%	Deep less SEO-polished
has_meta_description	90.2%	81.8%	Same direction
has_paywall_signals	7.5%	4.8%	Deep is less paywalled

Deep-cited pages are shorter, more content-dense per word, and less SEO-polished. The qualitative profile is closer to "content-marketing blog post" than "established publisher article." This is consistent with the H2 regression finding (Section 5.4) that schema markup and content-density features carry independent predictive weight even after controlling for Google rank.

5.3 Method A vs Method B agreement

On a stratified sample of 397 citation events (200 Google AI Mode + 197 ChatGPT, 159 evaluable after excluding events with zero captured fan-outs):

Platform	N	Binary κ (top-30 vs deeper)	Binary agreement	4-way κ
Google AI Mode (vs Gemini fan-outs)	197	**0.800** (95% CI 0.71–0.89)	90.4%	0.562
ChatGPT (vs ChatGPT API fan-outs)	159	0.419 (95% CI 0.28–0.56)	70.4%	0.247
Pooled	356	0.624	81.5%	0.424

Per Addendum 5, the strict pre-registered rule (κ < 0.70 → restrict ChatGPT analysis to fan-out subset) is not applied to ChatGPT, for two reasons: restricting would discard >99% of ChatGPT data with no analytic benefit, and Method B's own construct validity for ChatGPT is uncertain (API-routed `gpt-5.4-mini` fan-outs may systematically differ from ChatGPT UI retrieval). ChatGPT findings are reported with explicit "moderate κ" caveat.

5.4 H2: Cited-vs-uncited mixed-effects logistic regression

Primary specification on n=114,034 (URL, query) observations after Tier-5 exclusion:

Term	β	SE	OR (95% CI)	p
(Intercept)	−2.459	0.093	0.09 (0.07–0.10)	<0.0001
Tier 1 (1–3) vs Tier 3	**+2.056**	0.036	**7.82 (7.28–8.39)**	<0.0001
Tier 2 (4–10) vs Tier 3	+1.088	0.029	2.97 (2.81–3.14)	<0.0001
Tier 4 (31–100) vs Tier 3	−1.474	0.025	0.23 (0.22–0.24)	<0.0001
GEO Z-sum (per 1 SD)	**+0.059**	0.004	**1.06 (1.05–1.07)**	<0.0001
(Vertical fixed effects: 13 dummies, omitted)				

Tier 1 vs Tier 4 odds ratio = ~34. The GEO Z-sum effect is real (p<0.0001, 95% CI excludes 1.0) but small in absolute magnitude (6% odds increase per 1 SD). The PCA-1 specification produces a larger composite effect:

Composite	β	OR per 1 SD	Cronbach's α
Z-sum (equal-weighted)	+0.059	1.06	0.231
PCA-1 (first component)	+0.140	1.15	(Pearson r with Z-sum: 0.769)

The Z-sum vs PCA-1 difference is 0.081, ≈7 SE — well above the protocol §5.2 threshold (>0.5 SE) for reporting the discrepancy in main results. PCA-1 is the more sensitive measure here because its first-component axis re-weights toward statistics density and primary-source score, where the per-feature signal is concentrated; Z-sum dilutes that signal by averaging in features with weaker independent associations.

Interaction test: SEO_tier × GEO_Z-sum interaction is statistically significant (LRT χ^2=13.4, df=3, p=0.004) but small in effect size. The main implication: GEO has its largest positive slope at the reference tier (Tier 3) and a slightly attenuated slope at deeper tiers (Tier-4 × GEO interaction β=−0.028, p=0.0006). Translation: GEO content investment helps more on already-ranking pages than on far-SERP pages.

5.5 Per-feature regressions (the centerpiece sensitivity)

Each feature is also estimated alone in a model with SEO tier and vertical, to identify which composite components carry the predictive signal:

Feature (per 1 SD, alone)	β	OR (95% CI)	p
schema_presence	**+0.267**	**1.31 (1.28–1.33)**	<0.0001
primary_source_score	+0.113	1.12 (1.08–1.16)	<0.0001
answer_first_coverage	+0.087	1.09 (1.07–1.11)	<0.0001
comparison_signals	+0.061	1.06 (1.04–1.08)	<0.0001
list_structure	+0.035	1.04 (1.02–1.06)	0.0003
stats_density	+0.028	1.03 (1.01–1.05)	0.010
heading_density	−0.065	0.94 (0.90–0.98)	0.002

Six of seven features show positive associations with citation; schema markup is by an order of magnitude the strongest. Heading density shows a small negative association — this likely reflects that high-heading-density-per-word can signal thin / over-chunked / SEO-spam content as much as well-structured authoritative content; we discuss the interpretation in Section 6.

