Generative Engine Optimization: Concepts, Techniques, Benchmarks, and Open Problems

immediate

Abstract

The rapid adoption of large language model (LLM)-based generative search engines—including ChatGPT with browsing, Perplexity, Google AI Overviews, and Gemini—has fundamentally altered the information access landscape. Where traditional search returns a ranked list of links, generative engines synthesize a single narrative answer with inline citations, replacing ranking position with *inclusion* as the primary unit of visibility. This shift has given rise to *Generative Engine Optimization* (GEO): the study and practice of making content, evidence, products, webpages, and agents visible to, cited by, or absorbed into generative answers.

This paper presents a Systematization of Knowledge (SoK) for GEO, organizing a rapidly expanding but fragmented literature that spans information retrieval, natural language processing, security, and human–computer interaction. We make five contributions. First, we situate GEO within the broader landscape of AI-mediated information access, clarifying its relationship to traditional SEO, retrieval-augmented generation (RAG), and conversational and agentic search. Second, we propose a taxonomy of GEO optimization methods organized by the object of optimization—content structure, evidence and citation, query intent and role, multi-objective formulations, agentic strategy, and product visibility. Third, we consolidate the evaluation landscape, proposing a unified six-dimension metric framework (visibility, citation, absorption, ranking, robustness, and security), reviewing nine benchmarks and testbeds, and reporting a systematic reproducibility audit of 20 open-source GEO projects with unified benchmark results. Fourth, we develop a layered taxonomy of adversarial GEO—attacks targeting content, retrieval, citation, generation, ranking, and agentic pipeline layers—and analyze the benign–adversarial boundary that makes GEO dual-use. Fifth, we survey defenses and governance mechanisms and delineate the ethical boundaries between cooperative optimization and manipulation. Our audit finds that while evaluation-stage pipelines are broadly reproducible, generation-stage reproducibility is gated by GPU access and proprietary API availability, and that offline benchmark results diverge systematically from live-engine audits. We conclude by identifying nine open problems, including the need for a unified evaluation protocol, robust generalizable defenses, and a formal definition of the benign–adversarial boundary.

1 Introduction

The way people find information on the web is undergoing its most significant transformation since the advent of the modern search engine. For more than two decades, the dominant interface for information access has been the ranked list of hyperlinks pioneered by PageRank-era search engines (Brin and Page, 1998), and an entire discipline—Search Engine Optimization (SEO)—emerged to help content creators compete for visibility within that list. The rapid adoption of large language model (LLM) based search systems, including ChatGPT with browsing, Perplexity, Google's AI Overviews, Gemini, and Claude, has disrupted this equilibrium. These *generative engines* no longer return a list of links for the user to inspect; instead, they synthesize a single narrative answer, weaving together content drawn from multiple sources and embedding citations inline (Aggarwal et al., 2024). In this new regime, the unit of visibility is no longer a ranking position but inclusion—as a cited, paraphrased, or absorbed source—within a generated response.

This shift has given rise to *Generative Engine Optimization* (GEO): the study and practice of making content, evidence, products, webpages, and agents visible to, cited by, or absorbed into the answers produced by generative engines. Since the term was introduced by Aggarwal et al. (2024), the area has expanded with remarkable speed, encompassing empirical characterizations of how generative engines source information (Kirsten et al., 2026; Chen et al., 2026a; Grossman et al., 2026), content-optimization methods spanning content structure, evidence selection, query intent, and multi-objective

formulations (Aggarwal et al., 2024; Yu et al., 2026b; Chen et al., 2025d; Liu and Xu, 2026), agentic and self-evolving optimization systems (Yuan et al., 2026; Wu et al., 2026; Ye et al., 2026), dedicated benchmarks and measurement frameworks (Puerto et al., 2025; Bagga et al., 2025; Zhang et al., 2026b; Schulte et al., 2026), and a growing body of work on the security risks of adversarial GEO (Kumar and Lakkaraju, 2024; Nestaas et al., 2024; Tang et al., 2025; Jin et al., 2026).

This breadth, however, has come at the cost of fragmentation. Methods, metrics, benchmarks, and threat models are scattered across communities—information retrieval, natural language processing, security, and human–computer interaction—and across the boundary between academic preprints and proprietary industrial systems. Terminology is inconsistent: *visibility*, *citation*, *exposure*, and *absorption* are used interchangeably despite measuring different things. The line between legitimate optimization and adversarial manipulation is blurry. This is because with a change of intent, the very techniques that help a high-quality source earn a citation can be turned into ranking manipulation, retrieval poisoning, or product-visibility attacks (Wen et al., 2026; Nimase et al., 2026). And because generative engines are stochastic and continually updated, results reported on one engine at one point in time may not reproduce on another, raising deep questions about evaluation and reproducibility.

This paper presents a systematization of knowledge (SoK) for Generative Engine Optimization. Our goal is not merely to catalog the literature but to impose structure on it: to define the core concepts and their relationships, to organize methods and attacks into coherent taxonomies, to consolidate the evaluation landscape, and to delineate the security and ethical boundaries that distinguish cooperative optimization from manipulation. We are guided by the following research questions:

1. **RQ1 (Scope and relationships).** How does GEO relate to, and differ from, traditional SEO, retrieval-augmented generation, generative and conversational search, and agentic search and commerce?

2. **RQ2 (Methods).** What are content creators and platforms actually optimizing, and through which mechanisms? Can the space of GEO methods be organized into a principled taxonomy?

3. **RQ3 (Evaluation).** How is GEO effectiveness measured, what benchmarks and testbeds exist, and to what extent are results reproducible across engines and over time?

4. **RQ4 (Security).** How can GEO techniques be abused as attacks, what threat models and target layers do these attacks span, and how do they overlap with benign optimization?

5. **RQ5 (Defense and governance).** What defenses and platform mechanisms have been proposed, and where do the ethical and governance boundaries of GEO lie?

6. **RQ6 (Practice).** What do practitioners and platform operators actually do, measure, and claim about GEO, and where do those claims diverge from the academic evidence?

Contributions. This SoK makes the following contributions. First, we situate GEO within the broader landscape of AI-mediated information access, clarifying its relationship to SEO, RAG, and generative, conversational, and agentic search (Sections 2 and 3). Second, we propose a taxonomy of GEO methods organized by the object of optimization—content structure, evidence and citation, query intent and role, multiple objectives, agentic strategy, and product visibility—and compare representative methods along a common set of dimensions (Section 4). Third, we consolidate the evaluation landscape, organizing metrics, benchmarks, and testbeds, and surfacing the reproducibility challenges inherent to stochastic generative engines (Section 5). Fourth, we develop a taxonomy of adversarial GEO that maps attacks onto a layered model of the generative-search pipeline and analyzes the gray zone where benign optimization shades into manipulation (Section 6). Fifth, we survey defenses and platform governance mechanisms (Section 7), and articulate the security and ethical boundaries of the field together with a consolidated set of open problems (Sections 8 and 10). Sixth, we survey GEO as it is practiced outside academia, covering the commercial tooling market, platform guidance and first-party measurement, and agentic commerce, and locate where practitioner and platform accounts diverge from the research literature (Section 9).

Scope. We focus on the optimization of, and attacks against, the visibility of content in generative engines, together with the evaluation and governance of that visibility. We treat traditional SEO, RAG evaluation, and general LLM safety as adjacent fields that we reference where they inform GEO, but we do not survey them exhaustively. Throughout, we draw a clear line between *benign* GEO—cooperative,

quality-preserving optimization—and *adversarial* GEO—manipulation that degrades answer quality, fairness, or trust—while recognizing that the two occupy a continuum rather than a dichotomy.

Organization. Section 2 traces the evolution from SEO to GEO across six dimensions of divergence. Section 3 positions GEO relative to RAG and generative, conversational, and agentic search. Section 4 presents the methods taxonomy. Section 5 consolidates evaluation, benchmarks, and reproducibility. Section 6 develops the adversarial GEO taxonomy, Section 7 surveys defenses and platform mechanisms, and Section 8 examines security and ethical boundaries. Section 10 consolidates open problems, and Section 11 concludes. Section 9 turns to GEO in practice, Section 10 consolidates open problems, and Section 11 concludes.

2 From SEO to GEO: Evolution of Search Engine Optimization

Generative engines often build on, rather than wholly replace, the retrieval infrastructure that traditional search engine optimization was built for, layering a synthesis stage on top of it. Several studies observe that the technical properties SEO has long optimized, such as crawlability, indexability, and machine-readable page structure, remain consequential for GEO (Kumar and Palkhouski, 2025). The reason is architectural: a generative engine may reformulate or decompose a query, issue it to a conventional search backend, rerank the candidate sources it returns, and only then summarize and narrate them (Aggarwal et al., 2024). The technical fundamentals of SEO therefore remain preconditions for appearing in a generated answer at all (Martinez, 2026). The differences begin after retrieval. This section addresses the SEO-facing component of RQ1 by asking which established SEO practices remain effective at that additional stage, and where they break down.

Empirically, traditional rank turns out to be a weak floor rather than a reliable predictor. A high organic rank does improve the odds of being cited, but it does not confer a correspondingly high standing inside the generated answer. Kirsten et al. (2026) report that roughly 60% of AI-cited links fall within Google's top-100 and that just under half of cited domains appear in its top-10, so rank does carry signal; the complement, however, is nearly as large, since more than 40% of the links Google AI Overviews cites fall outside Google's own top-100 entirely. At the level of individual engines the decoupling is sharper still: across a thousand product-ranking queries, mean domain overlap with Google's top-10 runs from 4.0% for GPT-4o to 15.2% for Perplexity Sonar (Chen et al., 2026a). Rank and citation are therefore better treated as two loosely coupled outcomes than as one signal observed twice, since improving the first moves the second only weakly and unpredictably. The coupling is also asymmetric in direction, since the largest GEO gains accrue to sources that rank *poorly* in the underlying results, with visibility improvements of up to 115% for content at position five (Aggarwal et al., 2024). Where SEO returns are smallest, in other words, GEO returns are largest, so effort spent on one does not simply substitute for effort spent on the other. None of this means that ordering has stopped mattering. What these overlap figures compare is an engine's cited set against *Google's* ranking, whereas Jin et al. (2026) find, on the same kind of product query, that the order of the results an engine retrieves for itself strongly shapes the order it goes on to recommend. An engine can depart sharply from Google's ranking while remaining faithful to its own. Taken together, these results indicate that SEO largely establishes eligibility for retrieval without reliably determining selection, and that the two paradigms act as complements rather than successors.

What the added synthesis stage changes is the unit of visibility itself. That shift began before generative engines: once search providers started answering common queries on the results page, the unit moved from the ranked link to the extracted answer, and Answer Engine Optimization (AEO) formed around it. We treat AEO as the transitional case and examine it next, before turning to five dimensions along which SEO and GEO diverge.

2.1 Answer Engine Optimization: A Transitional Concept

AEO is the transitional case because it severed visibility from the click while leaving the underlying system intact. Search providers had begun surfacing answers directly through featured snippets and knowledge panels so that users could resolve common queries without leaving the results page, and voice assistants later read their spoken responses out of that same slot. AEO developed as the practitioner response to these platform changes: it optimizes content to be lifted directly into a snippet, knowledge panel, or spoken response, so that a source can be used without ever being visited. The system object

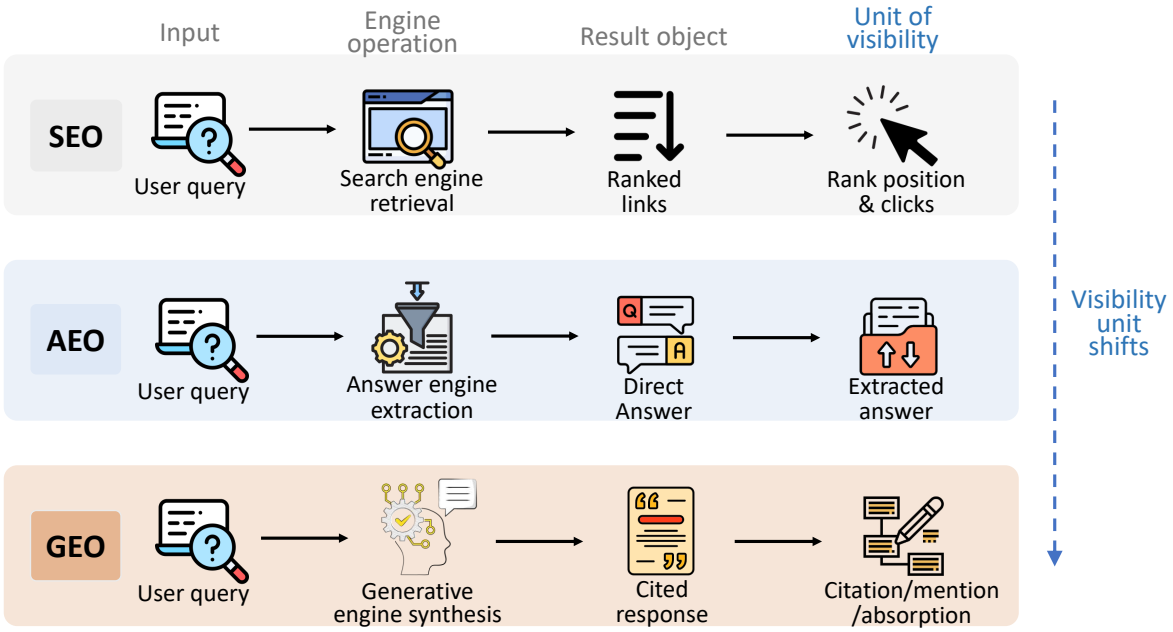

Figure 1: The unit of visibility across the SEO, AEO, and GEO regimes. Each row decomposes one regime into the user input, the engine operation, and the result object that operation produces, ending in the unit of visibility (rightmost column) that content must compete for. The input is the same in all three rows; the engine operation changes from retrieval to extraction to synthesis, and the unit of visibility changes with it.

is still a conventional search interface, but the unit of visibility has already moved from the ranked link to the extracted answer. Chen et al. (2026a) later generalize this logic to *answer engines* that blend pre-trained knowledge with live retrieval, recasting the objective from "being found and clicked" to "being synthesized and cited."

