AI Search Readiness Score and LLM Citation Frequency: An Empirical Study of 485 Domains Across 30 Intent-Based Queries

Version: 1.0 **Date:** March 2026 **Study Design:** Pre-registered (v1.1), observational cross-sectional **Data & Code:** Available upon request **Conflict of Interest:** The AI Search Readiness Score is developed by the authors. Full methodology and raw data are disclosed.

---

0. Executive Summary

We tested whether AI Search Readiness Score — a composite metric of structural website characteristics — is statistically associated with citation frequency in LLM-generated responses.

We collected 658 citations from 90 LLM query runs (Perplexity `sonar-reasoning-pro`, temperature=0, 3 replicates × 30 queries) and 573 search engine results across three verticals (SaaS, E-commerce, Services). The resulting dataset contained 485 unique domains, of which 441 were successfully scored.

Primary finding: No statistically significant association was detected between AI Search Readiness Score and LLM citation frequency (Pearson $r = 0.009$, $p = 0.849$). The null hypothesis (H0) was not rejected at the $\alpha = 0.05$ significance level. This result held across all domain authority segments and both linear and logistic regression specifications.

Moz Domain Authority was the only statistically significant predictor of citation rate in the OLS model ($\beta = 0.00019$, $p = 0.002$), though effect size was small ($R^2 = 0.022$). This result borderline survived Bonferroni correction for multiple comparisons (adjusted $p = 0.047$).

A post-hoc power analysis confirmed that the study was adequately powered (80% power) to detect correlations $|r| \geq 0.133$, meaning any practically meaningful effect would have been detected. The observed $r = 0.009$ is well below this threshold.

The Trust & Entity Signals sub-score showed a nominally significant negative correlation with citation rate ($r = -0.101$, $p = 0.034$), but this result does not survive multiple comparison correction (Bonferroni-adjusted $p = 0.237$) and is confounded by the negative correlation between Trust & Entity scores and Domain Authority ($r = -0.144$, $p = 0.002$). After controlling for DA in the sub-score regression, Trust & Entity is not significant ($p = 0.105$).

Multiple robustness checks — including a two-part hurdle model, sensitivity analysis with Google Rank, and exclusion of high-authority "giants" — all confirm the primary null finding.

> **Correlation does not imply causation.** These results describe statistical associations in a single cross-sectional snapshot using one LLM model.

---

1. Research Question

1.1 Problem Statement

As large language models increasingly mediate information retrieval, website owners face a new challenge: being cited in AI-generated answers. Despite growing industry discussion around "AI

search optimization" and "Generative Engine Optimization" (GEO), no publicly validated, open-methodology study has tested whether measurable structural website characteristics predict LLM citation probability.

1.2 Research Question

Is AI Search Readiness Score statistically associated with citation frequency in LLM-generated answers, after controlling for domain authority and domain age?

1.3 Hypotheses

- **H0 (Null):** No statistically significant association exists between AI Search Readiness Score and citation frequency.
- **H1 (Alternative):** Higher AI Search Readiness Score is associated with higher citation frequency, controlling for authority-related variables.

1.4 Pre-Registration

This study was pre-registered before data collection began (Pre-Registration Document v1.0, subsequently updated to v1.1 to document two forced deviations — see Section 11.3). The score formula, query list, statistical analysis plan, and publication commitment were frozen prior to the first data collection run.

---

2. Background & Context

The emergence of AI-powered search interfaces — including Perplexity, ChatGPT with web browsing, Google AI Overviews, and Microsoft Copilot — has created a new layer between users and websites. When an LLM generates an answer, it selects a small number of sources to cite from a potentially vast pool. The factors driving this selection are not publicly documented by any provider.

Industry practitioners have proposed various optimization strategies under labels such as "GEO" (Generative Engine Optimization) or "LLMO" (Large Language Model Optimization). These strategies typically focus on structured data, content clarity, authoritativeness signals, and technical accessibility. However, empirical evidence for these claims remains scarce.

The AI Search Readiness Score (developed at getaisearchscore.com) attempts to quantify a website's structural preparedness for AI-powered retrieval across four dimensions: Machine Readability, Extractability, Trust & Entity Signals, and Offering Readiness. This study tests whether this composite metric — or any of its sub-components — predicts actual citation behavior.

We emphasize that this is an exploratory correlational study, not a causal investigation. LLM citation decisions are influenced by retrieval mechanisms, model training data, prompt specifics, and factors that cannot be observed externally.

---

3. Definitions

3.1 AI Citation

A domain is considered "cited" for a given query if it appears in the structured `citations` array returned by the Perplexity API response. This is a binary determination per domain per query run.

For aggregation across 3 replicates per query, a domain is marked as cited for a query if it appeared in 2 or more of the 3 replicate runs (majority vote).

3.2 Citation Rate

For each domain *d*:

```
Citation_Rate(d) = (Number of queries where d is cited via majority vote) / 30
```

Continuous variable [0, 1]. A domain cited in 3 out of 30 queries has a Citation Rate of 0.10.

3.3 AI Search Readiness Score

A composite metric (0–100) computed as the direct sum of four sub-scores, each with a different maximum:

Component	Max Points	Implicit Weight	Focus
Machine Readability (MR)	25	25%	Semantic HTML, Schema.org, crawlability, performance
Extractability (EX)	30	30%	Answer-ready blocks, structured data, heading hierarchy, content density
Trust & Entity Signals (TE)	25	25%	Organization/person schema, authorship, NAP, external trust links
Offering Readiness (OR)	20	20%	Product/offer schema, visual evidence, social proof

```
Total_Score = MR + EX + TE + OR
```

No explicit multipliers are applied — each raw point contributes equally to the total. However, the different maximum values create an implicit differential weighting: Extractability contributes up to 30% of the maximum possible score, while Offering Readiness contributes up to 20%.

The full formula with individual check criteria is documented in Appendix A (frozen before data collection).

3.4 Domain Authority (Moz DA)

Domain Authority as provided by the Moz Links API v2. Continuous variable (1–100). Used as a proxy for overall domain strength and brand recognition.

---

4. Methodology

4.1 Sample Construction

The domain sample was constructed organically from two sources:

1. **LLM citations:** All domains cited by Perplexity across 90 query runs (30 queries × 3 replicates).
2. **Search engine results:** Top-10 results from Google, Bing, and DuckDuckGo for each query, retrieved via SearXNG meta-search.

All unique domains from both sources were included. No manual selection, filtering, or stratification was applied (except exclusion of 44 domains that could not be scored due to timeouts).

Final sample: 485 unique domains (441 with valid scores).

4.2 Query Design

Thirty intent-based queries were designed across three verticals (10 per vertical):

- **SaaS:** Software comparison and selection queries (e.g., "best AI-powered CRM for remote sales teams in 2026")
- **E-commerce:** Product comparison and recommendation queries (e.g., "best long-distance running shoes for marathon training in 2026")
- **Services:** Professional service selection queries (e.g., "how to choose a reliable immigration lawyer for digital nomad visas in Europe")

Design principles: - No brand names in queries (to avoid navigational intent) - Year "2026" included where natural (to prompt freshness-seeking behavior) - Multi-criteria complexity (forcing the LLM to synthesize across multiple dimensions)

The full query list is provided in Appendix B.

