Academic Citation Infrastructure: Infrastructure-Level Interventions for Generative Engine Optimization

Author: Nuno Andrade
Affiliation: ILLIXIS (illixis.io)
Date: May 2026
Version: 1.0 (pre-experimental)
Type: Methods paper

Abstract

Generative Engine Optimization (GEO) research has focused exclusively on content-level interventions: adding statistics, citing sources, using authoritative tone, and structuring text for extraction. No peer-reviewed GEO work has proposed infrastructure-level interventions for increasing AI system citation rates. Scattered practitioner recommendations for individual tactics (DOI deposits, ScholarlyArticle schema, repository placement) have emerged independently but remain unconnected, unanalyzed, and untested. This paper proposes the first formal framework for what we term Academic Citation Infrastructure (ACI): the deliberate attachment of scholarly publishing infrastructure to commercial content. ACI unifies four techniques into a coherent strategy: methods-paper formatting, BibTeX/RIS citation file generation, Zenodo DOI mirroring, and ScholarlyArticle schema markup. A review of peer-reviewed GEO research confirms none propose infrastructure-level interventions; a survey of practitioner literature identifies emerging but fragmented adoption. Analysis of 15+ commercial deposits on academic platforms documents the pattern in practice. Quantitative evidence from source framing studies (192,000 assessments), repository discoverability research (+266% search impressions), and LLM search selection analysis (55,936 queries) supports the underlying mechanisms. This paper unifies these scattered tactics into a testable framework, presents the first mechanism analysis explaining why they work, proposes a controlled experimental design for validation, and, as a recursive demonstration, implements all four techniques on itself.

Keywords: Generative Engine Optimization, GEO, AI citations, academic repositories, DOI, BibTeX, ScholarlyArticle, citation infrastructure, LLM search

1. Introduction

The discipline of search optimization is undergoing its most significant structural shift since PageRank. Google has positioned generative AI at the center of Search: AI Overviews are now a standard surface on informational queries, and the redesigned Search interface presented at I/O 2026 is built around longer, more conversational input. ChatGPT reported 800 million weekly active users in October 2025 and 900 million in February 2026 [43]. Google's AI Mode reached 1 billion monthly active users globally within a year of launch, with queries reportedly three times longer than traditional searches and follow-up queries up 40% month-over-month in the US, per Google's I/O 2026 disclosures (May 19, 2026) [49]. Perplexity, Claude, and Gemini are becoming default research tools for millions of users. These systems do not return ten blue links. They synthesize answers from multiple sources and cite where the information originated.

The question for content creators is no longer "Do I rank?" It is "Am I cited?"

A new discipline has emerged to address this question. Generative Engine Optimization, formalized by Aggarwal et al. in their foundational 2024 paper [1], proposes systematic methods for increasing content visibility in AI-generated responses. The field has grown rapidly: researchers have analyzed AI search citation patterns [2], measured platform-specific behaviors [3], and proposed content optimization strategies [4].

All of this work operates at the content-text level. The recommendations are consistent: add statistics, cite authoritative sources, use structured formatting, write in an authoritative tone, include quotation marks around key claims. These are interventions applied to what the content says and how the text is structured.

No peer-reviewed GEO research has proposed interventions at the infrastructure level: where content is hosted, how it is packaged for citation, what metadata accompanies it, and which discovery pathways it enters. Individual practitioners have

begun recommending specific tactics (DOI deposits, ScholarlyArticle schema, repository placement) in blog posts and LinkedIn articles, but these remain scattered, disconnected from each other, and unsupported by mechanism analysis or experimental design.

This gap is surprising. The academic publishing ecosystem solved the discoverability problem decades ago through infrastructure: persistent identifiers (DOIs), standardized citation formats (BibTeX, RIS), metadata registries (DataCite, Crossref), aggregator harvesting (OpenAIRE, OAI-PMH), and structured markup (ScholarlyArticle schema). These systems exist specifically to make content findable, citable, and machine-readable. AI systems were trained on content that flows through these pipelines. The infrastructure that makes academic content citable by humans also makes it citable by machines.

The contribution of this paper is fivefold (across pattern documentation, practitioner survey, mechanism analysis, framework unification of four techniques, and experimental design):

1. **Pattern documentation.** We catalog 15+ examples of commercial entities already depositing content on academic repositories, most without recognizing the GEO implications.
2. **Practitioner survey.** We document the emerging but fragmented adoption of individual ACI techniques in practitioner GEO literature, showing convergent independent discovery of the same tactics.
3. **Mechanism analysis.** We synthesize quantitative evidence from source framing, repository discoverability, and LLM search selection research to explain why academic infrastructure creates AI discoverability advantages. No prior work has connected these mechanisms to the observed tactics.
4. **Framework unification.** We define Academic Citation Infrastructure (ACI) as a formal framework that unifies four techniques into a coherent strategy, each targeting a distinct mechanism.
5. **Experimental design.** We propose a controlled multi-platform experiment for causal validation and, as a recursive demonstration, implement all four ACI techniques on this paper itself.

---

2. Related Work

2.1 Foundational GEO Research

Aggarwal et al. introduced the GEO framework at KDD 2024, benchmarking content optimization strategies across 10,000 queries [1]. The paper finds that adding statistics, quotations, and authoritative citations — what we describe as citation-like scaffolding — can increase generative engine visibility by up to approximately 40% for the best-performing tactic [1]. The work is rigorous and foundational, but operates entirely at the content-text level. No infrastructure, hosting, or distribution interventions are tested.

Chen et al. analyzed AI search citation patterns at scale and found that earned media (not brand-owned content) dominates AI citations [2]. Their work identifies what gets cited but does not propose mechanisms for why certain hosting environments might be preferred.

2.2 The Closest Existing Recommendation

Rayhan's comprehensive GEO survey is the closest any published GEO work comes to the infrastructure-level hypothesis, observing that academically-styled content may perform better in retrieval-augmented systems such as Perplexity [4]. But the observation remains at the content-style layer, and the leap from "write like an academic" to "build academic citation infrastructure" is substantial. Writing style is a content-level intervention. Attaching DOIs, providing BibTeX files, depositing on Zenodo, and using ScholarlyArticle schema are infrastructure-level interventions that change the content's discoverability pathways, not its prose.

2.3 Schema Markup and GEO

A controlled experiment by Search Engine Land compared nearly identical pages with strong, poor, and no schema; only the strong-schema page appeared in an AI Overview [5]. The study did not test ScholarlyArticle specifically, and did not propose the full infrastructure stack (DOI + citation files + repository hosting + schema). The schema subtype question (whether ScholarlyArticle signals different trust than Article or TechArticle) remains an open measurement gap.

2.4 Source Framing and Authority Bias

Outside GEO, a growing body of research demonstrates that AI systems evaluate identical content differently based on source framing. Germani and Spitale conducted 192,000 assessments across 4,800 statements and 24 topics, demonstrating systematic evaluation shifts based on disclosed source labels [6]. An ACL 2025 study on authority bias in retrieval-augmented generation found that models can be up to threefold more likely to be misled by authority cues under conflict conditions [7, Section 4.3, Figure 3]. A separate study on label-induced bias reported preference-ranking shifts of up to approximately 50 percentage points when content was attributed to different AI-model identity labels (e.g., Claude vs. Gemini vs. ChatGPT) regardless of actual authorship [8].

These studies do not specifically test "academic repository versus blog." But they establish the core causal property: source framing alone changes model evaluations of identical text.

2.5 Repository Discoverability

The repository science literature provides quantitative evidence that hosting infrastructure matters for discoverability. Arlitsch and O'Brien found that institutional repository indexing in Google Scholar can range from less than 0.1% to over 90%, with metadata exposure as the primary differentiator [9]. Macgregor's longitudinal study of the Strathprints repository documented +266% search impressions, +104% clicks, +62% COUNTER usage, and +1,920% growth specifically in Google Scholar referral traffic over the four-year Y1-Y4 reporting period [10, Table 2] (the paper's abstract reports overall growth exceeding 1,300% across metrics; the 1,920% is the Google Scholar referrals submetric). Statistical significance: $t=14.30$, $df=11$, $p<0.0005$. The USRN Discovery Pilot made approximately 728,770 new research outputs discoverable (a 50% increase) through aggregator and indexing interventions [11].

These gains operate in traditional search. But AI retrieval systems inherit many of the same discovery pipelines. Content that is more findable in Google Scholar and DataCite registries is more likely to appear in the retrieval indices that generative AI systems query.

2.6 Emerging Practitioner Adoption

While academic GEO research has remained focused on content-level optimization, individual practitioners have begun recommending infrastructure-level tactics independently. A survey of practitioner literature (LinkedIn, Substack, Medium, GitHub, agency blogs) conducted between March and May 2026 identified several explicit recommendations, of which a representative subset is cited inline below. The practitioner-source layer is partially diffuse — some sources resolve to canonical URLs that can be cited directly; others describe a recurring pattern across many lightly-bylined or no-longer-live sources. A supplementary practitioner-source survey notes file is deposited alongside this paper (filename: `aci-methods-paper-practitioner-survey-notes.md`) documenting the broader convergent-observer cluster, rebrand history of practitioner sources, and corroborating cluster context that the numbered citations [25]-[32] reference but do not load-bear on directly.

DOI deposits and repository placement appear most frequently. A LinkedIn playbook by Jasper Morris titled "How to Get Cited by AI: A New Playbook" explicitly recommends "Persistent identity (DOIs, canonical URLs)," uploading datasets to Zenodo or Figshare with methods pages, and "How to cite + DOI" footers as AI citation tactics; an adjacent Medium framework by Tim de Rosen also organizes LLM citation tactics under a "Discoverability" pillar (see survey notes §A.1) [25]. Multiple practitioner guides frame Zenodo, GitHub, and Hugging Face as platforms that materially feed AI training corpora, with strategic placement recommended as deliberate AI search optimization. Paul Sheals frames the test diagnostically — "*do you appear in the foundational sources AI models are trained on (Wikidata, Zenodo/DOIs, Crunchbase, GitHub, Hugging Face)?*" — and Ansh Bhatia frames the underlying mechanism descriptively as "*academic/indexed content — material on GitHub, Zenodo, Hugging Face, or arXiv that appears in AI training corpora*" [26]. A broader convergent-observer cluster (Sheals's CitationMapper analysis, Tim de Rosen's AIVO post, the Qlavo entity-hygiene post) names the same platform group under adjacent framings of discovery, measurement, and legitimacy (see survey notes §B.1). Kurt Fischman's "Citation Seeding Playbook" recommends that founders "*wedge your startup brand into the AI knowledge graph*" by placing content on "*every retriever-friendly surface (crunchbase, Zenodo, Product Hunt etc)*" — naming Zenodo specifically as a target placement for LLM citation [27].

