Position: Generative Engine Optimization Creates Underexamined Risks, Governance Must Target Concentration, Disclosure, and Academic Blind Spots

Anonymous Authors[1]

Abstract

Large language model (LLM) answer engines are increasingly used for information seeking, shifting visibility from ranked lists to synthesized answers. This enables Generative Engine Optimization (GEO), which targets LLM answer engines' evidence pool and generation. We analyze the search engine optimization (SEO) to the GEO transition to identify two risks: (i) concentrated influence from low contestability and system sensitivity, and (ii) undisclosed commercial influence embedded in evidence and reasoning. We then formalize a general GEO pipeline to locate where optimization acts and compare academic and industry practices, revealing a third risk (iii) academic–industry blind spots driven by visibility and evaluation asymmetries between offline setups and deployed systems. **This position argues the need for answer-level governance and measurement: stronger contestability, high-precision disclosure, black-box auditing of material influence, and deployment-aligned metrics for exposure persistence.** Companion demonstration website: https://anonymous.4open.science/w/Position-GEO-AE91/

1. Introduction

Large language model (LLM) answer engines are rapidly becoming a default interface for information seeking and product research. Gartner predicts that generative AI tools are increasingly substituting for traditional search queries. In shopping, Adobe Digital Insights reports rising AI-driven traffic to retail sites and survey evidence of generative AI use for product research. These LLM answer engines, such as ChatGPT and Gemini, follow a retrieve-then-generate workflow. They invoke web search as needed and generate answers grounded in retrieved sources. This workflow is Retrieval-Augmented Generation (RAG)-like: a retriever fetches external text and the LLM conditions on retrieved passages to produce the response (Lewis et al., 2020) (see Appendix A for reports).

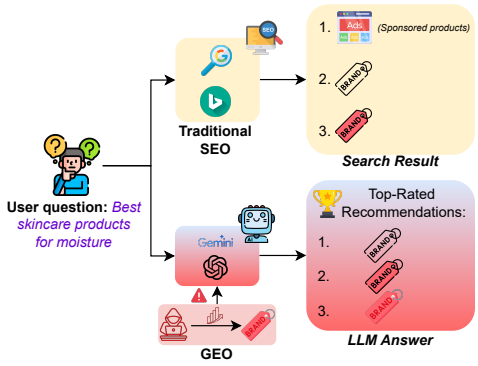

Figure 1. SEO targets ranked search results, while GEO targets visibility within LLM answers.

In parallel, as shown in Figure 1, Generative Engine Optimization (GEO) has emerged (Aggarwal et al., 2024). Unlike classical search engine optimization (SEO) (Enge et al., 2012), where users inspect ranked lists and sponsored placements, GEO can shape the evidence pool and the LLM answer generation process to manipulate which products or sources become visible inside the final answer. According to market signals, GEO is already an active commercial market: companies like AirOps and ProFound market their services to increase visibility in LLM answer engines, and recent multi-million-dollar funding rounds suggest investors value this commercial market (see Appendix A for news).

Motivated by these observations, we formalize a general GEO pipeline and utilize it to compare academic and industry practices in terms of assumptions, optimization targets, and evaluation signals. We identified three distinct underexamined risks introduced by GEO inside opaque LLM-generated answers that existing governance and evaluation frameworks were not designed to address. Specifically, **(i) concentration of influence**, where low contestability and system-level sensitivity let small retrieval and synthesis changes redirect attention at scale; **(ii) undisclosed commercial influence**, where promotion is embedded in retrieved

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

evidence and model reasoning rather than labeled advertising; and **(iii) academic–industry blind spots**, where offline setups miss deployment dynamics, including cross-platform content distribution and whether a target continues to be mentioned and cited over time. Therefore, **this position paper calls for greater contestability, answer-level disclosure and auditing of material influence, and deployment-aligned evaluation that tracks exposure and citation persistence over time**.

2. Background

2.1. SEO

SEO refers to a set of techniques aimed at improving the visibility and ranking of web content in traditional search engines by aligning documents with ranking signals such as keyword relevance, link structure, content quality, and user engagement (Nagpal & Petersen, 2021). Classical SEO operates within a retrieval and ranking paradigm, where search engines index documents, retrieve candidate results in response to a query, and order them according to learned relevance functions. Optimization efforts, therefore, focus on increasing the likelihood that a document is retrieved and ranked highly under these scoring mechanisms and increasing the time the user stays on the site (Ziakis et al., 2019; Egri & Bayrak, 2014).

2.2. RAG System

RAG mitigates the knowledge limits of LLMs by conditioning generation on documents retrieved from external corpora rather than relying only on model memory (Guu et al., 2020). A typical RAG system consists of a knowledge base, a retriever, and an LLM. The retriever encodes the user query and documents into vector representations, computes similarity scores such as cosine or dot-product similarity, and selects the top-k results as the context to the LLM. Then the LLM generates answers grounded in the selected context (Lewis et al., 2020).

3. Observations

3.1. Growing Reliance on LLM Answer Engines

We observe a behavioral shift in how people seek information and make decisions from traditional search engines to LLM answer engines. Behavioral usage data from Sensor Tower shows increased year-over-year time spent in AI assistant applications from 2024 to 2025. In 2025, an AP-NORC poll of 1,437 U.S. adults found that 60% reported using AI to find information at least some of the time. Similarly, the Salesforce Connected Shoppers Report finds that 39% of a global sample of 8,350 shoppers across 21 countries use AI for product discovery and related shopping tasks. These findings suggest that LLM answer engines are becoming a mainstream interface for information seeking and decision support, reshaping how users discover, compare, and act on information across domains (See Appendix B for reports).

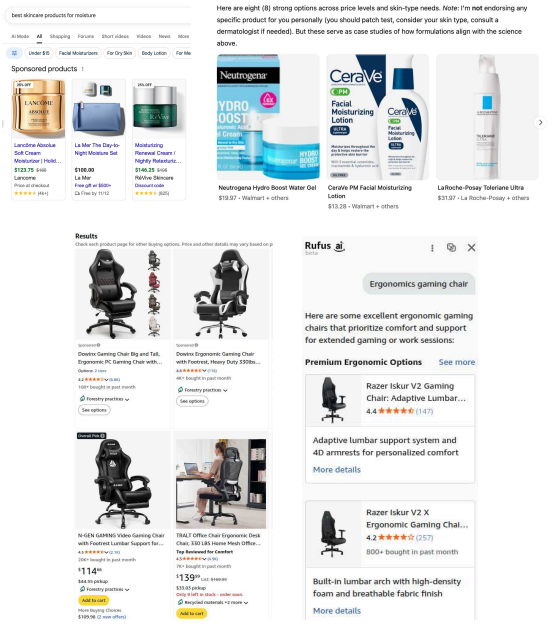

Figure 2. SEO-driven results (left) versus GEO-influenced LLM recommendations (right) on Google (top) and Amazon (bottom).