4.3 LLM Data Collection

- **Model:** Perplexity `sonar-reasoning-pro` (see Deviation D-1 in Section 11.3)
- **Temperature:** 0 (deterministic output)
- **Replicates:** 3 per query (90 total runs)
- **Inter-request delay:** 2 seconds
- **Citation extraction:** Structured `citations` array from API response
- **Result:** 90/90 successful runs, 658 citation entries

4.4 Search Engine Data Collection

- **Engine:** SearXNG meta-search (Google, Bing, DuckDuckGo aggregation)

- **Format:** JSON API, top-10 results per engine per query
- **Result:** 573 search result entries across 265 unique domains

4.5 Domain Normalization

All URLs were normalized to registrable domains using `tldextract` (Python). For example, `https://www.example.co.uk/page` → `example.co.uk`.

4.6 Scoring

Each domain was scored using the production API of getaisearchscore.com with the frozen formula (Appendix A). The scoring pipeline crawled up to 5 representative pages per domain and computed all four sub-scores.

- **Successfully scored:** 441 / 485 (90.9%)
- **Timeouts:** 44 domains (9.1%) — excluded from regression analysis

4.7 Enrichment

- **Moz Domain Authority:** Retrieved via Moz Links API v2 for all 485 domains.
- **Domain Age:** Retrieved via WHOIS for 456 / 485 domains (94.0%). Missing for 29 domains due to WHOIS privacy or unavailability. Of these, 415 had both age data and valid scores (used in regressions with age as a covariate; remaining domains used age=0 imputation).
- **Wikipedia Presence:** Checked via Wikipedia API. Due to a methodological limitation in brand-name extraction (using only the first subdomain segment), this variable returned 0 for all domains and was excluded from analysis.

4.8 Data Collection Window

All data was collected on March 12, 2026, within a single session (~8 hours). While a single-session design ensures snapshot consistency (no index changes between queries), it also introduces temporal bias: if Perplexity updates its retrieval index multiple times per day, our 8-hour window may not cover a full index refresh cycle. Results should be interpreted as reflecting retrieval state at one specific point in time.

---

5. Dataset Description

5.1 Variables

Variable	Type	Description	N
Citation_Rate	Continuous [0, 1]	Fraction of 30 queries where domain was cited (majority vote)	485
Total_Score	Continuous [0, 100]	AI Search Readiness Score	441
Machine_Readability	Continuous [0, 25]	MR sub-score	441

Variable	Type	Description	N
Extractability	Continuous [0, 30]	EX sub-score	441
Trust_Entity	Continuous [0, 25]	TE sub-score	441
Offering_Readiness	Continuous [0, 20]	OR sub-score	441
Moz_DA	Continuous [1, 100]	Moz Domain Authority	485
Domain_Age	Continuous [0.1, 39.1]	Years since registration	415

5.2 Descriptive Statistics

Variable	Mean	SD	Min	Median	Max
Citation_Rate	0.014	0.022	0.000	0.000	0.300
Total_Score	50.4	13.3	5	49.0	95
Machine_Readability	19.5	4.4	0	21.0	25
Extractability	15.8	6.1	0	15.0	30
Trust_Entity	7.0	4.5	5	5.0	25
Offering_Readiness	8.1	8.3	0	7.0	20
Moz_DA	43.0	23.6	1	39.0	100
Domain_Age (years)	15.5	9.5	0.1	—	39.1

5.3 Citation Distribution

- **Cited domains** (Citation_Rate > 0): 194 / 485 (40.0%)
- **Uncited domains** (Citation_Rate = 0): 291 / 485 (60.0%)

The citation rate distribution is heavily right-skewed: most domains are never cited, while a small number are cited across multiple queries (maximum: 30% of queries). The median citation rate is 0.000.

5.4 Topic Distribution

Each of the 485 domains was evaluated against all 30 queries, yielding a balanced design with 4,850 domain-query pairs per vertical (SaaS, E-commerce, Services). Note that most domains are topically relevant to only one vertical (~10 queries), so the effective signal window per domain is narrower than the full 30-query set. The citation rate denominator of 30 therefore includes ~20 queries for which most domains have near-zero citation probability regardless of structural quality.

---

6. Statistical Analysis

All analyses were conducted using Python 3.14 with `scipy.stats` (v1.15) and `statsmodels` (v0.14). Significance threshold: $\alpha = 0.05$.

6.1 Statistical Power

A post-hoc power analysis was conducted to determine the minimum effect size detectable in this sample. At $n = 441$, $\alpha = 0.05$ (two-sided), and 80% power, the minimum detectable Pearson $|r|$ is **0.133**. This means the study was adequately powered to detect any correlation explaining more than ~1.8% of variance. Effects smaller than $|r| = 0.133$ cannot be reliably distinguished from zero in this sample, but such effects would also be of negligible practical significance.

6.2 Bivariate Correlations

Seven bivariate correlations were computed (Table 1). Because multiple comparisons inflate the family-wise error rate, we report both uncorrected p-values and Bonferroni-adjusted p-values ($\alpha\_adjusted = 0.05 / 7 = 0.007$).

Table 1: Correlation between predictor variables and Citation Rate (n = 441)

Variable	Pearson r	p (raw)	p (Bonferroni)	Spearman ρ	p (raw)
Total_Score	+0.009	0.849	1.000	+0.079	0.097
Machine_Readability	+0.008	0.870	1.000	+0.079	0.096
Extractability	–0.013	0.790	1.000	+0.061	0.200
Trust_Entity	–0.101	0.034	0.238	–0.114	0.017
Offering_Readiness	+0.082	0.084	0.588	+0.087	0.069
Moz_DA	**+0.129**	**0.007**	**0.047**	**+0.109**	**0.022**
Domain_Age	+0.026	0.593	1.000	+0.028	0.571

Key observations:

1. **Total AI Score shows no significant correlation with citation rate** ($r = 0.009$, $p = 0.849$). This is a near-zero effect, well below the minimum detectable threshold of $|r| = 0.133$.
2. **Moz Domain Authority is the only predictor surviving Bonferroni correction** ($r = 0.129$, raw $p = 0.007$, adjusted $p = 0.047$).
3. **Trust & Entity Signals shows a nominally significant negative correlation** ($r = –0.101$, raw $p = 0.034$), but this **does not survive Bonferroni correction** (adjusted $p = 0.238$). Given 7 simultaneous tests, the expected number of false positives at $\alpha = 0.05$ is 0.35, making this result indistinguishable from chance. Furthermore, Trust & Entity is negatively correlated with Domain Authority ($r = –0.144$, $p = 0.002$), suggesting confounding (see Section 9.2).
4. Neither Machine Readability nor Extractability reach significance.
5. Offering Readiness approaches nominal significance ($p = 0.084$) with a positive direction, but does not survive correction.