The same move that severed visibility from the click also created a quality problem that GEO now carries. Once a source is quoted rather than clicked, visibility depends on whether the engine treats the content as verifiable, and the reader has no ranked list to fall back on. Audits of the first generation of generative search engines show how poorly that dependence is served: citation support proved unreliable well before any adversarial pressure was applied, a problem Section 8 examines in detail. More consequentially for optimization practice, the same audit found citation precision and perceived utility to run in opposite directions across the four engines it studied ($r = -0.96$), a pattern its authors trace to engines copying closely from the pages they cite, which keeps citations well supported but often beside the point (Liu et al., 2023). This tension between what reads well and what is faithfully supported recurs throughout GEO, where content must earn the trust of a language model rather than a ranking algorithm.

With all three regimes now in view, Figure 1 places them side by side. Read across a row, each regime starts from the same user query and diverges at the engine operation applied to it: retrieval returns a ranked list, extraction returns a direct answer, and synthesis returns a cited response. Read down the rightmost column, the unit that content competes for moves from rank position and clicks, to the extracted answer, to citation, mention, or absorption. We read this progression as cumulative rather than substitutive. Retrieval remains the common substrate across all three rows, and each regime inserts one more stage between it and the user, so the point at which visibility is decided moves further from the ranked list each time. The connection from AEO to GEO is not one of shared machinery, since synthesis does not build on snippet extraction; what carries over is the condition AEO first created, that a source can be used without being visited, together with the verifiability problem that follows from it.

2.2 Five Dimensions of Divergence Between SEO and GEO

SEO and GEO differ systematically along five dimensions:

1. the core objective for which content is optimized;

2. the system being optimized;
3. the source ecology from which that system retrieves information;
4. the strategies that influence the desired outcome; and
5. the metrics used to evaluate that outcome.

Table 2 summarizes these distinctions, which the following subsections examine in turn.

2.2.1 Core Objectives: From Being Found to Being Recommended

SEO and GEO differ first in what counts as success. Under SEO the target is rank, since a higher position on the results page translates into more clicks and more traffic (Aggarwal et al., 2024). Under GEO the first hurdle is inclusion: content competes to appear at all in a single synthesized answer, and the same authors formalize that target as an impression function over how prominently a source is cited within the answer. Chen et al. (2025b) put the contrast as a move from content that can "be found" to content that gets "recommended" as a source worth citing.

Inclusion is close to binary, since a source is either present in the answer or absent from it, but that does not make ordinal position irrelevant. Where a generative engine returns a recommendation list rather than continuous prose, the order within that list is itself an optimization target. Jin et al. (2026) append optimization content to the documents an engine retrieves and report an average top-1 promotion success rate of 80.3% across four commercial engines and fifteen product categories, and ranking manipulation has become an object of study and of benchmarking in its own right (Pfrommer et al., 2024; Nimase et al., 2026). What changes is what a position *is*. Under SEO it is one public list, stable enough to be monitored over weeks; under GEO it exists only inside an individual answer, is conditional on the source having been retrieved and admitted to the context window at all (Chen et al., 2026a), and can differ between identical queries issued moments apart, so it has to be estimated from repeated observations rather than read off a single result (Schulte et al., 2026). The objective is therefore two-tier rather than one-dimensional: content must first clear retrieval and admission to the context window, and only then compete on placement within the answer that gets generated.

2.2.2 System Objects: From Linear Lists to Narrative Synthesis

SEO and GEO also act on different machines. A conventional search engine returns a fixed-length list of hyperlinks, each at a well-defined ordinal position that any observer can read off the page. A generative engine returns a variable-length narrative with citations embedded in it; Aggarwal et al. (2024) characterize its architecture as a pipeline in which a query rewriting module reformulates the user's input, a search engine retrieves candidate sources, a summarization model condenses them, and a response generation model produces the final answer with inline citations. Prominence does not disappear in that answer, but it stops being an ordinal fact and becomes a matter of where a source is cited, how much of the text is allotted to it, and how it is framed there. The process also stops being inspectable: a ranked list exposes its own ordering, whereas a synthesized answer exposes only its output, leaving the selection and weighting of sources hidden from the content creator.

Generative engines are not one system, either. Retrieval depth alone spans nearly the whole available range. Comparing organic Google search with four generative engines from Google and OpenAI, Kirsten et al. (2026) report a median of zero external links per query for GPT-4o with search as a tool, which answers from parametric knowledge, four for GPT-4o Search, and nine for Google AI Overviews, against the ten fixed results of organic search. For an optimizer this is not a difference of degree: an engine that rarely retrieves cannot be reached by editing a web page at all, whereas an engine that retrieves widely rewards exactly that. A single GEO strategy therefore does not transfer across platforms, and the dimensions that follow have to be read per engine rather than for generative search as a whole.

2.2.3 Source Ecology: From Uniform to Engine-Specific Ecosystems

Generative engines draw on different kinds of sources than conventional search does, and also on different kinds of sources from one another. Taking Google Organic as their reference point, Chen et al. (2026a) find that AI engines lean heavily on Earned Media, meaning third-party reviews and journalistic coverage, whereas Google devotes roughly a third of its results to Social Media, a category Claude

Table 1: System-Level Comparison of Search Engines Across Key Dimensions

Dimension	Google Organic	GPT-4o	Claude 4.5 Sonnet	Gemini 2.5 Flash	Perplexity Sonar
Output format	Ranked list	Narrative + citations	Narrative + citations	Narrative + citations	Narrative + citations
Earned Media share	41%	57%	65%	46%	50%
Social Media share	34%	8%	1%	8%†	11%†
Mean domain overlap with Google top-10	(reference)	4.0%	12.6%	11.1%	15.2%

All figures are from Chen et al. (2026a) and were measured on 1,000 product-ranking queries across ten consumer topics, so they characterize these engines on that query type rather than on web search in general. Overlap values are means; the paper reports no medians. Retrieval depth is not tabulated here because the study that measures it covers a different set of engines (Section 2.2.2). †Earned and brand shares reported in the text account for 92% of Gemini's sources and 89% of Perplexity's, which is consistent with the social shares read from the paper's source-category figure.

4.5 Sonnet almost never cites; Table 1 gives the shares. Brand-owned content is weighted differently again: Gemini 2.5 Flash gives it 46% of its sources and Perplexity Sonar 39%, against smaller shares for GPT-4o and Claude. Yang (2025) observe the same divergence at much larger scale in the news domain: analyzing over 24,000 conversations with systems from OpenAI, Perplexity, and Google, they find that models from different providers cite distinct news outlets while sharing a concentrated citation pattern. Age is weighted differently as well: median article ages for AI-cited material run at roughly a third to two thirds of Google's in the same product verticals (Chen et al., 2026a), so the evergreen content that SEO has long treated as the safe investment is not what these engines reach for. A page that suits one engine's editorial preference can therefore go uncited by another, independently of its quality.

The baseline these comparisons are drawn against is itself not uniform. Conventional engines disagree with one another about which sources to surface: Google and Bing overlap on under a third of their top-10 results (Yagci et al., 2022), and Google stands apart from both Bing and DuckDuckGo, which are largely indistinguishable from each other (Dritsa et al., 2020). A divergence measured against Google is therefore a divergence from Google, and how far it extends to conventional search in general is not something the present evidence settles.

The mix is not a fixed property of each engine, either. Chen et al. (2026a) report that generative engines shift their source composition much more sharply across query intents, from informational to consideration to transactional, than Google does, so the ecology an optimizer faces depends on the kind of question being asked as well as on the engine answering it. Geography adds a further variation, though a smaller one: Chen et al. (2025b) find Earned Media shares for the same product verticals differing by a few percentage points between the United States and Canada. Each engine thus behaves less like a window onto one web than like its own information ecosystem, with its own editorial preferences and its own sensitivity to context.

2.2.4 Optimization Strategies: From Keyword-Driven to Evidence-Driven

The techniques that produce visibility change with the target. Traditional SEO divides into on-page work, covering keyword placement, meta tags, and site structure, and off-page work, covering backlinks and domain authority (Lewandowski et al., 2021). Both rest on keyword-driven matching against an index, with link-based authority ordering whatever matches (Brin and Page, 1998). GEO keeps the same division but changes what fills each half, because the reader it has to satisfy is a language model working over retrieved text rather than an index.

On the page, the currency shifts from keyword density to evidence. Aggarwal et al. (2024) give the first systematic test, evaluating nine strategies in three groups. Content enhancement performs best: adding quotations from trusted sources raises visibility by up to 40%, and replacing qualitative claims with statistics is comparably effective. Style rewriting helps less but more evenly, with fluency optimization gaining 15–30%. Lexical edits fail outright, and keyword stuffing, the canonical SEO tactic, *reduces* visibility by roughly 10% on Perplexity.ai. That reversal is the clearest single piece of evidence that SEO technique does not carry over on its own: the same page-level edit that once helped now hurts, because what the model rewards is what a passage appears to establish rather than which terms it repeats. Combinations drawn from different groups, such as fluency optimization together with statistics addition, outperform any strategy applied alone.

The rethinking extends past the page as well. Chen et al. (2025b) organize GEO around four commitments: structuring content so that a machine can scan it, cultivating third-party coverage rather than

owned media, adapting to the retrieval behavior of each engine as documented in Section 2.2.2, and working against the bias these engines show toward large established brands. Chen et al. (2026a) add a distinction with no SEO counterpart: for entities well covered in pre-training, rankings are driven by parametric knowledge, so the target becomes a consistent long-run footprint, whereas for sparsely covered entities everything turns on being retrieved into the context window. Where SEO applies one ranking algorithm to every page, the right GEO move depends on which engine is answering and on how much the model already knows about the entity in question. Strategy therefore carries a specification burden that SEO did not: because engines differ in what they retrieve and in what they already know, no single playbook transfers intact, and each tactic has to be matched to the engine it targets and to the entity it is meant to serve.

2.2.5 Evaluation Metrics: From Deterministic Ranking to Probabilistic Visibility

SEO's metrics rest on one assumption, that a single reading suffices. Average position, click-through rate, organic traffic, and bounce rate are all point measurements, and they work because a rank is directly observable and holds still long enough to be checked (Schulte et al., 2026). Neither condition holds for GEO. Prominence inside a generated answer is not written on the page the way a rank is, so it has to be defined before it can be recorded, and the answer itself changes between identical queries, so any one reading is a sample rather than a value.

The first difficulty is met by defining new quantities. Aggarwal et al. (2024) propose three: Word Count Impression, the share of the response devoted to a source; Position-Adjusted Word Count Impression, which discounts that share by an exponential decay in citation position, on the analogy of click-through decay by rank (Craswell et al., 2008); and Subjective Impression, in which a language model scores a citation on qualitative dimensions such as relevance, uniqueness, and whether the source is framed as an authority or as a counterpoint. Later work names properties SEO never had to measure: retrieval footprint and temporal stability (Kirsten et al., 2026), coverage-adjusted freshness and citation miss rate, the share of ranked entities absent from every retrieved snippet (Chen et al., 2026a), and, in e-commerce, rank change, which Bagga et al. (2025) value at tens of thousands of dollars in annual revenue per position gained. These are not competing measures of one quantity. Impression scores describe how a source fares inside an answer; retrieval footprint and temporal stability describe how an engine behaves irrespective of any particular source; rank change describes a commercial outcome downstream of both. No convention yet fixes which of them a GEO result should report, so figures from different studies are seldom comparable, a gap Section 5 takes up.

The second difficulty is met by measuring repeatedly. Running four engines for roughly six weeks, Schulte et al. (2026) find same-day Jaccard stability of 0.32–0.43 at the source level, close to the 0.34–0.42 observed from day to day, which places the variation in the model's own sampling rather than in index updates. The contrast with the layer underneath is direct: across a two-month interval Kirsten et al. (2026) measure 45% Jaccard overlap in the links Google Organic returns against 18% for Google AI Overviews, so the instability belongs to the synthesis stage rather than to the web it draws on. Their operating recommendations follow: at least seven repeated queries per prompt per day for brand-level monitoring and eight for source-level, a rolling window of about 24 days to bring the standard error below 0.05, and a portfolio of query phrasings rather than a single template. Figure 2 shows what this asks of practice. The online path along which sources are retrieved, synthesized, and cited has to be paired with an offline harness in which query variants, engines, and repeated runs form a test grid whose logged outputs are scored and fed back into revision. Repeated measurement also exposes how unevenly citations are distributed, with a Gini coefficient averaging 0.715 across engines.

