Does Schema Markup Predict AI Citation?

A Cross-Platform Empirical Study of Structured Data and Generative Engine Optimization

Kurt Fischman, Growth Marshal (growthmarshal.io)

with Claude (Anthropic) — AI research contribution acknowledged

Preprint — February 2026 | Not yet peer-reviewed

Abstract

This study examines whether JSON-LD schema markup independently predicts the probability that a web page will be cited in AI-generated responses. We collected 730 AI citations from ChatGPT (GPT-4o with web browsing) and Gemini (1.5 Pro with search grounding) across 75 commercial queries spanning five categories: SaaS and Technology, Health and Medical, Finance and Insurance, Professional Services, and How-To and DIY. Google top-10 organic results for the same queries were collected via SerpAPI as a control set, yielding 1,006 total unique pages analyzed for schema characteristics and domain authority (Ahrefs DR).

Initial pooled analysis produced a significant negative association between schema presence and AI citation (OR = 0.546, $p < .001$) — suggesting schema actively reduced citation probability. This finding proved to be a methodological artifact: Google's ranking algorithm systematically enriches top-10 organic results for schema-bearing pages, inflating schema prevalence in the non-cited control population. A within-Google diagnostic revealed that schema prevalence among AI-cited and non-cited Google pages was statistically indistinguishable (43.1% vs. 44.8%), collapsing the apparent effect entirely. Corrected models using Generalized Estimating Equations with query-clustered standard errors produced a null result for schema presence (OR = 0.678, $p = .296$), entity richness score (OR = 1.001, $p = .833$), and schema-to-query alignment (OR = 1.068, $p = .626$).

The dominant predictor of AI citation was Google organic rank position (OR = 0.762 per position, $p < .001$). Position-1 pages were cited in 43% of queries in which they appeared, declining to 5% at position 7. This gradient implies that each rank position reduces citation odds by approximately 24%, and that AI citation behavior is substantially mediated by the search backend ranking that precedes AI-level content evaluation.

One significant exception emerged: pages implementing Product or Review schema with populated concrete attribute fields — pricing, aggregateRating, specifications — were cited at substantially higher rates than pages implementing generic schema types such as Article, Organization, or BreadcrumbList (61.7% vs. 41.6%, $p = .012$). This attribute-rich advantage was most pronounced among lower-authority domains (DR ≤ 60), consistent with the interpretation that factual payload in structured data partially compensates for weak authority signals. Sophisticated entity-linking techniques — Wikidata sameAs links, genuine @id cross-referencing — appeared on fewer than 4% of schema-present pages and could not be evaluated statistically.

These findings support a more precise version of the schema-helps hypothesis than the practitioner consensus has articulated: attribute-rich schema that provides extractable factual content may confer modest citation advantages for lower-authority domains, while generic schema provides none. The dominant practical implication is that traditional organic rank position remains the primary lever for AI visibility, and that GEO-specific optimization efforts are most productive when directed at content quality and authority rather than generic structured data implementation.

Keywords: AI search optimization, schema markup, JSON-LD, structured data, large language models, AI citation, generative engine optimization, GEO, ChatGPT, Gemini, entity-graph, rank position

---

1. Introduction

The optimization advice is everywhere, and much of it originates from the systems it claims to describe. Ask ChatGPT how to improve your visibility in AI-generated answers and it will recommend structured data implementation. Ask Gemini the same question and it will recommend schema markup. Ask Perplexity, and it will cite SEO publications that were themselves informed by AI-generated summaries of SEO best practices. The result is a self-referential loop: practitioners ask AI platforms for optimization advice, AI platforms reproduce the accumulated SEO consensus of their training data, practitioners implement that advice, and nobody tests whether it works.

This study tests it.

The specific claim under examination is that JSON-LD schema markup — the structured data vocabulary embedded in web pages to help machines parse their content — increases the probability that an AI platform will cite a given page in its generated answers. This claim has achieved near-consensus status in the emerging field of Generative Engine Optimization (GEO). It appears in agency white papers, practitioner playbooks, SEO publication guides, and, notably, in the responses AI platforms themselves generate when asked how to optimize for AI visibility. Its mechanistic logic is coherent. Its empirical basis is essentially nonexistent.

The study reported here began, in part, as a challenge to our own assumptions. Growth Marshal's core methodology — the Modular Knowledge Asset (MKA) architecture — is built on the premise that entity-dense, semantically structured content is more parseable and therefore more citable by

AI retrieval systems. Schema markup is a central element of that framework. When we recognized that the evidentiary chain supporting this premise ran back, ultimately, to AI platforms telling us that schema helps AI platforms, we initiated a study designed to test the hypothesis against observed citation behavior rather than theoretical plausibility. What we found forced us to revise that hypothesis substantially — and to identify the conditions under which a more precise version of it may still hold.

1.1 The Practitioner Claim

Across the GEO practitioner landscape, the schema-helps hypothesis is stated with a confidence that the underlying evidence does not support. Industry publications describe schema markup as essential infrastructure for AI visibility, drawing an analogy to its established role in Google rich results eligibility. Agency frameworks position structured data as a primary optimization lever alongside content quality and domain authority. AI visibility tools score pages partly on schema implementation, implying a causal link that has not been empirically demonstrated.

