I Rank on Page 1 -- What Gets Me Cited by AI? Position-Controlled Analysis of Page-Level and Domain-Level Predictors of AI Search Citation

Anthony Lee AI+Automation Research ORCID: 0009-0002-4815-6373

Abstract

Generative Engine Optimization (GEO) aims to improve content visibility in AI-generated search responses. Prior observational studies have failed to isolate page-level signals because domain identity alone predicts AI citation at AUC = 0.975, confounding every between-domain comparison. We introduce a position-band matching design that controls for Google ranking position, asking: among equally-ranked pages, what page-level features predict AI citation? Using 250 queries across a balanced grid (5 intent types, 10 verticals), we collected citations from ChatGPT, Perplexity, and Google AI Mode and crawled 10,293 unique pages with 66 structural, semantic, and content-quality features. Within position bands, content features and domain identity provide comparable predictive power (content AUC = 0.673, domain AUC = 0.687 with enriched representations, combined AUC = 0.697), a convergence that contrasts sharply with the domain dominance observed without position control (AUC = 0.975). The top actionable predictors are comparison structure (d = 0.43, significant across all five intent types), query-term coverage (d = 0.42), subheading depth, statistical data density, and the absence of first-person/blog tone. Content structure provides the largest marginal lift beyond rank position (+0.021 AUC). In a second contribution, five domain-level tests reveal that SERP co-occurrence (topical breadth) is the strongest domain trust predictor (rho = 0.341, p = 2.6 x 10^-70), that cited domains are *less* lexically unique than their SERP competitors, and that a combined domain model achieves AUC = 0.921, with SERP presence accounting for 63% of importance. High-SERP-presence domains are cited more per appearance (2.04 citations per slot for 8+ appearances vs. 0.665 for single-appearance domains), confirming this is not merely an artifact of increased exposure. Data and code are publicly available.

---

1. Introduction

The emergence of generative AI search engines (ChatGPT, Perplexity, Google AI Mode, Claude) has created a new optimization problem for content creators: how to get cited in AI-generated responses. This problem, formalized as Generative Engine Optimization (GEO) by Aggarwal et al. (2024), differs from traditional search engine optimization because AI platforms synthesize answers from multiple sources rather than ranking pages in a list.

The central challenge for GEO research is separating what matters about a *page* from what matters about the *domain*. Our prior experiments demonstrated the severity of this confound:

- **Experiment K** (n = 4,350): Domain identity alone predicted citation at AUC = 0.975. Page speed, which appeared to be the dominant page-level signal (65.8% of model importance), was proven to be a domain-level proxy. Within the same domain, cited pages were actually *slower* than uncited pages (r = -0.221, p < 0.000001). The prescriptive advice "make your site faster" was wrong. The real signal was "be on a domain that gets cited" (Lee, 2026a).
- **Experiment L** (n = 2,478 domains): External authority metrics (PageRank, Wikipedia links, domain age) predicted *not* being cited. High-authority domains were less likely to receive AI citations. Content-role features (primary source score, technical depth) achieved AUC = 0.709 between domains but collapsed to AUC = 0.573 within domains.
- **Experiment C** (n = 3,205, 4 phases): The Princeton GEO paper's claim of 30-40% visibility improvement from content features (statistics, citations, quotations) did not replicate on production AI platforms. Princeton-

only features achieved AUC = 0.544 on balanced data, essentially chance level. Only statistics density showed a consistent positive effect, and only for ChatGPT. Citations and quotations either did not matter or went the wrong direction.

The core problem: every between-domain comparison confounds page content with domain identity. The only way to isolate page-level effects is to hold domain quality constant.

Experiment M solves this by using Google SERP position as the matching variable. All pages in a comparison set ranked for the same query at similar positions. They are equally "relevant" by Google's assessment. The question becomes: among these equally-qualified candidates, which ones do AI platforms select?

Contributions. (1) We introduce position-band matching as a method for isolating page-level citation predictors and identify six features that predict citation across all four rank bands. Within position bands, content and domain provide comparable predictive power -- a convergence that contrasts with the domain dominance observed in uncontrolled comparisons. (2) We identify SERP co-occurrence (topical breadth) as the strongest domain-level trust predictor and show that cited domains are *less* unique than their competitors, not more. (3) We release the full dataset of 10,293 pages with 66 features across 250 queries for replication.

---

2. Related Work

GEO Framework. Aggarwal et al. (2024) introduced GEO and the GEO-bench benchmark, demonstrating that optimization strategies (adding citations, statistics, quotations) can boost visibility by up to 40% in a custom generative engine. Our Experiment C replication found this result does not transfer to production platforms (ChatGPT, Perplexity, Google AI Mode). The Princeton result was obtained on a controlled system, not the black-box platforms content creators actually optimize for. Bagga et al. (2025) extended GEO to e-commerce with E-GEO, finding a potentially domain-agnostic optimization pattern. Tian et al. (2025) introduced diagnostic GEO (AgentGEO), achieving 40% citation improvement through targeted repairs versus 25% for generic approaches. Wen et al. (2025) raised concerns about adversarial dynamics in GEO optimization.

AI Citation Behavior. Lee (2026a) analyzed 19,556 queries across 8 verticals, finding that query intent is the strongest aggregate predictor of citation source type (Cramer's V = 0.258), while page-level features predict individual citation selection (AUC = 0.594). The study established that Google rank does not predict AI citation at the URL level (Spearman rho = -0.02 to 0.11, all non-significant). Chen et al. (2025) conducted a large-scale comparative analysis confirming AI search platforms' systematic preference for earned media over brand-owned content.