A note on the per-feature signs and reconciliation with prior literature. An initial H2 run revealed negative coefficients for `stats_density` and `list_structure`, in apparent conflict with our own prior Experiment M finding that cited pages contain more inline statistics and list elements than uncited peers. Forensic review identified the source: an ad-hoc regex pipeline in our feature post-processor that matched any four-digit year unconditionally (inflating density scores on "Best X 2026" listicles) and missed Princeton's stat-term, sample-size, ratio, and figure-table patterns; `primary_source_score` had been replaced with a regex for first-person research language rather than the protocol-specified Princeton-derived composite. Reverting to the pre-registered Princeton extractors flipped both coefficients to positive and reconciled our H2 results with prior Experiment M findings. The full flawed-vs-corrected coefficient comparison and the regex drift table are documented in Addendum 9 of the pre-registration log; the flawed run's outputs are archived for audit.

When all seven features are included simultaneously alongside SEO tier and vertical:

Feature (all-together)	β	OR	p
schema_presence	+0.256	1.29	<0.0001
primary_source_score	+0.070	1.07	0.001
answer_first_coverage	+0.058	1.06	<0.0001
comparison_signals	+0.044	1.04	<0.0001
list_structure	+0.028	1.03	0.006
stats_density	−0.013	0.99	0.32 (NS)
heading_density	−0.054	0.95	0.009

Schema markup dominates the multivariate model (OR=1.29 even in the presence of all other features). Primary-source score, answer-first coverage, and comparison signals retain independent positive

contributions. Stats density's marginal effect attenuates to near-zero in the presence of the other six features.

5.6 Per-platform H2

Platform	Tier 1 OR	Tier 4 OR	GEO Z-sum OR
ChatGPT	5.16	0.28	1.05
Perplexity	8.61	0.20	1.07
Google AI Mode	7.26	0.22	1.04
Claude (reduced-n)	4.94	0.26	1.11

All four platforms show the same dominant SEO-tier pattern. GEO Z-sum effects vary from near-zero (Google AI Mode, ~4% odds increase per 1 SD) to modest (Claude, ~11%). Perplexity has the steepest SEO tier gradient; ChatGPT is the flattest — both qualitatively consistent with their reported retrieval architectures. No platform shows the "Outcome C" platform-specific-floors pattern pre-registered as a possible outcome; per-platform GEO and tier signals are qualitatively aligned with the pooled estimates.

5.7 Per-vertical H2

All thirteen verticals with N≥150 obs (per protocol §6.5) yield positive and significant GEO Z-sum effects, ranging from OR=1.03 (E-commerce, Food & Cooking) to OR=1.16 (Pet). SEO-tier effects are uniformly massive across verticals. The GEO effect tends to be largest in technical / specialized consumer verticals (Pet, Consumer Electronics, Home & Home Improvement) and smallest in commodity verticals (E-commerce, Beauty, Food). Pet's reduced-n status (per §6.5 the threshold is met but N is the smallest at ~4,500) does not alter the qualitative pattern.

5.8 Sensitivity battery

Sensitivity	GEO Z-sum OR	SEO tier pattern
3-tier boundaries (1–10 / 11–30 / 31+)	1.06	Top-10 OR=4.03; deep OR=0.23
Domain leverage (top-20 cited domains excluded)	1.05	Tier 1 OR=8.61 (slightly stronger)
UGC exclusion (Reddit, YouTube, etc. removed)	1.06	Tier 1 OR=8.71
PCA-1 composite	1.15	Tier 1 OR=7.79

Direction and approximate magnitude of all primary findings replicate across all sensitivity specifications. No analysis flips a sign or moves a confidence interval across zero.

5.9 H3: Tier 5 citation activity

Of 810 confirmed Tier 5 URLs (cited by AI in our corpus AND `site:URL` returns zero on Google AND CommonCrawl confirms the URL exists on the open web), **678 (83.7%, 95% CI 80.9–86.1%) appeared in citation events within the most recent 30 days of the corpus** (after 2026-03-22).

The pre-registered "persistence rate" metric — same URL cited at T and at T+N — is mathematically undefined on this corpus because the 10-week span left zero URLs straddling the 30-day window cutoff. Per Addendum 6, we report H3 as a point-in-time statement of contemporaneous Tier-5 citation activity rather than a temporal-decay estimate.