The hypothesis is not unreasonable on its face. Schema markup was designed precisely to reduce machine parsing uncertainty — to tell automated systems not just what a page says, but what kind of thing it is, what entities it mentions, and how those entities relate to one another. If AI citation systems must classify pages, extract answers, resolve ambiguous entities, and assess source reliability before generating a response, schema theoretically assists at every step. This mechanistic argument has intuitive force. It has also propagated through the field largely untested, amplified by the feedback dynamic described above: LLMs trained on SEO content recommend schema, practitioners implement schema, and the recommendation is reinforced without outcome measurement.

1.2 The Mechanistic Argument

The strongest version of the schema-helps hypothesis rests on the architecture of retrieval-augmented generation (RAG), the technical pipeline through which AI platforms like ChatGPT and Gemini produce web-grounded answers. When a user submits a query that triggers web retrieval, the platform must perform several distinct operations: retrieve candidate pages from a search backend, rank and filter those candidates by relevance and authority, extract the specific information most relevant to the query, resolve any ambiguous entities to verify provenance, and generate a response that cites sources with appropriate confidence.

Schema markup is theoretically relevant to at least three of these stages. At the extraction stage, JSON-LD structured data provides machine-readable field labels — @type, name, price, ratingValue, description — that allow an automated parser to locate specific facts without relying on natural language inference. A Product schema block that explicitly labels a product's price is easier to extract than a sentence that mentions the price in running prose. At the entity resolution stage, sameAs links to canonical identifiers (Wikidata entries, Google Knowledge Graph nodes) allow AI systems to resolve ambiguous entity references and connect a local brand mention to a globally recognized entity. At the provenance stage, schema types like Organization with @id identifiers, address, and foundingDate allow AI systems to verify that a source is a real, established entity rather than a recently created page with no corroborating context.

This mechanistic chain is the basis of the practitioner consensus. It is also, notably, the explanation that ChatGPT itself generates when asked why schema markup should improve AI citation probability — a fact worth noting not as evidence but as illustration of how the feedback loop operates. An AI system trained on content that endorses schema markup will reproduce that endorsement, regardless of whether the training data included empirical tests of the underlying claim.

1.3 Research Questions

This study examines four research questions derived from the practitioner consensus and its underlying mechanistic assumptions.

RQ1. Does JSON-LD schema markup presence independently predict AI citation probability, after controlling for domain authority and retrieval rank position?

RQ2. Does schema quality — measured through entity richness metrics including @id cross-references, sameAs disambiguation links, nesting depth, and property completeness — show a stronger association with AI citation than mere schema presence?

RQ3. Do specific schema types, particularly attribute-rich implementations (Product, Review, aggregateRating) versus generic implementations (Article, Organization, BreadcrumbList), behave differently with respect to AI citation probability?

RQ4. What is the relative predictive power of schema implementation compared to domain authority and Google organic rank position in determining which pages AI platforms cite?

The remainder of this paper proceeds as follows. Section 2 reviews prior work on schema markup, retrieval-augmented generation, and the emerging GEO literature. Section 3 describes our methodology, including query design, citation collection, schema extraction, and statistical approach. Section 4 presents results, beginning with the naïve analysis and its correction, and proceeding through the primary findings and the attribute-rich schema exception. Section 5 addresses limitations. Section 6 discusses implications for practitioners and researchers. Section 7 concludes.

2. Related Work

2.1 Schema Markup and Traditional Search

Structured data markup has a well-documented relationship with traditional search engine behavior. Google's Search Central documentation explicitly links JSON-LD implementation to eligibility for rich results — enhanced SERP features including star ratings, price displays, FAQ dropdowns, and event cards — and has encouraged schema adoption as a means of improving machine readability since the founding of Schema.org in 2011, a collaborative vocabulary standard developed jointly by Google, Microsoft, Yahoo, and Yandex (Guha et al., 2016).

The SEO literature has examined schema's impact on traditional search outcomes with mixed results. Studies have found that rich result eligibility correlates with improved click-through rates for certain query types, particularly for Product and Recipe schema, though the causal relationship between schema implementation and organic rank improvement remains contested. Google has explicitly stated that schema markup is not a direct ranking factor, a position that has not prevented widespread practitioner belief to the contrary. Schema adoption rates across the web remain relatively low, with JSON-LD implementations concentrated among larger, higher-authority domains (Guha et al., 2016).

This adoption pattern has a direct consequence for empirical research into schema's effects. Because schema implementation correlates strongly with domain authority — larger organizations with technical SEO resources are more likely to implement structured data — isolating schema's independent contribution requires careful confound control. As we demonstrate in Section 4, failing to account for this correlation produces spurious findings that misrepresent the true relationship.

2.2 AI Retrieval-Augmented Generation

The AI platforms examined in this study — ChatGPT and Gemini — generate web-grounded answers through retrieval-augmented generation (RAG), a technical architecture in which a base language model is supplemented by real-time web retrieval during inference (Lewis et al., 2020). Rather than relying exclusively on knowledge encoded in model weights during training, RAG systems retrieve candidate documents from a search backend, extract relevant passages, and incorporate that content into the generated response with citation attribution.

The citation pipeline this creates involves several discrete stages, each of which could theoretically be influenced by structured data implementation. At the retrieval stage, candidate pages are identified and ranked by a search backend whose signals include domain authority, content relevance, and technical page quality. At the extraction stage, the system must identify and pull the specific information most relevant to the query — a task that structured field labels could theoretically simplify. At the generation stage, the model synthesizes retrieved content into a coherent response and determines which sources to attribute explicitly.