6.3 OLS Regression

Model: `Citation_Rate ~ AI_Score + Moz_DA + Domain_Age`

Note: Google Rank was not available at the domain level due to a data aggregation limitation. A sensitivity analysis including Google Rank on the subset of domains with rank data is reported in Section 6.6. Topic was not included as all domains were evaluated across all topics (balanced design).

Table 2: OLS Regression Results (n = 441)

Predictor	Coefficient	*p*-value	95% CI
Intercept	+0.0074	0.114	[−0.0018, +0.0167]
AI_Score	+0.000046	0.557	[−0.000107, +0.000199]
Moz_DA	**+0.000189**	**0.002**	**[+0.000068, +0.000311]**
Domain_Age	−0.000200	0.130	[−0.000459, +0.000059]

- **R^2 = 0.022** (Adjusted R^2 = 0.016)
- **F-statistic = 3.33** (p = 0.020)
- **AI Score is not a significant predictor** (p = 0.557)

The model explains only 2.2% of variance in citation rate. Moz DA is the only individually significant predictor. Domain Age shows a negative direction but does not reach statistical significance (p = 0.130). Variance Inflation Factors (VIF) for all three predictors were below 1.9, indicating no multicollinearity concern.

Methodological note on zero-inflated dependent variable. The citation rate distribution is heavily zero-inflated (60% zeros, median = 0). Standard OLS assumes a continuous, approximately normal residual distribution, which is violated here. Results should be interpreted with this caveat. A two-part hurdle model is reported in Section 6.7 as a robustness check.

6.4 Sub-Score Regression

Model: `Citation_Rate ~ MR + EX + TE + OR + Moz_DA + Domain_Age`

Decomposing the composite score into its four sub-components does not materially improve explanatory power.

Table 3: Sub-Score Regression Results (n = 441)

Predictor	Coefficient	*p*-value
Intercept	+0.0123	0.034
Machine_Readability	−0.000170	0.516
Extractability	+0.000133	0.448
Trust_Entity	−0.000385	0.105
Offering_Readiness	+0.000229	0.096
Moz_DA	**+0.000177**	**0.005**
Domain_Age	−0.000220	0.125

- **R^2 = 0.035**

No individual sub-score reaches statistical significance at $\alpha = 0.05$. Trust & Entity does not reach significance (p = 0.105) and its negative direction is consistent with the confounding interpretation (high-TE domains have lower DA; see Section 9.2). Offering Readiness also approaches significance (p = 0.096) in the positive direction but does not cross the threshold.

6.5 Logistic Regression

Model: `is_cited ~ AI_Score + Moz_DA + Domain_Age`

Table 4: Logistic Regression Results (n = 441)

Predictor	Coefficient	*p*-value	Odds Ratio
Intercept	−1.284	0.004	0.277
AI_Score	+0.013	0.071	1.013
Moz_DA	+0.009	0.144	1.009
Domain_Age	−0.009	0.487	0.991

- **Pseudo $R^2 = 0.008$**

No predictor reaches significance in the logistic specification. AI Score has $p = 0.071$ — marginally non-significant. The Pseudo R^2 of 0.008 indicates the model has negligible discriminative power for predicting citation vs. non-citation.

OLS vs. Logistic divergence. A notable pattern emerges when comparing the OLS and logistic specifications: DA is significant in OLS ($p = 0.002$) but not in the logistic model ($p = 0.144$), while AI Score is further from significance in OLS ($p = 0.557$) than in logistic ($p = 0.071$). This divergence suggests that DA primarily predicts citation *intensity* (how often a domain is cited, conditional on being cited at all) rather than citation *selection* (the binary event of being cited vs. not). The hurdle model in Section 6.7 confirms this interpretation: DA is highly significant in the intensity part ($p = 0.001$) while the cited-vs-uncited mean comparison shows no significant DA difference ($p = 0.248$). This has a practical implication: Domain Authority may function as an "amplifier" — domains that enter the retrieval pool for other reasons get cited more frequently if they have higher DA — rather than as a "gate" determining pool entry.

6.6 Sensitivity Analysis: Including Google Rank

Google Rank data was available at the domain-query pair level for 265 domains that appeared in search engine results. To test whether Google Rank is a confounding variable — potentially mediating the DA–citation relationship — we re-ran the OLS regression on the subset of 224 scored domains with rank data.

Table 5: OLS with Google Rank (n = 224)

Predictor	Coefficient	*p*-value
Intercept	+0.006866	0.334
AI_Score	−0.000015	0.902
Moz_DA	**+0.000337**	**0.001**
Domain_Age	−0.000464	0.048
Google_Rank	**−0.001177**	**0.044**

- **$R^2 = 0.068$** (Adjusted $R^2 = 0.051$)

Key observations: - **AI Score remains non-significant** ($p = 0.902$), even further from significance than in the primary model. - **DA remains significant** ($p = 0.001$) and is not attenuated by including

Google Rank. - **Google Rank is statistically significant** (p = 0.044) with a negative coefficient (higher rank number = lower position = lower citation rate, which is the expected direction). This suggests Google search ranking independently contributes to LLM citation probability. - **DA and Google Rank are not correlated** in this subsample (r = 0.010, p = 0.885), meaning DA is not simply a proxy for search visibility in this dataset. This may reflect the selection effect: all domains in this subsample already appeared in search Top-10, limiting rank variance.

6.7 Robustness Check: Two-Part Hurdle Model

Given the zero-inflated dependent variable (60% zeros), we estimated a two-part hurdle model to separate (a) whether a domain is ever cited (selection) from (b) how frequently it is cited given that it is cited (intensity):

- **Part 1 (Selection):** Logistic regression on is_cited (n = 441) — reported in Section 6.5.
- **Part 2 (Intensity):** OLS regression on Citation_Rate, restricted to cited domains with valid scores only (n = 179 of 194 cited; 15 cited domains lacked scores due to crawl timeouts).

Table 6: Hurdle Part 2 — OLS on cited domains only (n = 179)

Predictor	Coefficient	p-value
Intercept	+0.036062	<0.001
AI_Score	–0.000194	0.097
Moz_DA	**+0.000312**	**0.001**
Domain_Age	–0.000274	0.183

- **R^2 = 0.088** (Adjusted R^2 = 0.072)

Among cited domains, AI Score has a *negative* coefficient (p = 0.097) — the opposite of the hypothesized direction. Moz DA remains the dominant predictor and is significant. The hurdle model explains more variance (R^2 = 0.088 vs. 0.022) in the cited-only subsample, but **this improvement comes entirely from DA, not from AI Score**.

Comparison of cited vs. uncited domains: - Mean AI Score: cited 51.7 vs. uncited 49.6 (t-test p = 0.103, not significant) - Mean DA: cited 43.1 vs. uncited 40.6 (t-test p = 0.248, not significant)

Neither AI Score nor DA significantly differentiate cited from uncited domains in a simple mean comparison, suggesting that the selection mechanism is not well-explained by any measured variable.