Industry Studies. Sellm (2025) analyzed 400,000 ChatGPT-cited pages, finding cited pages average 13.75 list sections. Semrush (2025) studied AI Overviews source patterns. BrightEdge (2026) characterized Claude's citation preferences. These industry reports provide useful descriptive data but lack the methodological controls needed to distinguish correlation from confounded association.

The Domain Confound. The central methodological problem in this literature is inadequate control for domain identity. Prior page-level analyses (including our own Lee, 2026a) compared cited pages against not-cited pages drawn from different domains. When domain identity alone predicts at AUC = 0.975 (Lee, 2026a, Experiment K), any page feature correlated with domain quality appears significant regardless of its independent contribution. The position-band matching design introduced here addresses this gap.

---

3. Methodology

3.1 Research Question

Among pages that already rank in Google's top 20 for a given query, what page-level features predict whether an AI platform will cite that page?

3.2 Query Design

We constructed 250 queries across a fully balanced grid: 5 intent types (INFORMATIONAL, DISCOVERY, VALIDATION, COMPARISON, REVIEW_SEEKING) by 10 verticals (Technology, Health, Finance, Travel, Home Improvement, Education, Food, Fitness, Legal, Ecommerce), with 5 queries per cell. 150 queries were reused from a validated prior set; 100 were newly generated. This design ensures no intent type or vertical dominates the results.

3.3 Data Collection

Search engine results (candidate pool): Google CSE API returned top-20 organic results for all 250 queries (5,000 results, 100% coverage). A Bing SERP scraper (Playwright) returned top-20 for 237/250 queries (2,210 results, 95% coverage).

AI platform citations: All citations were collected via Playwright connecting to Chrome debug sessions in live web UIs.

- ChatGPT (GPT-5.3): 250 queries, 2,860 total citations, 82% of queries produced citations. Citations captured via JavaScript fetch interceptor reading SSE streams.
- Perplexity: 250 queries, 2,460 citations, 100% produced citations. Same JS interceptor approach.
- Google AI Mode: 250 queries, 4,818 citations, 86% produced citations. Citations extracted from DOM elements with favicon-encoded domain URLs.

Page crawling: 10,293 unique pages crawled via headless Playwright with 97% success rate. 92% (9,749 pages) yielded extractable visible text. Each page was measured for 60+ structural features with full visible text stored for semantic analysis.

3.4 Feature Extraction

66 features per page across six categories:

Structural (14 features): word count, visible text length, content-to-HTML ratio, H1/H2/H3 counts, heading density, internal/external link counts, page size, load time, canonical tag, mobile viewport.

Schema/Trust (12 features): schema presence and type (Product, FAQ, Review, Article), author attribution, meta description, date signals, paywall signals, affiliate links, popup/modal elements.

Princeton Content Density (7 features): statistics, citations, quotations, technical terms, authoritative language, readability (Flesch-Kincaid), and list items, all per 1,000 words.

Content-Role (4 features): data production ratio, technical depth, aggregator score, primary source score (all computed from Princeton features).

Semantic (3 features): query-term coverage (fraction of query content words in page), query-term first position (normalized 0-1), title-query overlap.

Content Classification (6+ features): content type (product/comparison/review/blog/article/general), current-year mention, recency, comparison signal count, first-person pronoun density.

3.5 Analysis Approach

Position-band matching. Google SERP results grouped into bands (1-3, 4-7, 8-12, 13-20). Within each band, Spearman rank correlations computed for each feature against binary citation status. This controls for rank position, though not for all domain-level confounds (see Section 6.4).

Blocked model comparison. Cumulative Random Forest models adding feature blocks sequentially (position, semantic, structure, quality, trust, technical) to measure each block's marginal contribution.

Cross-platform consensus. Feature profiles compared for pages cited by 0, 1, 2, or 3+ platforms independently.

Domain control. Models compared with and without domain identity features, using both sparse one-hot encoding and enriched domain representations (leave-one-out citation rate, aggregated content quality, domain frequency).

Stability. All RF/GBM models run with 10 random downsamples (balanced cited/not-cited) and 5-fold cross-validation per run, reporting mean AUC across runs.

---

4. Results: Page-Level Features

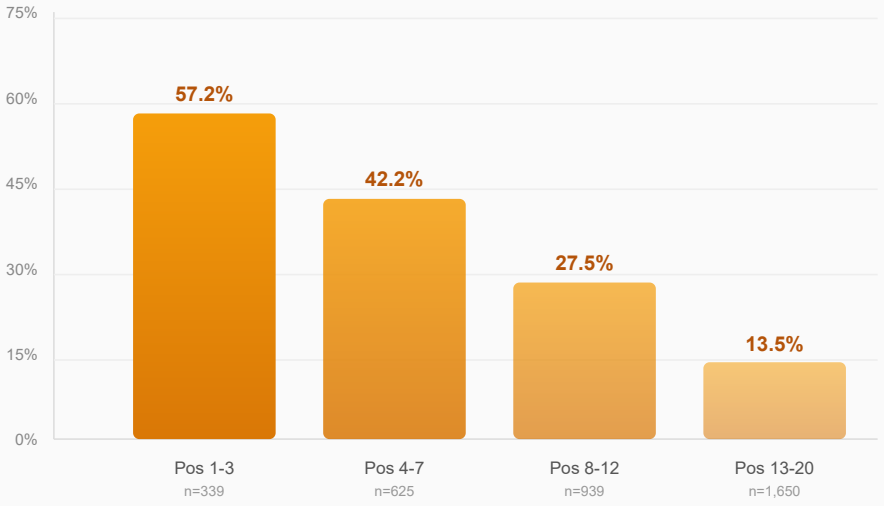
4.1 Position Gradient (Test 1)