Between them these two lines cover the mechanics of measurement, but neither settles what the resulting numbers stand for. The impression metrics are proxies without an observable referent to check against: Word Count Impression assumes that more text means more visibility, and the position discount is carried over from click-through behavior on ranked lists (Craswell et al., 2008) without evidence that attention inside a synthesized answer decays the same way. Click-through rate had a click behind it, whereas a GEO score so far has only the answer text. The repetition results cut the other way, setting a cost floor that a single benchmark run does not clear, so a reported gain means little unless the metric definition and the sampling discipline are stated together. What is missing is therefore less a metric than a protocol, and Section 5 takes up that question directly.

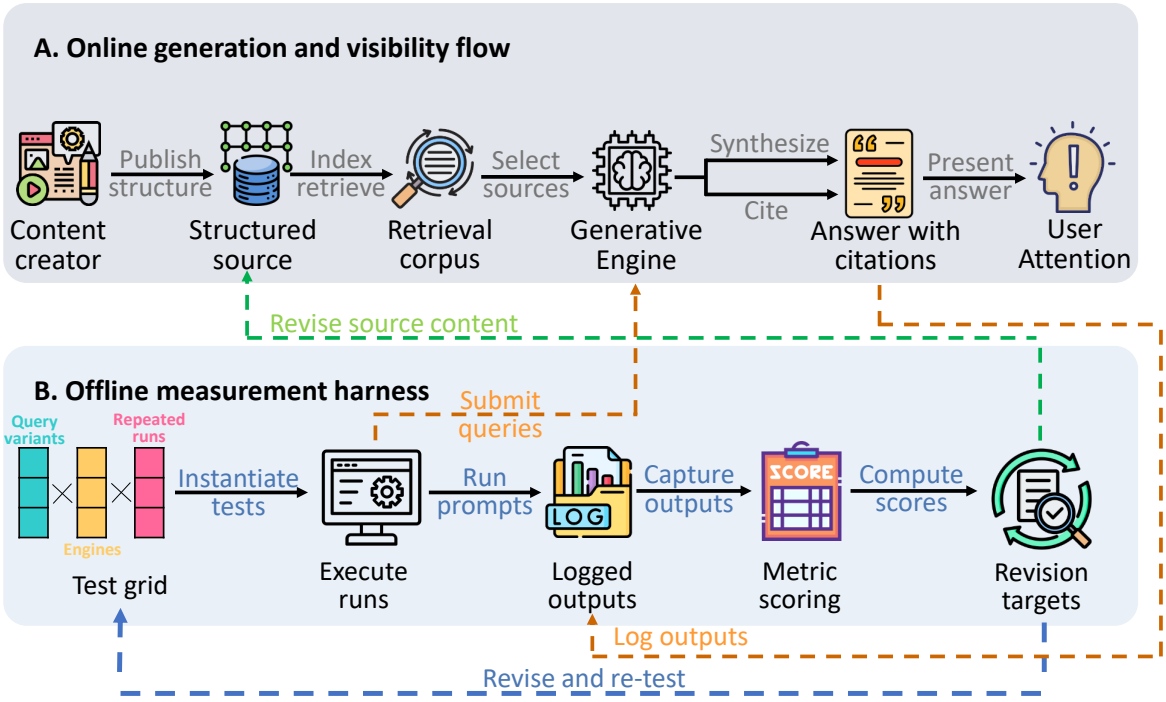

Figure 2: Online generation flow and offline measurement harness for GEO visibility. The online flow shows how structured content is retrieved, synthesized, cited, and presented to users. The offline harness shows how GEO visibility is measured through query variants, engines, repeated runs, logged outputs, metric scoring, and iterative content revision.

2.3 Summary: SEO versus GEO Across Five Dimensions

Table 2 sets the five dimensions side by side. Taken together they show something no dimension shows alone: under GEO each of them has become conditional. SEO could give one answer per question, a single target, a single system, a single source pool, a single playbook, a single number. GEO answers each question only after an engine, a query intent, an entity, and a particular run have been fixed.

Two things follow. The first concerns the transition itself, which is easy to read as a replacement. Nothing in these five dimensions shows SEO being displaced: retrieval eligibility still decides what can be cited, and the synthesis stage decides what is. Layering rather than succession is what the evidence supports, and layering is the more demanding arrangement of the two, since the base has to be maintained without paying off directly while the layer above it refuses to hold still. The second concerns evidence. Once every answer is conditional, the conditions of a measurement become part of what was measured, so a visibility result reported without its engine, its query intent, and its sampling regime gives another group nothing to build on. The rest of this survey is organized around that constraint: methods are grouped by what they act on rather than by the engine they were tested against, and evaluation is consolidated around a shared metric framework and a reproducibility audit rather than around individual platforms.

3 Positioning GEO: Relationships to RAG, Generative, Conversational, and Agentic Search

Existing surveys of the LLM–search intersection largely adopt the complementary system-builder perspective—surveying how search improves LLMs and how LLMs improve search (Xiong et al., 2024)—rather than the optimizer-side question of how external creators gain visibility; we compare this SoK to prior surveys and position papers in Section 3.1.

GEO and the generative-engine pipeline. A generative engine, in the sense formalized by Aggarwal et al. (2024), is a system that answers a query by retrieving sources, summarizing them, and generat-

Table 2: Comparison of SEO and GEO Across Five Dimensions

Dimension	SEO	GEO
Core Objective	Maximize position in one public ranked list; drive click-through and organic traffic	Clear inclusion in the answer first, then compete on placement within it; position exists only inside a single answer
System Object	Search engines returning a fixed-length ranked list of about 10 links, with the ordering visible on the page	Generative engines returning variable-length narrative with inline citations; selection and weighting of sources are not observable
Source Ecology	Balanced across Brand, Earned, and Social Media; Social Media about 34% of Google's top results	Skewed toward Earned Media, Social largely excluded; mix varies by engine, query intent, region, and content age
Optimization Strategy	Keyword-driven on-page and off-page work: keyword density, meta tags, backlinks, domain authority	Same division, different currency: quotations, statistics, fluency, earned coverage, engine- and entity-specific adaptation; keyword stuffing is counterproductive
Evaluation Metric	Deterministic: average position, click-through rate, organic traffic; one reading suffices, and Google Organic's retrieved links overlap 0.45 across two months	Probabilistic: word count and position-adjusted impression, LLM-scored subjective impression; estimates require repeated runs, with same-day source-level Jaccard as low as 0.32

ing a narrative response with inline citations. Architecturally, most deployed generative engines are instances of retrieval-augmented generation (RAG): they couple a retriever over a web-scale corpus with an LLM that conditions its generation on the retrieved evidence. The crucial distinction is one of *perspective*. RAG research asks how a system builder can construct a pipeline that produces accurate, well-grounded answers; GEO asks how an external content creator—who controls neither the retriever nor the generator—can make their content visible within the answers that pipeline produces. GEO is therefore best understood as the study of the generative-engine pipeline *from the outside*, treating the engine as a black box whose retrieval and citation behavior is to be characterized and influenced rather than designed. This black-box, third-party stance is what separates GEO from RAG system design, even though the two share the same underlying machinery. Existing surveys of the LLM–search intersection adopt the complementary system-builder perspective—surveying how search improves LLMs and how LLMs improve search (Xiong et al., 2024)—but do not address the optimizer-side question of how external creators gain visibility, which is the gap this SoK fills.

GEO and generative versus conversational search. Generative search engines return synthesized answers to single queries; conversational search engines extend this to multi-turn dialogue, where context accumulates across turns. The optimization target differs accordingly: in single-turn generative search the goal is inclusion in one answer, whereas in conversational search it is sustained presence across a dialogue whose intent drifts over turns. Much of the foundational GEO work targets single-turn generative search (Aggarwal et al., 2024; Wu et al., 2025), while a distinct line of work studies ranking manipulation and visibility specifically in the conversational setting (Pfrommer et al., 2024; Puerto et al., 2025). We treat both under the GEO umbrella but flag the turn structure as a dimension along which methods and threat models vary.

GEO and agentic search and commerce. The frontier of the field is the shift from a single generative engine answering a query to *agentic* systems that autonomously browse, follow links, issue follow-up queries, and synthesize across multi-step trajectories (Ye et al., 2026; Zhong et al., 2026). In this setting, visibility is no longer a property of a single page but of an *evidence ecosystem* encountered along an agent's browsing trajectory (Ye et al., 2026). Agentic commerce—where shopping agents select and purchase products on a user's behalf via protocols such as emerging agentic-commerce and universal-commerce standards—pushes this further, making product visibility in an agent's decision process an economic variable of direct commercial consequence (Kumar and Lakkaraju, 2024; Jin et al., 2026; Bagga et al., 2025). We return to agentic GEO as a methods category in Section 4 and as an attack surface in Section 6.

A note on terminology. Throughout this survey we use *visibility* as the umbrella term for the degree to which a source influences a generated answer, and reserve more specific terms for measurable sub-concepts: *citation* for the presence of an explicit reference to a source, *absorption* for the incorporation of a source's language, evidence, or structure into the answer text whether or not it is cited (Zhang et al.,

2026b), and *exposure* for the user-facing prominence of a cited source (Alipour et al., 2026). We make these distinctions precise in Section 5.

3.1 Related Surveys and Position Papers

Although GEO itself is young, several efforts have begun to consolidate the surrounding landscape, and situating this SoK against them clarifies what it adds. Surveys of the LLM–search intersection, most prominently Xiong et al. (2024), systematize how search services and LLMs enhance one another—retrieval-augmented generation on one side, LLM-based query understanding, ranking, and summarization on the other—but remain squarely on the system-builder side of the pipeline and do not treat content-creator visibility as an object of study. In the same system-side vein, Dai et al. (2025) diagnose the collapse of the web's feedback ecosystem under generative search and propose redesigned user-feedback loops; their concern is the sustainability of the ecosystem that GEO operates in, not optimization itself. Complementing these, the position paper of Wen et al. (2026) argues that GEO creates underexamined risks and that governance should target visibility concentration, disclosure, and academic blind spots; it advocates a research agenda rather than systematizing the existing literature.

The work closest to ours is the contemporaneous survey of Chen et al. (2026b), which reviews what it terms G-SEO: it formalizes the task and its assumptions over RAG-based generative search engines, organizes methods into retrieval-side manipulation (retrieval poisoning and hijacking) and generation-side manipulation (content optimization and prompt injection), reviews metrics, benchmarks, and experimental protocols, and closes with challenges such as long-context, black-box, and cross-system optimization. We differ in three principal respects. First, in taxonomy: their two-way, pipeline-side split places benign content optimization and overtly adversarial techniques such as poisoning and prompt injection within a single method space, whereas we deliberately separate a taxonomy of benign methods organized by the *object of optimization* (Section 4) from a layered adversarial taxonomy with explicit threat models (Section 6), precisely because the benign–adversarial boundary is intent-dependent and, in our view, must be analyzed rather than assumed—the gray zone between the two is itself a central object of this systematization. Second, in security depth: defenses appear in their survey as a single future-work direction (anti-G-SEO), whereas we devote full treatment to defenses and platform mechanisms (Section 7) and to the ethical and governance boundaries of the field (Section 8). Third, in scope: their analysis centers on text-centric, single-turn RAG pipelines, whereas we additionally cover conversational and agentic search, product-visibility and e-commerce GEO, and agentic commerce (Sections 3 and 4), consolidate the field's inconsistent terminology (visibility, citation, absorption, exposure), and analyze the reproducibility of results across stochastic, continually updated live platforms (Section 5). To our knowledge, this paper is the first security-oriented systematization that spans benign optimization, adversarial manipulation, defenses, and governance of generative-engine visibility end to end.

4 Methods: A Taxonomy of Generative Engine Optimization

Having positioned GEO, we turn to RQ2 and examine what GEO methods optimize and how they do so. We group existing methods into six clusters: content and structure optimization, evidence and citation optimization, query intent and role modeling, optimization through multiple objectives and feature abstractions, agentic GEO with strategy learning, and product visibility in e-commerce. The categories identify each method's main focus, although some methods combine more than one form of optimization. Section 6 examines how these methods can be abused.

4.1 Content and Structure Optimization

Content optimization edits a webpage so that generative engines are more likely to use and cite it. The first GEO study, Aggarwal et al. (2024), tested nine strategies that added evidence, changed presentation, or altered wording. Quotations and statistics improved visibility, while keyword insertion reduced it. This finding established a basic result for later work: edits that add useful information are more effective than edits that merely repeat query terms. Section 2 discusses this result in detail.