Critically, the retrieval stage is mediated by a search backend whose ranking behavior is largely independent of the AI platform itself. ChatGPT's web retrieval and Gemini's grounding both rely on underlying search infrastructure that applies its own relevance and authority judgments before AI-level processing begins. This architectural reality has a direct empirical consequence, documented in our results: if AI citation behavior is primarily governed by what the search backend surfaces, then any factor that predicts search ranking will appear to predict AI citation, regardless of whether it influences AI-level processing at all. Distinguishing between search-mediated effects and AI-native effects requires methodological care that prior practitioner work has not consistently applied.

2.3 The Emerging GEO Literature

Generative Engine Optimization as a formal research area was introduced by Aggarwal et al. (2024), who demonstrated that targeted content modifications — adding statistics, citing authoritative sources, including fluent and quotable prose — could improve a page's representation in AI-generated responses by up to 40%. Their work established that AI citation behavior is not random and is responsive to content characteristics, providing the empirical foundation for the field. Importantly, however, their study focused on content-level modifications rather than structured data, and did not examine schema markup as an independent variable.

Lee (2026) advanced the GEO literature through a multi-study investigation of AI citation behavior across four platforms — ChatGPT, Claude, Perplexity, and Gemini — examining 6,699 citations across 120 product recommendation queries. His work documented cross-platform citation divergence, identified query intent as a stronger predictor of citation behavior than previously recognized, and introduced the concept of the shadow corpus to describe sources whose influence on AI outputs operates through training data absorption rather than real-time retrieval. Lee's methodological framework — correlating independently constructed ranking systems against AI citation outcomes, with careful attention to confound structure — informs the statistical approach adopted in this study.

Neither Aggarwal et al. nor Lee directly examined schema markup as a citation predictor. The GEO practitioner literature has moved faster than the academic literature on this question, with agencies and consultancies publishing schema optimization recommendations that rest on mechanistic plausibility rather than empirical evidence. This gap between practitioner consensus and empirical validation is the primary motivation for the present study.

2.4 The MKA Framework and the Feedback Loop Problem

Growth Marshal's Modular Knowledge Asset (MKA) architecture represents one practitioner framework that assigns significant weight to structured data implementation as a pathway to AI visibility. The MKA approach emphasizes entity-first content architecture, answer-dense prose, and schema implementations that link content entities to canonical external identifiers — the kind of sophisticated entity-graph structure discussed in Section 1.2. Like most GEO frameworks, MKA was developed through observation of AI platform behavior and reasoning about retrieval architecture, rather than through controlled empirical testing.

We disclose this institutional context directly because it is relevant to interpreting this study's origins and its findings. The research reported here began as an internal challenge to Growth Marshal's own assumptions — an attempt to test empirically whether the schema component of the MKA methodology was producing measurable citation outcomes or merely reproducing an untested industry consensus. This kind of adversarial self-examination is uncommon in practitioner-generated research, where confirmation bias and commercial incentive typically push toward studies designed to validate existing frameworks rather than stress-test them.

The feedback loop that makes this self-examination necessary deserves explicit description. Large language models used by practitioners to generate optimization recommendations were trained on corpora that include SEO publications, marketing agency content, and practitioner forums — all of which reflect the accumulated consensus that schema markup improves AI visibility. When practitioners ask these same models for advice, the models reproduce that consensus regardless of its empirical basis. The consensus then reinforces itself through implementation without outcome measurement. Breaking this loop requires exactly what this study attempts: collecting observed citation outcomes and testing the consensus claim against behavioral data rather than theoretical plausibility.

3. Methodology

3.1 Experimental Design Overview

This study correlates structured data characteristics of web pages against observed AI citation behavior across two platforms and five commercial query categories. The core design is comparative: for each query in our sample, we collected the pages that ChatGPT and Gemini actually cited alongside the pages that Google ranked in its top-10 organic results for the same query. By comparing the schema characteristics of cited pages against this Google-sourced control set, we can assess whether structured data implementation predicts AI citation independent of the retrieval ranking that mediates it.

The design has one significant limitation that we address directly in the analysis: Google's ranking algorithm itself rewards certain technical page characteristics, including schema implementation, meaning that the control set is not schema-neutral. The methodological consequences of this enrichment, and the corrections we applied, are described in detail in Section 3.6 and revisited in Section 4.2.

3.2 Query Design

We constructed a query set of 75 commercial queries spanning five categories: SaaS and Technology, Health and Medical, Finance and Insurance, Professional Services, and How-To and DIY. Fifteen queries per category were selected to represent the range of intent types most common in commercial AI search behavior: discovery queries ("best project management software for remote teams"), validation queries ("is HubSpot worth it for small businesses"), informational queries ("how does term life insurance work"), and transactional queries ("QuickBooks Online pricing plans").

Category selection was motivated by the commercial orientation of Growth Marshal's client base and by the practical observation that AI citation behavior varies substantially by domain. Commercial categories generate the retrieval-augmented responses most relevant to GEO practitioners — responses in which AI platforms actively retrieve and cite external sources rather than answering from parametric knowledge alone. Queries were constructed to be representative of real user behavior rather than optimized for schema-rich results, and were reviewed to ensure no systematic bias toward domains where schema implementation rates are known to be unusually high or low.

3.3 AI Citation Collection

Each query was submitted three times each to ChatGPT (GPT-4o with web browsing enabled) and Gemini (Gemini 1.5 Pro with Google Search grounding), yielding 450 total platform runs across the 75-query set. All URLs cited in generated responses were extracted, deduplicated within each run, and compiled into a master citation dataset. Across all runs, we collected 730 total citations representing 460 unique URLs. Cross-platform URL overlap was approximately 4%, consistent with Lee's (2026) finding of near-random cross-platform citation agreement and confirming that ChatGPT and Gemini are drawing from meaningfully different retrieval pools despite both using web-grounded generation.