Google rank position predicts AI citation with a clear monotonic gradient:

Position Band	Total Pages	Cited	Citation Rate
1-3	339	194	57.2%
4-7	625	264	42.2%
8-12	939	258	27.5%
13-20	1,650	223	13.5%

Figure 1. AI Citation Rate by Google SERP Position Band

n = 3,553 pages across 250 queries | 3 AI platforms (ChatGPT, Perplexity, Google AI Mode)



Even at position 1, 31% of pages are NOT cited. Even at positions 13-20, 13.5% still ARE.

Position 1 achieves a 68.8% citation rate; position 20, 11.9%. This confirms that search engine ranking is the primary gateway to AI citation (Figure 1). But even at position 1, 31% of pages are not cited. Even at positions 13-20, 13.5% still are. Something beyond rank determines selection.

4.2 Within-Position-Band Signals (Test 2)

Six features are significantly associated with citation in all four position bands (consistent direction, $p < 0.05$ in each band):

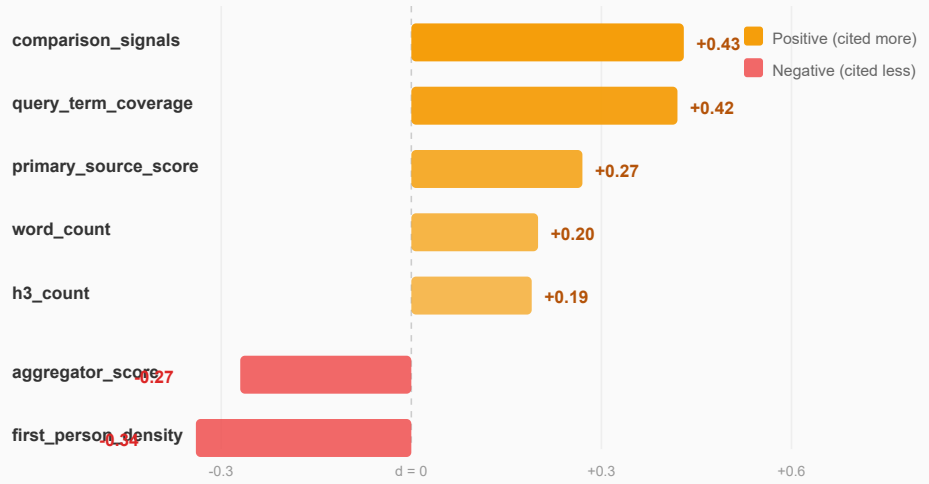
Feature	Direction	Cohen's d (band 4-7)	Interpretation
comparison_signals	+	0.43	Comparison structure ("vs", side-by-side)
query_term_coverage	+	0.42	Page contains the query's key terms
h3_count	+	0.19	More subheadings
word_count	+	0.20	Longer pages (median 2,150 vs 1,415)
primary_source_score	+	0.27	Produces data rather than aggregating
aggregator_score	-	-0.27	Primarily cites/quotes others

Features significant in 3 of 4 bands include statistics density (stats_per_1k), first-person density (negative, strongest in lower bands), FAQ schema, current-year mention, and internal link count.

Features NOT significant within position bands, despite prior experiments suggesting importance: load time (p > 0.39 in all bands), author attribution, content-to-HTML ratio.

Figure 2. Within-Position-Band Effect Sizes (Cohen's d, band 4-7)

Features significant in all four position bands | p < 0.05 in each band independently



Comparison structure and query-term coverage are the strongest actionable signals.
Blog/opinion tone (first-person density) is the strongest negative predictor.

The multiple comparisons concern (38 features x 4 bands = 152 tests) is addressed by the all-four-band criterion. The probability of a null feature reaching p < 0.05 in all 4 independent tests is 0.05^4 = 6.25 x 10^-6.

Comparison_signals robustness check. Because 50 of 250 queries are COMPARISON intent, the comparison_signals effect could reflect query-content alignment rather than an independent content quality signal. We tested comparison_signals separately for each intent type:

Intent Type	Cohen's d	p-value	Significant?
COMPARISON	0.480	< 0.001	Yes
DISCOVERY	0.459	< 0.001	Yes

INFORMATIONAL	0.128	0.040	Yes
VALIDATION	0.265	< 0.001	Yes
REVIEW_SEEKING	0.387	< 0.001	Yes

Comparison_signals predicts citation across all five intent types, not just comparison queries. The effect is nearly as strong for DISCOVERY queries (d = 0.459) as for COMPARISON queries (d = 0.480). This is not a query-type confound.