Later work changes both the unit of editing and the way edits are produced. Structural methods treat the organization of a document as an optimization target. GEO-SFE separates structural features from semantic content and tests document architecture, information chunking, and visual emphasis across six engines, increasing citation rate by 17.3% and subjective quality by 18.5% (Yu et al., 2026b). Other

methods automate the editing process. AutoGEO infers engine preferences and turns them into rules that guide a prompt optimizer and train a lightweight model (Wu et al., 2025), while Lüttgenau et al. (2025) train BART on synthetic optimization pairs. These approaches replace isolated lexical edits with coordinated changes to the document. Locating their effect within retrieval, source selection, or answer construction requires measures beyond aggregate visibility.

4.2 Evidence and Citation Optimization

Evidence and citation methods target a specific stage of generative search: whether an engine attributes information to a source. This focus separates citation from broader measures of contribution. A page may influence an answer without being cited, and a selected source may contribute little language or evidence to the final response. Zhang et al. (2026b) formalize this distinction as *citation selection* and *citation absorption*. Building on the same pipeline view, Tian et al. (2026) identify distinct citation failures and introduce AgentGEO, which diagnoses the failure before choosing a repair. This targeted process improves citation rate by more than 40% while changing 5% of the content, compared with 25% for baseline methods. The evidence study of Wan et al. (2024) helps explain these results: LLMs respond strongly to relevance but give little weight to stylistic signals of credibility. Citation is therefore a more direct target than aggregate visibility when the goal is source attribution. Factual support and source diversity require separate measures.

4.3 Query Intent and Role Modeling

Methods based on query intent treat an effective edit as conditional on the information a user seeks and the role a document can play in the answer. Chen et al. (2025d) infer search intent through reflective refinement across informational roles and use it to guide content revision. When the same document serves multiple queries, intent alignment becomes a conflict problem. IF-GEO mines preferences for individual queries, combines them into one revision plan, and measures when an improvement for one query harms another (Zhou et al., 2026). The shared problem is to optimize for a set of queries without letting gains on a few queries hide losses elsewhere. Evaluation therefore needs query sets that reflect the queries and conversational contexts expected after deployment.

4.4 Optimization with Multiple Objectives and Feature Abstractions

Directly optimizing a draft makes it difficult to see which document properties produce a gain and how those properties affect competing objectives. FeatGEO addresses this problem by representing a webpage through interpretable structural, content, and linguistic features (Liu and Xu, 2026). It searches over feature configurations before an LLM renders the selected configuration as text. Separating feature search from text generation makes the optimization easier to inspect and allows visibility and quality to be controlled explicitly. Its analysis also finds that properties of the whole document explain citation behavior better than word-level edits, and that learned configurations transfer across LLM sizes. Feature design and objective weights determine what the optimizer can pursue. Goals such as provenance or fairness require corresponding features and objectives.

4.5 Agentic GEO with Strategy Learning

Agentic GEO replaces a single revision with a cycle of planning, editing, evaluation, and learning. Within a document, AgenticGEO and MAGEO both learn strategies that can be reused across later edits. AgenticGEO searches for compositional strategies with a MAP-Elites archive and trains a lightweight critic to approximate costly engine feedback (Yuan et al., 2026). MAGEO assigns planning, editing, and fidelity checking to separate agents, then stores successful editing patterns as skills for particular engines and scenarios (Wu et al., 2026). Other work widens the unit of optimization. TRACE coordinates several pages so that they shape the evidence an agent encounters during browsing (Ye et al., 2026), while Zhao et al. (2026) model confidence under changes in context and use multiple agents for intent routing. These methods optimize both editing strategy and the path through which an engine gathers evidence. They require more computation and repeated feedback, and they give optimizers more control over the information presented to the engine, which creates the security concerns discussed in Section 6.

4.6 Product Visibility in E-Commerce

Product search adds multimodal inputs and direct commercial stakes to GEO. Results from E-GEO show that fifteen hand-designed strategies have little effect across more than 7,000 consumer queries, whereas prompt-based meta optimization finds more reliable promotion strategies (Bagga et al., 2025). Product pages also depend on images. Du et al. (2026) jointly optimize product images and descriptions for vision language rankers and obtain larger gains than either modality produces alone. Pinterest applies GEO at industrial scale by training vision language models to predict user queries and linking collections according to authority across billions of assets, with a reported 20% increase in organic traffic (Zhang et al., 2026a). These studies connect strategy discovery, multimodal optimization, and deployment under real traffic. The same methods can surface a relevant product or give an inferior product more exposure, a distinction examined in Sections 6 and 8.

4.7 Cross Cutting Comparison

Table 3 compares representative methods by optimization target, access to live engines, evaluation platform, and code availability. GEO methods now optimize more than page text. Newer methods target citation events, query-dependent behavior, and the evidence paths followed by browsing agents. The search process has changed as well. Fixed heuristics remain common, but E-GEO, AutoGEO, and AgenticGEO find effective strategies through model or engine feedback (Bagga et al., 2025; Wu et al., 2025; Yuan et al., 2026). Across their evaluated settings, these methods find recurring engine responses that can guide later edits. Code availability is uneven. Several research systems release code, while industrial systems with the largest reported traffic effects remain proprietary, as discussed in Section 5.

Table 3: Taxonomy of GEO methods by object of optimization. "BB query" indicates whether the method requires querying a live generative engine during optimization without internal access. "OSS" indicates whether code is openly released (as reported by the authors).

Category / Method	Object optimized	Platform(s) evaluated	BB query	OSS
Content structure optimization				
GEO (Aggarwal et al., 2024)	Page text (quotes, stats, style)	Generative engines / GEO-Bench	Yes	Partial
GEO-SFE (Yu et al., 2026b)	Document structure (macro/meso/micro)	Six generative engines	Yes	No
AutoGEO (Wu et al., 2025)	Page text via learned preference rules	AI Overview, ChatGPT / GEO-Bench	Yes	Yes
Evidence and citation optimization				
AgentGEO (Tian et al., 2026)	Citation failure repairs	Generative engines (held out queries)	Yes	No
Absorption framework (Zhang et al., 2026b)	Citation selection vs. absorption	ChatGPT, Google, Perplexity	Yes	Yes
Query intent and role modeling				
Role-Augmented G-SEO (Chen et al., 2025d)	Intent/role aligned content	Bing Chat, Perplexity	Yes	No
IF-GEO (Zhou et al., 2026)	Multi query instruction fusion	GPT-4o-mini, Gemini-2.0	Yes	No
Multiple objectives / feature abstractions				
FeatGEO (Liu and Xu, 2026)	Interpretable feature configuration	Three generative engines	Yes	No
Agentic GEO with strategy learning				
AgenticGEO (Yuan et al., 2026)	Compositional strategies (self evolving)	Two engines, three datasets	Yes	Yes
MAGEO (Wu et al., 2026)	Reusable engine specific skills	Three engines	Yes	Yes
TRACE/EcoGEO (Ye et al., 2026)	Trajectory level evidence ecosystem	OPR-Bench (agent search)	Yes	No
Product visibility in e-commerce				
E-GEO (Bagga et al., 2025)	Product listing text	E-GEO testbed (7k+ queries)	Yes	Yes
MGEO (Du et al., 2026)	Product image + text	VLM rankers	Yes	Yes
Pinterest GEO (Zhang et al., 2026a)	Collection pages + interlinking	Pinterest (industrial)	N/A	No

5 Evaluation, Benchmarks, and Reproducibility

This section addresses RQ3 by consolidating how GEO effectiveness is measured. We organize the discussion into three parts: the metrics used to quantify visibility and related concepts; the benchmarks and testbeds that operationalize these metrics; and the reproducibility challenges that arise from evaluating stochastic, continually updated, and often proprietary engines.

Table 4: GEO evaluation metrics organized by dimension. Each metric is listed with its primary reference, the type of engine access required (offline / live), and whether it has been adopted by multiple independent studies ("Adopted").

Dimension	Metric	Definition	Access	Adopted
Visibility	Word Count Impression	Fraction of answer words from source	Offline	Wide
	Position-Adjusted WCI	WCI × exponential position decay	Offline	Wide
	Subjective Impression	LLM-judged 7-dim qualitative score	Offline	Wide
Citation	Citation Rate	Fraction of queries source is cited	Live/Offline	Wide
	Citation Breadth	Unique sources cited per answer	Live	Moderate
	Citation Influence	Weighted coverage across citation depth	Live	Limited
Absorption	Citation Absorption	Fraction of source language absorbed	Offline	Limited
	Absorption Depth	Evidence type (def./fact/procedure) coverage	Offline	Limited
Ranking	Rank Change (ΔRank)	Mean change in ordinal product rank	Live/Offline	Moderate
	Promotion Success Rate	Fraction reaching top-k after optimization	Live	Moderate
Robustness	Jaccard Stability	Source-set overlap across repeated queries	Live	Moderate
	Temporal Stability	Jaccard overlap of citation sets over time	Live	Limited
	Gini Coefficient	Citation concentration across sources	Live	Limited
Security	Attack Success Rate	Fraction queries target reaches rank 1	Offline	Wide
	BadWordRatio	Keyword-violation rate (stealth proxy)	Offline	Moderate
	ELO / Competitive Score	Multi-adversary arena ranking	Offline	Limited

5.1 Metrics

GEO evaluation has produced a layered family of metrics that measure increasingly fine-grained notions of visibility. Table 4 consolidates the full metric landscape across six evaluation dimensions used in this survey.

Impression and visibility metrics. The foundational metrics were introduced by Aggarwal et al. (2024): *Word Count Impression*, the normalized fraction of answer text devoted to a source; *Position-Adjusted Word Count Impression*, which weights that fraction by an exponential decay over citation position, mirroring click-position bias in traditional search (Craswell et al., 2008); and *Subjective Impression*, an LLM-judged score across seven qualitative dimensions (relevance, influence, uniqueness, diversity, follow-up likelihood, subjective position, and subjective count). These remain the most widely reused GEO metrics and underpin most method papers surveyed in Section 4.

Citation, absorption, and exposure. Recognizing that impression conflates several phenomena, later work disaggregates it. Zhang et al. (2026b) separate *citation selection* from *citation absorption* and introduce citation breadth and depth and a citation-influence measure, finding for instance that ChatGPT cites fewer sources but with higher average influence than Perplexity or Google. Alipour et al. (2026) introduce *exposure* as the user-facing prominence of a cited source and show, in an audit of 44 enterprises, that exposure is biased toward already-prominent voices with larger follower bases—linking visibility metrics to fairness. The distinction among citation, absorption, and exposure is, in our view, the single most important conceptual refinement the field has made, and we adopt it as standard terminology (Section 3).

Ranking and product metrics. In product and recommendation settings, ordinal metrics are more natural. Jin et al. (2026) measure *Promotion Success Rate* at top-k, reporting 91.4% at top-5, 86.6% at top-3, and 80.3% at top-1 across fifteen product categories. Bagga et al. (2025) argue for *rank change* as an

economically interpretable metric, estimating that a one-position average improvement corresponds to tens of thousands of dollars in annual revenue—grounding GEO evaluation in business value.

Robustness, stability, and security metrics. Because generative engines are stochastic, a distinct family of metrics quantifies the *reliability* of visibility rather than its level. Kirsten et al. (2026) propose retrieval footprint (external links per query) and temporal stability (Jaccard overlap of citation sets over time). Schulte et al. (2026) formalize visibility as a *distribution* and derive operational guidelines—at least 7–8 repeated queries per prompt per day and a roughly 24-day window to drive standard error below 0.05—while reporting same-day source-level Jaccard stability as low as 0.32–0.43 and a citation-concentration Gini coefficient averaging 0.715. Security-oriented evaluations add attack success rate, stealth (perplexity ratio and keyword-violation rate), and, for the multi-adversary setting, competitive effectiveness (Nimase et al., 2026; Chen et al., 2025a); we treat these in Section 6.

5.2 Benchmarks and Testbeds

The benchmark landscape spans four methodological settings, which differ in realism and reproducibility.

Static benchmarks. GEO-Bench (Aggarwal et al., 2024) and its successors provide fixed query–source collections enabling offline, repeatable evaluation. C-SEO Bench (Puerto et al., 2025) extends this to conversational search across two tasks, multiple domains, and—crucially—varying numbers of competing actors, revealing that most C-SEO methods are ineffective or harmful and that gains shrink as adoption rises, exposing a zero-sum congestion dynamic. ConflictBank (Su et al., 2024) and the convincingness study of Wan et al. (2024) supply controlled testbeds for the knowledge-conflict and evidence-selection phenomena that underlie citation behavior.

Domain-specific testbeds. E-GEO (Bagga et al., 2025) targets e-commerce with intent-rich consumer queries, and ProductBench (Jin et al., 2026) collects real top-10 product recommendations across fifteen categories. These trade generality for ecological validity in a commercially important domain.

Realistic simulation arenas. SAGEO Arena (Kim et al., 2026) provides a reproducible pipeline-level environment spanning retrieval, reranking, and generation over a large web corpus, enabling stage-level analysis; it finds that many existing GEO/SEO methods degrade retrieval and reranking under realistic conditions. For agentic settings, AgentWebBench (Zhong et al., 2026) benchmarks multi-agent coordination across ranked-retrieval and open-ended synthesis tasks, finding that decentralized agent access concentrates traffic toward a few sites.