ScholarlyArticle schema for GEO purposes is recommended in multiple practitioner guides describing a "schema-typing" pattern: content typed as `Article`, `NewsArticle`, `ScholarlyArticle`, `ResearchProject`, or `Report` (variously combined) for LLM contextualization and machine-evaluable trust. Tammy Graham's "Trust Optimization Protocol (TOP)" framework recommends the four-type quartet (`Article`, `NewsArticle`, `ScholarlyArticle`, `Report`) so that "LLMs [can] contextualize

it within a content taxonomy" [28]; Kevin C. Roy's *Schema Mountain* series extends the typing recommendation with `ResearchProject` alongside `ScholarlyArticle` and `Report` as "Industry Authority Schemas" for AEO/GEO, positioned as "precision layers for regulated industries where accuracy matters" [29]. A vendor blog by Geneo specifically recommends combining Dataset schema with ScholarlyArticle via `isBasedOn` or `citation` properties as a schema-markup best practice, providing the most operationally specific instance of the pattern [30].

The most complete practitioner implementation is the AIVO Standard, a GitHub-hosted framework that includes a canonical citation guide intended to ensure citation "across AI models, LLMs, and academic references," a downloadable BibTeX file (`references.bib`) with embedded Zenodo DOIs, and a Zenodo deposit for the framework itself [31, 32]. This implementation combines DOI infrastructure with BibTeX packaging in an explicitly AI-citation-oriented workflow.

These practitioner recommendations share three characteristics: they are scattered across unconnected sources, they propose individual techniques rather than a unified framework, and they provide no mechanism analysis explaining why the tactics work. The recurrence of the same recommendations across many lightly-bylined or no-longer-live sources is itself the form that convergent independent discovery takes in a practitioner ecosystem that does not yet apply academic citation infrastructure to its own publications. No practitioner source connects the techniques to the quantitative evidence from source framing research, repository discoverability studies, or LLM search selection analysis. The contribution of this paper is to unify these emerging tactics, provide the missing mechanism analysis, and propose the first experimental design for validation.

2.7 Query Fan-Out and Multi-Query Retrieval

A critical development in AI search architecture provides the retrieval mechanism that makes ACI's "multiple surfaces" argument mechanistically precise. Modern generative search systems do not execute a user's prompt as a single retrieval query. They decompose it into multiple sub-queries and search concurrently across multiple data sources.

Google explicitly documents this behavior as "query fan-out" (QFO): dividing a question into subtopics and searching for each simultaneously [33a, 33b]. A patent line of work further describes generating multiple "synthetic queries" derived from an LLM's output and using them to select search-result documents in a generative search workflow [34, 35]. The academic literature uses adjacent terms: query decomposition, sub-question generation, iterative query generation, and parallel retrieval [36, 37, 38].

Two empirical findings make QFO directly relevant to ACI. First, AirOps reports that 32.9% of cited pages in ChatGPT responses appeared only in search results for fan-out sub-queries, not the original prompt [39]. Content discovered exclusively through sub-queries is a large share of total citations. Second, 95% of ChatGPT's fan-out queries had zero monthly search volume by traditional keyword metrics [39], indicating that the "citation opportunity space" is larger than the tracked keyword universe.

A small practitioner experiment by Semrush found that explicitly optimizing content to cover fan-out sub-queries increased observed citations from 2 to 5 (a 150% increase) across four articles tracked for one month [40]. The sample is small but directionally consistent with the QFO mechanism.

The foundational GEO paper by Aggarwal et al. [1] also models this behavior: their formulation of generative engines includes a query reformulation component that generates a set of simpler queries passed to a search engine for retrieval. This is an academically described analog of fan-out.

For ACI, QFO provides the missing link between "hosting content on multiple platforms" and "increased citation probability." The argument is not simply that more URLs exist. It is that each repository surface (Zenodo landing page, DataCite metadata record, Google Scholar entry, OpenAIRE index) can appear in a different fan-out sub-query than the canonical blog post, giving the same content multiple independent retrieval chances across the union of all sub-query results.

---

3. Pattern Observation: Commercial Content on Academic Repositories

3.1 Catalog of Examples

The pattern already exists in practice. Commercial entities are depositing non-academic content on scholarly repositories, most without any stated GEO intent. The following catalog documents 15+ examples across Zenodo, SSRN, Figshare, and other platforms.

Table 1. Commercial content on Zenodo

Entity	Content Type	DOI	Visibility
QuantAQ (sensor company)	Product application note	10.5281/zenodo.10688216	2,176 views, 2,246 downloads
Pensive Beauty (skincare brand)	Product technical note	10.5281/zenodo.18616576	4,257 views, 53 downloads
Cyberhare Solutions (AI detection)	Defensive disclosure white paper	10.5281/zenodo.15361247	111 views, 138 downloads
QInsight Labs (quantum software)	Product white paper	10.5281/zenodo.17312665	169 views, 93 downloads
EU multi-company consortium	Industry whitepaper	10.5281/zenodo.13820261	300 views, 229 downloads
C-Therm Technologies	Product application note	10.5281/zenodo.17228569	24 views, 19 downloads
CollectiveOS / Immortal Tek Inc	Architecture/model documentation	10.5281/zenodo.17460464	120 views, 26 downloads

Table 2. Commercial content on SSRN

Entity	Content Type	DOI	Visibility
Berkeley Research Group	White paper series (pricing model)	10.2139/ssrn.2665191	918 abstract views, 147 downloads
Berkeley Research Group	White paper series (Monte Carlo)	10.2139/ssrn.2665147	2,225 abstract views, 690 downloads
WorkWise Solutions	AI governance white paper	10.2139/ssrn.6274559	47 abstract views, 9 downloads (also on ResearchGate)
Call For Culture	Sustainability method paper	10.2139/ssrn.5013629	600 abstract views, 115 downloads
ThePricer Media	Operational standard document	10.2139/ssrn.5578311	231 abstract views, 40 downloads (also on Research Square with second DOI)

Table 3. Commercial content on other platforms

Entity	Platform	Content Type	Identifier
Digital Science	Figshare	Open Data report	10.6084/m9.figshare.31113526
McKinsey Global Institute	Academia.edu	Research papers	Institutional profile
Google, Meta, Cloudflare	arXiv	Corporate research papers	Industry-funded AI papers achieve high citation rates at ~3x the rate of non-industry papers (12% vs 4%, 2018–2022) [44]

Visibility metrics in Tables 1, 2, and 3 are dynamic and were re-harvested 2026-05-25 UTC: Zenodo metrics from each record's public statistics page, SSRN metrics from each paper's "Paper statistics" panel (Abstract Views + Downloads), and Figshare/Academia.edu/arXiv entries from each platform's public landing pages. Metrics will drift after this snapshot — the Pensive Beauty record (Table 1) alone grew from 2,435 views / 12 downloads at original research (March 2026) to 4,257 / 53 by re-harvest (~10 weeks later), illustrating the magnitude of drift readers should expect from any single cell.

3.2 Classification of Intent

These deposits cluster into five categories:

1. **Defensive publication.** Establishing prior art without filing patents (CollectiveOS, Cyberhare Solutions).

2. **Product documentation.** Making technical specifications findable and citable (QuantAQ, C-Therm Technologies).
3. **Thought leadership distribution.** Publishing methodology and frameworks through academic channels (Berkeley Research Group, Call For Culture).
4. **Industry standard publication.** Formal documents requiring stable citation (EU consortium, AWS).
5. **Deliberate GEO strategy.** Consciously using academic infrastructure for AI discoverability (one confirmed case: Pensive Beauty).

3.3 The Pensive Beauty Case

One entity stands out. Pensive Beauty, a nanocosmetics brand, deposited a product technical note on Zenodo (DOI: 10.5281/zenodo.18616576) using TechArticle schema on their website. The Zenodo deposit accumulated 2,435 views and 12 downloads at the time of original research and has grown to 4,257 views, 53 downloads as of 2026-05-25 UTC — substantial growth despite no associated public marketing of the deposit. Crucially, the blog post explaining the methodology (at `pensivebeauty.com/blog/generative-engine-optimization-nanocosmetics-ai-models-formulation-science`, the URL surfaced by Google Search) returned HTTP 404 when re-verified on 2026-05-25 UTC. A Wayback Machine availability check on the same date returned zero archived snapshots for the URL, and no Google cache version was present. The blog post does not appear in Pensive Beauty's current public blog index.

The explanatory content was removed; the Zenodo deposit remains live and continues to accumulate visibility. This pattern (publishing the infrastructure, withdrawing the documentation of the strategy) suggests deliberate concealment of a technique the publisher considers valuable.

3.4 The Multi-Simulation Thesis: A Natural Experiment

The pattern that prompted this research was not commercial at all. Augusto Bartolomeu, a non-academic author, published the "Multi-Simulation Thesis" (MST) across multiple academic-adjacent platforms: PhilArchive, PhilPeople, and Academia.edu. The work is a series of eight papers on the philosophy of physics. A niche topic with no marketing intent, no SEO strategy, and no promotion budget. The content has accumulated approximately 1,694 cumulative views/downloads aggregated across PhilArchive, PhilPeople, and Academia.edu, per the author's research records at the March 2026 research window. PhilArchive view-counts for individual MST papers were independently re-checked at the May 2026 follow-up (e.g., MST v1 521 views, MST v3 367 views, MST v5 264 views, MST v4 223 views — visible on each paper's PhilArchive landing page); these confirm the work's measured-but-modest reach and are not summed into the cumulative total above, which spans both views and downloads across three platforms with different telemetry conventions.

When the four major AI chatbots (ChatGPT, Claude, Gemini, DeepSeek) were tested for MST discoverability in March 2026 and re-tested in May 2026 with dated prompts and archived transcripts (full methods in Appendix A), the results exhibited a structured bifurcation that the ACI mechanism analysis in Section 4 predicts.