A Zero-Inflated Poisson model was also attempted but produced numerically unstable results (singular Hessian, all p-values = NaN), likely due to the sparse count distribution (most non-zero counts are 1). This model specification was not interpretable and is not reported.

---

7. Segment Analysis

7.1 By Domain Authority

Table 7: AI Score × Citation Rate Correlation by DA Segment

Segment	n	Pearson *r*	*p*-value	Mean Score	Mean CR
DA < 30	143	+0.075	0.371	51.0	0.011
DA 30–50	159	+0.055	0.491	51.7	0.013
DA > 50	139	–0.024	0.782	48.3	0.019

No segment shows a statistically significant correlation. The direction in the DA > 50 segment is slightly negative, though not significant. Mean citation rate increases monotonically with DA, consistent with the bivariate correlation finding.

7.2 Giants Analysis (DA > 80)

Group	n	Mean Score	Mean CR
Giants (DA > 80)	35	43.9	0.016
Non-giants (DA ≤ 80)	406	51.0	0.014

Giants have a *lower* average AI Score (43.9 vs. 51.0) but a similar citation rate (0.016 vs. 0.014). This is consistent with the interpretation that brand authority — not structural readiness — drives citation for high-authority domains.

Removing giants from the analysis does not change the conclusion: - Without giants: Pearson *r* = +0.043, *p* = 0.391

---

8. Replicate Stability

Each query was run 3 times to measure the stability of LLM citation behavior at temperature=0.

8.1 Domain-Level Consistency

Of 297 unique domain-query citation instances: - **47.5%** appeared in all 3 replicates (high stability) - **23.2%** appeared in 2 of 3 replicates - **29.3%** appeared in only 1 replicate

8.2 Source Set Overlap (Jaccard Analysis)

A more granular analysis compared the full source sets returned by each replicate run using the Jaccard similarity index (|intersection| / |union|) at the domain level.

Table 8: Source Set Overlap Between 3 Replicates Per Query

Metric	Value
Mean sources per run	7.2 (SD = 1.4, range 3–10)
Mean 3-way Jaccard	0.512
Median 3-way Jaccard	0.500
Mean pairwise Jaccard	0.644

Metric	Value
Queries with identical sets (Jaccard = 1.0)	2 / 30 (6.7%)
Queries with high overlap (Jaccard ≥ 0.8)	4 / 30 (13.3%)
Queries with low overlap (Jaccard < 0.5)	14 / 30 (46.7%)

On average, each run returns ~7.2 sources. Per-run source composition is approximately 4.7 core (appear in all 3 replicates, ~65% of each run), 1.5 partial (appear in 2/3, contributing ~1.0 per run), and 1.0 singleton (appear in 1/3 only) — making ~35% of each run's sources variable rather than the ~50% suggested by set-level Jaccard. The discrepancy arises because Jaccard operates on set intersection/union, which penalizes non-overlap more heavily than per-source counting.

Caveat on Jaccard with small sets. With $|S| \approx 7$, Jaccard is sensitive to single-source changes: adding or removing one source shifts J by ~0.07–0.14. The variance in per-query Jaccard values is therefore partially a mathematical artifact of small set sizes, and individual J values should be interpreted with caution. The aggregate pattern across 30 queries is more reliable than any single query's value.

Selected per-query examples:

Query	Sources A/B/C	Intersection	Union	Jaccard
q9 (helpdesk software)	7/7/7	7	7	1.000
q14 (ergonomic chairs)	7/7/7	7	7	1.000
q22 (SEO agencies)	8/7/8	7	8	0.875
q7 (HR software)	8/7/8	3	14	0.214
q10 (design tools)	7/6/7	2	12	0.167

The variance is substantial: some queries produce highly deterministic source sets (Jaccard = 1.0), while others return almost entirely different sources across replicates (Jaccard < 0.2), despite identical prompts and temperature=0. This query-dependent heterogeneity is analyzed in Section 8.5.

8.3 URL-Level vs Domain-Level Overlap

To distinguish between domain-level retrieval variance and page-level canonical selection, we repeated the Jaccard analysis at the full URL level.

Level	Mean Jaccard	Median Jaccard
Domain	0.512	0.500
URL	0.457	0.455
Delta	0.055	0.045

The small delta (0.055) indicates that **the dominant source of instability is domain selection, not page selection within a domain**. When the retrieval layer "decides" to cite a domain, it typically selects the same page across replicates.

However, in 13 out of ~110 stable domain-query instances (12%), the LLM cited the same domain across all 3 runs but selected *different pages*. Inspection reveals two patterns:

1. **Semantically adjacent pages** (9 cases): The LLM oscillates between closely related pages on the same site — e.g., nerdwallet.com's `best/accounting-software` vs. `best/cloud-accounting-software`, or hermanmiller.com's `best-office-chairs-for-back-pain` vs. `ergonomic-chairs`. This suggests the retrieval layer returns multiple candidate documents per domain, and the final page selection among near-duplicates is stochastic.
2. **Catalog/video platforms** (4 cases, all YouTube): Different videos or catalog entries on the same topic. These represent platforms where the "document" unit is granular (individual videos/listings), making canonical selection inherently noisy.

This observation is consistent with a two-stage retrieval architecture: (1) domain-level retrieval followed by (2) intra-domain page ranking. The small delta suggests the primary variance source is domain selection rather than page selection. However, an alternative explanation — document-level retrieval with emergent domain clustering due to semantic proximity of pages within a domain — cannot be ruled out. The data are consistent with both models. Future research on LLM canonical document selection could leverage this signal with larger replicate counts.

8.4 Per-Query Source Stability

Decomposing each query's source set into core (appear in 3/3 runs), partial (2/3), and singleton (1/3) domains reveals extreme heterogeneity:

Table 9: Per-Query Source Stability (sorted by core%)

Query	Topic	Union	Core	Partial	Single	Core%
q9 (helpdesk software)	SaaS	7	7	0	0	100%
q14 (ergonomic chairs)	E-com	7	7	0	0	100%
q25 (cloud consulting)	Serv	9	8	1	0	89%
q22 (SEO agencies)	Serv	8	7	1	0	88%
q1 (AI CRM)	SaaS	9	7	0	2	78%
q4 (low-code platforms)	SaaS	8	6	1	1	75%
q15 (electric bikes)	E-com	8	6	1	1	75%
q13 (laptops)	E-com	4	3	0	1	75%
q10 (design tools)	SaaS	12	2	4	6	17%
q30 (architecture firms)	Serv	10	2	4	4	20%
q7 (HR software)	SaaS	14	3	3	8	21%
q20 (smart kitchen)	E-com	8	2	5	1	25%
Mean		**9.9**	**4.7**	**2.3**	**2.9**	**47%**

Core% ranges from 17% to 100% across queries. This variance is far larger than what would be expected if retrieval stochasticity were a uniform system property. Instead, stability appears to be a property of the query itself.

8.5 Information Entropy Hypothesis

The query-dependent stability pattern suggests a potential explanatory variable: the size of the relevant information space for each query. Queries with many competing sources (high "information entropy") should produce less stable citation sets because the retrieval layer has more equally-viable candidates to choose from.