The qualitative interpretation: AI chatbots actively retrieve Google-unindexed pages in current production data, not merely in residual training signals from earlier model versions. A longer-panel follow-up would be required to estimate the time-to-citation-decay function for de-indexed URLs; this is queued as a Study A extension.

---

6. Discussion

The headline

Google rank is the dominant predictor of AI citation probability. A page in Google's top 3 for the keyword-intent version of an eliciting query is roughly 34× more likely to be cited than a page ranked 31–100 for the same query. This is the Lily-Ray-aligned position empirically supported.

GEO content features add small but real and statistically robust predictive power above Google rank (OR=1.06 per 1 SD on equal-weighted Z-sum; OR=1.15 on PCA-1). Schema markup is the dominant single-feature lever (OR=1.31 alone, OR=1.29 controlling for the other six features and SEO tier). Primary-source score, answer-first coverage, comparison signals, and list structure each contribute small positive effects (3–12% odds increases per 1 SD). Heading density shows a small negative effect after rank conditioning.

Why does heading density show a negative effect?

Heading density (H2 + H3 count per 1,000 words) is the only one of the seven pre-registered features with a negative coefficient in the H2 regression (β=−0.065 alone, β=−0.054 in the multivariate model). This appears to contradict Experiment M's published finding [4] that cited pages have 2× the H3 subheadings of uncited peers.

Two reconciling considerations: 1. Experiment M measured heading *count* (raw numbers of H3s on the page); H2 measures heading *density* (count per 1,000 words). Cited pages can have more H3s in absolute terms while having fewer per word — i.e., be longer pages with proportionally similar heading structure. 2. Within-rank-tier conditioning may be picking up a "thin / chunked content" signal: pages with 30 subheadings in 800 words are stylistically distinct from pages with 30 subheadings in 4,000 words, and the former pattern is more common in low-quality SERP-spam-style content that nonetheless gets indexed.

We report the heading-density coefficient direction transparently and recommend follow-up work that decomposes heading effects into "absolute count" and "density per word" specifications.

What the deep-tier sprawl finding means

The 75.4% Tier-4+ aggregate is not evidence that AI is "going deep" into a curated set of authoritative low-rank sources. As shown in Section 5.2, the deep-tier citations are dominated by:

1. URLs Google ranks beyond #100 (90% of "Tier 4" events), not URLs in the 31–100 band.
2. One-hit-wonders cited by a single platform for a single query (77% of deep-tier cited URLs).
3. A long tail of 43,000+ unique URLs across 16,000+ domains.
4. Heavy UGC reliance, with sharp platform divergence (Claude 0.6% UGC vs Perplexity 24% UGC in deep tier).

The implication for SEO/GEO practice is that deep-tier citation is not a meaningfully gameable target. There is no consistent feature pattern at the URL level for deep-tier-cited pages — most are cited once, by one chatbot, and never again. The actionable target for publishers seeking AI citation at scale is the top-30 SERP, where the per-page citation probabilities are an order of magnitude higher and citations tend to repeat across platforms.

Platform divergence as architecture revelation

Claude's near-zero UGC rate in the deep tier (0.6%) is the most striking platform-specific finding. Anthropic has not publicly disclosed the retrieval pipeline behind Claude's web_search tool, but the data suggest a strong publisher / non-UGC bias in the source pool — either by explicit filter or by training-data curation. Perplexity's opposite pattern (24% UGC in deep tier) is consistent with its public positioning as a "research-focused" engine that pulls from forum discussions, technical Q&A, and similar UGC long-tail sources.

These differences are not visible in the H1 aggregate or the H2 pooled regression. They suggest that future GEO research should stratify by platform earlier and more aggressively than this study did; the "pooled effect" framing risks averaging across genuinely different retrieval mechanisms.

The "GEO is just SEO" debate, settled

The Lily-Ray-aligned position — "indexation is a gate, ranking is the differentiator" — is empirically supported. Rank dominates. The "GEO is its own discipline" position requires more nuance:

- There is one content feature (schema markup) with a robust independent effect comparable to a within-tier rank improvement.
- There are 4–5 other content features with small but consistent independent effects (3–12% odds increases per 1 SD).
- There is no evidence for the strong-form GEO claim that content optimization can substantially compensate for poor SEO rank — the GEO Z-sum coefficient (+0.06 per 1 SD) is two orders of magnitude smaller than the Tier 1 vs Tier 4 SEO effect (+3.5 in log-odds).