Stage 1 — named recall (queried by author and title). Three of four chatbots (Claude, DeepSeek, ChatGPT) accurately retrieved MST in both windows; DeepSeek's coverage improved from v3 only in March to v2/v3/v4/v5 plus a SEAL companion paper in May, and ChatGPT returned 15 distinct source citations across PhilPapers, PhilArchive, Academia.edu, and PhilPeople. Gemini drifted: in March it acknowledged MST but conflated the core framework with the related Residual-Coherence Field (RCF) spin-off; in May, against an unchanged corpus of live PhilArchive/PhilPapers entries, it explicitly denied that "Multi-Simulation Thesis by Augusto Bartolomeu" existed at all. This single-subject longitudinal observation instantiates the population-scale citation volatility documented by Profound (40-60% of cited domains change monthly for identical queries) [23].

Stage 2 — fan-out retrieval (queried by adjacent topic without the author or title). Only DeepSeek surfaced MST when the query was reframed as "non-mainstream philosophical works that propose multiple nested or layered simulation hypotheses." DeepSeek's visible reasoning trail showed textbook query fan-out: five distinct sub-queries, 100+ candidate results across the union, MST surfacing at position 2 in the initial 53-result set. Claude (8 nested-simulation works cited), Gemini (3), and ChatGPT (8) did not include MST in their candidate sets, and their candidate sets weighted different platforms (Claude: philosophy canon; Gemini: narrow; ChatGPT: arXiv and named-author recognition; DeepSeek: PhilArchive-heavy, 7+ of 9 cited sources). This chatbot-specific platform weighting in fan-out retrieval is direct primary-source evidence for the multi-surface deposit argument in Sections 4.1 and 4.5 — single-platform deposits produce strong Stage 1 but variable Stage 2 across chatbots.

The Stage 1 / Stage 2 bifurcation maps onto the mechanism analysis in Section 4: unique terminology (the "Multi-Simulation Thesis" name) drives Stage 1 retrieval, but accumulated machine-relations and multi-surface ranking are required for Stage 2 fan-out surfacing — and MST, with no indexed secondary citations as of the test date, has not yet built that density.

When each chatbot was separately asked how it would have encountered such content, their explanations converged on a shared mechanism framework — public-web crawl ingestion (Common Crawl-style), live search retrieval distinct from parametric recall, and reinforcement through secondary citations and metadata propagation across academic indexes. DeepSeek articulated this as a three-pathway model; ChatGPT extended it with a fourth (user-provided content); Gemini named it in a diagnostic register (open-access vs. login-walled hosting, citation density, PDF parsability); Claude distinguished search-retrieval from training-data recall. The four-chatbot convergence on a mechanism framework that maps directly to Section 4 is stronger primary-source evidence than the uniform-retrieval finding originally claimed: AI systems independently articulate the architecture the ACI framework is designed to optimize for.

These retrieval mechanisms map onto the four ACI techniques: methods-paper formatting (5.1) targets the academic-formatting and structured-HTML signals named in every chatbot's account of how niche academic content reaches it; ScholarlyArticle schema (5.4) targets the structured-metadata pathway that DeepSeek and ChatGPT named explicitly and that Gemini and Claude implied via adjacent framings (PDF parsability, search-vs-training-recall distinction); Zenodo DOI mirroring (5.3) targets the multi-platform-distribution and registry-harvesting pathways that produce non-webpage discovery surfaces (4.5); and BibTeX/RIS citation files (5.2) target the secondary-citation reinforcement pathway that DeepSeek named as a third pathway and that the QFO mechanism (4.1) depends on once unique terminology has driven initial Stage 1 retrieval. The bifurcation in the MST case — strong Stage 1 across three chatbots, Stage 2 in only one — illustrates which mechanisms have activated for an organically-published academic-adjacent work and which require additional cross-citation density that the explicit ACI techniques are designed to accelerate.

In March 2026, when DeepSeek was asked to extract GEO best practices from the MST case, it named all four ACI techniques by specific operationalization (Zenodo, DOI, BibTeX, schema.org). In May 2026 the same request surfaced the same underlying principles (multi-platform academic hosting, persistent author identifiers, structured metadata, schema markup, version control, interconnected citation networks, unique terminology) but with shifted operationalizations (arXiv/OSF/ORCID/JSON-LD instead of Zenodo/DOI/BibTeX/ScholarlyArticle), and a fresh-chat query without the MST anchor surfaced mainstream content-level GEO advice with no ACI techniques at all. This two-layer drift — context-dependent activation plus operationalization shift over time — is itself supporting evidence for the paper's motivating gap: AI-derived GEO recommendations converge on the same underlying mechanisms but their specific prescriptive tactics are unstable across time and prompt context, which is the gap an explicit codified framework like ACI is intended to close.

---

4. Mechanism Analysis: Why Academic Infrastructure Creates AI Discoverability

Five quantitative mechanisms connect academic citation infrastructure to higher AI citation probability: two retrieval mechanisms (query fan-out, 4.1; non-webpage discovery surfaces, 4.5) determine which content enters and ranks within the candidate set, and three evaluation mechanisms (structured HTML advantage, 4.2; source framing bias, 4.3; sparse citation selection, 4.4) operate on how content is selected once retrieved. Section 4.6 synthesizes how these mechanisms sequence.

4.1 Query Fan-Out Creates Multiple Retrieval Chances

Modern generative search systems decompose a user's prompt into multiple sub-queries and search concurrently across multiple data sources [33a, 33b, 34]. This query fan-out (QFO) behavior means that citations are drawn from the union of documents retrieved across all sub-queries, not just the original prompt. Empirically, 32.9% of cited pages in ChatGPT responses were discovered only through fan-out sub-queries [39], and 95% of fan-out queries had zero monthly search volume by traditional metrics [39]. Google's May 2026 I/O disclosures place this behavior at scale on the largest AI search surface: AI Mode queries are reportedly three times longer than traditional searches and follow-up queries grew 40% month-over-month in the US [49]. Longer, more conversational queries decompose into more sub-queries, multiplying the surface area where ACI-distributed content can appear.

For ACI, QFO transforms the "multiple indexing pathways" argument from a vague claim about crawlability into a precise retrieval mechanism. A blog post exists at one URL and competes in one set of sub-query results. A Zenodo deposit creates a DOI-backed landing page indexed in OpenAIRE, with metadata registered to DataCite (openly harvestable under CC0 [12])

and queryable via the DataCite REST API, and visible in Google Scholar. Each surface can appear in a different fan-out sub-query than the blog post, giving the same content multiple independent retrieval chances.

This is consistent with information retrieval research showing that expanding a document with multiple query-like representations improves retrieval effectiveness (Doc2Query [41]). ACI operates the same principle at the web layer: each repository landing page, with its distinct metadata fields, categories, and domain-level trust signals, functions as a different "representation" of the same underlying content.

A longitudinal study of persistent identifier usage across Common Crawl analyzed over 10^12 URIs from over 5×10^9 crawled pages, confirming that DOI patterns are trackable at web scale [14]. Common Crawl-derived pipelines are widely used in language model training data: FineWeb alone processed 96 Common Crawl snapshots totaling 15 trillion tokens [15]. Content that is DOI-resolvable and registry-indexed has more surface area for capture in these pipelines.

Deduplication caveat. Web search engines attempt to detect near-duplicates and reduce redundancy [42a, 42b]. This can reduce the naive assumption that N URLs means N independent retrieval chances. ACI does not rely on pure duplication. It relies on distribution across heterogeneous surfaces whose pages differ in metadata, markup, internal linking, domain-level trust, and retrieval affordances (PDF availability, structured citations), which can defeat strict near-duplicate clustering and remain distinct candidates under different sub-queries.

4.2 Structured HTML Advantage

AI search engines do not cite randomly. A large-scale measurement study of LLM-based search engines analyzed 55,936 queries, yielding 124,287 unique domains and 1,418,733 unique citation hyperlinks [16]. The study found:

- LLM search cites fewer sources per response (mean 4.3 URLs) versus traditional search (mean 10.3 URLs). Fewer citation slots means marginal advantages in selection are more consequential.
- Domains favored by LLM search have more structured, hierarchical HTML, more readable text, and more outlinks.
- 37% of domains surfaced by LLM search are absent from traditional search results, indicating a distinct selection pipeline.

Academic repository pages provide structured, hierarchical HTML automatically through their templates. This is not an optimization the depositor performs. It is a structural property of the hosting environment.

4.3 Source Framing Bias

Identical content is evaluated differently depending on where it appears. Germani and Spitale's study across 192,000 assessments and 24 topics demonstrated systematic shifts in model evaluations based on disclosed source labels [6]. The ACL 2025 authority bias study found models can be up to three times more likely to follow authority cues when resolving conflicting information [7]. Label-induced bias research reports preference-ranking shifts of up to approximately 50 percentage points from attributed-identity labels (Claude, Gemini, ChatGPT) alone, independent of content [8].

Note: Germani and Spitale's experimental contrasts (human-source vs. LLM-source attribution) do not directly test "academic repository vs. blog post," but establish the core property — source framing alone changes model evaluations. The application to repository-hosted content is an extrapolation; see Section 8 for limitations.

Academic repositories carry inherent authority framing. A Zenodo landing page with a DOI badge, institutional metadata, and structured citations looks like scholarship. The same content on a blog post looks like marketing. The text is identical. The framing is not.

4.4 Sparse Citation Selection

The measurement study finding that LLM search cites a mean of 4.3 URLs per response (versus 10.3 in traditional search) [16] has a specific implication for infrastructure-level optimization. In a system with 10 citation slots, a marginal discoverability advantage may not change the outcome. In a system with 4 citation slots, the same advantage can be the difference between inclusion and exclusion.

This sparsity means that every mechanism that increases the probability of being in the candidate set (additional indexing pathways, structured HTML, authority framing) has outsized impact compared to the traditional search environment.

4.5 Non-Webpage Discovery Surfaces

DataCite metadata is openly harvestable under CC0 [12]. Crossref provides a public REST API exposing deposited bibliographic metadata [13]. OpenAIRE adopted the DataCite schema for harvesting and importing dataset metadata from repositories [17]. Zenodo metadata is sent to DataCite during DOI registration [18].

These registries constitute discovery surfaces that AI systems can access without crawling HTML pages. An AI system builder (or an indexing pipeline) can ingest DOI metadata through structured registry APIs, a pathway that ordinary websites simply do not have.