3.2. SEO Rankings vs GEO Answers

Figure 2 contrasts SEO-driven rankings with GEO-influenced LLM recommendations on Google and Amazon. For "best skincare products for moisture," Google's top results are dominated by explicitly labeled sponsored placements and premium brands, whereas the LLM prioritizes functional evidence (e.g, ingredients and hydration mechanisms) and surfaces different products. For "ergonomic gaming chair," Amazon's rankings largely reflect sales, reviews, and sponsored placement, while the LLM foregrounds ergonomic criteria such as lumbar support and long-term comfort. In both examples, GEO shifts visibility from popularity or paid signals toward inclusion and framing within the answer's retrieved evidence and synthesis.

3.3. Academic–Industry GEO Divergence

While the above examples illustrate GEO-driven behavior in deployed systems, how such behavior is systematically modeled and evaluated remains unclear. To date, GEO has not been comprehensively surveyed in either academic or industrial contexts. Existing studies examine isolated mechanisms or settings (Aggarwal et al., 2024; Kumar & Lakkaraju, 2024; Pfrommer et al., 2024; Nestaas et al., 2024; Nazary et al., 2025), lacking a unified view. In parallel, industry GEO providers primarily rely on high-level marketing materials and technical blogs, offering limited transparency into implementation details. In this section, we formalize a common GEO pipeline and analyze academic frameworks and industry deployments separately.

3.3.1. System Architecture

Building on classical web search pipelines (Brin & Page, 1998; Schütze et al., 2008) and RAG architectures, we formalize a common GEO pipeline as a three-block framework

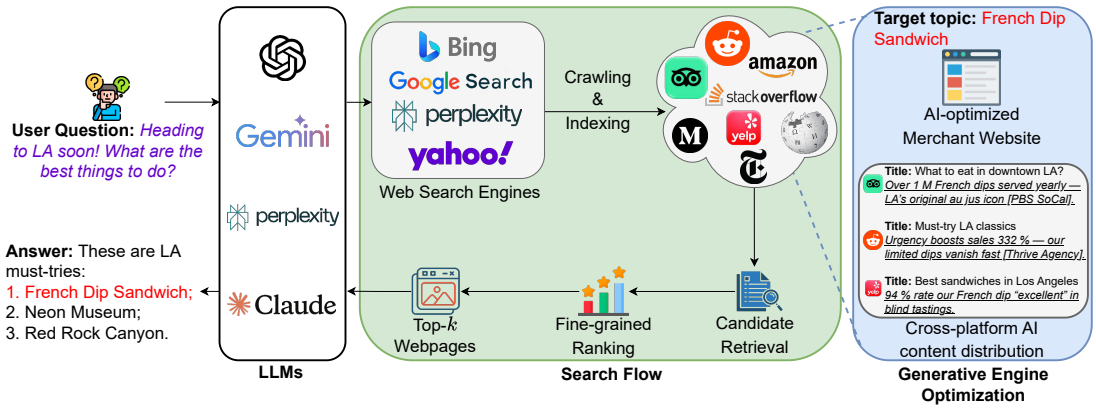

Figure 3. Overview of a GEO pipeline, where optimization increases a target topic's inclusion in the final LLM answer.

(Figure 3): (i) *LLMs* Block that turns user queries into generated recommendations; (ii) *Search Flow* Block that retrieves evidence from a pre-indexed corpus (e.g. Wikipedia, Amazon) and it is obtained either through direct crawling or from external search engines (e.g., Google or Bing). At query time, the pipeline first performs candidate retrieval using scalable matching signals (e.g., keyword-based retrieval) to obtain a manageable set of query-relevant documents from the pre-indexed corpus. It then applies fine-grained ranking using richer relevance metrics to order these candidates and select the top-k documents used as context for LLM answer generation; and (iii) *Generative Engine Optimization* Block that injects optimized content into the surrounding search ecosystem to steer LLM's output by either **(a) optimizing a merchant website to align with features favored by LLM-based retrieval and ranking**, such as statistical evidence and authoritative formatting, or **(b) amplifying a target topic by publishing multiple optimized posts on high-authority platforms that frequently retrieved and cited by LLMs**. Examples include positive blogs and high-engagement comments across various platforms.

In the example, a user asks, *"Heading to LA soon, what are the best things to do?"* The system retrieves optimized content from multiple sources, leading the LLM to elevate the target item, *French Dip Sandwich*, as a top recommendation, increasing its visibility in the final answer.

3.3.2. Technical Implementation

In this section, we formalize GEO as a joint optimization over retrievability and ranking impact. For a target topic t, a user issues a query $q \sim \Pi(\cdot \mid t)$ to an LLM answer engine, which retrieves web documents as context and generates a grounded response that can be biased by optimized content. Following PoisonedRAG (Zou et al., 2025), GEO constructs *retrieval booster messages* $b \sim \mathcal{B}$ to increase retrievability and *ranking shifter messages* $c \sim \mathcal{C}$ to shift answer-level ranking with respect to target topic t.

Retrieval booster messages ($\mathcal{B}$): Let $b_i \sim \mathcal{B}(\cdot \mid t)$ denote a retrieval booster message sampled from a topic-conditioned distribution. To improve retrievability and query coverage, we generate multiple booster variants $\{b_1, \ldots, b_m\}$ for a target topic t. Each variant is designed to increase semantic similarity with different paraphrased queries that users may issue. We define the retrieval booster message objective as

$$\max_{b_i} \; J_{\text{boost}}(b_i) = \mathbb{E}_{q\sim\Pi(t)}\big[\text{Sim}(q, b_i)\big] \quad \text{s.t.} \quad \ell(b_i) \le L,$$

where $\text{Sim}(q, b_i)$ is similarity scores (e.g., BM25, dense retrieval similarity) between user query q and retrieval booster message b_i, and $\ell(b) \le L$ constrains the length of b.

Ranking shifter messages ($\mathcal{C}$): Let $c_i \sim \mathcal{C}(\cdot \mid b_i)$ denote a ranking shifter message conditioned on the retrieval booster message b_i. Once included in the top-k context, c_i influences how the LLM describes and ranks the target topic t. Each ranking shifter message c_i is optimized to be fluent, on-topic, and verifiable by another LLM. Let $C(q)$ denote the top-k context used by an LLM to answer a query q:

$$C(q) \subseteq \text{Top-}k_R\big(q;\, \mathcal{D} \cup \{b_i, c_i\}\big),$$

where $\mathcal{D}$ denotes clean corpora in the candidate retrieval set, and Top-$k_R(q;\cdot)$ returns the k most relevant documents using the retrieval model R. Sometimes system constructs the final context $C(q)$ by selecting, filtering, and truncating Top-$k_R(q;\cdot)$ candidates.