Live-platform audits. The most ecologically valid but least reproducible studies query commercial engines directly. Kirsten et al. (2026) audit roughly 4,706 queries across six engines; Chen et al. (2026a) and Grossman et al. (2026) compare Google, Gemini, and AI Overviews over thousands of real-user queries, the latter reporting that AI Overviews are generated for 51.5% of queries with under 0.2 average Jaccard source overlap across engines.

Table 5 compares these benchmarks and testbeds by setting, domain, scale, and reproducibility characteristics.

5.3 Reproducibility Challenges

GEO evaluation faces reproducibility obstacles more severe than most empirical ML subfields, for three structural reasons. **Stochasticity:** as Schulte et al. (2026) demonstrate, identical queries issued moments apart yield substantially different citation sets, so single-shot results are unreliable and repeated measurement is mandatory. **Non-stationarity:** commercial engines are updated continuously and opaquely, so a result obtained on a dated model version may not replicate, and reported engine versions are essential metadata. **Opacity and access asymmetry:** the most impactful systems are proprietary, and live-audit studies cannot be exactly reproduced once an engine changes; Wen et al. (2026) characterize this as an academic "blind spot" arising from visibility and evaluation asymmetries between offline setups and deployed systems. These challenges motivate three best practices that the strongest studies already follow: report engine versions and collection dates, use repeated-measurement protocols with

Table 5: Comparison of GEO benchmarks and evaluation testbeds across methodological settings. "Repro." summarises reproducibility (High/Medium/Low). "Key limitation" identifies the principal gap for cross-study comparison.

Benchmark	Setting	Domain	Scale	Repro.	Key limitation
GEO-Bench (Aggarwal et al., 2024)	Static	Multi-domain	1k queries	High	Single-shot; no repeated measure
C-SEO Bench (Puerto et al., 2025)	Static, multi-actor	QA+products	Multi-task	High	Offline only; no live-engine results
E-GEO (Bagga et al., 2025)	Domain testbed	E-commerce	7k+ queries	High	E-commerce only; no general-domain
ProductBench (Jin et al., 2026)	Domain testbed	E-commerce	15 cats.×200	Medium	Narrow domain; no code released
SAGEO Arena (Kim et al., 2026)	Simulation	Web (general)	Pipeline-level	Med–High	Corpus snapshot; no live update
AgentWebBench (Zhong et al., 2026)	Agentic sim.	Web agents	4 tasks, 7 LLMs	High	Agentic only; no single-query GEO
CONFLICTBANK (Su et al., 2024)	Static	Knowledge QA	553k pairs	High	No GEO optimization task; KQ only
Kirsten et al. (Kirsten et al., 2026)	Live audit	General	4,706 q. 6 engines	Low	Live engines; not reproducible
Grossman et al. (Grossman et al., 2026)	Live audit	General	11,500 queries	Low–Med	Engine version uncontrolled

reported variance, and, where possible, complement live audits with reproducible static or simulation benchmarks. We list reproducibility as a first-order open problem in Section 10.

5.4 Open-Source Code Availability and Unified Benchmark Evaluation

To ground the survey in empirical practice, we audited the code and data availability of all open-source GEO projects covered in this survey, attempted to run each project, and evaluated a representative subset under a unified benchmark. This subsection reports the results of that exercise and draws cross-cutting lessons about what can and cannot be reproduced with current tooling.

Code availability. Table 6 summarizes the code, data, prompt, and evaluation-script availability of 20 open-source GEO projects. Of these, 18 release runnable code and 17 provide public datasets or pre-computed results; however, access to prompts and experiment scripts is less consistent. Critically, three projects (AutoGEO-Mini, StealthRank, LLM Rank Optimizer) require local GPU-hosted LLMs for their generation stages and cannot be reproduced on consumer hardware. Two projects (AgentGEO, GEO/AEO Tracker) depend on third-party search APIs (ChatNoir, commercial engine APIs) that require separate registration. This pattern—readily available code paired with hardware- or API-gated execution—represents the dominant reproducibility barrier across the field.

Table 6: Code and data availability for 20 open-source GEO projects. "Code": runnable implementation released. "Data": dataset or pre-computed results available. "Prompt": optimization prompts or templates released. "Script": end-to-end experiment scripts provided. "GPU req.": generation stage requires GPU-hosted LLM. "GPU free": full pipeline reproducible on CPU or via API.

Project	Code	Data	Prompt	Script	GPU free	Primary metric
GEO optimization methods						
GEO (Aggarwal et al., 2024)	✓	✓	✓	✓	✓	GEO score
AutoGEO (API) (Wu et al., 2025)	✓	✓	✓	✓	✓	GEO-pos/word/wordpos
FeatGEO (Liu and Xu, 2026)	✓	✓	✓	✓	✓	GEO score, quality
MAGEO (Wu et al., 2026)	✓	✓	✓	✓	✓	ΔGEO score
AgenticGEO (Yuan et al., 2026)	✓	✓	✓	✓	✓	Strategy execution rate
E-GEO (Bagga et al., 2025)	✓	✓	✓	✓	✓	Rank change
GEO/AEO Tracker	✓	–	–	✓	✓	Citation monitoring
Attack and security						
GASLITE (Ben-Tov and Sharif, 2024)	✓	✓	✓	✓	✓	NDCG@1, rank
PoisonArena (Chen et al., 2025a)	✓	✓	partial	partial	✓	ELO, ASR
StealthRank (Tang et al., 2025)	✓	✓	✓	✓	✗	AvgRank, BadWordRatio
LLM Rank Optimizer (Kumar and Lakkaraju, 2024)	✓	✓	✓	✓	✓	Rank advantage
FeatGEO (adversarial) (Liu and Xu, 2026)	✓	✓	✓	✓	✓	GEO score
AgentGEO (Tian et al., 2026)	✓	✓	✓	✓	partial	Citation rate
Benchmarks and measurement						
C-SEO Bench (Puerto et al., 2025)	✓	✓	✓	✓	✓	ΔRank
AmazonCOREBench (Jin et al., 2026)	✓	✓	partial	✓	✓	Top-1 rate
AI Product Bench	✓	✓	partial	✓	✓	Jaccard, Top-1
AIO Benchmark (Grossman et al., 2026)	✓	✓	–	✓	✓	Citation rate, RBO, Jaccard
RAG-Convincingness (Wan et al., 2024)	✓	✓	partial	✓	✓	Stance, conflict rate
geo-citation-lab (Zhang et al., 2026b)	✓	✓	partial	✓	✓	72-feature influence
CONFLICTBANK (Su et al., 2024)	✓	✓	partial	partial	✓	Accuracy, conflict rate
When Attention Becomes Exposure (Alipour et al., 2026)	✓	✓	–	✓	✓	Exposure uplift

Reproduction log and known issues. Table 7 records, for each project, the primary environment dependency, whether the end-to-end pipeline reproduced cleanly, the main obstacle encountered, and how we adapted the pipeline. This table is Task 2's running-log artifact and serves as a guide for researchers who wish to replicate or build on these projects.

Unified benchmark evaluation. To enable cross-method comparison, we evaluated a subset of projects under a unified query set drawn from GEO-Bench (Aggarwal et al., 2024) (100 queries, stratified by tag, `random_seed=42`) using `gpt-4o-mini` as the judge LLM throughout. We organize results by measurement dimension.

GEO visibility (Dimension A). Table 8 compares optimization methods on GEO-wordpos (position-adjusted word count), ΔRank, and the fraction of queries with a positive GEO score change. The rule-based Authoritative rewriting from the original GEO paper achieves positive gains on all 100 queries with no API cost, confirming its value as a strong, reproducible baseline. AutoGEO-API and

Table 7: Reproduction log for 20 open-source GEO projects. "Reproduced": ✓ = full end-to-end pipeline; ∼ = evaluation stage only (generation stage skipped); ✗ = blocked. "Adaptation" summarises the workaround applied.

Project	Key dependency	Repro.	Main obstacle	Adaptation
GEO optimization methods				
GEO	datasets, nltk	✓	Bing API needed for live queries	Offline GEO-Bench analysis
AutoGEO (API)	OpenAI API	✓	core.py hard-coded Gemini call	1-line fix: route by engine_llm param
FeatGEO	OpenAI API	✓	SAMPLING_MODE config required	Set SAMPLE_LIMIT=2, disable GA
MAGEO	Python stdlib	✓	None	Direct run
AgenticGEO	Python stdlib	✓	None	Direct run
E-GEO	OpenAI API	∼	3 path bugs; numpy JSON serialisation	Path fixes + default=str serialiser
GEO/AEO Tracker	Python stdlib	✓	Full run needs 6 API keys	Demo mode (no key required)
Attack and security				
GASLITE	PyTorch, BEIR	✓	5 hard-coded `.cuda()` calls	Replace with `.to(model.device)`
PoisonArena	Python stdlib	∼	LLM judge requires A100 GPU	Re-implement ELO from pre-computed scores
StealthRank	Local LLM	∼	7B LLM generation not GPU-free	Read pre-computed JSON results
LLM Rank Optimizer	OpenAI API	∼	GCG STS gen. requires GPU	Use provided STS; run evaluate.py
FeatGEO (adv.)	OpenAI API	✓	Same as FeatGEO above	Same fix
AgentGEO	OpenAI API	∼	ChatNoir search API requires registration	LLM optimisation pipeline runs; search disabled
Benchmarks and measurement				
C-SEO Bench	datasets, API	✓	None	Direct run
AmazonCOREBench	Python stdlib	✓	None	Direct run
AI Product Bench	OpenAI API	✓	None	Direct run
AIO Benchmark	pandas, scipy	✓	Filename encoding (shell escape)	Quote path in shell call
RAG-Convincingness	pandas	✓	Data on Google Drive (manual DL)	Download pkl; run analysis
geo-citation-lab	pandas	✓	API keys for data collection	Use pre-computed features CSV
CONFLICTBANK	datasets	✓	Dataset 5.79 GB	Use streaming=True (200-item sample)
When Attention Becomes Exposure	pandas, scipy	✓	None	Direct run

FeatGEO improve on the baseline in the majority of queries; MAGEO's ΔRank gain of $+1.850$ reflects the benefit of multi-agent strategy reuse. The E-GEO prompt optimizer achieves an 80% improvement rate in training queries but shows more variance in validation rounds (50–80%), consistent with the meta-optimization dynamics described in Bagga et al. (2025).

Attack effectiveness (Dimension B). Table 9 compares adversarial methods. GASLITE achieves NDCG@1 of 1.0 after a single gradient step on CPU (after adapting five CUDA calls for CPU/MPS), confirming the extreme retrieval vulnerability documented by Ben-Tov and Sharif (2024). StealthRank with Mistral-7b achieves AvgRank 1.85 across 53 product categories with BadWordRatio 0.130, demonstrating that effective ranking manipulation can be imperceptible to keyword-based filters. The LLM Rank Optimizer shows negligible rank advantage when the STS generated for GPT-4o is transferred to GPT-4o-mini (0% advantage, 60% tie, 40% disadvantage), consistent with the cross-model transfer limitations noted by Kumar and Lakkaraju (2024). PoisonArena's ELO ranking (GASLITE = 1978, content_poison = 307) reproduces the key finding that competitive multi-adversary settings invert single-attacker ASR rankings.

Measurement and bias (Dimension C). Table 10 summarises the key findings. Among the measurement-oriented projects, the AIO Benchmark finds an overall AI Overview trigger rate of 65.6% across 11,500 queries, with substantial variation by dataset (ELI5: 94.6%; Amazon Retail: 17.4%), and AIO-vs-SERP Jaccard of 0.178, consistent with Grossman et al. (2026). The AI Product Bench records a cross-model Jaccard of 0.042 and a ChatGPT Top-1 consistency rate of 3.8%, confirming the extreme platform-specificity of product visibility. The geo-citation-lab analysis covers 23,745 citation records across 72

Table 8: GEO optimization results under the unified GEO-Bench-100 query set (`gpt-4o-mini` judge, `seed=42`). GEO-wordpos is position-adjusted word-count impression (higher is better). ΔRank is mean rank change. "Impr. rate" is the fraction of queries with positive GEO score change. "API" indicates whether an LLM API is required.

Method	GEO-wordpos ↑	ΔRank ↑	Impr. rate	API
GEO Authoritative	0.2065	+0.0087	100%	No
AutoGEO API	0.1070	+0.0085	100%	Yes
FeatGEO (citing_credible)	0.1398	+0.0012	20%	Yes
MAGEO DSV-CF	–	+1.850	100%	No
E-GEO Authoritative	–	+1.600	80%	Yes
AgenticGEO	–	–	–	No

Table 9: Attack effectiveness under unified evaluation. ASR@1: fraction of queries where the target reaches rank 1. AvgRank: mean rank of target (lower is better). BadWordRatio: fraction of tokens flagged as keywords (lower indicates greater stealth). ELO: competitive score in PoisonArena arena (higher is better). Cross-model: whether the STS/adversarial suffix transfers to a different LLM.