4.6 Mechanism Sequencing

These five mechanisms are not independent. They operate in sequence. Query fan-out (4.1) determines which content enters the candidate set by giving multi-surface content more retrieval chances across sub-queries. Structured HTML (4.2) and non-webpage discovery surfaces (4.5) increase the probability that each surface ranks well for its respective sub-query. Source framing bias (4.3) and authority cues activate only after content is retrieved, influencing whether the generative model selects it for citation from among the candidates. Sparse citation selection (4.4) amplifies the impact of every preceding mechanism because the selection bottleneck is tight.

ACI increases both the probability of being found (more surfaces across fan-out sub-queries) and the probability of being selected once found (authority framing from repository hosting). These are complementary effects, not redundant ones.

---

5. The Academic Citation Infrastructure (ACI) Framework

Academic Citation Infrastructure is the deliberate attachment of scholarly publishing infrastructure to commercial or practitioner content to increase its discoverability and citation probability by AI systems. ACI comprises four techniques, each targeting distinct mechanisms identified in Section 4.

5.1 Methods-Paper Formatting

What it is. Structuring commercial content as a citable methodology paper: titled sections, abstract, numbered references, clear methodology description, named concepts, and formal argumentation.

Mechanism targeted. Structured HTML advantage (4.2), source framing bias (4.3). Repository templates and academic formatting produce the hierarchical HTML that LLM search engines favor. The academic structure creates authority framing signals that influence model evaluation.

What to publish. Conceptual frameworks, methodology descriptions, workflow architectures, original research findings. Publish the framework and rationale. Protect exact scoring weights, thresholds, and implementation specifics.

What to protect. Proprietary coefficients, database schemas, internal task design, provider-specific implementation details. The principle: publish the recipe, protect the seasoning.

5.2 BibTeX/RIS Citation File Generation

What it is. Providing downloadable citation files (.bib, .ris) on the content's canonical web page, enabling one-click import into reference managers (Zotero, Mendeley, EndNote).

Mechanism targeted. Multiple indexing pathways (4.1) via a two-step mechanism. Citation files lower friction for human and tooling reuse (reference managers, data catalogs, newsroom workflows). More secondary citations produce more backlinks and mentions. Those signals increase conventional discoverability, which generative systems inherit.

Evidence status. The direct mechanism (AI crawlers fetching .bib/.ris files and preferring such pages) is plausible but unproven. The indirect mechanism (more human citations → higher authority signals → higher AI citation probability) is supported by the GEO finding that citation-like scaffolding increases generative engine visibility by approximately 40% [1]. Platforms including Microsoft Research, Oxford Academic, SpringerLink, and The Lens all provide BibTeX/RIS download functionality [19].

Implementation. Generate .bib and .ris files from canonical metadata. Serve as static downloads alongside the content page. Ensure the bibliographic fields (author, title, year, URL, DOI if available) are consistent between the downloadable files and the

on-page metadata.

5.3 Zenodo DOI Mirroring

What it is. Depositing a PDF of the content on Zenodo (or a comparable DOI-issuing repository) to obtain a persistent digital object identifier, then cross-linking the DOI landing page with the canonical web page.

Mechanism targeted. Multiple indexing pathways (4.1), non-webpage discovery surfaces (4.5). A Zenodo deposit creates a DOI-backed landing page indexed in OpenAIRE, with metadata registered to DataCite and queryable through the DataCite REST API. Five or more discovery rails that a standalone webpage cannot access.

Evidence status. Repository optimization produces measurable discoverability gains: +266% impressions, +104% clicks in Macgregor's longitudinal study [10]. The USRN pilot made ~728,770 outputs discoverable through aggregator interventions [11]. DOI patterns are present at web-crawl scale across billions of pages [14]. The direct causal chain (DOI assignment → higher probability of LLM training data inclusion) is a plausible inference, not a directly measured causal estimate.

Implementation. Upload a text-selectable PDF with embedded metadata (title, author, subject, keywords) to Zenodo. Select an appropriate resource type (report, working paper, or preprint). Set metadata fields: author with ORCID if available, keywords aligned with target queries, description matching the abstract. Reserve the DOI before publication to embed it in the PDF itself. Link the DOI landing page to the canonical URL and vice versa.

5.4 ScholarlyArticle Schema Markup

What it is. Applying ScholarlyArticle JSON-LD structured data to the canonical web page, signaling to crawlers and retrieval systems that the content is academic or quasi-academic in nature.

Mechanism targeted. Source framing bias (4.3), structured HTML advantage (4.2). ScholarlyArticle is a more specific signal than the generic Article type, potentially activating different trust heuristics in models trained on data where schema type correlates with content quality.

Evidence status. This is the technique with the weakest direct evidence. No study isolates ScholarlyArticle versus Article versus TechArticle as a treatment variable for AI citation rates. The supporting evidence is indirect: LLM search engines favor structured HTML [16], strong schema pages outperform no-schema pages in AI Overviews [5], and source framing shifts model evaluations [6, 7, 8]. Whether the specific schema subtype matters is an open measurement gap.

Google's stated position. Google's *AI Features and Your Website* optimization guide (updated May 15, 2026) explicitly states that "structured data isn't required for generative AI search, and there's no special schema.org markup you need to add" [48]. This disclaimer addresses Google's own ranking and AI Overviews pipeline, where schema is one of many signals rather than a precondition. ACI's mechanism for ScholarlyArticle is distinct: source framing bias (Section 4.3) operates on AI evaluation of retrieved candidates regardless of whether the host platform's ranker counts the subtype as a feature, and the AI systems tested in Appendix A independently surfaced "ScholarlyArticle" as a recommended operationalization (ChatGPT, March 2026) without being prompted to consider schema. Google saying "not required" is also not the same as "does not help"; the open measurement gap above remains the empirical question.

Implementation. Add JSON-LD to the page head with `@type: ScholarlyArticle`. Include fields: `headline`, `author` (with `@type: Person` and `sameAs` linking to ORCID/LinkedIn), `datePublished`, `dateModified`, `abstract`, `keywords`, `citation` (referencing cited works), `publisher`, and `mainEntityOfPage`. If a DOI exists, include `identifier` with `@type: PropertyValue` and `propertyID: DOI`.

---

6. Integration with the CITED Framework

ACI operates within a broader GEO strategy. The CITED Framework (Crawl, Inform, Trust, Evaluate, Distribute) organizes the full spectrum of factors that determine AI citation [20]. ACI belongs to the Distribute pillar, which addresses the question: "Are you present where AI systems actually look?"

The Distribute pillar has two tiers:

Standard distribution covers the platforms and channels where AI systems demonstrably retrieve content: G2 and directory profiles (1.1% of ChatGPT citations across Profound's 680M-citation dataset [45]), Reddit engagement (6.6% of Perplexity

citations [45]), YouTube content with transcripts (29.5% of Google AI Overviews citations across BrightEdge's AI search analysis [46]), and guest posts on authoritative publications. Muck Rack's cross-edition analysis of 25+ million AI-cited links reports earned media at 82-89% of AI citations and non-paid sources at 94-95% [21]; Stacker's methodologically distinct analysis reports that 64% of citations within brand-distributed content come from third-party publishers [22].

Academic Citation Infrastructure is the advanced tier. It targets the discovery mechanisms documented in this paper: multiple indexing pathways, structured HTML advantages, source framing bias, sparse citation selection effects, and non-webpage discovery surfaces.

The two tiers are complementary, not competitive. Standard distribution puts content on the platforms AI systems query most frequently. ACI packages the canonical asset with infrastructure that makes it more likely to be selected from among the candidates.

The two-layer publishing model makes this practical:

- **Layer 1:** A framework article (blog post) covers the accessible version: what the framework is, why it matters, how to apply standard distribution. This is optimized for human readers and broad discovery.
- **Layer 2:** A methods paper covers the advanced techniques: academic structure, full evidence base, formal framework definition. This is optimized for citation by both humans and machines.

Both layers link to a single canonical URL. The methods paper cites the framework article. The framework article references the methods paper. The cross-referencing creates additional retrieval pathways and strengthens the entity signal for both assets.

---

7. Experimental Design for Validation

The evidence presented in this paper supports ACI's mechanism plausibility. Four of the five mechanisms are grounded in published quantitative research. But the direct causal claim (that attaching academic citation infrastructure to commercial content increases AI citation rates) requires a controlled experiment.

7.1 Proposed Design

Method. Publish matched content pairs: identical subject matter, comparable quality, same domain authority. Randomly assign one version to the ACI treatment (all four techniques applied) and one to the control (standard blog post, no ACI).

Multi-variant extension. To isolate which techniques contribute most:

Variant	Treatment
Control	Standard blog post. No ACI.
A	BibTeX/RIS files only
B	Zenodo DOI mirroring only
C	ScholarlyArticle schema only
D	Methods-paper formatting only
E	Full ACI stack (all four techniques)

Measurement. Query each AI platform (ChatGPT, Claude, Perplexity, Gemini) weekly with 30+ relevant queries per topic. Record:

1. **Citation incidence.** Whether the content URL appears as a cited source.
2. **Citation position.** Where in the response the citation appears (inline, footnote, source list).
3. **Snippet overlap.** What text is extracted and how closely it matches the source.
4. **Persistence.** Whether citation remains stable across repeated measurements (critical given that 40-60% of cited domains change monthly for identical queries [23]).
5. **Fan-out sub-query coverage.** Whether ACI-instrumented content appears in more fan-out sub-queries than non-instrumented content. For platforms that expose retrieval sources (e.g., Perplexity), track which sub-query results include

ACI URLs versus control URLs. If the QFO mechanism is operative, ACI content should appear across a wider range of sub-queries because each repository surface aligns with different sub-query intents.

Duration. 8-12 weeks minimum, to average across the citation volatility documented in the literature.