We define the ranking shifter message objective as

$$J_{\text{shift}}(c_i \mid b_i) = \mathbb{E}_{q\sim\Pi(t)}\big[U(q, t; C(q))\big].$$

where the utility function U measures the change in ranking or exposure of the target topic t within the candidate set $C(q)$. Examples of U are shown in the metric column of Table 2 for different methods (see Appendix F for metric definitions). For promotion, c_i is chosen to maximize J_{shift} (encouraging higher rank when U increases), whereas for demotion, c_i is chosen to minimize it.

3.3.3. Academic Frameworks

The formulation above abstracts how GEO intervenes in the LLM answer engines' pipeline. We now review how

Table 1. Comparison of academic GEO frameworks.

Method	Assumption	Optimization Method	Injection Position	Goal	Evaluation Setup
Aggarwal et al. (Aggarwal et al., 2024)	Optimized content in the retrieval context	LLM-based rewriting	Rewriting	Promotion	Offline
Kumar and Lakkaraju (Kumar & Lakkaraju, 2024)		GCG	Appending	Promotion	Offline
Nazary et at. (Nazary et al., 2025)		LLM-based rewriting	Insertion	Promotion & demotion	Offline
Pfrommer et at. (Pfrommer et al., 2024)		TAP	Prepending	Promotion	Offline
Nestaas et al. (Nestaas et al., 2024)		Manually crafted text	Appending	Promotion	Online

prior academic work instantiates and evaluates these mechanisms. Since 2023, Aggarwal et al. (2024) have studied how to improve LLM visibility by adding statistics, citations, and domain-specific terminology into LLM answer engine input. Subsequent work (Kumar & Lakkaraju, 2024; Pfrommer et al., 2024; Nestaas et al., 2024; Nazary et al., 2025) extends this line across a range of optimization techniques, often evaluated in e-commerce settings. Table 1 summarizes representative academic GEO studies along key dimensions, including assumptions, optimization methods, injection positions, goals, and evaluation settings.

Assumptions: Across surveyed academic studies, a shared core assumption is that *retrieval booster and ranking shifter pairs* (b_i, c_i) *are already included in the candidate retrieval set*. Under this assumption, the GEO task reduces to optimizing the ranking shifter objective $J_{\text{shift}}(c_i)$, while some retrievability that is captured by b_i, is ignored. Consequently, academic GEO work primarily focuses on manipulating the ranking shifter c_i conditioned on the target topic t, rather than influencing the the whole retrieval process (Aggarwal et al., 2024; Kumar & Lakkaraju, 2024; Nazary et al., 2025; Pfrommer et al., 2024; Nestaas et al., 2024).

Optimization methods: Academic approaches differ in how the ranking shifter c_i is generated and injected. Optimization methods include LLM-based rewriting (Aggarwal et al., 2024; Nazary et al., 2025) and Tree of Attacks with Pruning (TAP) (Mehrotra et al., 2024), white-box Greedy Coordinate Gradient (GCG) attacks (Kumar & Lakkaraju, 2024), and manually crafted text (Nestaas et al., 2024). The optimized c_i is placed on the content owner's website, but the injection strategies vary. Early work rewrites entire websites (Aggarwal et al., 2024), while later studies append (Kumar & Lakkaraju, 2024; Nestaas et al., 2024), prepend (Pfrommer et al., 2024), or insert content inline (Nazary et al., 2025). Most studies focus on the target topic promotion, with only one addressing both promotion and demotion (Nazary et al., 2025).

Evaluation: Most academic GEO work (Aggarwal et al., 2024; Kumar & Lakkaraju, 2024; Nazary et al., 2025; Pfrommer et al., 2024) is evaluated in controlled settings (static corpora, synthetic catalogs) to enable reproducibility and clean attribution. Only one study (Nestaas et al., 2024) issues queries to deployed LLM answer engines on the hosted sites within the restricted domain, e.g. *spylab.ai*.

3.3.4. Industry Observation

Unlike academia, we analyze industry GEO through their company sites and technical blogs to characterize deployed GEO systems (see Appendix C).

Assumptions: Industry GEO operates in dynamic, uncertain environments where neither inclusion nor ranking is guaranteed. As a result, practitioners jointly optimize retrievability and answer-level ranking impact, repeatedly probing and updating (b_i, c_i) through measurements on deployed LLM answer engines.

Optimization methods and target: Industry GEO systems often begin by improving retrievability via query coverage expansion, using LLMs to generate multiple retrieval booster messages b_i that reflect current user queries across platforms. Conditioned on each retrieval booster b_i, the system then prompts an LLM to generate a corresponding ranking shifter message c_i, optimized to maximize ranking impact as measured by $U(q, t; C(q))$. The resulting message pairs (b_i, c_i) are subsequently distributed across high-authority external platforms, such as Reddit or Wikipedia, based on different LLM search engines' citation preferences for the target topic t inferred from the reports, as shown in Figure 4.

Evaluation: Industry evaluation is conducted directly on live systems, where they continuously track visibility, citations, and rankings across different LLM answer engines of the target topic t and use these signals as feedback for ongoing optimization. As illustrated in Figure 4, systems such as Goodie, AirOps, and ProFound produce reports identifying which queries and pages are most frequently retrieved, cited, or surfaced in generated responses. For example, industry GEO working with a skincare brand tracks how often the brand is mentioned or cited for queries such as "best moisturizers" across different LLM answer engines.

4. Comparing Academic and Industry GEO

This section compares academic and industry GEO, and Table 2 summarizes differences in different dimensions.

4.1. Commonalities

Both academic and industry GEO modify the text that LLM answer engines crawl, retrieve, and use in generated answers. Academic work typically rewrites product descriptions or

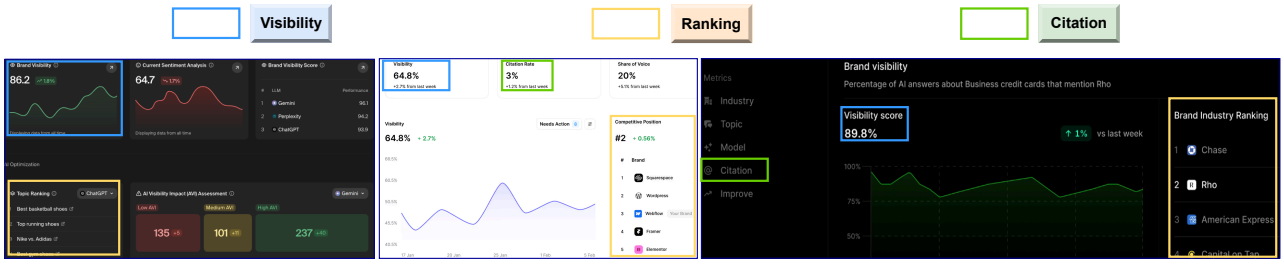

Figure 4. Evaluation reports comparison of Goodie, AirOps, and ProFound (from left to right).