Method	Target layer	Key metric	Value	GPU-free
GASLITE	Retrieval	NDCG@1	1.000	✓
StealthRank (Mistral-7b)	Ranking	AvgRank / BadWordRatio	1.85 / 0.130	✗
LLM Rank Optimizer (cross-model)	Ranking	Rank advantage	0%	✓
PoisonArena (GASLITE)	Retrieval	ELO	1978	✓
PoisonArena (content_poison)	Retrieval	ELO	307	✓

features spanning three platforms. RAG-Convincingness finds that 81.8% of queries contain conflicting stances, while Yes- and No-stance documents have nearly identical lengths (316.8 vs. 319.0 tokens), suggesting that LLM citation preference is driven by content properties other than length. When Attention Becomes Exposure confirms that top-10 creators receive 4.6× more citations than tail creators on Grok, a finding consistent with the Gini coefficient of 0.715 reported by Schulte et al. (2026).

Reproducible vs. black-box experiments. A consistent pattern across all three dimensions is the gap between what can be reproduced from open-source code and what has been demonstrated on live engines. All 20 projects support reproducible evaluation of their *evaluation* stages (scoring, ranking, statistical analysis). However, the *generation* stages of five projects (AutoGEO-Mini, StealthRank, LLM Rank Optimizer STS generation, PoisonArena LLM judge, AgentGEO search integration) require either a GPU-hosted LLM or a live third-party API unavailable without registration, and were reproduced using pre-computed results or partial pipelines. The most ecologically valid results—GASLITE's 100% attack success rate, CORE's 91.4% promotion rate (Jin et al., 2026), and AutoGEO's gains on live AI Overviews—were each obtained under conditions (specific model versions, corpus snapshots, or live engine states) that cannot be exactly replicated. We flag this offline–online gap as the central reproducibility challenge for the field and return to it as an open problem in Section 10.

6 Adversarial GEO: A Taxonomy of Attacks and Risks

The techniques that make GEO effective for legitimate content creators are dual-use: with a change of intent or a relaxation of quality constraints, the same mechanisms that earn a deserved citation can manipulate rankings, poison retrieval, fabricate authority, or suppress competitors. This section addresses RQ4 by organizing adversarial GEO into a taxonomy structured by the *layer of the generative-search pipeline* that an attack targets, analyzing the threat models involved, and—critically for a systematization—mapping the gray zone where benign optimization shades into manipulation.

6.1 A Layered Model of the Attack Surface

We model the generative-search pipeline as a stack of six layers, each of which constitutes a distinct attack surface:

1. **Content layer:** the text, structure, and metadata of a source document.

Table 10: Measurement and bias results (Dimension C). Key findings from benchmark and measurement-oriented projects under unified evaluation. "Platform" indicates the AI search system(s) studied. "Key metric" is the primary reported statistic.

Project	Platform	Key metric	Value
AIO Benchmark (Grossman et al., 2026)	Google AIO	AIO trigger rate (overall)	65.6%
		AIO trigger rate (ELI5)	94.6%
		AIO trigger rate (Amazon Retail)	17.4%
		AIO vs. SERP Jaccard	0.178
AI Product Bench	ChatGPT, Gemini, Claude	Top-1 consistency rate	3.8%
		Cross-model Jaccard	0.042
geo-citation-lab (Zhang et al., 2026b)	ChatGPT, Google, Perplexity	Citation records analysed	23,745
		Feature dimensions	72
RAG-Convincingness (Wan et al., 2024)	GPT-4	Queries with conflicting evidence	81.8%
		Yes vs. No token length diff.	316.8 vs. 319.0
CONFLICTBANK (Su et al., 2024)	Multiple LLMs	QA pairs (total)	553,117
		Answer option balance (A/B/C/D)	23.5 / 25.5 / 27.0 / 24.0%
When Attention Becomes Exposure (Alipour et al., 2026)	Grok, Perplexity	Top-10 vs. tail exposure (Grok)	4.6×
		Top-10 vs. tail exposure (Perplexity)	2.0×

2. **Retrieval layer:** the (often dense embedding-based) retriever that selects candidate sources from a corpus.
3. **Citation layer:** the mechanism that decides which retrieved sources are explicitly cited.
4. **Generation layer:** the LLM that synthesizes the answer and can itself be steered via injected instructions or poisoned training data.
5. **Ranking layer:** the ordering of products or sources in recommendation-style outputs.
6. **Agentic layer:** the multi-step browsing and tool-use trajectory of an autonomous search agent.

Table 11 maps representative attacks onto these layers together with their threat models, and we discuss each grouping below.

6.2 Content- and Ranking-Layer Manipulation

The most direct attacks inject adversarial text into a source the attacker controls in order to manipulate its selection or ranking. Kumar and Lakkaraju (2024) append a crafted *strategic text sequence* to a product page and show that it significantly raises the chance of the product becoming the LLM's top recommendation, framing GEO as the LLM-era analogue of SEO with the potential to distort fair competition. Pfrommer et al. (2024) formalize ranking in conversational search as an adversarial problem and use a tree-of-attacks jailbreaking technique to reliably promote low-ranked products, with attacks transferring to a production engine. Tang et al. (2025) introduce StealthRank, which uses energy-based optimization with Langevin dynamics to generate fluent adversarial strings (StealthRank Prompts) that boost a target's ranking while evading perplexity-based detection—explicitly trading off effectiveness against stealth. In the product setting, Jin et al. (2026) present CORE, a black-box method that appends string-, reasoning-, or review-based "optimization content" to steer output rankings, achieving a 91.4% promotion success rate at top-5 while preserving fluency. The benchmark study of Nimase et al. (2026) unifies these under a common evaluation, finding that black-box content rewriting matches or exceeds gradient-based attacks on rank promotion while producing more fluent, harder-to-detect text—and, strikingly, that the access model (black-box vs. white-box) does not predict attack strength.

6.3 Retrieval- and Generation-Layer Poisoning

A second family attacks the retriever or the generator's knowledge directly. Ben-Tov and Sharif (2024) present GASLITE, a gradient-based method that crafts adversarial passages which achieve high retrieval rank in dense embedding-based search at extremely low poisoning rates (as little as 10^{-4} of the corpus), generalizing across corpora and evading common defenses—a tightened, realistic SEO threat model. Chen et al. (2025a) study the *multi-adversary* setting via PoisonArena, showing that poisoning strategies effective against a single attacker degrade sharply under competition, exposing the inadequacy of single-attacker evaluation. At the deepest layer, Zhang et al. (2024) demonstrate *pre-training* poisoning: corrupting as little as 0.1% of pre-training data suffices for several attack objectives (including belief manipulation, the training-time analogue of product promotion) to persist through post-training alignment, with denial-of-service persisting at just 0.001%. These attacks are significant for GEO because they target the same web corpora that generative engines retrieve from and train on, blurring the line between content optimization and corpus poisoning.

6.4 Citation- and Agentic-Layer Attacks

The citation layer can be abused by fabricating the markers of authority—spurious quotations, fabricated statistics, or misattributed sources—that Wan et al. (2024) show LLMs find convincing, turning legitimate evidence-optimization (Section 4) into citation abuse. The multimodal attack of Du et al. (2026) extends manipulation to the citation/ranking of image-bearing content via coordinated image-and-text perturbations against VLM rankers. At the agentic layer, the trajectory-shaping technique of Ye et al. (2026), benign in intent, also describes an attack surface: an adversary who constructs a coordinated ecosystem of mutually reinforcing pages can manipulate what an autonomous agent encounters and concludes across its entire browsing trajectory. Zhao et al. (2026) note that as platforms move toward deterministic intent routing, the manipulation surface shifts to whoever controls the routing/brokerage layer—a forward-looking risk for agentic commerce.

6.5 Threat Models and the Benign–Adversarial Continuum

The attacks above span a spectrum of access assumptions: black-box content injection where the attacker controls only their own page (Kumar and Lakkaraju, 2024; Pfrommer et al., 2024; Jin et al., 2026); white-box or gradient access to a retriever or ranker (Ben-Tov and Sharif, 2024; Tang et al., 2025); and control over training data (Zhang et al., 2024). A recurring theoretical theme is that these dynamics constitute a *prisoner's dilemma*: Hu (2026) and Nestaas et al. (2024) both model competing publishers as incentivized toward manipulation, degrading answer quality for everyone in equilibrium.

The defining difficulty for a systematization is that there is no clean syntactic boundary between benign and adversarial GEO. Quotation and statistics addition (benign) becomes citation abuse when the quotes are fabricated; fluency rewriting (benign) becomes a stealthy adversarial suffix when optimized to evade detection while inverting true quality (Tang et al., 2025); evidence ecosystem construction (benign) becomes trajectory poisoning when the reinforcing pages are deceptive (Ye et al., 2026); and product-listing optimization (benign) becomes manipulation when it promotes an inferior product over a superior competitor (Kumar and Lakkaraju, 2024). The distinguishing variable is not the mechanism but the *intent and the effect on answer quality and fairness*—which is precisely why GEO cannot be governed by content filters alone and requires the answer-level, intent-aware analysis we take up in Sections 7 and 8.

7 Defenses and Platform Mechanisms

Addressing the first half of RQ5, this section surveys defenses against adversarial GEO and the platform mechanisms that govern visibility. The discussion follows the same six pipeline layers used in the attack taxonomy in Section 6. A governance layer then covers platform and ecosystem mechanisms that sit above the technical stack. A cross-cutting theme throughout is the *false-positive problem*: because benign and adversarial GEO share surface features—persuasive language, authority signals, structured evidence—defenses that catch manipulation also risk penalizing faithful optimization. Every layer must therefore be evaluated not only on attack reduction but on side effects to legitimate content.

Table 11: Taxonomy of adversarial GEO by target layer and threat model. Access: BB = black-box (attacker controls own content only), WB = white-box / gradient access, TD = training-data control.

Attack	Target layer	Access	Mechanism / key result
STS (Kumar and Lakkaraju, 2024)	Content→ranking	BB	Appended text sequence raises top-recommendation rate
Conv. ranking (Pfrommer et al., 2024)	Ranking/content	BB	Tree-of-attacks promotes low-ranked products; transfers to production
StealthRank (Tang et al., 2025)	Ranking/content	WB	Fluent, stealthy adversarial prompts via Langevin dynamics
CORE (Jin et al., 2026)	Ranking	BB	Optimization content; 91.4% promotion@top-5, fluency-preserving
GEO-Bench (Nimase et al., 2026)	Ranking	BB+WB	Unified benchmark; access model does not predict strength
GASLITE (Ben-Tov and Sharif, 2024)	Retrieval	WB	Adversarial passages, high rank at 10^{-4} poison rate
PoisonArena (Chen et al., 2025a)	Retrieval→gen.	BB	Multi-adversary poisoning; single-attacker metrics inadequate
Pre-train poison (Zhang et al., 2024)	Generation	TD	0.1% data poison persists through alignment
MGEO (Du et al., 2026)	Citation/ranking	WB	Coordinated image+text perturbation vs. VLM rankers
Trajectory shaping (Ye et al., 2026)	Agentic	BB	Coordinated page ecosystem steers agent conclusions

Content layer defenses. The central challenge at the content layer is distinguishing deceptive persuasion from honest persuasion when both can produce fluent, well-structured, authority-signaling text. The current state of the art, SCI-Defense (Yu et al., 2026c), addresses this through a multi-signal approach combining perplexity detection, semantic integrity scoring across authority attribution, narrative purposiveness, comparative claims, and temporal claims, and cross-candidate comparison—reporting strong results against string-, reasoning-, and review-based product attacks. The limits of this approach, however, are informative: relevance-based manipulation (use-case saturation) eludes detection, and stealth-optimized attacks (Tang et al., 2025) and fluent rewrites (Nimase et al., 2026) bypass perplexity- and keyword-based signals entirely. The implication is not that multi-signal detection is wrong in principle, but that single-signal baselines will be insufficient and that defenses must be evaluated on attack variants they were not trained to detect.

Retrieval layer defenses. Retrieval layer defenses face a fundamental asymmetry: a retrieval corpus can contain millions of documents, but adversarial passages need only outrank legitimate sources on the specific queries they target. GASLITE demonstrates that this can be achieved at poisoning rates as low as 10^{-4} of the corpus, generalizing across corpora and evading common defenses (Ben-Tov and Sharif, 2024). PoisonArena adds a further complication: strategies effective against a single attacker degrade under competition, meaning that single-attacker evaluation understates real-world robustness requirements (Chen et al., 2025a). These results point to a necessary combination of ranking robustness—making retrievers less sensitive to adversarial text—with corpus-level monitoring for anomalous newly indexed documents, rather than relying on local passage scoring alone. Neither component currently has a deployed, publicly evaluated solution in the GEO setting.

Citation layer defenses. Citation layer defenses examine whether an answer cites a source and whether the source supports the associated claim. Early commercial engines achieved citation recall of 51.5% and precision of 74.5%, showing that citation problems already exist without adversarial pressure (Liu et al., 2023). Citation verification and entailment checking test whether citations support the answer. Source diversity rules and diagnostics for unsupported attribution examine how engines select and use sources. Citation failure diagnostics identify why a source was or was not cited (Tian et al., 2026), while knowledge conflict benchmarks and evidence convincingness analyses assess whether an answer relies on manipulated or reliable evidence (Su et al., 2024; Wan et al., 2024).