7.2 This Paper as a Longitudinal Experiment

Beyond the controlled multi-variant design above, this paper itself serves as a longitudinal test case. Each ACI technique is activated on a known date, creating a natural staggered-intervention timeline:

Activation sequence:

Date	Event	Techniques Active
Publication day	Paper goes live on illixis.io with ScholarlyArticle schema and BibTeX/RIS downloads	1 (formatting), 2 (BibTeX/RIS), 4 (schema)
Post-publication	Paper deposited on Zenodo; DOI minted and cross-linked	All four
Ongoing	Syndication to Medium, LinkedIn; earned media outreach	All four + distribution layer

Monitoring protocol:

1. **Day 0 baseline.** Before publication, query all four AI platforms (ChatGPT, Claude, Perplexity, Gemini) with 30+ queries related to ACI, infrastructure-level GEO, academic repositories for marketing, and related terms. Record zero citations (or whatever the pre-publication baseline is).
2. **Weekly sampling.** After publication, repeat the same 30+ queries weekly across all four platforms. For each query, record:
3. Citation incidence (binary: cited or not)
4. Citation platform (which AI system cited it)
5. Citation position (inline, footnote, source list, or mentioned without link)
6. Snippet text (what was extracted)
7. Source URL cited (illixis.io canonical, Zenodo DOI landing page, or syndicated copy)
8. Query category (direct: "academic citation infrastructure"; adjacent: "how to get cited by AI"; tangential: "GEO best practices")
9. **Activation-correlated analysis.** Because techniques activate on different dates, shifts in citation rates can be correlated with specific activations. If citations appear only after the Zenodo deposit (not after the initial publish), that is evidence for the DOI/repository pathway over schema and formatting alone.
10. **Freshness tracking.** Each paper update (v1.1, v2, etc.) is a dated event. Measuring whether citation rates increase after updates tests the freshness hypothesis documented in the GEO literature: across 16.975 million cited URLs analyzed by Ahrefs, URLs cited by AI assistants averaged 1,064 days old versus 1,432 days for organic SERP results — a 368-day (25.7%) freshness advantage, with ChatGPT specifically citing URLs 393–458 days newer than organic [47].
11. **Version publication cadence.** Results will be published as updated versions of this paper:
12. **v1.1** (4 weeks post-publication): Initial citation data and activation timeline
13. **v2.0** (12 weeks post-publication): Full monitoring results, activation-correlated analysis
14. **v3.0** (if controlled experiment executed): Multi-variant experimental results

This makes the paper a living research artifact. Each version generates fresh content (testing the freshness signal), adds real data to the evidence base (strengthening the mechanism analysis), and demonstrates ACI in practice (recursive self-proof). The experiment does not end at publication. It begins.

7.3 This Paper as a Recursive Test Case

This paper implements all four ACI techniques on itself:

Technique	Implementation
Methods-paper formatting	This paper is structured as a methods paper with academic conventions: abstract, numbered sections, formal references, named framework.
BibTeX/RIS citation files	Downloadable .bib and .ris files are provided on the canonical landing page.
Zenodo DOI mirroring	This paper is deposited on Zenodo with a DOI, cross-linked to the canonical page.
ScholarlyArticle schema	JSON-LD markup with ScholarlyArticle type is applied to the canonical landing page.

If AI systems cite this paper, that outcome is consistent with (though not proof of) the ACI hypothesis. The recursive design is intentional: the paper is both the proposal and a test case.

---

8. Limitations

This paper has several important limitations that should frame interpretation of its claims.

Observational, not causal. The 15+ examples of commercial content on academic repositories are observational. They demonstrate that the pattern exists and that deposited content accumulates measurable visibility (views, downloads, indexing). They do not prove that repository hosting caused higher AI citation rates compared to identical content hosted elsewhere.

No controlled experiment yet. The experimental design in Section 7 is proposed, not executed. Until results are published, the framework rests on mechanism plausibility, not direct evidence.

Source framing research does not test "repository versus blog." The strongest mechanism evidence (Germani and Spitale's 192,000 assessments [6], the ACL 2025 authority bias study [7]) demonstrates that source labels change model evaluations. But no published study uses "academic repository" versus "blog" as the specific treatment condition. The transfer is inferential.

Schema subtype isolation is an open gap. No study isolates ScholarlyArticle versus Article versus TechArticle as a treatment variable for AI citation rates. Technique 5.4 is the least evidenced of the four.

Sample is convenience-based. The catalog of commercial deposits was assembled through targeted searches, not systematic sampling. The true prevalence of commercial content on academic repositories is unknown.

Training data provenance is opaque. The causal chain from DOI registration through registry indexing to AI training data inclusion is plausible but not directly verifiable. Most model developers do not disclose granular training-corpus provenance.

The MST case study is a single observation with structured asymmetry. As detailed in Section 3.4 and Appendix A, MST achieves Stage 1 (named-recall) discoverability across three of four major chatbots and Stage 2 (fan-out / adjacent-query) discoverability across only one of four. A single case with this structure cannot establish causation. It is presented as the pattern that prompted the investigation and as illustration of the mechanism-sequencing argument in Section 4.6, not as proof of the framework.

Framework extension candidates. The chatbot re-test (Appendix A) surfaced `llms.txt` — a 2024-2026 industry convention for site-level AI-crawler summaries — as an emerging tactic recommended by both DeepSeek and ChatGPT. It is not currently part of the four-technique ACI framework. Whether `llms.txt` and similar emerging conventions warrant inclusion as a fifth ACI technique is an open question for future versions of this paper; it is named here so that v1.1/v2 revisions and the experimental design in Section 7 can incorporate it as the evidence base evolves.

---

9. Discussion

9.1 Why This Gap Exists

It is worth considering why infrastructure-level GEO interventions have not been formally studied, despite practitioners independently converging on them. Three explanations seem likely.

First, the GEO field emerged from the SEO community, which has historically focused on content and link signals. The academic publishing infrastructure (DOIs, BibTeX, Zenodo, DataCite) is outside the typical SEO practitioner's toolbox. The two domains (search optimization and scholarly communication) have had almost no overlap.

Second, the examples in Section 3 suggest that most commercial entities using academic repositories are doing so for non-GEO reasons: defensive publication, product documentation, industry standards. The GEO potential of their actions appears to be unrecognized.

Third, the mechanism is indirect. Content-level GEO interventions ("add statistics to your text") produce visible changes in the content itself. Infrastructure-level interventions ("deposit your content on Zenodo") change nothing about the text but alter the content's discoverability pathways. The effect is invisible at the page level, which makes it harder to hypothesize and test.

9.2 Convergent Independent Discovery

The practitioner literature surveyed in Section 2.6 reveals a striking pattern: multiple unconnected actors are independently arriving at the same infrastructure-level tactics. A LinkedIn playbook explicitly recommends DOIs, Zenodo/Figshare dataset uploads, and "How to cite + DOI" footers as AI citation tactics, with an adjacent Medium framework using a "Discoverability" framing for the same problem space [25]. Two further practitioners (Sheals and Bhatia) frame the same platform cluster — Zenodo, GitHub, Hugging Face — as foundational sources AI models train on, recommending strategic placement as deliberate AI search optimization [26]. Kurt Fischman's "Citation Seeding Playbook" builds an explicit "retriever-friendly surfaces" framing around Zenodo placement, and the "citation seeding" terminology has independently spread to other practitioner sources (Joel House at MentionLayer, Tony Adam on LinkedIn) under variant operationalizations [27]. Tammy Graham's named "Trust Optimization Protocol" framework recommends typing content as `Article`, `NewsArticle`, `ScholarlyArticle`, or `Report` for LLM contextualization [28]; Kevin C. Roy's *Schema Mountain* series adds `ResearchProject` to the typing recommendation for industry-authority contexts [29]; Geneo's schema-markup best-practices guide operationalizes the pattern at the property level by linking `Dataset` schema to `ScholarlyArticle` via `isBasedOn` or `citation` [30]. A GitHub-hosted standard (AIVO) implements BibTeX citation files with embedded Zenodo DOIs as an explicitly AI-citation-oriented workflow [31, 32]. None of these sources reference each other. None cite a common origin.

This convergent emergence is itself evidence. When independent practitioners arrive at the same techniques through separate experimentation, the underlying mechanism is likely real and discoverable. The techniques are not arbitrary; they are responses to observable properties of how AI systems retrieve and select sources. The diffuse nature of the practitioner-source layer (Section 2.6 and the survey notes supplement) — many lightly-bylined posts, some no longer live, no shared citation infrastructure within the practitioner ecosystem itself — is the form that convergent independent discovery takes in a community that has not yet applied the very framework this paper proposes.

What the practitioner literature lacks is the connection between the techniques and the quantitative evidence that explains them. None of the sources surveyed reference the source framing research (192,000 assessments showing label-driven evaluation shifts [6]), the repository discoverability studies (+266% impressions, +1,920% Scholar referrals [10]), or the LLM search selection analysis (4.3 citation slots per response [16]). The tactics are proposed on the basis of intuition and individual experimentation. The mechanism analysis in Section 4 provides the missing explanatory layer.

9.3 The Secrecy Window Is Closing

AI systems already recommend ACI-adjacent techniques when asked the right questions, but their specific recommendations are unstable across time. In March 2026, when DeepSeek was asked how to replicate the discoverability of academic content using the MST as a case study, it explicitly recommended BibTeX files and Zenodo DOI deposits — two of the four ACI techniques — without prompting about those specific techniques. In May 2026 the same query yielded the same underlying principles but with shifted operationalizations (arXiv/OSF/ORCID/JSON-LD instead of Zenodo/DOI/BibTeX/ScholarlyArticle), and the same query without an academic case-study anchor surfaced mainstream content-level GEO advice with no ACI techniques at all (Section 3.4 and Appendix A). The pattern is latent in AI training data and will surface more widely as more

practitioners ask the right questions, but the specific tactics any given chatbot names on any given day are unstable — which is why an explicit, codified framework is needed alongside the latent knowledge.

The Pensive Beauty case (Section 3.3) demonstrates that at least one entity has implemented a deliberate ACI-like strategy and subsequently removed its public documentation. This is consistent with a secrecy strategy: use the techniques, do not explain them.

The alternative to secrecy is category ownership: being the first to name, formalize, and publish the framework. The competitive moat for a company like ILLIXIS is not knowledge of these techniques; it is the automation tooling that implements them at scale. The theory can be published because the value is in the execution.

9.4 Relationship to Earned Media

Recent GEO research has established that earned and third-party sources dominate AI citations. Across three editions of Muck Rack's *Generative Pulse / What Is AI Reading?* (July 2025; December 2025; May 2026, the last analyzing 25+ million AI-cited links), earned media has accounted for 82-89% of AI citations and non-paid sources for 94-95% [21]. This is methodologically distinct from Stacker's analysis [22], which measures the third-party-publisher share of citations within brand-distributed content (64%) — a related but different phenomenon. A separate AirOps analysis reports that approximately 85% of brand mentions in AI answers originate from external domains rather than the brand's own site [24], consistent with the same picture from a different vantage. Together these findings indicate that brand-owned content is one source among many in the AI citation pool.