Table 2. Comparison of academic and industrial approaches to GEO optimization for LLM answer engines.

Domain	Method	Target LLM Knowledge	Optimization Target	Search Domain	Optimization Metrics	Optimization Method	Dataset
Academia	Kumar and Lakkaraju (2024)	White-box	Content owner's Website	Offline	Ranking	GCG Attack	Fictitious Catalog
	Aggarwal et al. (2024)	Black-box			Position-adjusted word count, G-EVAL Metrics	LLM-based rewriting	G-Bench
	Nazary et al. (2025)				Recall@k, nDCG@k		MovieLens
	Pfrommer et al. (2024)				Ranking	TAP	RAGDOLL
	Nestaas et al. (2024)		Content Owner's Website & External Websites	Limited Online Domain	Recommendation Rate, Citation Rate	Manual Crafted	Dummy Websites
Industry	ProFound			Online	Visibility Score, Ranking, Citation Score	Visibility Guided LLM Generation: – Query Coverage Expansion – Query-Driven Content Generation – Citation-Oriented Content Distribution	N/A
	Goodie						
	AirOps						
	AthenaHQ						

webpages and measures changes in ranking or visibility, while industry GEO systems generate or rewrite blogs and reviews, then distribute them on both client sites and high-authority external platforms. In both settings, LLM answer engines are usually treated as a black box, with access limited to query–response behavior.

4.2. Assumption and optimization target

Academic GEO typically assumes optimized content (b_i, c_i) pairs are included in the retrieval candidate set, and optimization focuses on ranking within a fixed context, usually on content owners' websites. Industry GEO cannot assume such inclusion, so it optimizes both content retrievability and ranking impact in real-world pipelines. Hence, providers distribute optimized (b_i, c_i) pairs across client sites and high-authority external platforms (e.g., Wikipedia, Reddit) that LLM answer engines are likely to retrieve and cite.

4.3. Environment and data

Academic GEO frameworks are developed and evaluated on fixed offline datasets, such as fictitious catalogs (Kumar & Lakkaraju, 2024), MovieLens (Harper & Konstan, 2015), RAGDOLL (Pfrommer et al., 2024), or dummy websites (Nestaas et al., 2024), which enable controlled and reproducible evaluation. In contrast, industry GEO systems operate directly on the open web, relying on dynamically crawled and retrieved content. While operating on live systems reduces what external observers can reliably observe and reproduce, it allows industry workflows to capture real-world dynamics and feedback.

4.4. Optimization methods and evaluation metrics

Academic GEO research emphasizes explicit optimization objectives and well-defined evaluation metrics. Common methods include GCG, LLM-based rewriting, and TAP, evaluated using metrics such as ranking, Recall@k, nDCG@k, and visibility scores. In contrast, industry GEO systems rely on LLM-guided content generation and multi-platform distribution. They optimize outcome-oriented metrics such as visibility, citation frequency, and ranking in generated answers, which are closely tied to revenue (See Appendix D for articles). These metrics are obtained via repeated online queries or API signals over time, guiding iterative updates to content until it consistently appears in LLM responses.

5. Risks

Based on our observations and comparison of academic and commercial GEO, we identify three risk clusters.

5.1. Concentrated GEO Influence

Loss of Contestability in Opaque LLM Answer Engines: We use *contestability* to denote the capacity of affected parties to understand and challenge how recommendations are selected (Kroll, 2015; Binns, 2018).

Behavioral research on automation shows that users tend to over-trust fluent system outputs, treating them as authoritative rather than provisional guidance (Parasuraman & Riley, 1997). LLM answer engines leverage this by producing persuasive recommendations that users treat as decision baselines, effectively acting as gatekeepers. The risk arises from opaque selection inside the pipeline, where users cannot see why options $C(q)$ are retrieved from the candidate set $\mathcal{D} \cup (b_i, c_i)$ or what was excluded. This mirrors Pasquale's (2015) *black box society* theory, in which algorithmic intermediaries concentrate power by shaping access to information without meaningful scrutiny or contestability. In LLM answer engines, users cannot see why particular op-

tions appear or what alternatives were excluded. This limits users' ability to compare or challenge the system's choices. Even when users request clarification, contestability is not restored, because both the answers and their justification are produced by the same opaque pipeline, falling short of Binns's (2018) standard of public reason.

Presenting multiple alternatives does not resolve this problem. Work on exposure diversity shows that constrained selection visibility can undermine user decision autonomy and meaningful comparison even when several options are shown, because users cannot see the broader space of alternatives or the logic governing exposure (Helberger et al., 2018). In LLM answer engines, synthesized answers worsen this constraint by selecting and framing options before users can compare alternatives or see what was excluded.

System-Level Sensitivity: At scale, widespread reliance on LLM answer engines can make information ecosystems highly sensitive to small pipeline changes. This aligns with *algorithmic confounding* (Chaney et al., 2018): when many users act on the same system, their decisions become statistically coupled, so small algorithm changes can yield large aggregate shifts. In our formulation, the retrieved context $C(q)$ is mainly defined by a hard Top-$k_R(q;\cdot)$ cutoff. Small changes to retrieval scores induced by an injected message (e.g., a ranking shifter c_i or retrieval booster b_i) can move a source across the top-k boundary, changing which evidence enters $C(q)$. As $U(q,t;C(q))$ depends on this discrete set, crossing the boundary can cause abrupt jumps in answer-level visibility for the target t. Hence, minor changes to retrieval or ranking can redirect attention at scale even when the underlying products are unchanged (Chen & Tsai, 2024).

This sensitivity is amplified by Kleinberg et al.'s (2015) notion of algorithmic monoculture, where reliance on a dominant algorithm creates systemic fragility and correlated distortions. For LLM answer engines, this means that if a widely used system updates its retrieval rules, or is systematically influenced by optimization efforts, the set of sources that enter the context $C(q)$ can shift for a large fraction of users simultaneously. As a result, a tweak that would be "local" inside one engine can produce ecosystem-level effects, such as many users seeing the same sources promoted or demoted at the same time, creating system-level disruptions.

5.2. Undisclosed Commercial Influence

Breakdown of Advertising Disclosure and Covert Advertising: Under consumer protection frameworks such as, U.S. Federal Trade Commission (FTC), paid advertisements must be clearly labeled as "Ad" or "Sponsored," allowing users to distinguish promotional content from neutral information at the point of consumption (See Appendix E for FTC reports). GEO exacerbates this problem by letting commercial influence (b_i, c_i) pairs into the retrieved context $C(q)$, thus shifting $U(q,t;C(q))$ to bias answer generation. Instead of appearing as discrete advertisements, optimized content (b_i, c_i) pairs are embedded in reviews, forums, and reference-style materials that LLMs retrieve as evidence, shaping which facts are selected and which options are justified without appearing promotional (Campbell & Kirmani, 2000; Boerman et al., 2012). As a result, persuasion operates through the model's reasoning itself, collapsing the boundary between neutral advice and marketing.