We operationalized information entropy as **union size** — the total number of unique domains cited across all 3 replicates — and tested its correlation with core%.

Table 10: Information Entropy Proxy vs Retrieval Stability

Test	r	p
Union Size vs Core% (Pearson)	−0.630	0.0002
Union Size vs Core% (Spearman)	−0.672	< 0.0001
Union Size vs Core Count (Pearson)	−0.194	0.305
Avg Run Size vs Core% (Pearson)	0.104	0.584

The correlation between union size and core% is strong and highly significant ($r = -0.630$, $p = 0.0002$). Queries with larger information spaces produce less stable citation sets.

Critically, the absolute number of core sources does not decrease as the information space grows ($r = -0.194$, $p = 0.305$). This means the core remains roughly constant at ~3–7 domains per query; what changes is the size of the variable tail. Low-entropy queries (union ≤ 8, n = 8) have mean core% = 71.9%; high-entropy queries (union ≥ 12, n = 8) have mean core% = 31.0%.

This finding reframes retrieval instability: rather than being a uniform system property (as the aggregate Jaccard = 0.512 might suggest), **citation stability is query-dependent and correlates with the competitiveness of the information space**. The retrieval layer appears to maintain a stable core of high-relevance sources while sampling variably from a larger pool of equally-viable candidates — and the size of that pool varies dramatically across queries.

Topic cluster does not explain the variance: SaaS (mean core% = 49.5%), E-commerce (56.3%), and Services (47.8%) are similar. The entropy appears to be query-specific rather than vertical-specific.

8.6 Core vs Singleton: Domain Authority Comparison

If the variable sources were systematically different from core sources (e.g., due to deliberate diversity injection), we would expect measurable differences in their characteristics. We tested whether core domains (cited in 3/3 replicates for a given query) differ from singleton domains (cited in 1/3) on Domain Authority:

Group	n	Mean DA	Median DA
Core (3/3 runs)	132	43.9	42.5
Singleton (1/3 runs)	85	44.6	41.0
Difference		**−0.7**	**+1.5**

T-test: $t = -0.215$, $p = 0.830$. Mann-Whitney U: $p = 0.821$.

There is no measurable difference. Core and singleton domains are statistically indistinguishable on DA. This argues against deliberate diversity injection (which would produce systematically different singletons) and is more consistent with stochastic sampling from a pool of similarly-authoritative candidates.

Domain citation frequency distribution. Across all 90 runs, 275 unique domains were cited 658 times. The distribution is moderately concentrated: the top 13.5% of domains account for 25% of all citations, and 51.6% of domains were cited ≤ 2 times. A log-log regression yields slope $\alpha = -0.448$ ($R^2 = 0.659$), indicating moderate power-law concentration. The most-cited domain (youtube.com, 33 citations) appears to be an outlier, but further analysis reveals this is driven by content-format relevance rather than retrieval privilege (see below).

YouTube as a case study: retrieval bias vs content relevance. YouTube's 33 citations (5.1% of total) across 9 of 30 queries could suggest a platform-level retrieval prior. However, the distribution across verticals tells a different story:

Vertical	YouTube citations	Total citations	YouTube %
E-commerce	13	99	13.1%
SaaS	3	104	2.9%
Services	0	111	0.0%

YouTube appears almost exclusively in E-commerce queries for physical products where video reviews are a natural content format: laptops for video editing (40% YouTube), electric mountain bikes (30%), running shoes (27%), mirrorless cameras (22%), smart kitchen appliances (22%). YouTube is entirely absent from Services queries (immigration lawyers, dental clinics, translation services, architectural firms) where video is not the natural content format.

This pattern suggests **modality-sensitive retrieval** rather than domain privilege: video platforms appear in citations primarily when the query context favors visual inspection (physical products where text is a less efficient communication format than video).

However, a confound remains: YouTube's dominance over other video platforms (Vimeo, TikTok, vendor-hosted video) may reflect **transcript availability bias** rather than pure content relevance. YouTube's machine-readable transcripts are likely more accessible to retrieval pipelines than video content on other platforms. The data cannot distinguish between "YouTube is cited because video is relevant" and "YouTube is cited because its transcripts are indexable." Both mechanisms may operate simultaneously — a format bias (favoring platforms with accessible text representations of video content) layered on top of content relevance.

Three limitations constrain interpretation: (1) the uniform comparison/recommendation query design provides no intent-type variance (informational vs tutorial vs definition) for a proper modality test; (2) 30 queries yield only 10 per vertical — too few for robust vertical-level inference; (3) results reflect a single platform (Perplexity) and cannot be generalized to ChatGPT, Gemini, or Copilot retrieval architectures.

8.7 Implications for the Dependent Variable

This finding has direct consequences for interpreting the regression results:

1. **Measurement noise in Citation Rate.** Approximately 35% of sources in any single run are variable (not present in all 3 replicates). Citation Rate (based on majority vote across 3 replicates) therefore contains measurement error. The theoretical maximum R^2 for any predictor is bounded by the reliability of the dependent variable.
2. **Majority vote as a partial fix.** The 2-of-3 majority vote filters out some noise (singleton domains are classified as uncited), but cannot eliminate it entirely. A domain with true 50% citation probability has a 50% chance of being classified as cited (2+ appearances) by majority vote — which is correct in expectation but adds binomial noise.
3. **Retrieval-layer stochasticity is query-dependent, not uniform.** The fact that temperature=0 (deterministic generation) still produces variable source sets confirms that the stochasticity originates in the retrieval (RAG) layer, not the generation layer. However, this stochasticity is not a fixed system property: core% ranges from 17% to 100% across queries and correlates strongly with information space size (r = –0.630, Section 8.5). Queries with many equally-viable candidate sources produce less stable citations, while queries with a clear set of authoritative sources produce highly deterministic results.
4. **Variable sources are not systematically different.** Core and singleton domains have identical DA distributions (p = 0.830), arguing against deliberate diversity injection and in favor of stochastic sampling from equally-qualified candidates (Section 8.6).
5. **Future studies should increase replicates.** Three replicates provide limited averaging. Increasing to 5–10 replicates per query would substantially reduce measurement noise and improve statistical power for detecting small effects.

---

9. Findings Summary

9.1 Primary Result

H0 is not rejected. AI Search Readiness Score does not show a statistically significant association with LLM citation frequency in this sample (Pearson r = 0.009, p = 0.849; OLS p = 0.557).

9.2 Secondary Findings

1. **Domain Authority (Moz DA) is the only predictor to borderline survive multiple comparison correction** (r = 0.129, Bonferroni-adjusted p = 0.047), though the effect is small (R^2 contribution ≈ 2%). DA remained significant across OLS specifications: primary OLS (p = 0.002), OLS with Google Rank (p = 0.001), hurdle model cited-only (p = 0.001), and sub-score regression (p = 0.005). However, DA was *not* significant in the logistic regression (p = 0.144) — see point 7 below.
2. **Trust & Entity Signals sub-score shows a nominally significant negative correlation** with citation rate (r = –0.101, raw p = 0.034), but this result has three important caveats: (a) it does not survive Bonferroni correction (adjusted p = 0.237); (b) Trust & Entity is negatively correlated with Domain Authority (r = –0.144, p = 0.002) — domains with high TE scores (≥ 15, n = 82) have substantially lower average DA (34.1 vs. 43.3) and are predominantly niche service sites (medical tourism, translation, event planning, coffee roasters); (c) after controlling for DA in the sub-score regression, TE is not significant (p = 0.105). The negative

bivariate association is therefore likely a confounding artifact: high-TE domains tend to be low-DA niche businesses, and low-DA domains are less cited.