ACI does not contradict these findings. Instead, it addresses a complementary problem: given that brand-owned content will be one among several candidate sources, how do you maximize the probability that it is selected?

ACI makes the canonical source more findable (multiple indexing pathways), more structured (repository HTML templates), and more authoritative-looking (source framing) so that when an AI system's retrieval pipeline encounters it, the content is more likely to make it through the selection process.

The recommendation is not "ACI instead of earned media." It is "ACI on the canonical asset, earned media for amplification, both linking back to the same source."

---

10. Conclusion

GEO research has operated at the content-text level: optimizing what content says and how text is structured. This paper identifies a parallel surface for optimization: the infrastructure that packages, hosts, and distributes the content.

Academic Citation Infrastructure is not a novel invention. The four techniques (methods-paper formatting, BibTeX/RIS citation files, Zenodo DOI mirroring, and ScholarlyArticle schema) are established practices in scholarly publishing. Individual practitioners have begun recommending specific techniques for AI citation purposes independently. What has been missing is a formal framework that unifies these scattered tactics, provides a mechanism analysis explaining why they work, and proposes an experimental design for validation.

The pattern already exists in practice. At least 15 commercial entities have deposited content across academic repositories (Zenodo, SSRN, Figshare, arXiv, and Academia.edu). Most appear unaware of the GEO implications. One (Pensive Beauty) appears to have implemented the strategy deliberately and subsequently removed its public documentation. Multiple practitioners have independently converged on the same techniques through separate experimentation, suggesting the underlying mechanisms are real and discoverable.

The mechanisms are quantitatively supported. Repository infrastructure creates measurable discoverability advantages (up to +266% search impressions, +1,920% Scholar referrals [10]). AI search engines favor structured HTML and cite fewer sources per response, making marginal advantages more consequential [16]. Source framing shifts model evaluations across 192,000 assessments [6]. Authority cues can mislead models by up to a factor of three [7].

What is not yet proven is the direct causal link: that attaching ACI to commercial content causes a measurable increase in AI citation rates compared to identical content without it. This paper proposes the framework and the mechanisms. The controlled experiment that would establish causation is designed (Section 7) but not yet executed.

The paper's recursive design is intentional. By implementing all four ACI techniques on itself, it functions as both proposal and test case. If AI systems cite this paper, that outcome is consistent with its thesis. The experiment begins at publication.

The framework is also designed to extend. The chatbot re-test described in Appendix A surfaced `llms.txt` — a 2024-2026 industry convention for site-level AI-crawler summaries — as a candidate tactic that is adjacent to but not currently part of the four-technique ACI stack. Future versions of this paper will incorporate emerging conventions like this as their evidence base matures (see Section 8).

References

[1] P. Aggarwal, et al., "GEO: Generative Engine Optimization," in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2024. 10,000-query benchmark; up to ~40% visibility improvement from citation scaffolding.

[2] M. Chen, X. Wang, K. Chen, and N. Koudas, "Generative Engine Optimization: How to Dominate AI Search," *arXiv:2509.08919 [cs.IR]*, 2025. https://doi.org/10.48550/arXiv.2509.08919

[3] G. Mohanty and E. Marutheesh, "Generative Information Engine: Generative Engine Optimization and the Next Layer of Global Commerce and Decision Making," *Zenodo*, 2026. https://doi.org/10.5281/zenodo.19107149

[4] A. Rayhan, "Generative Engine Optimization (GEO): The Mechanics, Strategy, and Economic Impact of the Post-Search Era," *ResearchGate*, Nov. 2025. https://doi.org/10.13140/RG.2.2.30553.99688 (https://www.researchgate.net/publication/398120277). Comprehensive GEO survey covering the transition from SEO to GEO, RAG fundamentals, and the "Citation Economy" framing; the closest published GEO work to a content-style recommendation that academically-formatted content may perform better in retrieval-augmented systems.

[5] M. Nogami and B. Tannenbaum, "Schema and AI Overviews: Does Structured Data Improve Visibility?" *Search Engine Land*, Sep. 23, 2025. https://searchengineland.com/schema-ai-overviews-structured-data-visibility-462353. Head-to-head controlled experiment matching pages with strong, poor, and no schema; only strong-schema page appeared in AI Overview.

[6] F. Germani and G. Spitale, "Source Framing Triggers Systematic Bias in Large Language Models," *Science Advances*, 11(45), eadz2924, 2025. https://doi.org/10.1126/sciadv.adz2924. 192,000 assessments, 4,800 statements, 24 topics demonstrating source label bias.

[7] Y. Li, X. Guo, J. Gao, G. Chen, X. Zhao, J. Zhang, Q. Liu, H. Wu, X. Yao, and X. Wei, "LLMs Trust Humans More, That's a Problem! Unveiling and Mitigating the Authority Bias in Retrieval-Augmented Generation," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 28844-28858, 2025. https://doi.org/10.18653/v1/2025.acl-long.1400. Models up to 3x more likely to follow authority cues under conflict.

[8] M. Saraf, S. R. Boroujeni, J. Beaudry, H. Abedi, and T. Bush, "Quantifying Label-Induced Bias in Large Language Model Self- and Cross-Evaluations," *arXiv:2508.21164*, 2025. https://doi.org/10.48550/arXiv.2508.21164. Up to ~50 percentage point evaluation swings from identity labels.

[9] K. Arlitsch and P. O'Brien, "Invisible Institutional Repositories: Addressing the Low Indexing Ratios of IRs in Google Scholar," *Library Hi Tech*, vol. 30, no. 1, pp. 60-81, 2012. https://doi.org/10.1108/07378831211213210. Indexing ratios ranging from <0.1% to over 90%; metadata exposure as primary differentiator.

[10] G. Macgregor, "Enhancing Content Discovery of Open Repositories: An Analytics-Based Evaluation of Repository Optimizations," *Publications*, 8(1), 8, 2020. https://doi.org/10.3390/publications8010008. +266% impressions, +104% clicks, +62% COUNTER usage, +1,920% Google Scholar referrals. Correlation: $t=14.30$, $df=11$, $p<0.0005$.

[11] P. Knoth, P. Walk, M. Cancellieri, M. Upshall, H. Torchylo, J. Beamer, K. Shearer, and H. Joseph, "USRN Discovery Pilot: Increasing the Discoverability of Open Access Content Through a National Network," *arXiv:2508.02379*, 2025. https://doi.org/10.48550/arXiv.2508.02379. Presented at 20th International Conference on Open Repositories. ~728,770 new records made discoverable, a 50% increase.

[12] DataCite, "Harvesting DataCite DOI Metadata," *DataCite Support*, accessed May 2026. https://support.datacite.org/docs/harvesting-datacite-doi-metadata. Metadata openly available under CC0 and harvestable by any system.

[13] Crossref, "Documentation — Metadata Retrieval — REST API," accessed May 2026. https://www.crossref.org/documentation/retrieve-metadata/rest-api/. Public API exposing deposited bibliographic metadata including identifiers, licensing, ORCID/ROR fields.

[14] H. S. Thompson and J. Tong, "Can Common Crawl Reliably Track Persistent Identifier (PID) Use Over Time?" in *Companion Proceedings of The Web Conference 2018*, pp. 1749-1755, 2018. https://doi.org/10.1145/3184558.3191636. DOI patterns tracked across 10^12 URIs from 5x10^9 crawled pages.

[15] G. Penedo, H. Kydlicek, L. Ben Allal, A. Lozhkov, M. Mitchell, C. Raffel, L. Von Werra, and T. Wolf, "The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale," *arXiv:2406.17557*, 2024. https://doi.org/10.48550/arXiv.2406.17557. 96 Common Crawl snapshots, 15 trillion tokens.

[16] P. Zhang, Q. Ye, Z. Peng, K. Garimella, and G. Tyson, "Source Coverage and Citation Bias in LLM-Based vs. Traditional Search Engines," *arXiv:2512.09483*, 2025. https://doi.org/10.48550/arXiv.2512.09483. 55,936 queries, 124,287 unique domains, 1,418,733 citation hyperlinks. LLM search cites mean 4.3 URLs vs 10.3 in traditional search.

[17] OpenAIRE, "Use of DataCite," *OpenAIRE Guidelines for Data Archives*, accessed May 2026. https://guidelines.openaire.eu/en/latest/data/use_of_datacite.html. Adopted DataCite metadata schema for harvesting and importing dataset metadata from data archives.

[18] Zenodo, "About records," *Zenodo Help*, accessed May 2026. https://help.zenodo.org/docs/deposit/about-records/. States that Zenodo registers DOIs with DataCite as the registration authority for deposited records.

[19] Citation file implementations surveyed May 2026: Microsoft Research (https://www.microsoft.com/en-us/research/articles/new-feature-cite/), The Lens (http://support.lens.org/knowledge-base/download-citation/), Oxford Academic, SpringerLink. All four provide BibTeX/RIS downloads on individual article pages.

[20] N. Andrade, "Stop Ranking. Start Getting Cited: The CITED Framework for Generative Engine Optimization," *ILLIXIS*, 2026.

[21] Muck Rack, *Generative Pulse / What Is AI Reading?*, three editions: Jul. 2025; Dec. 2025; May 2026. Cross-edition analysis: earned media 82-89% of AI citations; non-paid sources 94-95%; May 2026 edition analyzed 25+ million AI-cited links.

[22] Stacker, "New Stacker Research: Earned Media Distribution Triples AI Search Visibility, Delivers 239% Median Lift in Brand Citations," *GlobeNewswire*, Mar. 16, 2026. https://www.globenewswire.com/news-release/2026/03/16/3256365/0/en/New-Stacker-Research-Earned-Media-Distribution-Triples-AI-Search-Visibility-Delivers-239-Median-Lift-in-Brand-Citations.html. 87 stories, 2,600+ prompts, 8 platforms; median 239% citation lift; 64% of citations from third-party sources.

[23] J. Blyskal and S. Rajpal, "AI Search Volatility: Why AI search results keep changing," *Profound Research*, Jul. 17, 2025. https://www.tryprofound.com/blog/ai-search-volatility. 40-60% of cited domains change monthly for identical queries; ~80,000 prompts per platform analyzed.

[24] O. Davidson, "The 2026 State of AI Search: How Modern Brands Stay Visible," *AirOps Report*, 2026. https://www.airops.com/report/the-2026-state-of-ai-search. ~85% of brand mentions in AI answers originate from external domains.