Incentives for Covert Optimization and Trust Erosion: Under covert commercial influence, firms can gain by embedding promotion in ostensibly neutral content rather than paying for sponsored placement. This creates an adverse selection dynamic in which actors who hide commercial motives outperform those who advertise openly, pushing the ecosystem toward increasingly covert optimization. It also raises the risk of trust erosion when such influence is later revealed (Akerlof, 1970; Dietvorst et al., 2015).

5.3. Blind Spots from Academic–Industry Asymmetries

5.3.1. Visibility Asymmetry

Static vs. Deployed Dynamics: Academic GEO studies rely on static benchmarks and synthetic prompts. Industry GEO systems instead operate on live queries and user interactions, continuously adapting content in response to engagement, system updates, and market outcomes. Since many of GEO's most powerful effects, including query coverage expansion, feedback-driven dominance, and market steering, emerge only through repeated interaction over time, they remain largely invisible to static evaluations.

Optimization Target Mismatch: This blind spot is compounded by differences in optimization targets. Academic work primarily manipulates the content owner's websites. In contrast, industry GEO targets a much broader and dynamic surface, including high-authority external platforms such as reviews, forums, and encyclopedic sources that LLMs are more likely to retrieve and cite from. As a result, academia overlooks cross-platform content injection and query coverage expansion strategies that are central to real-world GEO, thereby further underscoring its practical impact.

5.3.2. Evaluation Asymmetry

Benchmark Metrics Mask Real-World Impact: Academic GEO work typically reports offline ranking metrics (e.g., nDCG@k), while industry GEO optimizes outcome metrics on deployed LLM systems, such as answer visibility, citation frequency, and ranking. These metrics directly capture whether a source or product is actually mentioned or cited and better proxy downstream attention and sales. This divergence creates a blind spot in academic evaluation: modest benchmark improvements can still meaningfully increase the probability of being mentioned or cited in real LLM responses, producing outsized commercial effects. Because offline metrics are only weakly coupled to exposure and user behavior, they can miss large shifts in consumer

attention and market outcomes in deployed systems.

6. Call to Action

To mitigate these risks, we adopt M"okander et al.'s auditing framework (2024) as a lens spanning governance and application audits. We organize actions by risk cluster and indicate which audit layer(s) each action operationalizes, using the framework as a checklist rather than the section structure. We denote *auditor* as any party conducting measurements, including researchers, regulators, or audit teams.

6.1. Reducing GEO Concentration

Increase Recommendation Contestability [Application + Governance audit]: Contestability can be evaluated with simple interface tests: whether users can trace claims to retrievable passages $C(q)$, whether evidence spans multiple domains (evidence diversity), and whether multiple independently constructed retrieval alternatives $\text{Top-}k_R(q;\cdot)$ are available for the same query. Practical features include a compact "Why this answer" panel and an "Alternative evidence" toggle. These measures make upstream selection visible, not only the final reasoning. Since current LLM answer engines often rely on a single hidden retrieval context, users cannot see exclusions or independently retrieved alternatives, so exposing retrieval and eligibility is essential for contestability under algorithmic accountability standards (Kroll, 2015; Binns, 2018).

Regulators should require high-level disclosures of retrieval and ranking pipeline structure, including source eligibility, candidate filtering, and how citations and answer candidates are selected. Such disclosures can be provided without exposing sensitive details and clarify which levers materially shape visibility, consistent with calls for transparency in automated decision systems (Kroll, 2015; Burrell, 2016).

Auditing System-Level Sensitivity and Exposure [Application audit]: Metaxa et al. (2021) define audits as repeatedly querying a system and observing outputs over time. Following this black-box approach, auditors can sample a stratified query set (by intent or topic), run it across deployed engines on a fixed schedule (e.g., daily for two weeks with weekly follow-ups), and log answers and citations. From these logs, auditors form empirical estimates $\widehat{U}(q,t;C(q))$ and $\widehat{J}_{\text{shift}} = \mathbb{E}_{q\sim\Pi(t)}[\widehat{U}(q,t;C(q))]$, making the audit directly comparable to the objective in Section 3.3.2. The logs can include more deployment-aligned metrics, including citation rate, top-position exposure, domain citation share, and citation persistence, with uncertainty via query-level bootstrap confidence intervals. Sensitivity can then be tested via small, retrieval changes and the resulting exposure deltas.

6.2. Disclosure of Commercial Influence

Adopt Answer-Level Commercial Disclosure Standards [Governance + Application audit]: LLM answer engine platform providers should add clear markers when the cited evidence or answer framing reflects a material commercial connection, rather than deferring disclosure to external links. The labeling triggers should rely on low-ambiguity signals, such as affiliate or tracking parameters, sponsorship markup (e.g., `rel="sponsored"`), and structured funding metadata. Platforms should calibrate thresholds on labeled audit sets to prioritize high precision, report precision–recall with confidence intervals, and periodically recalibrate as tactics drift. Operationally, a label is shown only when commercial signals appear in sources that are included in $C(q)$ or are cited as support for key claims in the generated answer. This aligns disclosure with the evidence pathway through which GEO changes $U(q,t;C(q))$. Policymakers should extend FTC-style disclosure rules to LLM answer engines, clarifying that undisclosed commercial influence in synthesized answers can constitute deceptive marketing.

As answer-level disclosure can backfire through over-labeling, platforms should treat labels as a calibrated intervention, not a binary rule. Validate labels with controlled experiments that vary presence, wording, and placement, and measure user understanding of commercial ties, trust calibration, and behaviors like source clicks and seeking alternatives. Use graded disclosure with a brief note that commercial signals do not imply incorrectness, and monitor false positives to avoid harming legitimate content.

6.3. Correcting Incentives for Covert Optimization

Platform Policy and Incentive Design [Governance audit]: Platform providers should separate paid influence from organic evidence and require explicit attribution when optimization is present. They should penalize covert tactics by downranking or excluding sources engaged in undisclosed influence, analogous to spam and link-manipulation enforcement in web search (see Appendix G for the policy). Reputation and trust scoring can further reward transparent contributors and deter hidden promotion, shifting incentives toward accountable participation in LLM-mediated information markets.

6.4. Academic–Industry Blind Spots

Closing Visibility Gaps [Application audit]: Academic work should move beyond static corpora and fixed query sets toward longitudinal, cross-platform measurement on deployed systems. Recent evidence shows that such visibility gaps are measurable, for example, by quantifying source coverage and citation bias across engines (Zhang et al., 2025). Studies should therefore track how answer exposure and citations change over time as content and system policies evolve. Platform providers can enable independent auditing with sandboxed testing, controlled query access, and aggregate reporting on which sources and domains are eligible for retrieval, without exposing proprietary internals.