3. **The overall model explains very little variance** ($R^2 = 0.022$). Even the best-performing specification (hurdle model, cited-only subsample) explains only 8.8% — and none of that explanatory power comes from AI Score. Most of what determines LLM citation behavior is not captured by any of the variables in this study.
4. **LLM citation stability is query-dependent, not uniform.** The mean 3-way Jaccard index is 0.512, but this masks extreme heterogeneity: core% ranges from 17% to 100% across queries. Retrieval stability correlates strongly with information space size (union size vs core%: $r = -0.630$, $p = 0.0002$). Low-entropy queries (union ≤ 8) show 71.9% core stability; high-entropy queries (union ≥ 12) show only 31.0%. The absolute number of core sources remains roughly constant (~3–7); what varies is the size of the stochastic tail. Core and singleton domains are indistinguishable on DA ($p = 0.830$), indicating the variable sources are not systematically different but rather equally-viable candidates sampled stochastically. This places a query-dependent upper bound on predictive power and reframes "retrieval instability" as a reflection of topic competitiveness rather than system noise.
5. **Giants (DA > 80) have lower average structural scores but are cited at comparable rates**, reinforcing that brand authority operates independently of structural readiness.
6. **The hurdle model reveals that AI Score has a negative (non-significant) coefficient** among cited domains (coef = −0.000194, $p = 0.097$), meaning that among domains that do get cited, higher structural scores are weakly associated with *lower* citation frequency. This further undermines the hypothesized positive relationship.
7. **DA predicts citation intensity, not citation selection.** DA is significant in OLS ($p = 0.002$) and the hurdle intensity model ($p = 0.001$), but not in the logistic regression ($p = 0.144$) or the cited-vs-uncited mean comparison ($p = 0.248$). This pattern suggests DA functions as an "amplifier" — increasing citation frequency for domains that are already in the retrieval pool — rather than as a "gate" determining pool entry. The mechanism driving initial selection into the retrieval pool remains unidentified by any variable in this study.

---

10. Limitations

This study has substantial limitations that must be considered when interpreting results:

1. **Single model, single point in time.** Results reflect Perplexity `sonar-reasoning-pro` behavior on March 12, 2026. Other models (ChatGPT, Gemini, Claude) may weigh structural factors differently. Model updates could change citation behavior overnight. The 8-hour data collection window may not cover a full retrieval index refresh cycle, introducing temporal snapshot bias.

2. **Retrieval layer opacity.** We observe the final citation output but cannot observe the retrieval process. The LLM's RAG pipeline likely has its own ranking logic, index freshness constraints, and source preferences that are independent of the scored structural factors.

3. **Content relevance not controlled.** The most important omitted variable is content relevance — the degree to which a domain's content actually answers the specific query posed. A perfectly structured but topically irrelevant site will never be cited. This study measures structural characteristics but does not measure how well each domain's content matches each query (e.g., via BM25 or semantic similarity). Content relevance is likely the dominant

driver of citation decisions and its absence as a control variable means the $R^2 = 0.022$ should not be interpreted as "structural factors explain only 2%" but rather as "structural factors, net of unmeasured content relevance, explain only 2%."

4. **Zero-inflated dependent variable.** The citation rate distribution has 60% zeros (median = 0.000). Standard OLS assumes approximately normal residuals, which is violated here. While we addressed this with a two-part hurdle model (Section 6.7) and the primary null finding was confirmed across all specifications, the OLS coefficient estimates and p-values should be interpreted with caution. A Zero-Inflated Poisson model was attempted but did not converge to stable estimates. Future studies should consider Tobit or hurdle specifications as the primary analysis.

5. **Google Rank partially available.** Google Rank was available at the domain-query pair level for 265/485 domains but was not properly aggregated to the domain level for the primary analysis. A sensitivity analysis on the subsample with rank data (Section 6.6) confirmed that including Google Rank does not change the null finding for AI Score ($p = 0.902$) and that DA remains significant independently of rank. Notably, Google Rank itself was significant ($p = 0.044$) in this specification, suggesting search ranking independently contributes to LLM citation probability. However, the subsample (n = 224) is not representative of the full dataset — it excludes domains that appeared only in LLM citations and not in search results.

6. **Wikipedia Presence variable failed.** The Wikipedia presence check used a naive brand extraction (first subdomain segment) that returned 0 for all 485 domains. This pre-registered control variable could not be used.

7. **Correlation ≠ Causation.** Even a significant result would not establish that improving one's AI Score causes higher citation rates. LLMs may cite based on content relevance, training data exposure, or retrieval index composition — factors orthogonal to structural readiness.

8. **Brand prior partially unobservable.** Domain Authority is an imperfect proxy for "brand strength" as perceived by LLMs. An LLM trained on text corpora will have implicit frequency-based priors about brand mentions that DA does not capture.

9. **Score formula limitations.** The AI Search Readiness Score is a heuristic composite. Domains with high scores may excel at characteristics that humans value but that LLMs do not query for during retrieval. The very low Trust & Entity scores across the sample (Mean = 7.0 / 25, 81.5% at minimum value of 5) suggest this sub-component has poor discrimination and may not be well-calibrated.

10. **Power for small effects.** The study was powered (80%) to detect $|r| \geq 0.133$. Effects between $|r| = 0$ and $|r| = 0.133$ fall below the detection threshold. However, effects this small would explain less than 1.8% of variance and would have negligible practical significance.

11. **Query set scope.** Thirty queries across three verticals provide breadth but not depth. Within each vertical, only 10 queries were used — a limited window into LLM citation behavior for any single domain.

12. **Scoring timeouts.** 44 / 485 domains (9.1%) could not be scored due to crawl timeouts and were excluded. If these domains systematically differ in structural quality, this introduces selection bias.

13. **Multiple comparisons.** Seven bivariate correlations were tested simultaneously. Without correction, the expected number of false positives at $\alpha = 0.05$ is 0.35. After Bonferroni

correction, only Moz DA survives (adjusted $p = 0.047$); the Trust & Entity finding (raw $p = 0.034$) does not (adjusted $p = 0.237$).

14. **Retrieval stochasticity and measurement noise.** Source set overlap between replicates (mean 3-way Jaccard = 0.512) reveals that ~35% of sources in any single run are variable, even at temperature=0. Crucially, this instability is query-dependent (core% range: 17–100%, $r = -0.630$ with information space size), meaning the measurement noise in Citation Rate is heteroscedastic across the query dimension. Three replicates provide limited averaging; 5–10 replicates would reduce noise and improve power. The Jaccard index itself is noisy on sets of size ~7 (±0.07–0.14 per source), so individual query-level J values should be interpreted cautiously. The observed low R^2 values may partly reflect measurement unreliability rather than true absence of effect.