Citation extraction followed a consistent protocol: inline citations and source panel links were both captured, and URLs were normalized to remove session parameters and tracking suffixes before deduplication. Pages that returned HTTP errors or were inaccessible at the time of scraping were excluded from schema analysis but retained in the citation count to preserve accurate citation rate calculations.

3.4 Control Set Construction

Google organic top-10 results were collected for all 75 queries via the SerpAPI service, yielding 644 unique URLs after deduplication across queries where the same page ranked for multiple queries. These Google SERP results serve as the comparison population: pages that a search-grounded AI platform could plausibly have retrieved but did not cite.

The combined dataset of 460 AI-cited URLs and 644 Google SERP URLs, after removing overlap between the two populations, yielded 1,006 unique pages submitted for schema extraction and domain authority scoring. We acknowledge upfront that this control set construction introduces a systematic bias that required methodological correction. Because Google's ranking algorithm rewards technical page quality — including schema implementation — the top-10 organic results are not a schema-neutral sample of the retrievable web. The implications of this bias for interpretation of the raw results, and the query-matched analytical approach used to correct it, are described in Section 3.6.

3.5 Schema Extraction and Scoring

Schema extraction was performed via static HTML scraping of all 1,006 pages. For pages returning valid HTML, we executed a two-pass extraction procedure.

The first pass assessed schema presence and type inventory. For each page, we detected all JSON-LD blocks embedded in script tags, parsed each block for @type declarations, and compiled a complete schema type inventory. Pages were classified as schema-present or schema-absent, and schema-present pages were further classified by primary type — Product, Review, Article, Organization, FAQPage, BreadcrumbList, or other — based on the highest-specificity type detected.

The second pass scored entity richness across seven dimensions for all schema-present pages: @id identifier presence and specificity, sameAs link count and target quality (canonical authority sources versus social media profile links), cross-reference density between schema blocks on the same page, nesting depth of the primary schema object, named property completeness relative to the Schema.org specification for the detected type, content alignment between schema field values and visible page text, and boilerplate detection for templated schema blocks that carry no page-specific information. These seven dimensions were combined into a composite entity richness score ranging from 0 to 100.

Several important measurement limitations apply. Schema extraction was performed on static HTML only, meaning JavaScript-rendered schema was not captured. Microdata and RDFa implementations were similarly excluded, as JSON-LD represents the dominant implementation format recommended by Google and most widely adopted across the web.

3.6 Domain Authority

Ahrefs Domain Rating (DR) was collected for all 1,006 pages via the Ahrefs API batch analysis endpoint. DR is a logarithmic 0–100 scale measuring the strength of a domain's backlink profile relative to all other domains in the Ahrefs index, and serves as the primary proxy for domain authority in this study. Complete DR coverage was achieved across the full dataset, with no pages excluded due to missing authority data.

DR was included in all models as a confound control. The correlation between DR and schema implementation in our dataset was positive and significant ($r = 0.31$, $p < .001$), confirming that higher-authority domains are more likely to implement schema — the adoption pattern documented in prior web-scale studies. Failing to control for DR would conflate schema effects with authority effects throughout the analysis.

3.7 Statistical Approach

Our statistical analysis proceeded in four stages, each motivated by the limitations of the preceding stage.

The first stage estimated a naïve pooled logistic regression predicting AI citation (binary: cited vs. not-cited) from schema presence, entity richness score, and DR across the full 1,006-page dataset. This model is presented not as a primary result but as a transparency measure — to show the raw pattern in the data before methodological correction and to document the artifact that correction was designed to address.

The second stage identified the control set enrichment problem through a within-Google diagnostic: comparing schema prevalence among Google top-10 pages that were subsequently cited by AI platforms against Google top-10 pages that were not. This within-Google comparison holds retrieval eligibility constant and isolates the AI-level citation decision from the search-backend ranking decision.

The third stage estimated the primary corrected models using Generalized Estimating Equations (GEE) with clustered standard errors by query. GEE accounts for the non-independence of observations within the same query — pages retrieved for the same query share retrieval context and are not statistically exchangeable with pages from different queries. All primary hypothesis tests use GEE estimates.

The fourth stage validated GEE results using mixed-effects logistic regression with query-level random intercepts, providing a complementary modeling framework that allows for query-level heterogeneity in baseline citation rates. All analyses were conducted in R (version 4.3.1; R Core Team, 2023) using the geepack (Halekoh et al., 2006) and lme4 (Bates et al., 2015) packages. Effect sizes are reported as odds ratios with 95% confidence intervals throughout.

4. Results

4.1 Descriptive Statistics

The final analysis dataset comprised 1,006 unique pages: 460 AI-cited URLs and 644 Google SERP URLs, with 98 pages appearing in both populations. Of the 1,006 pages submitted for scraping, 769 returned valid HTML and were successfully scored for schema characteristics; the remaining 237 were inaccessible due to HTTP errors, paywalls, or bot-detection responses at the time of scraping.

Citation rates varied substantially across categories. Finance and Insurance produced the highest AI citation density (mean 11.4 citations per query), followed by SaaS and Technology (10.2) and Health and Medical (9.8). Professional Services and How-To and DIY queries generated lower citation counts per query (8.1 and 7.6 respectively). ChatGPT produced slightly more citations per query than Gemini on average (10.4 vs. 9.4), though the difference was not significant at the query level.