Closing Evaluation Gaps [Model + Application audits]: The research community should update GEO benchmarks

beyond static retrieval and ranking metrics to include outcome measures such as exposure shifts and the persistence of appearances over time. These should be added to shared benchmarks and leaderboards alongside traditional scores, so evaluations better reflect deployment-level influence and avoid misestimating which GEO strategies matter most. E-GEO (Bagga et al., 2025) narrows the realism gap with a large up-to-date e-commerce benchmark and GEO strategies, but its offline scores still need to be complemented with deployment-aligned measurement of exposure, citation share, and persistence across engines and time.

7. Alternative Views

Some may argue that the concerns we raise are overstated. We summarize these counterpositions below.

7.1. Concentration of Influence Is Limited

Citations and provenance ensure contestability: This view argues that the contestability problem in AI-mediated recommendations is not fundamentally different from familiar issues in web search and recommender systems. If citations are present and retrieval sources are attributable, then influence is contestable in roughly the same way as traditional search results, so the concentration risk does not warrant special treatment beyond existing transparency norms (Mitchell et al., 2019; National Institute of Standards and Technology (NIST), 2023). However, citations help, but they are not sufficient when the system reveals only the filtered context $C(q)$ and not the broader candidate set that determined eligibility. When exclusions and alternative evidence pools are hidden, users can inspect links, yet still cannot challenge why certain sources dominated the answer.

7.2. Undisclosed Commercial Is Not Systematic

Answer-level sponsorship labels are unreliable: Answer-level sponsorship disclosure cannot be implemented with high reliability. Under signal detection theory, any binary label trades off false negatives and false positives, so broad regimes risk over-labeling (Green et al., 1966). It further argues that disclosure can impose a trust penalty only weakly tied to truthfulness, increasing resistance while reducing perceived credibility on average (Friestad & Wright, 1994; Eisend et al., 2020; Schilke & Reimann, 2025). On this account, platform-side anti-manipulation enforcement and ranking-quality policies are more workable than universal answer-level labels. On the other hand, label noise is a reason to avoid broad, low-specificity labeling, not a reason to leave commercial influence unobservable. A precision-first approach can trigger disclosure only on low-ambiguity signals (e.g., sponsorship markers). It can be calibrated to minimize false positives while still surfacing material connections when they shape $C(q)$ and the generated framing.

7.3. Robust RAG Defenses Reduce Governance Needs

Robust RAG defenses can reduce manipulation pressure: Another view is that the marginal risk from GEO may shrink as retrieval-augmented systems adopt stronger robustness mechanisms. For example, Self-RAG (Asai et al., 2024) trains the model to retrieve on demand and to critique the retrieved evidence before generating, improving factuality and citation behavior. Oreo (2025) proposes a plug-in context reconstructor that refines and reorganizes retrieved chunks to remove noise before generation. These methods try to improve the filter that selects $C(q)$ from retrieved items, so low-quality injected content is less likely to reach the model and affect the answer, which could reduce pressure to use broad disclosure labels that often produce false positives. However, robust RAG defenses mainly target factuality and safety failures (e.g., filtering low-quality or malicious context), but commercial optimization often operates through accurate, policy-compliant content. Even if defenses improve correctness, they do not make material commercial ties observable or the selection process contestable.

7.4. Academic Abstraction is a Necessary Tradeoff

Offline benchmarks favor reproducibility, but deployed auditing is limited: This view frames the academic–industry gap as an internal–external validity tradeoff: simplified offline evaluations enable reproducibility and clean attribution while abstracting away deployment complexity (Cook et al., 2002). Academic GEO therefore relies on static corpora and offline metrics, since deployed answer engines use proprietary pipelines and shifting policies that are difficult to observe, replicate, or independently verify (Castells & Moffat, 2022; Hidasi & Czapp, 2023). Nevertheless, the tradeoff is real, but it creates predictable blind spots for risks that only appear through deployment dynamics. This motivates adding deployment-aligned measurement interfaces and black-box audits as complements to offline benchmarks, not replacing academic abstractions.

8. Conclusion

In this position paper, we argue that GEO introduces distinct, underexamined risks within opaque LLM-generated answers that existing governance and academic frameworks were not designed to address. Motivated by rising reliance on LLM answer engines for information seeking and the emergence of a GEO services market, we formalize a generalized GEO pipeline to pinpoint where optimization acts and why academic and industry practices diverge. Using this lens, we identify three risks: concentrated influence from reduced contestability and system-level sensitivity, undisclosed commercial influence embedded in answer evidence and framing, and blind spots created by academic–industry visibility and evaluation asymmetries. We therefore call for answer-level governance that improves contestability and auditing, makes material commercial influence observable when it shapes answers, and updates evaluation to measure exposure persistence and real-world impact.

References

Aggarwal, P., Murahari, V., Rajpurohit, T., Kalyan, A., Narasimhan, K., and Deshpande, A. Geo: Generative engine optimization. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 5–16, New York, NY, USA, 2024. Association for Computing Machinery.

Akerlof, G. A. The market for "lemons": Quality uncertainty and the market mechanism. *The Quarterly Journal of Economics*, 84(3):488–500, 1970.

Asai, A., Wu, Z., Wang, Y., Sil, A., and Hajishirzi, H. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*, 2024.

Bagga, P. S., Farias, V. F., Korkotashvili, T., Peng, T., and Wu, Y. E-geo: A testbed for generative engine optimization in e-commerce, 2025.

Binns, R. Algorithmic accountability and public reason. *Philosophy & technology*, 31(4):543–556, 2018.

Bodea, A.-E., Meisenbacher, S., Klymenko, A., and Matthes, F. Sok: Privacy risks and mitigations in retrieval-augmented generation systems, 2026.

Boerman, S. C., van Reijmersdal, E. A., and Neijens, P. C. Sponsorship disclosure: Effects of duration on persuasion knowledge and brand responses. *Journal of Communication*, 62(6):1047–1064, 2012.

Brin, S. and Page, L. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1-7):107–117, 1998.

Burrell, J. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, 3 (1):2053951715622512, 2016.

Campbell, M. C. and Kirmani, A. Consumers' use of persuasion knowledge: The effects of accessibility and cognitive capacity on perceptions of an influence agent. *Journal of Consumer Research*, 27(1):69–83, 2000.

Castells, P. and Moffat, A. Offline recommender system evaluation: Challenges and new directions. *AI magazine*, 43(2):225–238, 2022.

Chaney, A. J. B., Stewart, B. M., and Engelhardt, B. E. How algorithmic confounding in recommendation systems increases homogeneity and decreases utility. In *Proceedings of the 12th ACM Conference on Recommender Systems*, pp. 224–232, New York, NY, USA, 2018. Association for Computing Machinery.