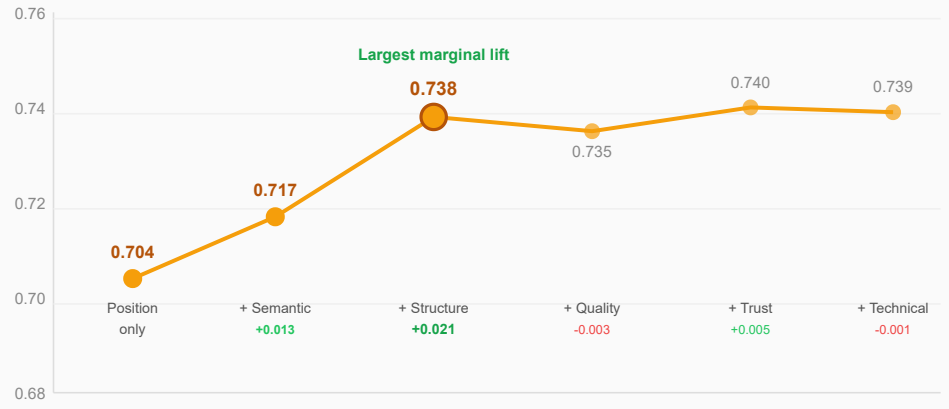
4.3 Blocked Model Comparison (Test 3)

Sequential RF models adding feature blocks:

Model	RF AUC	Marginal Lift
Position alone	0.704 (0.005)	--
+ Semantic relevance	0.717 (0.005)	+0.013
+ Content structure	0.738 (0.004)	+0.021
+ Content quality	0.735 (0.005)	-0.003
+ Trust signals	0.740 (0.004)	+0.005
+ Technical	0.739 (0.004)	-0.001

Figure 3. Blocked Model: Cumulative AUC by Feature Category

Sequential RF models adding feature blocks | 10 downsamples, 5-fold CV



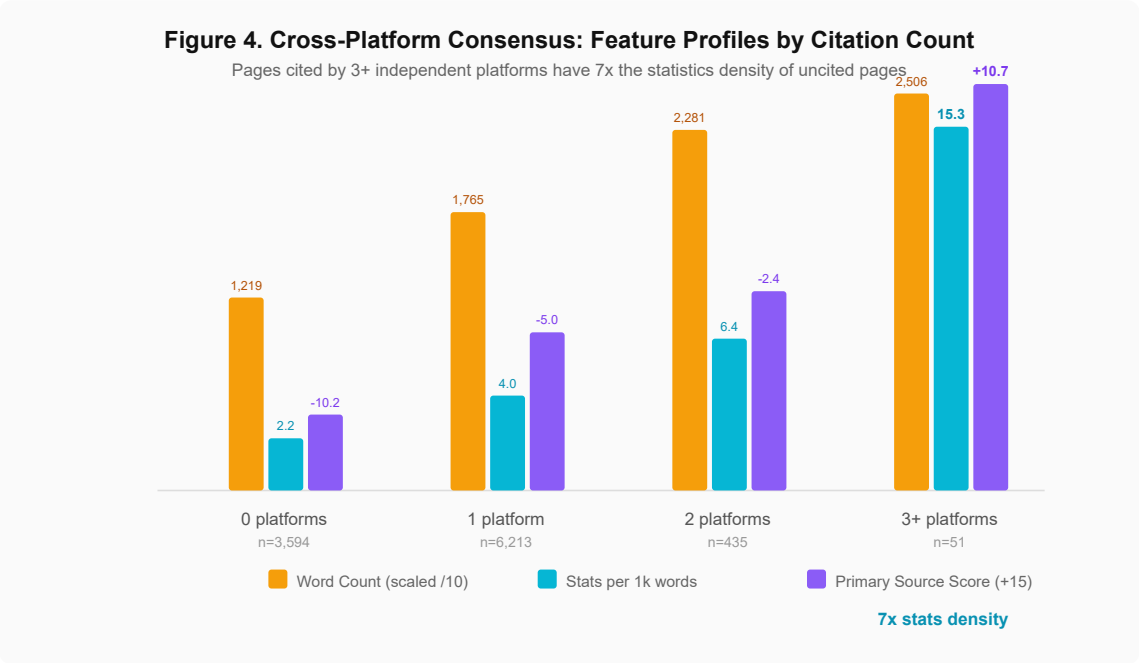
Content structure (headings, word count, lists) provides the largest lift beyond position.
Content quality and technical features add negligible predictive power.

Content structure provides the largest lift beyond position (+0.021 AUC, Figure 3). Content quality features add nothing beyond structure, suggesting structural features already capture the quality signal. Technical features (load time) add nothing.

4.4 Cross-Platform Consensus (Test 4)

Pages cited by more platforms independently show a clear quality gradient:

Cited By	N Pages	Word Count	Stats/1k	Primary Source Score	Query Coverage
0 platforms	3,594	1,219	2.2	-10.2	0.75
1 platform	6,213	1,765	4.0	-5.0	1.00
2 platforms	435	2,281	6.4	-2.4	1.00
3+ platforms	51	2,506	15.3	+10.7	1.00