Schema prevalence differed markedly between the AI-cited and Google SERP populations in the raw data. Among AI-cited pages, 41.3% implemented some form of JSON-LD schema. Among Google SERP pages that were not subsequently cited by AI platforms, 58.7% implemented schema. This raw differential produced the naïve negative association examined in Section 4.2.

Domain authority distributions reflected expected patterns. Google SERP pages showed a higher mean DR (67.4, SD = 18.2) than AI-cited pages (61.8, SD = 22.6). The subset of AI-cited pages not appearing in the Google top-10 — the breakthrough population — showed substantially lower mean DR (52.3, SD = 24.1).

4.2 Naïve Analysis and Its Correction

The pooled logistic regression predicting AI citation from schema presence, entity richness, and DR across all 1,006 pages produced a significant negative association between schema and citation (OR = 0.546, 95% CI [0.421, 0.708], $p < .001$). Entity richness showed a similarly negative association (OR = 0.544, $p < .001$). These findings, taken at face value, would suggest that schema markup actively reduces AI citation probability — a conclusion that is both counterintuitive and, as the within-Google diagnostic revealed, entirely artifactual.

The artifact is a consequence of control set composition. Google's ranking algorithm rewards technical page quality, including structured data implementation. Top-10 organic results are therefore systematically enriched for schema relative to the broader population of pages. When we pool AI-cited pages (schema prevalence 41.3%) against this schema-enriched Google control set (schema prevalence 58.7%), the mathematical result is a negative association that reflects the control set's elevated schema baseline rather than any AI-level suppression of schema-bearing pages.

The within-Google diagnostic made this artifact visible. Restricting the comparison to pages appearing in Google's top-10 results — holding retrieval eligibility constant — schema prevalence among AI-cited pages (43.1%) and non-cited Google pages (44.8%) was nearly identical. The apparent negative association collapsed entirely once the comparison was confined to a schema-comparable population.

4.3 Primary Corrected Analysis

The query-matched GEE models, controlling for DR and Google rank position, produced a consistent picture across all schema variables: no significant association between structured data implementation and AI citation probability.

Schema presence showed an odds ratio of 0.678 (95% CI [0.332, 1.384], $p = .296$) — directionally negative but consistent with a true null effect. Entity richness score showed OR = 1.001 (95% CI [0.990, 1.012], $p = .833$), indistinguishable from no effect. Schema-type alignment to query intent showed OR = 1.068 (95% CI [0.820, 1.392], $p = .626$). Mixed-effects models produced substantively identical estimates across all three variables.

Table 1 presents the complete model comparison across all schema variables and both modeling frameworks.

Table 1. Logistic Regression Models Predicting AI Citation: Naïve vs. Corrected

Predictor	Naïve OR	Naïve p	GEE OR	GEE p	Mixed OR	Mixed p
Schema presence	0.546	<.001	0.678	.296	0.691	.318
Entity richness	0.544	<.001	1.001	.833	1.002	.801
Schema alignment	—	—	1.068	.626	1.071	.614
Google position	—	—	0.762	<.001	0.758	<.001
Domain Rating	1.024	<.001	1.009	.041	1.010	.038

4.4 The Position Gradient

Google organic rank position emerged as the strongest and most consistent predictor of AI citation in our data, with an odds ratio of 0.762 per rank position ($p < .001$) in the primary GEE model. This estimate implies that each additional rank position reduces the odds of AI citation by approximately 24%.

Position-1 pages were cited in 43% of queries in which they appeared in the top-10 results. Citation rates declined sharply through the ranking: position 2 at 27%, position 3 at 20%, position 5 at 10%, and position 7 at 5%. Below position 7, citation rates were effectively negligible. There is a notable step-down between positions 3 and 4, suggesting AI retrieval systems may apply a soft cutoff that concentrates citation among the top three organic results before declining more gradually through the remainder of the top-10.

The position gradient carries a significant practical implication that reorients the entire schema optimization question. If AI citation probability is primarily governed by search rank, then the primary lever for AI visibility is not structured data implementation — it is whatever drives organic rank.

4.5 The Attribute-Rich Schema Exception

Against the backdrop of null results for generic schema, one finding achieved statistical significance. Pages implementing Product or Review schema with populated concrete attribute fields — pricing, aggregateRating, specifications, availability — were cited at substantially higher rates (61.7%) than pages implementing generic schema types such as Article, Organization, or BreadcrumbList (41.6%, $p = .012$). Pages with no schema at all occupied an intermediate position (59.8%).

The comparison structure here is important. The significant finding is not that Product and Review schema outperforms no schema — the 61.7% versus 59.8% difference is not significant ($p = .71$). The significant finding is that attribute-rich schema substantially outperforms generic schema, which itself performs worse than no schema. Generic schema appears to carry a modest citation penalty relative to schema-free pages, while attribute-rich schema recovers and slightly exceeds the no-schema baseline.

The attribute-rich advantage was most pronounced among lower-authority pages. Among pages with DR of 60 or below, Product and Review schema with concrete attributes was associated with a citation rate of 54.2% compared to 31.8% for generic schema — a gap of more than 22 percentage points. Among high-DR pages (DR > 75), the schema-type difference narrowed considerably, consistent with the interpretation that authority signals dominate citation decisions for established domains, while structured data provides relatively more leverage for pages that lack strong authority signals.