The honest synthesis is therefore: **SEO is the gate. Schema markup is the strongest content-feature predictor inside the gate. Other GEO features add marginal value. Most of the rest is noise.**

The schema causal-interpretation problem

Schema markup's OR=1.31 must be read as a *predictive association*, not a causal effect. Schema presence in a publisher's HTML is correlated with site budget, server-side rendering quality, fast time-to-interactive, high backlink-derived domain authority, and a dedicated technical-SEO function — every one of which independently predicts AI citation in ways our observational design cannot disentangle. From the H2 regression alone, we cannot determine whether (a) AI retrieval pipelines consume JSON-LD as structured data and weight pages with rich schema higher, or (b) AI is merely preferring well-resourced publishers, who happen to be the population that systematically deploys schema markup. The two hypotheses produce identical observational signatures. Disambiguating them requires an interventional design — adding schema to otherwise-equivalent pages and measuring downstream citation change — which is the target of a planned follow-up Study B (greenfield factorial on uncrawled domains). Until that experiment runs, the responsible reading of our schema finding is "schema markup is the strongest *predictive signal* among the seven pre-registered GEO features," not "schema markup

causes citation." We urge GEO practitioners and SEO commentators to preserve this distinction when citing this work.

---

7. Limitations

1. **Observational, not causal.** Study A measures association. Although the comparison-pool design eliminates collider bias on the citation outcome, unobserved publisher-side confounders (domain authority, backlink profile, site age, editorial investment) remain.
2. **Method A model dependence.** The keyword-reformulation pipeline depends on Gemini 3.1 Flash Lite's mapping from conversational query to search intent. Cohen's κ between Method A and Method B varies by platform (0.80 Google AI Mode → 0.42 ChatGPT). Per-platform validation levels are reported in Section 5.3.
3. **Temporal scope.** The corpus spans ~10 weeks (February through mid-April 2026). Findings may not generalize to longer-horizon AI behavior; H3's classical persistence metric is unestimable on this window (Addendum 6).
4. **Claude reduced-n.** Claude contributes ~5,000 analytic citation events vs ~25–30k for the other three platforms, due to a known web_search cache constraint that limits per-account scrape throughput. Claude-specific findings are reported with expanded CIs.
5. **Google-centric SEO proxy.** SEO tier is measured against Google rank. Bing parallel data is auxiliary; Bing's own SERP depth caps at ~20 results. ChatGPT's Bing-derived retrieval pipeline is not separately validated.
6. **Tier 5-ambiguous bucket.** 5.7% of citation events are to URLs absent from both Google's index and recent CommonCrawl snapshots. These likely include dead URLs, paywalled content, very recent publications, and noindex'd pages. They are reported as a separate bucket but not characterized further.
7. **Pre-registration deviations.** Nine addenda are documented at osf.io/w76y8. The most consequential are Addendum 7 (Time Epoch dropped), Addendum 8 (Wald CIs), and Addendum 9 (feature-pipeline correction). All are disclosed before publication.

8. Reproducibility

- **Pre-registration**: OSF DOI 10.17605/OSF.IO/FMSRD | Project osf.io/w76y8
- **Pre-registration log**: nine addenda public on the OSF project wiki.
- **Code**: tagged at git commit `study-a-v1.0.2`. Resolve to commit hash via `git rev-parse study-a-v1.0.2`. Earlier tags `study-a-v1.0` (initial submission) and `study-a-v1.0.1` (peer-review revisions) are also available for inspecting the pre-revision codebases. All `study_a_*` Python and R scripts in the reproducibility package.
- **Data package**: Zenodo DOI 10.5281/zenodo.19787328. Includes the locked 2,000-query analysis set, the citation corpus (analytic version), the comparison pool, the Playwright crawl features, the H2 analytic dataset, the regression coefficient tables (corrected), and the flawed-run archive for audit.
- **Replication**: any reader can re-run the H2 regression by checking out the tagged commit, downloading the Zenodo data package, installing R 4.5.3 + lme4 + jsonlite + dplyr + broom.mixed + optimx, and executing `scripts/run_h2_all.ps1` (Windows PowerShell) or its bash translation. Total wall time ~7 hours on commodity hardware.

9. What Study A does not test

- Whether SEO improvements *cause* citation changes (Study B, planned greenfield factorial).
- Whether GEO improvements *cause* citation changes (Study B).
- Whether AI platforms re-cite a page after its Google rank changes (planned de-ranking decay study).
- Whether AI citation drives traffic, conversions, or any downstream business outcome.
- Whether the 2,000-query set is representative of all queries sent to AI platforms globally. Sample is weighted toward e-commerce, SaaS, and consumer verticals.