---

11. Reproducibility & Transparency

11.1 Pre-Registration

The study was pre-registered (v1.0) before data collection. Hypotheses, score formula weights, query list, and statistical analysis plan were frozen in advance.

11.2 Open Data

- **Raw LLM responses:** 90 JSON files archived
- **Raw search results:** 30 JSON files archived
- **Study database:** SQLite with full provenance chain
- **Exported datasets:** `dataset_domain.csv` (485 rows) and `dataset_domain_query.csv` (14,550 rows)
- **Analysis code:** Python pipeline (collectors, normalizer, scorer, enricher, dataset builder)

11.3 Deviations from Pre-Registration

Two deviations from v1.0 were documented in v1.1:

Deviation D-1: Model Change. The pre-registered model `sonar-reasoning` was deprecated by Perplexity before data collection started (HTTP 400: "model has been deprecated"). The successor model `sonar-reasoning-pro` was used instead. This was a non-discretionary substitution — no alternative within the same model family was available.

Deviation D-2: Sample Size. The pre-registered estimate was 300–400 domains. The actual sample was 485 unique domains. All domains were collected automatically per protocol; no manual addition or removal occurred.

---

12. Practical Implications

12.1 For Site Owners

These results do not support the claim that optimizing structural website characteristics (as measured by AI Search Readiness Score) will increase LLM citation probability. Site owners should be cautious about investing in "AI SEO" or "GEO" services that promise improved AI visibility based solely on structural optimization.

However, this does not mean structural quality is irrelevant — it may contribute through pathways not captured by this study (e.g., improving traditional search ranking, which in turn feeds LLM retrieval indices). The study also cannot rule out that specific structural improvements (not captured by the composite score) may matter.

Domain Authority remains the single measurable factor most associated with citation probability in this sample, though even it explains very little variance.

12.2 For Researchers

This study highlights several areas for future work:

- **Longitudinal studies:** Does the Score–Citation relationship change over time as LLM retrieval pipelines evolve?
- **Multi-model replication:** Do ChatGPT, Gemini, and Claude show different structural preferences?
- **Content relevance controls:** Future studies should incorporate content relevance scoring (e.g., BM25 or semantic similarity to the query) as a control variable. Content relevance is likely the dominant unobserved factor.
- **Alternative dependent variables:** Citation count (rather than binary cited/uncited with majority vote) and citation position within responses may reveal effects masked by the binary operationalization.
- **Hurdle or zero-inflated models as primary specification:** The heavily zero-inflated citation distribution (60% zeros) violates OLS assumptions. Future studies should use two-part hurdle models or zero-inflated Poisson/Negative Binomial as the primary analysis, separating the retrieval-pool-entry mechanism from the citation-intensity mechanism.
- **Google Rank as pre-registered control:** Proper aggregation of query-level rank data to the domain level should be ensured before analysis. In this study, the subsample sensitivity analysis (Section 6.6) mitigated but did not fully resolve this gap.

12.3 For the AI Search Readiness Score

The null result does not invalidate the Score as a diagnostic tool for structural website quality. It does, however, indicate that the Score does not predict LLM citation behavior — at least not in its current form, for this LLM, at this point in time.

The Trust & Entity Signals sub-score showed a nominally negative bivariate correlation with citation rate, but this did not survive multiple comparison correction and was confounded by Domain Authority. Post-hoc investigation revealed that high-TE domains (≥ 15, $n = 82$) are predominantly niche service businesses (medical tourism, event planning, translation, specialty coffee) with lower average DA (34.1 vs. 43.3). The apparent negative TE–citation relationship is more parsimoniously

explained by the TE–DA confound than by any causal effect of trust signals on citation. Future iterations of the Score should consider whether the TE sub-component is measuring a meaningful independent construct or merely proxying for site type.

---

13. Conclusion

We tested whether AI Search Readiness Score predicts domain citation frequency in LLM-generated responses. Across 485 domains, 30 queries, and 90 LLM runs, we found no statistically significant association (Pearson $r = 0.009$, $p = 0.849$). The null hypothesis was not rejected. The study was adequately powered (80% power for $|r| \geq 0.133$), meaning any practically meaningful effect would likely have been detected.

This null finding was robust across multiple analytical specifications: OLS regression ($p = 0.557$), logistic regression ($p = 0.071$), hurdle model on cited-only domains (negative coefficient, $p = 0.097$), sensitivity analysis with Google Rank ($p = 0.902$), all three DA segments, and with/without high-authority "giants."

Domain Authority (Moz DA) was the only predictor to borderline survive Bonferroni correction for multiple comparisons (adjusted $p = 0.047$), though the effect was small (R^2 contribution $\approx 2\%$). Notably, DA predicted citation *intensity* (how often a domain is cited) but not citation *selection* (whether a domain is cited at all), suggesting it operates as an amplifier rather than a gate. The Trust & Entity Signals sub-score showed a nominally significant negative bivariate correlation, but this was confounded by Domain Authority and did not survive correction.

The overall model explained approximately 2% of variance ($R^2 = 0.022$), indicating that the vast majority of LLM citation behavior is driven by factors not captured in this study — likely including content relevance, retrieval index composition, model training data exposure, and other opaque internal mechanisms.

These findings are consistent with the possibility that LLM citation decisions are primarily content-driven and retrieval-pipeline-dependent, rather than structurally determined. Structural readiness may be a necessary but insufficient condition, or may operate through indirect pathways (e.g., improving traditional search ranking).

Further research with multiple LLM models, longitudinal designs, content relevance controls, and hurdle/zero-inflated model specifications is needed to develop a more complete understanding of what drives AI citation behavior.

Appendix A: Score Formula

See `Appendix A – Score Formula.md` for the full frozen formula specification.

Appendix B: Query List

See `Appendix B – Query List.md` for the complete set of 30 queries across 3 verticals.