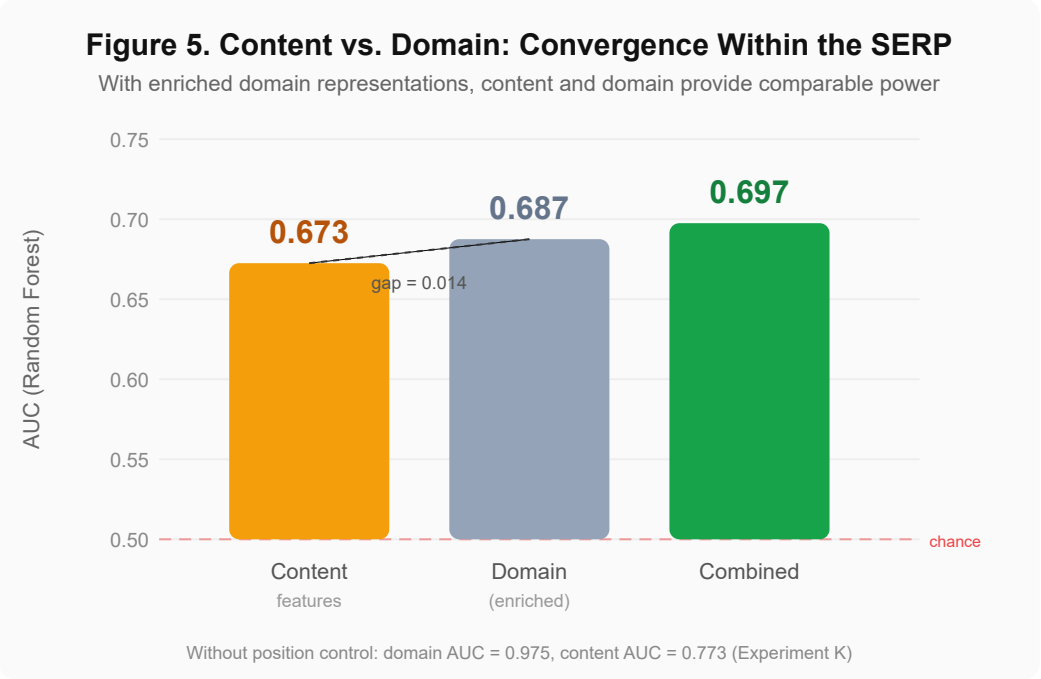
Pages cited by 3+ platforms have 7x the statistics density of uncited pages (15.3 vs 2.2), positive primary source score (+10.7 vs -10.2), 2x the word count, and 100% query term coverage (Figure 4). This is the strongest evidence that content quality matters: when three independent AI systems with different search backends all select the same page, that page has measurably different content characteristics.

4.5 Content vs. Domain (Test 5)

We compared content features against domain identity using both sparse and enriched domain representations:

Model	RF AUC (SD)
Content features alone	0.673 (0.005)
Domain identity (enriched)	0.687 (0.008)
Content + Domain	0.697 (0.006)

The enriched domain model uses leave-one-out citation rate, aggregated content quality metrics, and domain frequency rather than sparse one-hot encoding. With these richer representations, domain modestly outperforms content (+0.014 AUC), but the gap is narrow. The combined model gains only +0.010 over domain alone.



The key finding is the *convergence*: content and domain provide comparable predictive power within the SERP. This contrasts sharply with uncontrolled comparisons, where domain identity dominated at AUC = 0.975 (Experiment K). Without position control, domain explains nearly everything. With position control, content and domain are on equal footing. The position-band design does not eliminate the domain signal, but it reduces it from overwhelming to comparable, allowing content features to emerge as meaningful independent predictors.

4.6 Actionable Feature Ranking (Test 7)

Full model RF AUC: 0.753 (SD = 0.004). The top 15 features by importance:

Rank	Feature	Importance	Actionable	Direction
1	best_google_position	27.9%	No	-
2	first_person_density	8.1%	Yes	-
3	word_count	7.1%	Yes	+
4	comparison_signals	6.4%	Yes	+
5	h3_count	5.9%	Yes	+
6	primary_source_score	4.1%	Yes	+
7	heading_density	3.9%	Yes	+
8	query_term_first_position	3.8%	Yes	-
9	internal_link_count	3.7%	Yes	+
10	query_term_coverage	3.7%	Yes	+
11	content_to_html_ratio	3.4%	Yes	+

12	h2_count	3.1%	Yes	+
13	load_time_ms	2.9%	No	-
14	external_link_count	2.5%	Yes	+
15	stats_per_1k	2.5%	Yes	+

Actionable features account for 69.2% of total model importance. The strongest actionable signal is first-person density (negative): reducing blog/opinion tone increases citation probability. The strongest positive actionable signals are word count, comparison structure, and subheading depth.