4.6 The Breakthrough Population

Among the 460 AI-cited pages, 292 (63.5%) did not appear in the Google top-10 results for the query that surfaced them. These pages — the breakthrough population — represent cases where AI platforms cited sources that Google's organic ranking did not surface among the ten most relevant results, demonstrating that AI citation behavior is not simply a restatement of Google's top-10.

The breakthrough population skewed toward lower domain authority (mean DR 52.3 vs. 71.6 for Google top-10 pages). Within the 80 low-DR breakthrough pages examined for schema characteristics, entity richness scores for schema-present pages were statistically equivalent between cited and non-cited low-DR pages in the same query pool ($p = .82$). Product and Review schema with concrete attributes remained the consistent differentiator within this population, consistent with the attribute-rich finding in Section 4.5.

4.7 The Absence of Sophisticated Schema

The entity-linking techniques most prominently recommended in GEO practitioner literature — Wikidata sameAs links, true @id cross-referencing, nested entity structures — were functionally absent from our dataset. Wikidata sameAs links appeared on fewer than 4% of schema-present pages. The overwhelming majority of sameAs implementations that did exist pointed to social media profile URLs rather than to Wikidata, DBpedia, or other canonical knowledge graph sources.

This is not a finding about whether sophisticated entity-linking works. It is a finding about the current state of the web: the technique most theorized to improve AI citation is so rarely deployed that empirical evaluation is currently impossible. For practitioners, this means the most promising implementation territory remains completely uncontested. Early adopters of genuine entity-graph architecture are operating in a space where the competition is effectively zero.

5. Limitations

5.1 Control Set Selection Bias

The most consequential limitation of this study is also its most instructive methodological finding. Using Google's top-10 organic results as the non-cited control population introduced a systematic enrichment for schema that inverted the apparent relationship between structured data and AI citation in the naïve analysis. The query-matched GEE approach substantially mitigates this bias by restricting comparisons to pages within the same retrieval-eligible population, but it does not eliminate it entirely. A more methodologically conservative control set would draw from a random sample of pages ranking between positions 11 and 50 for the same queries. Future work should pursue this design.

5.2 Static Schema Extraction

Schema was extracted via static HTML scraping, which captures JSON-LD embedded directly in server-rendered HTML but misses structured data injected by JavaScript after initial page load. As single-page application frameworks have proliferated, the gap between static-scraped and fully-rendered schema has widened. Our extraction method may systematically undercount schema on technically sophisticated pages, potentially attenuating the true schema prevalence in both populations.

5.3 Temporal Misalignment

Citation data and schema data were collected in separate passes, with schema scraping occurring after the initial citation collection window. Pages can and do modify their schema implementations between collection events. While the collection window was short enough to make large-scale schema changes unlikely, we cannot rule out temporal misalignment as a source of measurement error for individual pages.

5.4 The Missing Content Variable

Schema characteristics and domain authority together explain a modest fraction of the variance in AI citation behavior. The most plausible explanation for the residual variance is content quality — the degree to which a page's prose provides clear, extractable, authoritative answers to the query that surfaced it. Answer-first heading structure, entity clarity in running text, factual density, and modular extractability are all candidate predictors that this study did not measure. A study that operationalizes content quality through automated scoring of extractability, entity clarity, and answerability will substantially advance the field beyond what schema-focused analysis alone can achieve.

5.5 Platform Scope and Temporal Stability

This study examined two AI platforms — ChatGPT and Gemini — during a single collection window in late 2025. Neither Perplexity nor Microsoft Copilot nor Google's AI Overviews were included in the primary citation analysis. Lee (2026) documented substantial cross-platform divergence in citation behavior, with cross-platform URL agreement near random at the query level — a finding consistent with our own 4% cross-platform URL overlap. The position gradient finding is likely to be more durable than specific effect size estimates, since it reflects a structural feature of retrieval-augmented generation. Schema-related findings should be interpreted as provisional pending replication.

5.6 The Untestable Hypothesis

Sophisticated entity-graph schema — Wikidata-linked sameAs identifiers, genuine @id cross-referencing across multiple schema blocks, nested entity structures — is the implementation approach that mechanistic reasoning most strongly supports and that practitioner frameworks like MKA most specifically recommend. It is also effectively absent from the current web. Our null results for entity richness therefore speak to schema as it is actually deployed — template-generated, CMS-default, semantically thin — rather than schema as it could be implemented with deliberate entity-graph architecture. The absence of evidence for entity richness effects among existing implementations is not evidence of absence for effects that might emerge from sophisticated implementations that virtually no one has yet built.

6. Discussion

6.1 What Generic Schema Does Not Do

The central finding of this study is a null result, and null results require the same careful interpretation as positive ones. Generic JSON-LD schema markup — the Article, Organization, BreadcrumbList, and WebPage implementations that populate the majority of schema-present pages on the commercial web — does not independently predict AI citation probability after controlling for domain authority and retrieval rank position. This finding is robust across two modeling frameworks, consistent across all five query categories, and survives the methodological correction that reversed the naïve analysis.

The practical implication is direct. The industry consensus that schema markup improves AI visibility is not supported by observed citation behavior for the implementations that dominate the current web. Practitioners who have prioritized generic schema implementation as a GEO strategy, or who have reported schema improvements to clients as AI visibility wins, are operating on a hypothesis that citation data does not validate. It is worth being precise: this finding establishes that schema presence, as currently distributed across commercial web pages, carries no detectable independent signal for AI citation systems. It does not establish that schema is invisible to AI retrieval pipelines or that all schema implementations are equivalent.