Generation layer defenses. The generation layer is the deepest and hardest to defend: attacks that reach this layer have already survived retrieval and citation selection, and training-time poisoning can persist through post-training alignment even at contamination rates as low as 0.1% (Zhang et al.,

2024). This persistence result is significant because it implies that post hoc alignment is not a reliable defense against corpus-level manipulation—the corrupted behavior is encoded in the model weights before alignment is applied. Effective generation-layer defense therefore requires upstream provenance tracking and data curation rather than downstream instruction following, alongside generation-time checks—grounded decoding, conflict resolution, output consistency verification—that verify claims remain anchored to trusted evidence. No currently deployed system addresses all of these components together.

Ranking layer defenses. Ranking layer defenses must solve the same false-positive problem as content defenses, but with higher commercial stakes: a ranking defense that incorrectly flags legitimate product optimization harms smaller sellers disproportionately, since they have fewer alternative channels. SCI-Defense provides a domain-specific example for product descriptions, but its evaluated setting is narrower than the full range of ranking attacks in Section 6. More general defenses need to monitor promotion patterns over time, compare ranking changes against independent quality and relevance signals, and assess robustness under adversarial product descriptions—while also reporting false positive rates against legitimately optimized content and measuring differential impact across seller sizes.

Agentic layer defenses. Agentic defenses cover the full browsing trajectory because an attacker can influence the evidence an agent encounters at several steps (Ye et al., 2026). Feedback during search provides one form of control. NExT-Search uses user debug and shadow user modes to surface manipulation that reduces answer quality during the search process (Dai et al., 2025). System architecture provides another control point. AgentWebBench shows that decentralized agent systems can concentrate traffic toward a few gatekeeping sites, so architecture can increase or reduce exposure to manipulation (Zhong et al., 2026). Bias aware retrieval addresses fairness in knowledge retrieval (Singh and Ngu, 2025). Current work lacks a shared framework for evaluating these defenses, and multi step trajectories remain difficult to audit.

Defense evaluation costs. A defense's reported attack reduction is necessary but not sufficient for evaluating its utility. At every layer, defenses risk penalizing legitimate GEO: content filters may flag honest persuasion; retrieval hardening may favor already-prominent domains; citation verification may benefit sources with machine-readable evidence over those without. Evaluation should therefore report, alongside attack reduction: false positive rates on legitimately optimized content, coverage of attack variants not seen during development, transfer across engines and model versions, and effects on source diversity and smaller publishers. The absence of these figures from most current defense papers is itself an open problem.

Governance layer mechanisms. Technical defenses operating layer by layer cannot address manipulations that are only visible in aggregate patterns—ranking concentration over time, systematic citation bias toward commercial partners, or ecosystem-level traffic concentration. Governance mechanisms operating above the technical stack are therefore a necessary complement. Wen et al. (2026) propose four: answer-level contestability, high-precision disclosure of material commercial influence, black-box auditing by independent researchers, and deployment-aligned exposure metrics. These are not alternatives to technical defenses but address the residual risk that remains after technical defenses are in place—the manipulations that are individually legal, individually undetectable, but collectively harmful.

8 Security and Ethical Boundaries

The techniques that make content citable are the techniques that make citation manipulable. The method taxonomy of Section 4 sorts interventions by what they optimize rather than by whether the result stays accurate. It therefore puts two providers in the same category: one who improves a page so that it can be cited accurately, and one who engineers a page so that it is cited regardless of accuracy. The two edits can look the same on the page, and what separates them is the intent behind the edit and its effect on the answer. This section addresses the second half of RQ5, together with the part of RQ4 that concerns the overlap between attacks and benign optimization, and asks where that difference lies when it cannot be read off the content itself.

Table 12: GEO Behaviors Across Mechanistic Loci and Normative Classes

Mechanistic locus	Cooperative	Benign	Adversarial
Content / product	Faithful, utility-aware rewriting (Wu et al., 2025; Bardas et al., 2025)	Authoritative rewriting; structuring for scannability (Nimase et al., 2026; Wen et al., 2026)	Ranking manipulation through the document text (Tang et al., 2025)
Citation / ranking channel	Accurate, verifiable attribution (Bardas et al., 2025)	Authority signals; earned-media placement (Wen et al., 2026)	Citation-slot capture; discreditation (Mochizuki et al., 2026; Nestaas et al., 2024)
Corpus / ecosystem	Preserves source diversity and provenance	Wide adoption drifts toward zero-sum competition (Puerto et al., 2025)	Corpus poisoning; opinion manipulation, with retrieval collapse downstream (Zou et al., 2024; Chen et al., 2025e; Yu et al., 2026a)

Table 13: Criteria Separating the Three Classes of GEO Behavior

Criterion	Cooperative	Benign	Adversarial
Objective	Raise visibility while preserving answer utility	Raise visibility, with utility unconstrained	Change the outcome by misrepresenting relevance, authority, or standing
System impact	Utility preserved by construction	Utility may fall as a side effect	Typically degrades answer utility
Transparency	Publicly declarable	Publicly declarable	Typically concealed
Target audience	Reader and engine, reader value enforced	Reader and engine, no constraint on reader value	The engine, reader value optional
Competitor impact	Not targeted	Not targeted, but aggregate drift toward zero-sum	May be targeted directly

That difference is better read as a range than as a line, and both ends of it are documented. At one end, faithful rewriting raises visibility while holding verifiability fixed (Bardas et al., 2025; Wu et al., 2025). At the other, attacks reach both production engines and the corpora they retrieve from, changing which sources are recommended and what the resulting answer says (Nestaas et al., 2024; Zou et al., 2024; Chen et al., 2025e). What lies between them has no comparable record. It holds practices that are aggressive without fabricating anything, such as keyword stuffing or sentences structured for a model rather than for a reader (Wen et al., 2026), and it is that middle this section takes up.

8.1 A Taxonomy of GEO Behaviors

We separate two questions the literature tends to run together: what an intervention is trying to achieve, and where in the pipeline it acts. The first question gives a normative axis, running from cooperative optimization through benign visibility improvement to adversarial manipulation; the second gives a mechanistic one, separating interventions that take hold in the content itself, in the citation and ranking channel, or in the corpus from which future answers are drawn. Both axes are needed because manipulation does not arrive from outside the space of legitimate content; it is produced inside it, so one surface form can belong to any of the three classes depending on intent and effect. Table 12 shows why the two axes do not collapse into one: all three classes appear at every mechanistic locus, so no stage of the pipeline is inherently benign or inherently adversarial. The classes are therefore ideal types, not labels a page carries on its face. Table 13 sets out what would have to be established before an intervention could be assigned to one. Its cooperative and benign columns agree on transparency and differ in every other row, and the differences all follow from one criterion, whether a utility constraint is enforced, which is a property of the process that produced the text rather than of the text itself.

Cooperative optimization accepts a constraint that neither benign nor adversarial practice enforces: answer utility must survive the edit. AutoGEO (Wu et al., 2025) measures that utility as it rewrites and still reaches an Overall GEO score of 43.76 on Researchy-GEO with Gemini, against 31.20 for a hijack-attack baseline. The same constraint can be imposed at the document level instead of in the objective: Bardas et al. (2025) require edits to stay faithful to the source so that visibility rises without verifiability falling. The more telling case is the one where nobody imposes it: in the e-commerce testbed of Bagga

et al. (2025), 12 of 15 prompts still carry a factual-preservation constraint after meta-optimization, which the authors read as the optimizer converging on genuine improvement rather than exploitation. Three measurements are not a law, but none of them shows that keeping an answer useful costs visibility, so the pressure to drop the constraint comes from competition rather than from the method itself.

Benign visibility improvement leaves the utility constraint unenforced without adopting deception. Its methods are publicly defensible and target no competitor, but nothing in them requires the answer to stay useful or the claims faithful. This is not a residual category. Nimase et al. (2026) evaluate ten white-hat strategies in GEO-Bench, a security-oriented benchmark distinct from the similarly named visibility dataset of Aggarwal et al. (2024). Authoritative rewriting reaches a Normalized Rank Gain of 0.83 at a keyword violation rate of 0.00: openly describable rewriting can move rank while leaving nothing for a keyword detector to catch. What that does not establish is utility preservation. CC-GSEO-Bench (Chen et al., 2025c) separates the two questions, scoring exposure apart from faithful credit and from trustworthiness. The class also has a collective failure mode: Puerto et al. (2025) find ranking gains shrinking on C-SEO Bench as more actors adopt the same tactics, so benign competition drifts toward zero-sum without anyone acting in bad faith. That is what makes this the class that governance has to reach. A cooperative optimizer knows it accepted a constraint, and an adversarial one knows it did not, but a benign optimizer has no occasion to ask, and the harm it contributes to surfaces in the aggregate, where no single participant can see it.

Adversarial manipulation is defined here by mechanism and target rather than by intent, which is rarely observable. It exploits the retrieval or generation pipeline to misrepresent how relevant or authoritative a source is, or to damage a competitor's standing, and it addresses the model that allocates citations rather than the reader. Nestaas et al. (2024) give the foundational taxonomy under the name Preference Manipulation Attacks, separating prompt injection, discreditation of competitors through false negative claims, and plugin optimization. In their black-box experiments, these attacks raise recommendation rates on production Bing and Perplexity. Section 6 organizes the rest of the class by the pipeline layer each attack targets. Defining the class this way buys a handle at a price: mechanism and effect can be recovered after the fact, whereas intent cannot, so an intervention joins this class once its damage can be demonstrated, and not before. That is a workable rule for a survey and a poor one for a platform, which has to decide before the answer is served.

The three classes are not equally hard to identify. Cooperative practice can be certified by the party doing it and by no one else, because the constraint lives in the objective rather than in the output. Adversarial practice can be established from outside, but not before its effect exists. Benign practice is what remains when neither of those can be shown, so a category that is not residual in its methods is residual in its evidence.

8.2 The Boundary Problem

The obstacles between these classes and any procedure that would apply them are not all of one kind. Two of them are epistemic: what can be observed about a page does not carry the information the classification needs. The third is not, and better observation would not touch it.

Technical indistinguishability. The detectors most often proposed for this problem measure the wrong thing. Fluency and perplexity scores answer whether a passage is well formed, not whether it is honest, and being well formed is what any competent writer aims at. Legitimate GEO makes the mismatch sharper still, because improving fluency is itself one of the sanctioned optimization strategies, worth 15–30% in visibility in the evaluation of Aggarwal et al. (2024) (Section 2). A detector that flags fluent, low-perplexity text is therefore flagging the intended output of legitimate optimization, and its score carries no information about intent. Attacks exploit the gap: Tang et al. (2025) optimize adversarial text that preserves fluency and defeats perplexity- and keyword-based detectors, and Pfrommer et al. (2024) produce adversarial documents at a cosine similarity of 0.93 to the originals they displace, leaving embedding-similarity filters little to work with. The clearest case involves no evasion at all: the Authoritative rewriting result above conceals nothing and still moves rank. One benchmark cannot settle how general that is, but the argument does not rest on its magnitude. Detectability cannot ground a normative boundary when the feature being detected is an objective both classes share.

Goal duality. If the page cannot be read, the objective behind it is the natural place to look next, and it is no more available. AutoGEO (Wu et al., 2025) and the strategic text sequences of Kumar and Lakkaraju (2024) both optimize toward a visibility score, and they differ in a single term: the first also measures answer utility, the second does not. Neither of the two ways of establishing that term is open. It cannot be inferred from the result, because an answer that stayed useful is equally consistent with an

Table 14: Behaviors in the Gray Zone Between Benign and Adversarial GEO

Behavior	Why the class is ambiguous	Relevant work
Content structuring for scannability	Improves human readability, but may also be tailored to generative-engine formatting preferences	Wen et al. (2026), Wu et al. (2025)
Strategic text embedding via GCG	Can be framed as visibility optimization, but is driven primarily by ranking gain rather than search utility	Kumar and Lakkaraju (2024)
White-hat authoritative rewriting	Improves ranking without overt manipulation, yet does not by itself establish utility preservation	Nimase et al. (2026)
Earned media distribution	Places content in public information channels that may serve both informational and promotional purposes	Wen et al. (2026)

optimizer that was scored on utility and with one that was not. That second case is not a remote one: nothing in the measurements above shows that keeping an answer useful costs visibility, so an optimizer indifferent to utility will often leave it intact anyway, and a useful answer becomes an outcome the two classes share. Nor can the term be read directly. The objective is nowhere in the output, and a provider who states it is asserting something about a run that leaves no record in anyone else's hands. What an outside observer has is a page compatible with either class, and what would change that is not a closer look at the page but a record of the run.