---

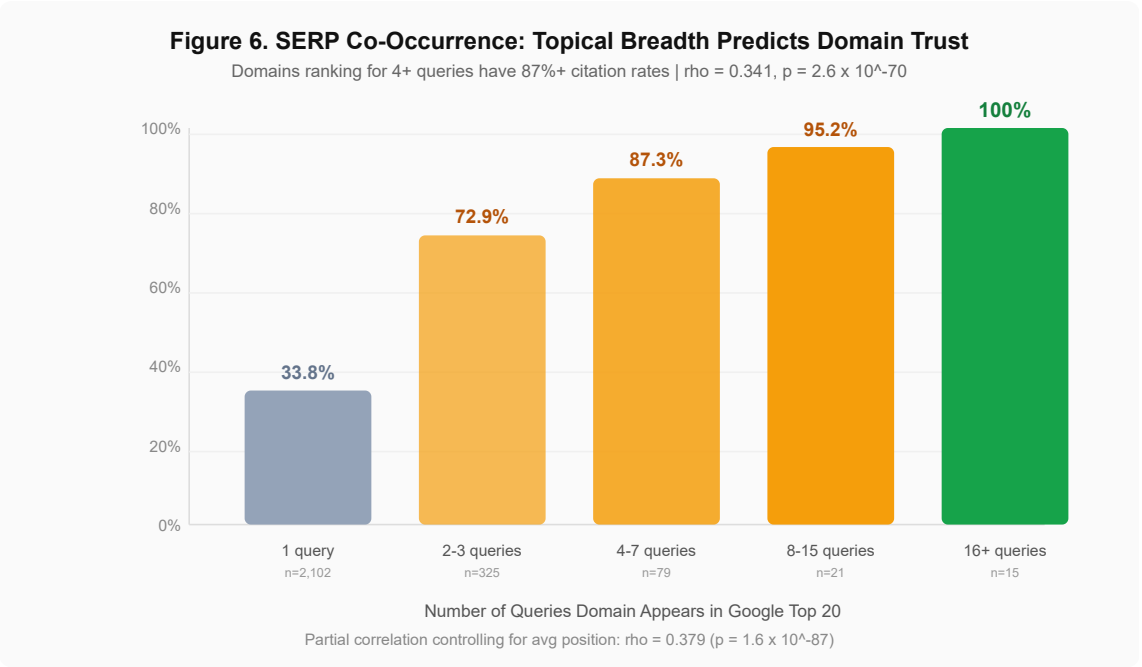
5. Results: Domain Trust

Experiment M's primary analysis (Section 4) identifies page-level levers within the SERP. But prior experiments established that domain identity accounts for approximately 97% of citation prediction globally. Five additional tests investigate what determines domain-level trust.

5.1 SERP Co-Occurrence (Test 7.1)

Among 2,542 domains appearing in Google's top 20 across 250 queries:

Queries Domain Appears For	Domains	Cited	Citation Rate
1 query	2,102	711	33.8%
2-3 queries	325	237	72.9%
4-7 queries	79	69	87.3%
8-15 queries	21	20	95.2%
16+ queries	15	15	100%



Spearman rho = 0.341 (p = 2.6 x 10^-70). Partial correlation controlling for average position: rho = 0.379 (p = 1.6 x 10^-87). A domain that ranks #10 for 5 queries is more likely to be cited than a domain that ranks #3 for 1 query (Figure 6). This is the clearest domain trust predictor in the entire research program.

Addressing the volume confound. A domain with more pages in the top 20 has more opportunities to be cited, raising the concern that n_pages (54.8% of the domain model, Section 5.5) merely reflects increased exposure rather than trust. To test this, we computed per-appearance citation rates:

SERP Appearances	Citations per SERP Slot	Domains
1	0.665	2,102
2-3	1.23	325
4-7	1.64	79
8+	2.04	36

High-presence domains are cited at higher rates *per appearance* (2.04 citations per slot for 8+ appearances vs. 0.665 for single appearances). This is not merely more lottery tickets. However, there are diminishing returns: among cited domains, additional SERP appearances correlate negatively with per-appearance rate (rho = -0.127), suggesting each additional page helps less. The effect is real but decelerating.

5.2 Cross-Citation Network (Test 7.2)

We tested whether cited domains reference each other more than expected, using text matching as a proxy for cross-linking (full URL extraction was not available). The directional finding (rho = 0.250 between inlinks from cited domains and citation volume) suggests a network effect, but the measurement is unreliable: common English words matched as false-positive domain names in the text-matching approach, inflating the apparent network density. We report the direction for completeness but do not draw conclusions from this test. A proper URL-based link extraction would be needed to validate any network effect.

5.3 Competitor Pair Analysis (Test 7.3)

For 1,718 matched pairs (cited vs. not-cited pages for the same query at similar Google positions):

Feature	Mean Difference	Wilcoxon p	Direction
h3_count	+3.9	6.0 x 10^-22	Cited more structured
internal_link_count	+58.1	5.6 x 10^-29	Cited more connected
comparison_signals	+1.0	5.6 x 10^-20	Cited more comparative
word_count	+482	2.3 x 10^-10	Cited longer
h2_count	+1.3	7.2 x 10^-15	Cited more organized
first_person_density	-2.0	3.7 x 10^-6	Cited less blog-like
stats_per_1k	+1.0	3.1 x 10^-6	Cited more data-rich
has_faq_schema	+0.04	7.1 x 10^-9	Cited more FAQ schema

Every major finding from the within-position-band analysis replicates in this matched-pair design. The competitor pair analysis serves as a quasi-experimental validation: for the same query at similar rank positions, the cited page is consistently more structured, more comparative, more data-rich, and less blog-like.

5.4 Content Uniqueness (Test 7.4)

Using TF-IDF cosine distance from the SERP centroid:

- Cited domains average uniqueness: 0.454
- Not-cited domains average uniqueness: 0.499
- Spearman rho = -0.147 (p = 1.4 x 10^-12)

Cited domains are *less* unique. They cover the same topics as their competitors. However, among already-cited domains, higher uniqueness predicts citation by more platforms (rho = 0.112, p = 5.1 x 10^-4).