6.2 What Attribute-Rich Schema Might Do

The significant finding that Product and Review schema with populated concrete attributes outperforms generic schema (61.7% vs. 41.6%, $p = .012$) points toward a more precise version of the schema-helps hypothesis than the practitioner consensus has articulated. The relevant distinction is not between schema-present and schema-absent pages but between schema that provides extractable factual payload and schema that provides machine-readable metadata without substantive informational content.

This distinction maps onto a specific mechanism. AI citation systems must make confidence assessments about whether a page can reliably answer the query that surfaced it. A page whose schema explicitly labels a product's price, aggregate rating, and key specifications gives the retrieval system concrete, verifiable claims that reduce the inferential burden of citation. A page whose schema declares itself an Article with a name and datePublished provides nothing that a competent HTML parser would not already infer from the title tag and byline. The former reduces extraction uncertainty; the latter does not.

The finding that attribute-rich schema performs best among lower-authority domains carries its own practical implication. For high-authority domains, authority signals are strong enough that AI retrieval systems can commit to citation regardless of schema specificity. For lower-authority domains, the additional confidence provided by extractable factual structure may constitute a meaningful tiebreaker. Attribute-rich schema, on this reading, is most valuable precisely where traditional SEO has the least leverage.

6.3 The Retrieval Position Finding and Its Implications

The dominance of Google organic rank position as a citation predictor — OR = 0.762 per position, position-1 pages cited 43% of the time versus 5% at position 7 — is this study's most practically consequential finding. It implies that AI citation behavior, for current-generation retrieval-augmented systems, is primarily mediated by search backend ranking rather than by AI-native content evaluation. The systems that practitioners are optimizing for AI visibility are, to a substantial degree, optimizing for the same signals that traditional search engines have rewarded for decades.

For GEO practitioners, this finding redirects the optimization calculus. The most reliable path to AI citation for most pages, on most queries, is to rank higher in Google's organic results. Schema optimization, content architecture improvements, and entity-linking techniques all operate downstream of this primary effect. A page that moves from position 5 to position 2 for a target query gains more expected AI citation probability than any structured data intervention this study can identify.

This does not render GEO-specific optimization irrelevant. The breakthrough population — 63.5% of AI-cited pages that did not appear in Google's top-10 — demonstrates that AI platforms exercise independent judgment beyond search ranking. Understanding what characterizes this population, and what content and structural signals drive AI-native citation decisions, is the central empirical question for the next generation of GEO research.

6.4 The LLM Feedback Loop and the Obligation to Test

Perhaps the most broadly applicable finding of this study is not about schema at all. It is about the epistemological hazard created when AI platforms become both the subject and the source of optimization advice. The schema-helps hypothesis achieved practitioner consensus not because it was tested and confirmed but because it was mechanistically plausible, consistent with established SEO wisdom, and endlessly reproduced by the AI systems that practitioners queried for guidance. No one tested it because the AI said it was true, and the AI said it was true because its training data contained the practitioner consensus, which existed because the AI had been saying it.

The obligation this creates is straightforward: GEO recommendations should be tested against observed citation behavior, not derived from AI-generated advice about AI systems. The methodology required is not complex — query design, citation collection, control set construction, and regression analysis are all accessible to practitioner-researchers with modest technical resources. What has been missing is not capability but the willingness to design studies that might falsify the recommendations being made to clients. This study is one attempt to model that posture. The field needs more of them.

6.5 Implications for the MKA Framework

The findings reported here require honest engagement with Growth Marshal's own methodology. The MKA framework assigns significant weight to schema implementation as a component of AI visibility architecture — a recommendation this study finds unsupported for generic implementations and only partially supported for attribute-rich ones. The entity-graph architecture that MKA most specifically recommends cannot be evaluated because it does not yet exist at testable scale on the commercial web.

What the data does support, within the MKA framework, is the emphasis on content architecture over structural metadata. The dominant citation predictor is retrieval rank, and rank reflects content relevance and authority. The MKA components with the strongest empirical basis are those that address content quality directly — answer-first heading structure, entity-clear prose, modular knowledge assets that AI systems can extract without surrounding context.

The honest summary for MKA practitioners is this: implement attribute-rich schema where applicable — particularly Product and Review schema with populated pricing, rating, and specification fields — and do not expect generic schema to provide AI citation advantages that the data does not support. Invest optimization effort into content architecture and organic ranking improvements. Treat sophisticated entity-graph schema implementation as an early-adopter bet on an untested frontier — one that the mechanistic argument supports and the empirical literature has not yet been able to evaluate, because virtually no one has built it yet.

7. Future Research

7.1 Content Quality as the Primary Variable

The most urgent research priority is operationalizing and measuring content quality as a citation predictor. Our models leave the majority of citation variance unexplained, and content quality is the most plausible candidate for that residual. Specific measurable dimensions include answer-first heading structure, entity clarity in running prose, factual density, and the modular extractability of individual content units. A study that combines the citation collection methodology used here with content quality scoring across the full page set would represent a substantially more complete account of AI citation determinants.

7.2 Controlled Schema Intervention

The observational design of this study permits correlation analysis but cannot establish causation. The definitive test of schema's contribution requires a controlled intervention: identifying matched pairs of pages similar in authority, content quality, and rank position, adding or modifying schema on one page in each pair, and measuring citation rate change over a defined observation window. Growth Marshal's client base and ongoing schema implementation work creates a natural laboratory for exactly this kind of study.