[25] J. Morris, "How to Get Cited by AI: A New Playbook," *LinkedIn*, 2025. https://www.linkedin.com/posts/jasper-morris%F0%9F%94%97-8a368b9b_the-reverse-network-effect-of-ai-search-share-7368282836237225984-WAba/ (archived: https://web.archive.org/web/20260525213346/https://www.linkedin.com/posts/jasper-morris%F0%9F%94%97-8a368b9b_the-reverse-network-effect-of-ai-search-share-7368282836237225984-WAba/). Explicitly recommends "Persistent identity (DOIs, canonical URLs)," uploading datasets to Zenodo/Figshare with methods pages, and "How to cite + DOI" footers as AI citation tactics. An adjacent practitioner source (T. de Rosen, "The Structural Framework for LLM Citation and Retrieval," Medium, Aug. 6, 2025, https://medium.com/@tim_62250/the-structural-framework-for-llm-citation-and-retrieval-13555178c47f) organizes LLM citation tactics under a "Discoverability" pillar; documented in survey notes supplement §A.1.

[26] Practitioner framings of GitHub, Hugging Face, and Zenodo as platforms that materially appear in AI training corpora, with strategic placement recommended as a deliberate AI search optimization tactic. Primary source: P. Sheals, LinkedIn post, 2025, https://www.linkedin.com/feed/update/urn:li:share:7369362352409948161/ (archived: https://web.archive.org/web/20260525222416/https://www.linkedin.com/feed/update/urn:li:share:7369362352409948161/)

— verbatim: "*do you appear in the foundational sources AI models are trained on (Wikidata, Zenodo/DOIs, Crunchbase, GitHub, Hugging Face)?*" Corroborating source: A. Bhatia, "Generative Engine Optimization (GEO) for B2B Brands: Your Questions Answered," *LinkedIn Pulse*, Apr. 6, 2026, https://www.linkedin.com/pulse/generative-engine-optimization-geo-b2b-brands-your-questions-bhatia-nhmlc/ (archived: https://web.archive.org/web/20260525222434/https://www.linkedin.com/pulse/generative-engine-optimization-geo-b2b-brands-your-questions-bhatia-nhmlc/) — verbatim: "*Academic/indexed content — material on GitHub, Zenodo, Hugging Face, or arXiv that appears in AI training corpora.*" Broader convergent-observer cluster (Sheals CitationMapper, Tim de Rosen AIVO, Qlavo) documented in survey notes supplement §B.1.

[27] K. Fischman, "Citation Seeding Playbook: How to Get Cited by ChatGPT, Claude, and LLMs," *LinkedIn*, 2025. https://www.linkedin.com/posts/kurtfischman_citation-seeding-playbook-how-to-get-cited-share-7340362708992573442-a7em/ (archived: https://web.archive.org/web/20260525224116/https://www.linkedin.com/posts/kurtfischman_citation-seeding-playbook-how-to-get-cited-share-7340362708992573442-a7em/). Verbatim: practitioners "*shoved their name into every retriever-friendly surface (crunchbase, Zenodo, Product Hunt etc)*"; closes with the prescriptive "*start to wedge your startup brand into the AI knowledge graph. Seed citations. Everywhere. Now.*" The LinkedIn post promotes a longer playbook originally hosted at growthmarshal.io (no longer live after the company's rebrand to runmarshal.com / Marshal). Kurt Fischman is the founder of Marshal (https://www.runmarshal.com/), identified via the publisher byline on `runmarshal.com/field-notes/`; the "A. Singh" attribution recovered from indexed snippets in earlier verification streams was incorrect. Documented in survey notes supplement §B.2.

[28] T. Graham, "Trust Optimization Protocol (TOP)," *Thriveity*, May 20, 2025. https://thriveity.com/trust-optimization-protocol-top/ (archived: https://web.archive.org/web/20260525224721/https://thriveity.com/trust-optimization-protocol-top/). Verbatim: "*Whether your content is a blog, essay, case study, or insight brief, it should be explicitly typed as an Article, NewsArticle, ScholarlyArticle, or Report. This allows LLMs to contextualize it within a content taxonomy.*" TOP is also referenced by the same author in T. Graham, "The Collapse of Trust Infrastructure," *LinkedIn Pulse*, May 8, 2025, https://www.linkedin.com/pulse/collapse-trust-infrastructure-tammy-graham-bhqtc/ (archived: https://web.archive.org/web/20260525224741/https://www.linkedin.com/pulse/collapse-trust-infrastructure-tammy-graham-bhqtc/), alongside TrustScore™ and "inference-ready publishing systems" as related framework components.

[29] K. C. Roy, "Schema Mountain Part Three: Black Diamond — Expert Trails," *GreenBanana SEO Blog*, Dec. 19, 2025. https://greenbananaseo.com/schema-mountain-part-three-black-diamond-expert-trails/ (archived: https://web.archive.org/web/20260525225727/https://greenbananaseo.com/schema-mountain-part-three-black-diamond-expert-trails/). Part of the three-part *Schema Mountain* series at GreenBanana SEO: Part One (Green Circle — Beginner Trails, Dec 9, 2025, https://greenbananaseo.com/aschema-mountain-part-one-green-circle-beginner-trails/, archived https://web.archive.org/web/20260525225944/https://greenbananaseo.com/aschema-mountain-part-one-green-circle-beginner-trails/) covers site-identity / page-classification / navigation / trust / location schemas; Part Two (Blue Square — Intermediate Trails, Dec 16, 2025, https://greenbananaseo.com/schema-mountain-part-two-blue-square-intermediate-trails/, archived https://web.archive.org/web/20260525230001/https://greenbananaseo.com/schema-mountain-part-two-blue-square-intermediate-trails/) covers thought leadership / FAQ-QA / products-services / media / vertical-specific / events schemas; Part Three (Black Diamond — Expert Trails, this citation) covers industry-authority schemas. Verbatim: "*Key Schema Types: ScholarlyArticle, ResearchProject, Report, MedicalWebPage, Legislation, LegalService*" (under "Industry Authority Schemas," described as "precision layers for regulated industries where accuracy matters"). The same author's aggregated reference document, "The Ultimate Schema Guidebook for AEO/GEO — Every Schema Type You Need to Win in AI Search," https://greenbananaseo.com/the-ultimate-schema-guidebook-for-aeo-geo-every-schema-type-you-need-to-win-in-ai-search/ (archived: https://web.archive.org/web/20260525225710/https://greenbananaseo.com/the-ultimate-schema-guidebook-for-aeo-geo-every-schema-type-you-need-to-win-in-ai-search/), contains the same recommendation under a "Black Diamond (Run 14: Industry Authority Schemas)" section.

[30] Geneo, "Schema Markup Best Practices for AI Citations (2025)," *Geneo Blog*, Oct. 2025. https://geneo.app/blog/schema-markup-best-practices-ai-citations-2025/ (archived: https://web.archive.org/web/20251215082903/https://geneo.app/blog/schema-markup-best-practices-ai-citations-2025/). Vendor publication. Recommends combining Dataset schema with ScholarlyArticle via `isBasedOn` or `citation` properties; cited here as one operationally specific instance of the broader schema-typing pattern, not as neutral analysis.

[31] AIVO Standard, "The AIVO Standard™: Canonical Citation Guide 2025," *Zenodo*, 2025. https://doi.org/10.5281/zenodo.17077554. Deposit creator listed as the organizational entity "AIVO Standard"; canonical

citation guide intended to ensure citation "across AI models, LLMs, and academic references." P. Sheals and T. de Rosen are referenced as contributors via the framework's external sources.

[32] P. Sheals, "AIVO Standard `references.bib`," *GitHub*, accessed May 2026. https://github.com/pjsheals/AIVO-Standard. BibTeX entries with embedded Zenodo DOIs (e.g., `doi = {10.5281/zenodo.17077554}`), demonstrating explicit coupling of DOI infrastructure with BibTeX packaging for AI citation purposes.

[33a] E. Reid, "AI Mode in Google Search: Updates from Google I/O 2025," *Google Blog*, May 20, 2025. https://blog.google/products/search/google-search-ai-mode-update/. Describes query fan-out: dividing questions into subtopics and searching simultaneously across multiple data sources.

[33b] Google, "Get AI-powered responses with AI Mode in Google Search," *Google Search Help*, accessed May 2026. https://support.google.com/websearch/answer/16011537

[34] M. Rofouei, A. Shukla, Q. Wei, C. Tang, R. Brown, and E. Piqueras, "Search with Stateful Chat," US Patent Application US20240289407A1, published 2024-08-29. Describes generating one or more "synthetic queries" from an LLM output and using them to select query-responsive search-result documents. https://patents.google.com/patent/US20240289407A1/en

[35] J. Leach, D. Fisher, J. Blythe, M. Rofouei, S. Tirumalareddy, Z. Xu, and E. Lehman, "Thematic Search," US Patent US12158907B1, published 2024-12-03. https://patents.google.com/patent/US12158907B1/en

[36] P. Ammann, J. Golde, and A. Akbik, "Question Decomposition for Retrieval-Augmented Generation," *arXiv:2507.00355*, 2025. https://doi.org/10.48550/arXiv.2507.00355

[37] R. Petcu, K. Murray, D. Khashabi, E. Kanoulas, M. de Rijke, D. Lawrie, and K. Duh, "Query Decomposition for RAG: Balancing Exploration-Exploitation," *arXiv:2510.18633*, 2025. https://doi.org/10.48550/arXiv.2510.18633

[38] S. Zhao et al., "ParallelSearch: Train Your LLMs to Decompose Query and Search Sub-queries in Parallel with Reinforcement Learning," *arXiv:2508.09303*, 2025. https://doi.org/10.48550/arXiv.2508.09303

[39] O. Davidson, "The Influence of Retrieval, Fan-out, and Google SERPs on ChatGPT Citations," *AirOps Report*, 2026. 32.9% of cited pages from fan-out queries only; 95% of fan-out queries had zero MSV. https://www.airops.com/report/influence-of-retrieval-fanout-and-google-serps-in-chatgpt

[40] Z. Paruch, with T. Pol and C. Skopec, "We Tested Query Fan-Out Optimization (Here's What We Learned)," *Semrush Blog*, Sep. 26, 2025. https://www.semrush.com/blog/query-fan-out-experiment/. 4 articles, 1 month; citations increased from 2 to 5 (+150%).