Appendix A — Author attribution: controlled regression and reconciliation with prior internal findings

A reviewer of an earlier draft asked us to verify whether crawler-detected `has_author_attribution` (a heuristic byline detector matching `[rel="author"]`, `.author`, `[class*="author"]`, `[itemprop="author"]`) shows the negative association reported in our prior internal work (Lee, "Updated Findings," March 2026; OR=0.632, p<0.000001 on n=4,658 real-website-only pages). We ran the controlled test and found a substantively different result; the reconciliation is mechanical (comparison-pool construction) rather than a sign reversal.

Controlled-regression specification: `cited ~ has_author_attribution + seo_tier + log(word_count) + geo_composite_zsum + vertical + (1 | qid)` on Study A's full sample (n=114,034) and again on the non-UGC subset (n=108,479).

Specification	OR (`has_author_attribution`)	95% CI	p
Full sample, full controls	1.137	1.090–1.187	<0.0001
Full sample, tier + length only	1.148	1.100–1.197	<0.0001
Full sample, tier only (no length)	1.285	1.232–1.339	<0.0001
Non-UGC subset, full controls	**1.122**	**1.074–1.173**	**<0.0001**
Non-UGC, tier + length	1.134	1.085–1.185	<0.0001

Per-platform with full controls:

Platform	OR	95% CI	p
ChatGPT	1.395	1.302–1.495	<0.0001
Claude	1.306	1.102–1.547	0.002
Perplexity	1.114	1.050–1.183	0.0004
Google AI Mode	1.086	1.026–1.149	0.005

In Study A's controlled comparison-pool design, `has_author_attribution` shows a small, robust positive association with citation, statistically significant in every specification including the non-UGC subset. Per-platform, the effect is largest for ChatGPT (OR=1.40) and Claude (OR=1.31) and smallest for Google AI Mode (OR=1.09).

Reconciliation with the prior internal finding (OR=0.632 negative, n=4,658, real-website-only). Study A reproduces the prior raw direction at the marginal level *if and only if* we replicate the prior comparison-pool construction. The prior study compared cited URLs to a pool of

3,031 uncited pages drawn from Google's organic Top-10, while Study A's comparison pool is the full Google **Top-100** for each of 2,000 queries (~100,000 uncited URLs). The raw cited-vs-uncited author-attribution rates in each pool are:

Pool construction	Cited author-rate	Uncited author-rate	Marginal OR
Prior (Top-10 uncited, real-website-only)	44.1%	**55.5%**	**0.632** (negative)
Study A (Top-100 uncited, full sample)	52.3%	45.6%	1.306 (positive)
Study A (Top-100 uncited, non-UGC)	50.7%	45.2%	1.249 (positive)

The cited-side rates differ modestly across studies (44.1% vs 50.7% non-UGC). The dramatic difference is on the **uncited side**: 55.5% in the prior Top-10 pool versus 45.2% in Study A's Top-100 non-UGC pool. Google's Top-10 organic results are dominated by premium publishers (NYT, Forbes, Healthline, Investopedia, Consumer Reports) that systematically use author bylines for E-E-A-T signaling. Restricting the uncited comparison to Top-10 inflates the uncited author-attribution baseline by ~10 percentage points relative to a Top-100 comparison, which inverts the marginal direction.

The two findings are not contradictory at the level of underlying behavior. They reflect different denominators answering different questions:

- *Prior question (Top-10 comparison):* "Among Google's top-10 organic publishers, are AI-cited pages more byline-y than uncited?" Answer: no, because top-10 uncited is exceptionally byline-heavy.
- *Study A question (Top-100 mixed-effects logit):* "Across the realistic retrieval consideration set defined by Google top-100, does author attribution increase per-page citation odds after controlling for length, tier, GEO composite, vertical, and query random effects?" Answer: yes, slightly (OR ~1.12).

Study A's Top-100 comparison is the broader of the two and represents the realistic retrieval-eligible URL set rather than the small premium-publisher subset. Within that broader set, the controlled effect of author attribution is small but positive across all four platforms.

A second methodological note: the prior study's cited pool included **ChatGPT API citations via `gpt-5.4-nano`**. Study A excludes API-derived citations a priori (Addendum 2 of the pre-registration log), per the pre-registered scope of UI/VPS-only data. ChatGPT API and ChatGPT UI behave differently in citation patterns (Lee, 2026, Study 3), so the API inclusion in the prior study may further contribute to the marginal-direction difference; we cannot decompose this from the data we have.