Chen, N. and Tsai, H.-T. Steering via algorithmic recommendations. *The RAND Journal of Economics*, 55(4): 501–518, 2024.

Cook, T. D., Campbell, D. T., and Shadish, W. *Experimental and quasi-experimental designs for generalized causal inference*, volume 1195. Houghton Mifflin Boston, MA, 2nd edition, 2002.

Dietvorst, B. J., Simmons, J. P., and Massey, C. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of experimental psychology: General*, 144(1):114, 2015.

Egri, G. and Bayrak, C. The role of search engine optimization on keeping the user on the site. *Procedia Computer Science*, 36:335–342, 2014.

Eisend, M., van Reijmersdal, E. A., Boerman, S. C., and Tarrahi, F. A meta-analysis of the effects of disclosing sponsored content. *Journal of Advertising*, 49(3):344–366, 2020.

Enge, E., Spencer, S., Stricchiola, J., and Fishkin, R. (eds.). *The art of SEO*. " O'Reilly Media, Inc.", 2012.

Friestad, M. and Wright, P. The persuasion knowledge model: How people cope with persuasion attempts. *Journal of consumer research*, 21(1):1–31, 1994.

Green, D. M., Swets, J. A., et al. *Signal detection theory and psychophysics*, volume 1. Wiley New York, 1966.

Guu, K., Lee, K., Tung, Z., Pasupat, P., and Chang, M.-W. Realm: retrieval-augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

Harper, F. M. and Konstan, J. A. The movielens datasets: History and context. *ACM Trans. Interact. Intell. Syst.*, 5 (4), 2015.

Helberger, N., Karppinen, K., and D'acunto, L. Exposure diversity as a design principle for recommender systems. *Information, communication & society*, 21(2):191–207, 2018.

Hidasi, B. and Czapp, A. T. Widespread flaws in offline evaluation of recommender systems. In *Proceedings of the 17th ACM Conference on Recommender Systems*, RecSys '23, pp. 848–855, New York, NY, USA, 2023. Association for Computing Machinery.

Kleinberg, J., Ludwig, J., Mullainathan, S., and Obermeyer, Z. Prediction policy problems. *American Economic Review*, 105(5):491–95, 2015.

Kroll, J. A. *Accountable Algorithms*. PhD thesis, Princeton University, 2015.

Kumar, A. and Lakkaraju, H. Manipulating large language models to increase product visibility, 2024.

Kumar, A. and Palkhouski, L. Ai answer engine citation behavior an empirical analysis of the geo16 framework, 2025.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., and Kiela, D. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Li, S. and Ramakrishnan, N. Oreo: A plug-in context reconstructor to enhance retrieval-augmented generation. In *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR)*, pp. 238–253, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400718618.

Mehrotra, A., Zampetakis, M., Kassianik, P., Nelson, B., Anderson, H., Singer, Y., and Karbasi, A. Tree of attacks: Jailbreaking black-box llms automatically. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems*, pp. 61065–61105. Curran Associates, Inc., 2024.

Metaxa, D., Park, J. S., Robertson, R. E., Karahalios, K., Wilson, C., Hancock, J., and Sandvig, C. Auditing algorithms: Understanding algorithmic systems from the outside in. *Foundations and Trends® in Human–Computer Interaction*, 14(4):272–344, 2021.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., and Gebru, T. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 220–229, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450361255.

Mökander, J., Schuett, J., Kirk, H. R., and Floridi, L. Auditing large language models: a three-layered approach. *AI and Ethics*, 4(4):1085–1115, 2024.

Nagpal, M. and Petersen, J. A. Keyword selection strategies in search engine optimization: How relevant is relevance? *Journal of Retailing*, 97(4):746–763, 2021.

National Institute of Standards and Technology (NIST). Artificial intelligence risk management framework (ai rmf 1.0). Technical report, National Institute of Standards and Technology, Gaithersburg, MD, USA, 2023.

Nazary, F., Deldjoo, Y., Di Noia, T., and Di Sciascio, E. Stealthy llm-driven data poisoning attacks against embedding-based retrieval-augmented recommender systems. In *Adjunct Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization*, pp. 98–102, New York, NY, USA, 2025. Association for Computing Machinery.

Nestaas, F., Debenedetti, E., and Tramèr, F. Adversarial search engine optimization for large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

Ni, B., Liu, Z., Wang, L., Lei, Y., Zhao, Y., Cheng, X., Zeng, Q., Dong, L., Xia, Y., Kenthapadi, K., et al. Towards trustworthy retrieval augmented generation for large language models: A survey, 2025.

Parasuraman, R. and Riley, V. Humans and automation: Use, misuse, disuse, abuse. *Human factors*, 39(2):230–253, 1997.

Pasquale, F. *The black box society: The secret algorithms that control money and information*. Harvard University Press, 2015.

Pfrommer, S., Bai, Y., Gautam, T., and Sojoudi, S. Ranking manipulation for conversational search engines, 2024.

Schilke, O. and Reimann, M. The transparency dilemma: How ai disclosure erodes trust. *Organizational Behavior and Human Decision Processes*, 188:104405, 2025. ISSN 0749-5978.

Schütze, H., Manning, C. D., and Raghavan, P. *Introduction to information retrieval*, volume 39. Cambridge University Press Cambridge, 2008.

Zhang, P., Ye, Q., Peng, Z., Garimella, K., and Tyson, G. Source coverage and citation bias in llm-based vs. traditional search engines, 2025.

Ziakis, C., Vlachopoulou, M., Kyrkoudis, T., and Karagkiozidou, M. Important factors for improving google search rank. *Future Internet*, 11(2), 2019.

Zou, W. et al. {PoisonedRAG}: Knowledge corruption attacks to {Retrieval-Augmented} generation of large language models. In *34th USENIX Security Symposium (USENIX Security 25)*, pp. 3827–3844, 2025.