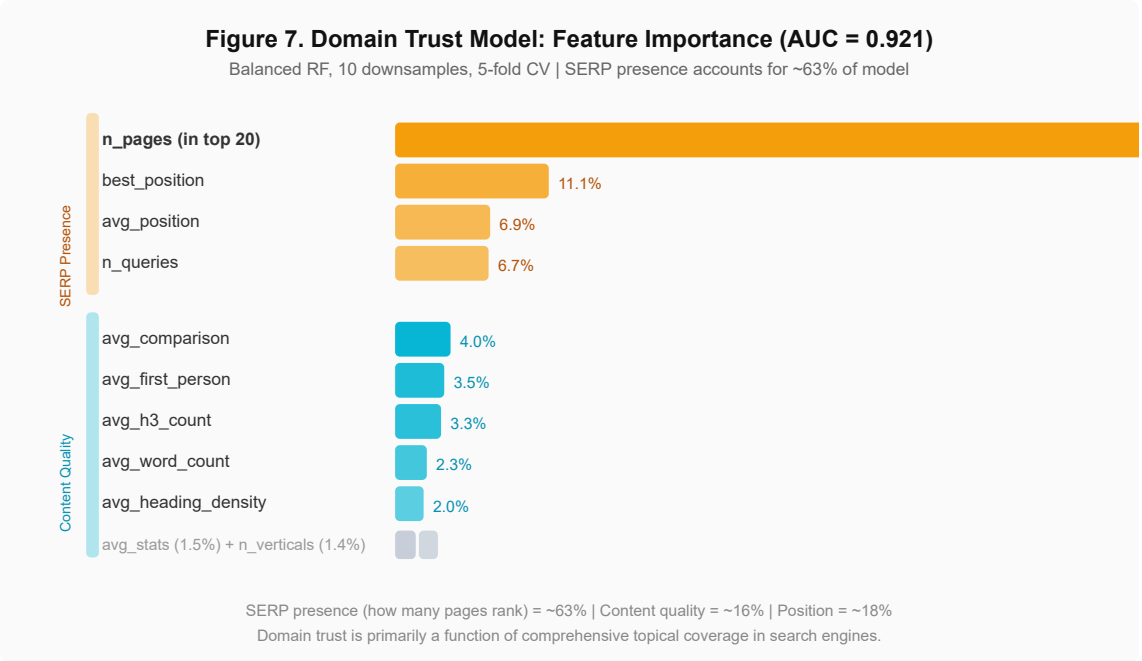
This reveals a two-stage process: (1) match the baseline coverage (gets you cited by one platform), (2) add unique value (gets you cited by multiple platforms). The gateway to AI citation is comprehensive, well-structured coverage, not originality.

5.5 Combined Domain Model (Test 7.5)

Aggregating features to domain level and training a balanced RF:

Feature	Importance
n_pages (in Google top 20)	54.8%
best_position	11.1%
avg_position	6.9%
n_queries	6.7%
avg_comparison	4.0%

avg_first_person	3.5%
avg_h3_count	3.3%
avg_word_count	2.3%
avg_heading_density	2.0%



AUC = 0.921 (SD = 0.007). SERP presence (n_pages + n_queries + n_verticals) accounts for approximately 63% of the model (Figure 7). Content quality features account for approximately 16%. The practical implication: to become a trusted domain for AI citation, build comprehensive topical coverage that ranks across many related queries.

5.6 The Domain Trust Equation

Three tiers emerge:

Tier 1 -- Gateway: SERP Presence (~63%). Rank in Google's top 20 for multiple queries. Domains ranking for 4+ queries have 87%+ citation rates. The per-appearance analysis (Section 5.1) confirms this reflects genuine trust, not merely increased exposure.

Tier 2 -- Content Quality (~16%). Among domains with similar SERP presence, the page-level features from Section 4 (comparison structure, subheading depth, statistics, objective tone) aggregate to differentiate.

Tier 3 -- Network Effects (~5-10%). Being referenced by other AI-cited domains correlates with more citations, though this may be a consequence of Tiers 1 and 2.

Traditional authority metrics (PageRank, Wikipedia links, domain age) do not explain domain trust (Experiment L). Content uniqueness does not explain it either (Test 7.4). The predictors are topical breadth and content quality.

---

6. Discussion

6.1 Implications for Practitioners

For a site owner who already ranks in Google's top 20, the actionable priority list is:

1. **Write comparison content** (d = 0.43): Include "vs" structures, comparison tables, side-by-side analyses. This effect holds across all five intent types, not just comparison queries.
2. **Cover the query's terms explicitly** (d = 0.42): Ensure the page contains the words people search for.
3. **Answer the question early**: Query terms should appear in the first few percent of the page.
4. **Structure with deep subheadings** (d = 0.19-0.40): Use H3 subheadings liberally.
5. **Write 2,000+ words** (d = 0.20): But moderate length, not the "5,000+ word guides" some GEO advice recommends.
6. **Include statistics and data**: Pages cited by 3+ platforms have 7x the statistics density.
7. **Do not write in first person** (d = -0.30 to -0.37): Blog/opinion tone is the strongest negative predictor.
8. **Add FAQ schema**: Significant in all four position bands.
9. **Build topical authority**: Domains ranking for 4+ queries in a niche have 87%+ citation rates.

6.2 What Does Not Matter

Within position bands, the following show no significant effect: page load speed ($p > 0.39$ in all bands), author bylines, content-to-HTML ratio, readability scores, Review schema, and Product schema. This contradicts several widely circulated GEO recommendations.

6.3 Relationship to Prior Work

The Princeton GEO framework (Aggarwal et al., 2024) emphasized statistics, citations, and quotations as key optimization levers. Our findings partially validate this: statistics density matters (significant in 3/4 bands, 7x gradient in cross-platform consensus). But citations and quotations do not replicate as predictors, consistent with our Experiment C replication failure. The comparison structure signal (our strongest predictor) was not identified in the Princeton work, likely because their custom generative engine did not model the comparison-seeking behavior of production platforms. Notably, comparison_signals predicts citation across all five intent types (Section 4.2), not just comparison queries. We hypothesize this is because comparison structure serves as a proxy for information density and extractability: pages that compare alternatives provide structured, citable claims (specific attributes, direct contrasts, ranked recommendations) regardless of whether the user's query was explicitly comparative. A page structured as "Product A vs Product B" contains discrete, attributable assertions that AI systems can extract and cite more readily than narrative prose making the same points in flowing paragraphs.