Practical interpretation. For publishers asking whether to keep author bylines on substantive content, Study A's finding supports keeping them. The effect is small (≈12% odds increase per binary flag) but robust across specifications, larger on ChatGPT/Claude than on Google AI/Perplexity, and not driven by UGC contamination. Schema-specific Person markup (Schema.org `Person` JSON-LD specifically) was not isolated in this analysis and remains untested; the present finding refers to the broader byline-detection heuristic.

Appendix B — ChatGPT regression on the Method-B-validated subset

The strict pre-registered rule for Method B validation ($\kappa < 0.70 \rightarrow$ restrict analysis to the fan-out subset) was not applied to ChatGPT in the main analysis ($\kappa_{binary} = 0.42$), for reasons documented in Addendum 5. To address the natural reviewer concern that this deviation conveniently retained data we

should have discarded, we re-fit the per-platform ChatGPT regression on the κ-validation subset itself. If Method A's reformulated queries had hallucinated the SEO-floor pattern, the κ-restricted regression should fail to reproduce the gradient.

Subsetting: 159 evaluable ChatGPT events from the κ-validation sample (197 events with `fanouts_checked > 0`, retaining those with non-trivial Method B comparison) yield 139 unique queries. The full Google top-100 SERP pool for those 139 queries gives 7,960 (URL, query) observations with 718 cited by ChatGPT. The same H2 specification (`cited_chatgpt ~ seo_tier + geo_composite_zsum + vertical + (1 | qid)`) is fit on this subset.

Result: the model converged without warnings. Comparison of Wald fixed-effect coefficients to the full-data ChatGPT model (Section 5.6):

Term	Full-data (n = 114,034; 4,932 ChatGPT cited)	Method-B subset (n = 7,960; 718 ChatGPT cited)
(Intercept)	OR 0.03 (0.02–0.04)	OR 0.12 (0.06–0.23)
Tier 1 (1–3) vs Tier 3	**OR 5.16** (4.67–5.70)	**OR 6.05** (4.61–7.94)
Tier 2 (4–10) vs Tier 3	OR 2.26 (2.07–2.48)	OR 2.33 (1.85–2.93)
Tier 4 (31–100) vs Tier 3	OR 0.28 (0.26–0.30)	OR 0.17 (0.14–0.21)
GEO Z-sum (per 1 SD)	OR 1.05 (1.04–1.06)	OR 1.04 (1.02–1.06)

Standard errors widen by roughly 3× across the SEO tier coefficients (e.g., Tier 1 SE 0.051 → 0.139), as expected at 1/14 the sample size. **Point estimates retain the same direction and similar magnitude in every coefficient that matters for the H2 hypothesis.** Tier 1 vs Tier 3: $p < 10^{-38}$ even on the restricted sample. Tier 4 vs Tier 3 is actually somewhat more negative in the restricted subset (OR 0.17 vs 0.28).

The κ-restricted subset confirms the SEO-floor pattern observed in the full ChatGPT analysis. Method A's keyword-reformulation pipeline is not generating the tier gradient by retrieval-misalignment artifact; the gradient is present in the data even when we restrict to the events for which Method B validation was directly performed. This is the strongest internal-validity check on the ChatGPT findings that the available data permits.

References

1. Aggarwal, P., Murahari, V., Rajpurohit, T., Kalyan, A., Narasimhan, K., & Deshpande, A. (2024). *GEO: Generative Engine Optimization*. Proceedings of KDD 2024. ↩↩
2. Seer Interactive. (2024). *AI Search Cites What Bing Ranks*. Industry report. ↩↩
3. SurferSEO. (2025). *36M AI Overviews dataset analysis*. Industry report. ↩↩
4. Lee, A. (2025). *Experiment M: Page-feature associations with AI citation*. aiplusautomation.com internal research, December 2025. ↩↩

5. Lee, A. (2026). *Experiment J v2: Is GEO just repackaged SEO?* aiplusautomation.com internal research, March 2026. ↩
6. Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48. ↩
7. Ray, L. (2025–2026). Public commentary on AI citation and SEO via Threads (@lilyray) and direct correspondence. ↩
8. Lee, A. (2026). *Study 3: AI Platform Fan-Out Capture and Taxonomy*. Zenodo DOI 10.5281/zenodo.19554329. ↩