7.3 Sophisticated Entity-Graph Architecture at Scale

Testing whether deliberate entity-graph schema — Wikidata-linked sameAs identifiers, genuine @id cross-referencing, nested entity structures — influences AI citation requires building pages that implement this architecture and measuring their citation outcomes against matched controls. A deployment study across 50 to 100 matched page pairs, observed over three to six months, would provide sufficient observations to detect effect sizes consistent with the attribute-rich advantage documented here. No such study has been published. The first group to execute it will produce the most practically consequential finding in the GEO structured data literature.

7.4 Cross-Platform and Longitudinal Replication

The findings reported here are specific to ChatGPT and Gemini during a single collection window. Perplexity, Microsoft Copilot, and Google AI Overviews each operate with distinct retrieval architectures that may produce materially different schema effect patterns. Establishing a citation monitoring infrastructure that tracks schema effects across platforms and over time would transform GEO research from a series of point-in-time snapshots into a longitudinal field capable of detecting how AI citation behavior changes as the technology develops.

8. Conclusion

JSON-LD schema markup, as currently deployed across the commercial web, does not predict AI citation probability. This is the central finding of this study, and it is worth stating without qualification before the caveats that appropriately follow. The generic schema implementations produced by standard CMS plugins, applied without deliberate content-specific customization, represent an untested default that AI citation systems neither reward nor penalize in any statistically detectable way. Practitioners who have positioned schema implementation as a primary GEO lever are recommending an intervention that observed citation behavior does not support.

The dominant predictor of AI citation is retrieval rank position. Position-1 pages are cited at 43% the rate for queries in which they appear; position-7 pages at 5%. This gradient is steep, consistent, and robust across modeling specifications. It reflects the architectural reality of retrieval-augmented generation: AI platforms that rely on search backends for candidate retrieval largely inherit those backends' ranking judgments, and any factor that drives organic rank will appear to drive AI citation.

These findings do not collapse GEO into SEO, and they do not render structured data irrelevant. The attribute-rich exception — Product and Review schema with populated concrete attributes cited at 61.7% versus 41.6% for generic schema, $p = .012$ — demonstrates that schema type and informational content matter in ways that schema presence alone does not capture. The mechanism appears to be uncertainty reduction: schema that provides extractable, verifiable facts gives AI retrieval systems the confidence to cite pages that uncertainty would otherwise exclude. This advantage is largest precisely where traditional SEO has the least leverage — among lower-authority domains where authority signals are too weak to carry citation decisions independently.

The near-total absence of sophisticated entity-graph architecture from the current web means the upper bound of schema's potential contribution remains genuinely unknown. Fewer than 4% of schema-present pages in our sample implemented anything resembling deliberate entity-linking, and the implementations that did exist were overwhelmingly CMS-generated artifacts with no semantic depth. This study establishes what template-generated schema does not do. What carefully architected, entity-rich, Wikidata-linked structured data might do — when built deliberately rather than defaulted into existence — is a question the field has not yet tested and this study cannot answer.

The broader implication extends beyond schema to the epistemology of the GEO field itself. Optimization recommendations that circulate through practitioner communities without empirical validation, amplified by AI platforms that reproduce unverified consensus as confident advice, create a self-reinforcing loop that serves no one well — least of all the clients implementing recommendations on that basis. The methodology required to break this loop is not inaccessible. What has been missing is the willingness to design studies that might return uncomfortable results. This study returned several. The field needs more researchers willing to do the same.

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 (pp. 5–16). Association for Computing Machinery. https://doi.org/10.1145/3637528.3671900

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. Journal of Statistical Software, 67(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Guha, R. V., Brickley, D., & Macbeth, S. (2016). Schema.org: Evolution of structured data on the web. Communications of the ACM, 59(2), 44–51. https://doi.org/10.1145/2844544

Halekoh, U., Højsgaard, S., & Yan, J. (2006). The R package geepack for generalized estimating equations. Journal of Statistical Software, 15(2), 1–11. https://doi.org/10.18637/jss.v015.i02

Lee, A. (2026). Reddit doesn't get cited (through the API): Training data influence, access-channel divergence, and the shadow corpus in AI brand recommendations. *aiXiv*. https://aixiv.science/abs/aixiv.260218.000005

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In Advances in Neural Information Processing Systems (Vol. 33, pp. 9459–9474). https://arxiv.org/abs/2005.11401

Paulheim, H. (2017). Knowledge graph refinement: A survey of approaches and evaluation methods. Semantic Web, 8(3), 489–508. https://doi.org/10.3233/SW-160218

R Core Team. (2023). R: A language and environment for statistical computing (Version 4.3.1). R Foundation for Statistical Computing. https://www.R-project.org/

SerpAPI. (2025). Google Search API documentation. https://serpapi.com

Author Note

Kurt Fischman is the founder of Growth Marshal (growthmarshal.io), an AI search agency specializing in generative engine optimization. This research was conducted independently and was not supported by external funding. Growth Marshal provides schema optimization and AI visibility services to commercial clients; readers should consider this context when interpreting findings related to the commercial value of structured data implementation.

Claude (Anthropic) contributed to research design discussion, statistical interpretation, and manuscript drafting. AI contribution is acknowledged in accordance with emerging norms for transparency in human-AI collaborative research. All empirical data collection, analysis decisions, and conclusions are the responsibility of the human author.

The authors have no conflicts of interest to declare beyond the commercial context disclosed above. Data and analysis code are available upon request to the corresponding author at kurt@growthmarshal.io.

Correspondence concerning this article should be addressed to Kurt Fischman, Growth Marshal.

Email: kurt@growthmarshal.io