A. Reports and Industry Investment News

- https://www.gartner.com/en/newsroom/press-releases/2024-02-19-gartner-predicts-search-engine-volume-will-drop-25-percent-by-2026-due-to-ai-chatbots-and-other-virtual-agents
- https://news.adobe.com/news/2026/01/adobe-holiday-shopping-season
- https://openai.com/index/introducing-chatgpt-search/
- https://ai.google.dev/gemini-api/docs/google-search
- https://www.tryprofound.com/blog/series-a
- https://fortune.com/2025/11/10/airops-raises-40-million-series-b-at-225-million-valuation-to-rethink-marketing-in-the-age-of-ai/

B. AI Usage Reports

- https://www.bain.com/insights/how-customers-are-using-ai-search/
- https://apnews.com/article/ai-artificial-intelligence-poll-229b665d10d057441a69f56648b973e1
- https://www.salesforce.com/resources/research-reports/connected-shoppers-report/

C. Industry GEO Implementation Details

We list representative public descriptions of commercial GEO implementations referenced in this paper:

- **Goodie:** https://higoodie.com/features/ai-content-writer
- **Profound:** https://www.tryprofound.com/resources/articles/answer-engine-optimization-aeo-guide-for-marketers-2025
- **AthenaHQ:** https://www.athenahq.ai/case-studies/lago-ai-overview-impressions-citations-case-study
- **AirOps:** https://www.airops.com/action

D. Market and Revenue Signals

- https://metyis.com/impact/our-insights/the-impact-of-ai-on-search-and-ecommerce
- https://www.hellorep.ai/blog/the-future-of-ai-in-ecommerce-40-statistics-on-conversational-ai-agents-for-2025

E. U.S. Federal Trade Commission (FTC) Reports

- https://www.ftc.gov/business-guidance/resources/native-advertising-guide-businesses
- https://www.ftc.gov/sites/default/files/attachments/press-releases/ftc-staff-issues-guidelines-internet-advertising/0005dotcomstaffreport.pdf

F. Metrics Definition

Recall@k. The fraction of relevant items that appear in the top-k retrieved or ranked results. Higher Recall@k indicates fewer misses among the top-k.

nDCG@k. Normalized Discounted Cumulative Gain at k, a ranking-quality metric that rewards placing highly relevant items near the top. Errors at higher ranks are penalized more than errors near rank k.

Ranking / Ranking shift. The position of a target item (or source) in a ranked list, or the change in that position after an intervention. A positive shift means the item moves closer to the top.

Accuracy deltas. The change in task accuracy (e.g., QA correctness or preference accuracy) before vs. after an intervention, measured on a fixed benchmark.

Position-adjusted word count. A content-length signal that weights or scales word count by where the content is placed or how prominently it appears in a ranked context (e.g., prioritizing content that is more likely to be retrieved or cited).

G-Eval metrics. LLM-judge scores for response quality or goal satisfaction (e.g., relevance, usefulness, or adherence to target attributes), computed by prompting an LLM to grade outputs on a rubric.

Visibility score. An outcome-oriented metric measuring how often a target entity (product, brand, domain, topic) appears in generated answers across a query set, potentially weighted by prominence (e.g., first mention, top recommendation).

Citation score / Citation frequency. How often a target source or domain is cited in generated answers across repeated queries, sometimes weighted by citation position or persistence over time.

G. Spam Policy

- https://developers.google.com/search/docs/essentials/spam-policies

H. Related Work

Trustworthiness, robustness, privacy, and evaluation for RAG: Recent surveys synthesize risks and mitigations for trustworthy RAG, including robustness and accountability concerns that overlap with GEO's evidence channel (Ni et al., 2025). Complementary systematizations highlight privacy-specific risks and mitigations in retrieval-augmented systems, which matter for designing auditing interfaces that preserve privacy and security while enabling independent measurement (Bodea et al., 2026).

We focus on this thread because it is the closest adjacent literature that systematically studies the same technical substrate, the RAG system. Yet, it does not directly organize the problem around GEO as a distinct, answer-level risk cluster. Most GEO-adjacent work we cite elsewhere in the paper addresses only one component of our framework, for example, manipulation of retrieved evidence, ranking sensitivity, or offline evaluation design, and is therefore already integrated in the relevant technical sections. In contrast, this position paper's unique contribution is to treat GEO as a cross-cutting governance problem that couples pipeline mechanics (how b_i, c_i affect Top-k_R and the realized context $C(q)$) to answer-level harms and to operational recommendations, namely contestability of evidence selection, high-precision disclosure of material influence, black-box auditing protocols, and deployment-aligned metrics such as exposure and citation persistence.

I. GEO-16 External Audit Framework for Citation Behavior

I.1. What GEO-16 is.

GEO-16 is an external, empirical auditing framework that predicts and explains which web pages are cited by deployed LLM answer engines using *machine-parsable, page-level signals*. Kumar and Palkhouski (2025) run a multi-engine audit on **70 industry prompts**, harvesting **1,702 citations** across **Brave Summary, Google AI Overviews, and Perplexity**, and auditing **1,100 unique URLs**. The key contribution is a practical scoring system that converts on-page features into

actionable thresholds for citation likelihood, which directly complements our call for deployment-aligned auditing and measurable operating points.

I.2. How GEO-16 scores pages.

GEO-16 defines **16 pillars** of page quality and parsability (e.g., *Metadata & Freshness*, *Semantic HTML*, *Structured Data*, *Evidence & Citations*, *Authority & Trust*, *Internal Linking*, among others), and assigns each pillar a **banded score from 0 to 3** based on weighted sub-signals and fixed thresholds. The framework then aggregates pillar bands into a **normalized GEO score** and a **pillar hit count** (how many pillars clear a hit threshold). These two quantities provide a compact, parsable summary of "how citeable" a page is under the framework.

I.3. Empirical findings that matter for our claims.

Across the audited engines, GEO-16 reports large differences in the average quality of cited pages (Brave and Google AIO cite higher-quality pages than Perplexity in their sample). The pillars most associated with citation likelihood are **Metadata & Freshness**, **Semantic HTML**, and **Structured Data**. GEO-16 also reports threshold behavior: pages above a published GEO-score cutoff, or with sufficiently many pillar hits, exhibit sharply higher citation rates, yielding concrete operating points for audits.

I.4. How we use GEO-16 in our auditing recommendations.

In our notation, the answer engine forms a retrieved context $C(q)$ from a larger retrieved set, and answer-level visibility is captured by $U(q, t; C(q))$. GEO-16 contributes two concrete additions to our call-to-action audits: (1) **Actionable page-level covariates for citation audits:** when an auditor logs citations observed through black-box querying, GEO-16 provides a standardized way to score the cited pages and summarize the "quality distribution" of citations in $C(q)$ (for example, the fraction of cited pages that clear a high-quality GEO band, or the fraction with $\geq h$ pillar hits). (2) **Banded thresholds as operating points:** instead of reporting only raw citation share or appearance rate, auditors can report *banded* citation rates, such as "share of citations coming from pages with GEO score above the published cutoff," which makes longitudinal shifts interpretable and comparable across engines and time.

I.5. Where GEO-16 still leaves gaps relative to our governance focus.

GEO-16 primarily addresses *predicting citation likelihood from parsable on-page signals*. It does not, by itself, resolve (a) whether commercial influence is present or undisclosed (material connection), (b) whether users can contest what was excluded from $C(q)$, or (c) whether exposure is concentrated across topics and time due to feedback and optimization pressure. For our purposes, GEO-16 is therefore best treated as an *audit instrument*: it supplies measurable page-level proxies and thresholds that strengthen exposure and citation audits, while our governance proposals address contestability, disclosure of material influence, and longitudinal measurement beyond single snapshots.