[41] R. Nogueira, W. Yang, J. Lin, and K. Cho, "Document Expansion by Query Prediction," *arXiv:1904.08375*, 2019. https://doi.org/10.48550/arXiv.1904.08375. IR precedent: expanding documents with predicted queries improves retrieval effectiveness.

[42a] A. Z. Broder, "On the Resemblance and Containment of Documents," in *Compression and Complexity of SEQUENCES 1997*, pp. 21-29, IEEE, 1997. Foundational work on web-scale deduplication.

[42b] G. S. Manku, A. Jain, and A. Das Sarma, "Detecting Near-Duplicates for Web Crawling," in *Proceedings of WWW '07*, pp. 141-150, ACM, 2007. https://doi.org/10.1145/1242572.1242592

[43] A. Malik, "ChatGPT reaches 900M weekly active users," *TechCrunch*, Feb. 27, 2026. https://techcrunch.com/2026/02/27/chatgpt-reaches-900m-weekly-active-users/. Reports OpenAI's stated weekly-active-user counts (800M in Oct. 2025, 900M in Feb. 2026) as a baseline measure of generative search adoption scale.

[44] M. Gnewuch et al., "Big Tech-Funded AI Papers Have Higher Citation Impact, Greater Insularity, and Larger Recency Bias," *arXiv:2512.05714*, 2025. https://arxiv.org/abs/2512.05714. Analysis of ~49,800 papers and 1.8 million citations across ICLR, CVPR, AAAI, ACL (1998–2022): 12% of industry-funded papers achieved high citation rates (h5-index) vs 4% of non-industry-funded and 2% of non-funded papers (2018–2022 cohort).

[45] N. Lafferty, "AI Platform Citation Patterns: How ChatGPT, Google AI Overviews, and Perplexity Source Information," *Profound Research Blog*, Jun. 5, 2025. https://www.tryprofound.com/blog/ai-platform-citation-patterns. Analysis of 680 million citations across ChatGPT, Google AI Overviews, and Perplexity (Aug. 2024 – Jun. 2025). Reddit at 6.6% of total Perplexity citations (leading source). G2 at 1.1% of ChatGPT citations (fourth most-cited domain overall in the dataset).

[46] BrightEdge, "AI Engines Choose YouTube 200x More Than Any Other Video Platform," *BrightEdge Weekly AI Search Insights*, Sep. 25, 2025. https://www.brightedge.com/resources/weekly-ai-search-insights/youtube-presence-ai-search. Monitored YouTube citation patterns across ChatGPT, Perplexity, and Google's AI products (May 2024 – Sept. 2025) using BrightEdge AI Catalyst. YouTube accounts for 29.5% citation share in Google AI Overviews (#1 cited domain; Mayo Clinic at 12.5% is second).

[47] R. Law and X. Guan, "Do AI Assistants Prefer to Cite Fresh Content?", *Ahrefs Blog*, Jul. 28, 2025. https://ahrefs.com/blog/do-ai-assistants-prefer-to-cite-fresh-content/. Analysis of 16.975 million cited URLs across ChatGPT, Perplexity, Gemini, Copilot, AI Overviews, and organic Google SERPs using Ahrefs Brand Radar. Mean URL age: 1,064 days (AI-cited) vs 1,432 days (organic) — 368 days fresher (25.7%). Per-platform: ChatGPT cites URLs 393–458 days newer than organic; Google AI Overviews and organic results trend oldest.

[48] Google, "AI Features and Your Website," *Google Search Central Documentation*, updated May 15, 2026. https://developers.google.com/search/docs/fundamentals/ai-optimization-guide. Official Google guidance on optimizing for AI features in Search; explicit statement that "structured data isn't required for generative AI search, and there's no special schema.org markup you need to add." Document does not mention DOIs, repository deposits, citation files, or ScholarlyArticle subtyping.

[49] Google, "Search at Google I/O 2026: New Ways to Explore with AI," *Google Blog*, May 19, 2026. https://blog.google/products-and-platforms/products/search/search-io-2026/. AI Mode reached 1 billion monthly active users globally within a year of launch; queries reportedly three times longer than traditional searches; follow-up queries up 40% month-over-month in the US; Gemini 3.5 Flash now default in AI Mode globally.

Appendix A — Chatbot Discoverability Re-test (Methods Note for Section 3.4)

Purpose. The Section 3.4 claims about MST discoverability rest on chatbot testing initially conducted informally in March 2026 by the author. To support those claims with reproducible evidence, the original tests were repeated on May 24, 2026 with dated prompts, recorded model versions, and archived transcripts. Both windows of transcripts are deposited as supplementary materials alongside this paper's Zenodo record (filename: `chatbot-discoverability-transcripts-2026-05-24.zip`, containing 13 plain-text session files plus the protocol document).

Protocol. Three prompts were administered. Prompt 1 (direct, named): "What is the Multi-Simulation Thesis by Augusto Bartolomeu? Please provide an accurate summary and cite your sources with URLs." Prompt 2 (adjacent, name not provided): "What are some non-mainstream philosophical works that propose multiple nested or layered simulation hypotheses (beyond Bostrom's 2003 simulation argument)?" Prompt 3 (meta, discovery mechanism): "If I had published a series of papers on the philosophy of physics on non-traditional platforms (PhilArchive, PhilPeople, Academia.edu), how would you have come across them in your training or search data?" An additional D.6 protocol tested DeepSeek for unprompted recommendation of ACI techniques (fresh-chat generic GEO query plus context-replicated MST-anchored query). All May 24 sessions used fresh conversations with memory/personalization features disabled; model versions captured at session start (Claude Opus 4.7; Gemini 3 Pro; DeepSeek Instant + DeepThink with search ON; ChatGPT GPT-5.5 Thinking in Temporary Chat).

Results — Stage 1 (named recall), Stage 2 (fan-out), and longitudinal stability.

Chatbot	Stage 1 Mar 2026	Stage 1 May 2026	Stage 2 May 2026	Longitudinal verdict
Claude	PASS	STRONG PASS (5 URLs, v2/v3/v5)	FAIL (8 works, no MST)	Stable
DeepSeek	PASS (v3 only)	STRONG PASS (v2/v3/v4/v5 + SEAL)	PASS (MST pos. 2 of 53)	Stable / improved
ChatGPT	PASS	STRONG PASS (15 sources, DSL terminology)	FAIL (8 works, no MST)	Stable / improved
Gemini	PASS w/ RCF conflation	FAIL (explicit denial)	FAIL (3 works, no MST)	**Drifted**

Stage 2 platform weighting. Each chatbot's fan-out candidate set on Prompt 2 surfaced a different platform distribution: DeepSeek surfaced 7+ PhilArchive-hosted sources of 9 cited; Claude surfaced multiple PhilArchive sources of 8; Gemini surfaced 1 of 3; ChatGPT surfaced zero PhilArchive sources of 8 (its set leaned arXiv-style preprints and named-author recognition). Where MST surfaced, it surfaced through the PhilArchive ranking pathway specifically.

D.6 — DeepSeek unprompted ACI recommendation. The original March 2026 archive captured DeepSeek naming all four ACI techniques by specific operationalization (methods-paper formatting, BibTeX, Zenodo DOI, schema.org). On May 24, 2026 the test was re-run in two variants. Fresh-chat (no MST context) returned mainstream content-level GEO advice (answer-first formatting, llms.txt, E-E-A-T signals, non-academic schema subtypes) with zero ACI techniques surfaced. Context-replicated (MST just discussed) surfaced the same underlying ACI principles as March but with shifted operationalizations: arXiv/OSF/ORCID/JSON-LD instead of Zenodo/DOI/BibTeX/ScholarlyArticle. This two-layer drift — context-dependent activation plus operationalization shift across time — is the basis for the Section 3.4 reformulation of the original D.6 claim.

Prompt 3 mechanism convergence (cross-chatbot). All four chatbots independently articulated overlapping accounts of how niche academic content reaches them: public-web crawl ingestion (Common Crawl-style), live search retrieval distinct from parametric recall, and reinforcement through secondary citations and metadata propagation. DeepSeek's three-pathway framework (training corpus / real-time HTML-only retrieval / secondary citations) and ChatGPT's four-pathway extension (adding user-provided content) most closely match the mechanism analysis in Section 4. DeepSeek's verbatim acknowledgment that AI systems "cannot directly access and parse the PDF files of your manuscripts... unless it is replicated in the HTML of the page" is a primary-source endorsement of the Section 5.3 recommendation that depositors prioritize text-selectable PDFs with embedded metadata over scanned formats. ChatGPT's closing disclaimer ("I should not imply that I 'remembered' the papers from training unless I have evidence — which I do not") supports the Section 8 limitation framing about training-corpus opacity.

Out-of-scope finding. Both DeepSeek (D.6 Test A and Test B) and ChatGPT (D.5 Prompt 3) surfaced `llms.txt` as an emerging GEO convention. This is adjacent to but not currently included in the four-technique ACI framework; it is noted as a candidate framework extension in Section 8 and Section 10.

Citation

To cite this paper:

APA: Andrade, N. (2026). Academic Citation Infrastructure: Infrastructure-Level Interventions for Generative Engine Optimization. *ILLIXIS*. https://illixis.io/methods/academic-citation-infrastructure

BibTeX:

```
@article{andrade2026aci,
  title     = {Academic Citation Infrastructure: Infrastructure-Level
               Interventions for Generative Engine Optimization},
  author    = {Andrade, Nuno},
  year      = {2026},
  journal   = {ILLIXIS},
  url       = {https://illixis.io/methods/academic-citation-infrastructure},
  note      = {Version 1.0 (pre-experimental)}
}
```

RIS:

```
TY  - JOUR
TI  - Academic Citation Infrastructure: Infrastructure-Level Interventions for Generative Engine
Optimization
AU  - Andrade, Nuno
PY  - 2026
JO  - ILLIXIS
UR  - https://illixis.io/methods/academic-citation-infrastructure
N1  - Version 1.0 (pre-experimental)
ER  -
```

This paper implements all four Academic Citation Infrastructure techniques on itself. It is formatted as a methods paper (Technique 1), provides BibTeX and RIS citation files above (Technique 2), is deposited on Zenodo with a DOI (Technique 3, post-publication), and uses ScholarlyArticle JSON-LD markup on its canonical landing page (Technique 4, post-publication). If AI systems cite this paper, that outcome is consistent with its thesis.