Appendix C: Regression Diagnostics

C.1 Full OLS Output

```
Model: OLS
Dependent Variable: Citation_Rate
N = 441
R² = 0.022, Adjusted R² = 0.016
F-statistic = 3.33 (p = 0.020)

               Coef       p-value    95% CI
Intercept     +0.007358    0.114    [-0.001786, +0.016503]
AI_Score      +0.000046    0.557    [-0.000107, +0.000199]
Moz_DA        +0.000189    0.002    [+0.000067, +0.000310]
Domain_Age    -0.000218    0.130    [-0.000500, +0.000064]
```

C.2 Full Logistic Output

```
Model: Logistic Regression
Dependent Variable: is_cited (binary)
N = 441
Pseudo R² = 0.008

               Coef       p-value    Odds Ratio
Intercept     -1.283891    0.004      0.277
AI_Score      +0.013405    0.071      1.013
Moz_DA        +0.008574    0.144      1.009
Domain_Age    -0.009460    0.487      0.991
```

C.3 Sub-Score Regression Output

```
Model: OLS (sub-score decomposition)
Dependent Variable: Citation_Rate
N = 441
R² = 0.035

                     Coef       p-value
Intercept          +0.012279    0.034
Machine_Readability -0.000170   0.516
Extractability     +0.000133    0.448
Trust_Entity       -0.000385    0.105
Offering_Readiness +0.000229    0.096
Moz_DA             +0.000177    0.005
Domain_Age         -0.000220    0.125
```

C.4 OLS with Google Rank (Sensitivity)

```
Model: OLS (subsample with Google Rank data)
```

```
Dependent Variable: Citation_Rate
N = 224
R² = 0.068, Adjusted R² = 0.051
F-statistic = 4.00 (p = 0.004)

                Coef       p-value
Intercept     +0.007811    0.321
AI_Score      -0.000015    0.902
Moz_DA        +0.000337    0.001
Domain_Age    -0.000464    0.048
Google_Rank   -0.001177    0.044

DA-GoogleRank correlation: r = 0.010 (p = 0.885)
```

C.5 Hurdle Model Part 2 (Cited-Only OLS)

```
Model: OLS (cited domains only)
Dependent Variable: Citation_Rate
N = 179
R² = 0.088, Adjusted R² = 0.072

                Coef       p-value
Intercept     +0.036062    <0.001
AI_Score      -0.000194    0.097
Moz_DA        +0.000312    0.001
Domain_Age    -0.000274    0.183

Cited vs Uncited comparison:
  Mean Score: 51.7 vs 49.6 (t-test p = 0.103)
  Mean DA:    43.1 vs 40.6 (t-test p = 0.248)
```

C.6 Power Analysis

```
N = 441, alpha = 0.05 (two-sided), power = 0.80
Minimum detectable |r| = 0.133
Observed r = 0.009
Post-hoc power at r = 0.009: 0.054 (essentially zero)
Conclusion: Study adequately powered; effects |r| > 0.133 can be ruled out
```

C.7 Bonferroni Correction (7 bivariate tests)

```
Threshold: 0.05 / 7 = 0.0071

Variable              Raw p    Adjusted p   Significant?
Moz_DA                0.007    0.047        Yes (borderline)
Trust_Entity          0.034    0.237        No
Offering_Readiness    0.084    0.588        No
Domain_Age            0.593    1.000        No
```

Extractability	0.790	1.000	No
Total_Score	0.849	1.000	No
Machine_Readability	0.870	1.000	No

C.8 Trust & Entity Confounding Analysis

TE value distribution: 5 (n=359), 15 (n=74), 25 (n=8)

TE >= 15 (n=82): mean_cr=0.010, cited=29.3%, mean_da=34.1
TE = 5 (n=359): mean_cr=0.016, cited=43.2%, mean_da=43.3

TE–DA correlation: r = −0.144 (p = 0.002)
After controlling for DA (sub-score regression): TE p = 0.105 (not significant)

C.9 Replicate Source Set Overlap (Jaccard per query)

Query	Sources A/B/C	A∩B∩C	Union	Jaccard
q1	7/8/8	7	9	0.778
q2	6/6/6	3	9	0.333
q3	8/7/6	3	11	0.273
q4	7/7/7	6	8	0.750
q5	7/7/6	3	11	0.273
q6	7/7/8	5	10	0.500
q7	8/7/8	3	14	0.214
q8	8/7/8	6	9	0.667
q9	7/7/7	7	7	1.000
q10	7/6/7	2	12	0.167
q11	7/7/8	6	9	0.667
q12	8/8/9	6	10	0.600
q13	3/4/3	3	4	0.750
q14	7/7/7	7	7	1.000
q15	7/7/7	6	8	0.750
q16	9/9/10	7	13	0.538
q17	8/7/7	3	12	0.250
q18	6/6/5	3	8	0.375
q19	8/8/8	5	11	0.455
q20	6/4/7	2	8	0.250
q21	7/8/7	6	9	0.667
q22	8/7/8	7	8	0.875
q23	7/8/8	5	10	0.500
q24	8/9/6	3	12	0.250
q25	9/9/8	8	9	0.889
q26	6/6/6	3	9	0.333
q27	10/10/10	5	16	0.312
q28	6/6/7	3	12	0.250
q29	9/9/9	6	12	0.500
q30	6/7/5	2	10	0.200

```
Summary:
  Mean sources/run: 7.2 (SD=1.4)
  Mean 3-way Jaccard: 0.512
  Median 3-way Jaccard: 0.500
  Mean pairwise Jaccard: 0.644
  Identical sets (J=1.0): 2/30
  High overlap (J≥0.8): 4/30
  Low overlap (J<0.5): 14/30

Per-run composition (mean):
  Core (3/3 replicates): 4.7 domains (~65% of each run)
  Partial (2/3): 2.3 domains (~1.5 per run)
  Singleton (1/3): 2.9 domains (~1.0 per run)
  Total unique per query (union): 9.9

Information Entropy Hypothesis:
  Union Size vs Core%: r = -0.630, p = 0.0002 (Spearman ρ = -0.672, p < 0.0001)
  Union Size vs Core Count: r = -0.194, p = 0.305 (n.s.)
  Low entropy (union ≤ 8, n=8): mean core% = 71.9%
  High entropy (union ≥ 12, n=8): mean core% = 31.0%
  By cluster: SaaS 49.5%, E-commerce 56.3%, Services 47.8%

Core vs Singleton DA Comparison:
  Core (3/3): n=132, mean DA=43.9, median=42.5
  Singleton (1/3): n=85, mean DA=44.6, median=41.0
  T-test: t=-0.215, p=0.830
  Mann-Whitney U: p=0.821

Domain Citation Frequency Distribution:
  275 unique domains, 658 total citations
  Top 13.5% of domains → 25% of citations
  51.6% of domains cited ≤ 2 times
  Power law fit (log-log): slope=-0.448, R²=0.659
  Most-cited: youtube.com (33), g2.com (10), clutch.co (8)

URL-Level vs Domain-Level Overlap:
  Domain Jaccard: 0.512 (median 0.500)
  URL Jaccard: 0.457 (median 0.455)
  Delta: 0.055
  Canonical divergence cases: 13/~110 (12%)
    - Semantically adjacent pages: 9 cases
    - YouTube (different videos): 4 cases
```

C.10 Segment Correlations

```
DA < 30  (n=143): r = +0.075, p = 0.371
DA 30-50 (n=159): r = +0.055, p = 0.491
DA > 50  (n=139): r = -0.024, p = 0.782
```

C.11 Data Collection Metadata

```
Collection date: 2026-03-12
LLM model: Perplexity sonar-reasoning-pro
Temperature: 0
Replicates: 3 per query
LLM runs: 90 (90 successful)
LLM citations: 658
Search results: 573 (via SearXNG)
Unique domains: 485
Scored domains: 441
Enriched domains: 485 (DA), 456 (Age, of which 415 have valid scores)
```