The content uniqueness paradox (Test 7.4) challenges the intuition that "unique, original content" is the path to AI citation. Cited domains are less unique, not more. They match the baseline coverage, then differentiate through structure and depth. This aligns with Lee (2026a)'s finding that intent matching is the primary determinant of citation source type.

6.4 Limitations of the Position-Band Design

Position-band matching controls for rank but not for all confounds. Pages at the same rank position may still differ systematically: Forbes at position 5 and a niche blog at position 5 are in the same band but are not "equally qualified" in all respects. Domain reputation still leaks through, as pages from trusted domains may have different content features because of better editorial processes, not because those features independently cause citation.

The enriched domain control test (Section 4.5) partially addresses this: with richer domain representations, domain AUC rises to 0.687, modestly exceeding content's 0.673. The honest interpretation is convergence, not content dominance. A within-domain, within-position analysis (comparing pages from the same domain at different positions) would provide stronger causal evidence, but sample sizes within individual domains were insufficient for this analysis in the current dataset.

---

7. Limitations

1. **Observational, not interventional.** We have shown correlations, not causal effects. An A/B test modifying pages and measuring citation changes would provide stronger evidence. However: (a) the within-position-band design controls for the specific confound (domain identity) that killed prior observational work, (b) cross-platform consistency is hard to explain by any single unidentified confound, and (c) the competitor pair analysis (Test 7.3) serves as a quasi-experimental matched-pair validation.
2. **250 queries across 10 verticals.** Larger than prior GEO studies but still a sample of the infinite query space. Different query distributions may produce different results.
3. **Single time point.** AI platforms change citation behavior with model updates. ChatGPT data reflects GPT-5.3 behavior (April 2026).
4. **SERP-conditional findings.** All results apply to "pages that already rank in Google's top 20." We do not claim generalization to pages that do not rank.
5. **Platform scraper differences.** ChatGPT citations came from SSE stream interception, Perplexity from JS fetch interception, Google AI Mode from DOM scraping. Cross-platform consensus (Test 4) mitigates this concern.
6. **Semantic feature coverage.** Query-term coverage was computed for 43% of pages (those with query mappings).
7. **Content uniqueness measured via TF-IDF.** This captures lexical similarity but not conceptual originality.
8. **SERP co-occurrence measured within our 250-query sample.** A domain's true SERP footprint may differ.
9. **n_pages volume confound partially addressed.** Per-appearance citation rates (Section 5.1) demonstrate the effect is not purely arithmetic, but diminishing returns (rho = -0.127) suggest the relationship is more complex than a linear trust signal.

---

8. Conclusion

We introduced position-band matching as a method for isolating page-level citation predictors in AI search, addressing the domain confound that invalidated prior approaches. Within equally-ranked pages, content features and domain identity provide comparable predictive power (AUC 0.673 vs. 0.687), a convergence that contrasts with the domain dominance observed without position control (AUC = 0.975). Position alone achieves AUC 0.704; adding content features lifts this to 0.753, a +0.049 improvement. The top actionable predictors are comparison structure (significant across all five intent types), query-term coverage, subheading depth, and statistical data density. Blog/opinion tone is the strongest negative predictor. At the domain level, SERP co-occurrence (topical breadth) accounts for 63% of a domain trust model (AUC = 0.921), with per-appearance citation rates confirming this reflects genuine trust. Cited domains are less lexically unique than competitors, suggesting the gateway to AI citation is comprehensive baseline coverage, not originality.

These findings provide practitioners with evidence-based priorities for AI search optimization and researchers with a position-controlled methodology for future GEO studies.

Data Availability

The full dataset (10,293 pages, 66 features, 250 queries, citations from 3 AI platforms, Google and Bing SERP positions) is available at https://doi.org/10.5281/zenodo.19398158.

References

- Aggarwal, P., Murahari, V., Rajpurohit, T., Kalyan, A., Narasimhan, K., & Deshpande, A. (2024). GEO: Generative Engine Optimization. *KDD 2024*. DOI: 10.48550/arXiv.2311.09735
- Bagga, P. S., Farias, V. F., Korkotashvili, T., & Peng, T. Y. (2025). E-GEO: A Testbed for Generative Engine Optimization in E-Commerce. *Preprint*.
- BrightEdge (2026). The Ultimate Guide to Claude Search. *Industry Report*.
- Chen, M. L., Wang, X., Chen, K., & Koudas, N. (2025). Generative Engine Optimization: How to Dominate AI Search. *Preprint*.
- Lee, A. (2026a). Query Intent, Not Google Rank: What Best Predicts AI Citation Behavior. *Preprint v5*. DOI: 10.5281/zenodo.18653093
- Sellm (2025). ChatGPT Citation Analysis. *Industry Report (400K+ pages analyzed)*.
- Semrush (2025). AI Overviews Study; AI Search SEO Traffic Study. *Industry Report*.
- Tian, Z., Chen, Y., Tang, Y., & Liu, J. (2025). Diagnosing and Repairing Citation Failures in Generative Engine Optimization. *Preprint*.
- Wen, Y., Zhang, N., Yuan, H., & Chen, X. (2025). Position: On the Risks of Generative Engine Optimization in the Era of LLMs. *Preprint*.