Strategic incentives toward manipulation. Even a boundary that could be read reliably would not hold, because the payoff structure rewards crossing it. Nestaas et al. (2024) model competing providers as a prisoner's dilemma in which manipulation is the strictly dominant strategy in one-shot play, and Hu (2026) show that repetition does not by itself rescue cooperation: in their infinite-horizon game, cooperation survives under specific parameter conditions, and outside that region rational providers converge on manipulation whatever their own preferences. The cost side of the same calculus is falling. Kumar and Lakkaraju (2024) take Greedy Coordinate Gradient optimization from safety bypass to product ranking, and Liu et al. (2025) automate strategy discovery and carry the discovered strategies across models. Attacks that are cheaper to mount and that transfer between systems move the equilibrium further from the cooperative region. This obstacle differs from the first two in what would remove it. Better evidence, about the page or about the process behind it, would settle whether a given intervention is cooperative; it would not change what anyone stands to gain by not being. That lies beyond the reach of any classification scheme, however well evidenced.

Together the three mechanisms rule out the obvious governance move. No content-level test can certify a class that cannot be read off the page, cannot be read off the objective, and would not stay fixed even if it could be read, which is why the governance discussion turns to disclosure and answer-level contestability rather than to better detectors. That move answers the two epistemic obstacles and leaves the third alone, so the remainder is a question about incentives rather than about evidence. Table 14 collects the practices that sit on the line, keeping to those whose class is genuinely open rather than those that are hard to detect but not hard to classify.

8.3 Ethical and Societal Implications

Each harm below has the same shape: the party that gains from an optimization is not the party that pays for it. What changes across the four is who pays, and how readily they can tell that they are paying.

8.3.1 Fairness and Concentration

Concentration in generative search has two sources, and a small publisher can act on neither. One sits in the model: Pfrommer et al. (2024) find that GPT-4's brand rankings are shaped more strongly by its training prior than by the retrieval evidence in front of it, so an entity that was prominent when the model was trained stays prominent regardless of what the retriever supplies. The other sits in the citation channel, where Schulte et al. (2026) measure a Gini coefficient of citation distribution averaging 0.715 across engines. Audits of deployed systems keep finding the same asymmetry, in the sentiment and commercial tilt of generated answers, in preference for large brands, and in the way agentic architectures concentrate traffic further (Li and Sinnamon, 2024; Chen et al., 2025b; Alipour et al., 2026; Zhong et al., 2026). This is the distribution GEO acts on, before any optimization takes place.

GEO can push against that distribution as well as with it. Wen et al. (2026) argue that LLM-mediated recommendation can broaden exposure when a better but less known brand is surfaced over a famous incumbent, which would make optimization corrective rather than exploitative. The disadvantage being corrected is structural: Jin et al. (2026) show that an LLM's final recommendations remain largely determined by the initial retrieval order, so products buried in retrieval, often those of small businesses and independent creators, stay invisible in the synthesized answer whatever their listings say. Wu et al. (2025) report the corresponding upside, with AutoGEO's gains largest for low-visibility documents, raising the Overall metric from 9.46 to 35.83 on Researchy-GEO with Gemini. That gain is measured for one document at a time against a corpus nobody else is optimizing, which is the setting in which optimization is bound to look corrective.

The same machinery reinforces concentration just as well. The CORE method of Jin et al. (2026) lifts a bottom-ranked item to the top recommendation at an average Promotion Success Rate at Top-1 of 80.3%, which leaves the distributional meaning of GEO undetermined: correction and concentration run on the same technique, and which of the two happens depends on who can afford to run it repeatedly. Wen et al. (2026) make that point directly, arguing that well-resourced brands can sustain continuous GEO investment and crowd out smaller creators. Architectural correctives exist, such as the bias-aware retrieval agent of Singh and Ngu (2025) (Section 7), but they take no effect unless a platform chooses to deploy them.

Competition then compounds the resource gap. The prisoner's-dilemma dynamics discussed above assume symmetric players, but real participants differ in resources, technical capacity, and tolerance for short-term losses, and those differences decide who can stay in a reciprocal-optimization race at all (Nestaas et al., 2024; Hu, 2026; Puerto et al., 2025). A smaller actor is disadvantaged twice over, starting from lower visibility and being less able to sustain the continuous optimization that competitive pressure demands. The two readings of GEO are therefore not symmetric, and the evidence for them is not gathered under comparable conditions. The corrective reading holds for a single actor optimizing against a corpus at rest, which is what a benchmark measures; the concentrating reading is what the same technique does once everyone has it, which is what a benchmark of that kind cannot show. Nothing currently in the ecosystem rewards the first case over the second, so the default outcome is the one that needs no coordination.

8.3.2 User Trust and Information Integrity

User trust in generative search rests on two conditions: that a reader can tell earned visibility from manipulated visibility, and that answers stay faithful to the sources they cite. The first fails under measurement. Nestaas et al. (2024) report that 80% of external injection attacks are invisible to end users, adversarially optimized text passes both fluency and keyword checks (Tang et al., 2025), and Jin et al. (2026) find that human annotators judged review-style CORE content manipulated in 18.4% of cases against 12.0% for unmodified content, with fluency rated 4.6 and 4.7 on a five-point scale. The same invisibility appears in public-interest domains, where opinion manipulation reverses the stance of an answer without leaving easily detectable traces (Chen et al., 2025e). The one study here that put the question to human annotators found a margin too small for a reader to act on, so the first condition fails not at the edges but in the ordinary case.

The second condition fails even without an adversary. Averaged across four commercial generative search engines, Liu et al. (2023) find 51.5% of generated sentences fully supported by their citations, a measure they call citation recall; at much larger scale, Xu et al. (2026a) decompose Google AI Overviews into 98,020 verifiable atomic claims and find 11.0% inconsistent with the sources cited for them, most often through silent omission rather than contradiction. Manipulation does not have to open the gap between what an answer asserts and what its sources support. Widening one that is already there is enough.

A reader who cannot separate earned visibility from manipulated visibility still has one move left, which is to check the answer against the sources it cites. The second failure removes it: that check is unreliable with no adversary in the picture at all, so a manipulated answer has nothing to stand out from. The two conditions therefore do not fail side by side but in sequence, and the cost of both lands on a reader with no way to register that either has failed.

The stakes are highest in health and medicine, where a wrong answer costs more than a misranking, and the baseline there is weaker than tests of the models alone suggest. In a randomized study, Bean et al. (2026) find that the tested models identified the relevant medical conditions in 94.9% of cases when evaluated on their own, but that members of the public using those same models did so in fewer than

34.5% of cases, no better than an unaided control group. That failure lies in the interaction rather than in the model, which is the point: the baseline is already fragile before anyone manipulates the content feeding it.

Disclosure is the standard response, and its effect on users is untested. Wen et al. (2026) propose requirements analogous to the FTC's native advertising guidelines, but no study yet shows whether such a label helps a reader calibrate trust in a synthesized answer. The main proposed remedy for this harm therefore rests on an assumption about user behavior that nobody has checked.

8.3.3 Attribution and Citation Integrity

A citation in a generated answer does two jobs at once, and manipulation breaks them separately. It signals that the answer is supported, and it assigns credit and traffic to whoever produced the source. Wen et al. (2026) identify the opacity that undermines both, since a creator often cannot find out whether their work was used, why it was cited, or how to contest a misattribution. The evidentiary job fails on its own terms as well: Mochizuki et al. (2026) find that easily injectable sources such as community-editable platforms are cited often but weakly reflected in the answer text, and Nestaas et al. (2024) show that adversarial techniques can steer which sources a model cites at all. Once placement is manipulable and placement no longer tracks support, a citation functions less as evidence than as synthetic authority.

The credit job fails in money. Xu et al. (2026a) find that at least 50.6% of the pages cited in Google AI Overviews carry display advertising, a conservative floor because user-generated-content references went unanalyzed and were counted as ad-free. A reader who takes the synthesized answer without visiting the source removes the impression that would have paid for it, while the sponsored slots on the results page keep earning. The engine keeps the informational value and returns neither the traffic nor the revenue, which weakens the incentive to produce the material it depends on.

The two failures are mirror images. One set of sources gets placement without being used; another gets used without being paid. A citation is supposed to be a single act that both attests and credits, and in generative search the two halves have come apart in opposite directions, which is why fixing one of them does nothing for the other.

8.3.4 Information Ecosystem Integrity

The last harm has no identifiable victim, which is why nothing corrects it. An optimizer captures the rank and the citation share; what it spends is source diversity and evidentiary trust, and neither belongs to anyone in particular. Widespread optimization can therefore degrade the shared substrate that publishers and generative engines both draw on, while no participant faces a reason to stop.

The degradation is invisible while it is happening, which is what makes it hard to arrest. Yu et al. (2026a) call the process *retrieval collapse* and split it in two. First, fluent synthetic content captures the top retrieval positions and squeezes out source diversity: in controlled experiments on MS MARCO, a 67% contamination rate in the candidate pool produces 80.95% exposure contamination under BM25 and 79.98% under an LLM ranker, while surface-level answer accuracy holds steady. The system reports health at the point where provenance is already gone. Not until the second stage, when lower-quality or adversarial content follows the synthetic content in, do users meet the harm directly, with lexical baselines such as BM25 exposing them to roughly 19% harmful passages at the same contamination level. By the time the harm is visible the first stage has already been paid for, and nothing in the second stage returns the source diversity the first one consumed.

Contamination also feeds back through training. Shumailov et al. (2024) show that models trained on data produced by earlier generations of models can suffer *model collapse*, in which the tails of the original content distribution progressively disappear. That loop has already closed at the retrieval end: auditing four major generative search engines, Allaham and Diakopoulos (2026) classify roughly 16% of unique cited sources as AI-generated by the Pangram detector, ranging from 7.3% for ChatGPT to 27.8% for Copilot. Engines are already grounding answers in machine-produced text, and GEO adds to that text on purpose. Retrieval collapse and model collapse then feed each other, and each pass makes the previous one harder to undo.

Figure 3 traces the four harms back through the boundary problem to the surfaces they enter through. They are not four independent problems, since each persists for the same reason: optimized content stays fluent and useful, so no filter placed on the content can separate the harm from the ordinary case.

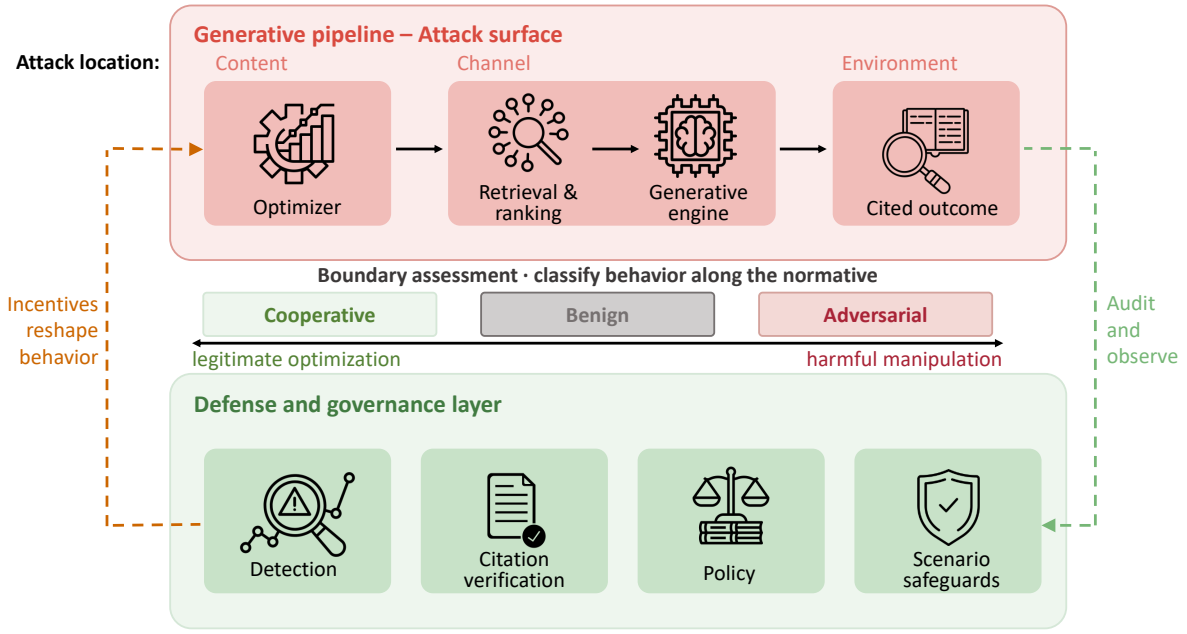

Figure 3: From GEO attack surfaces to governance responses. Ranking manipulation, corpus poisoning, citation exploitation, and jailbreak-derived optimization methods pass through a boundary problem in which benign and adversarial behaviors can appear technically similar. The resulting harms motivate platform defenses, regulatory clarification, and scenario-specific governance.

8.4 Governance and Regulation

If the class cannot be read from the content, governance is left with two handles: the process that produced the content, which its author alone can attest to, and the effect it had, which does not become available until after the fact. What follows asks how far platform mechanisms and existing law reach with those two, and why the answer changes with the deployment setting.

8.4.1 Platform Responsibility and Defense

The platform is the sole party that can see the whole pipeline, which makes it the natural site of intervention and the hardest one to check from outside. Proposals cluster in three places, filtering manipulated content, hardening the retrieval and generation architecture, and measuring how well a system holds up under attack (Wen et al., 2026), with GEO-Bench (Nimase et al., 2026) and a broader RAG security taxonomy (Xu et al., 2026b) filling in the last two. The first of the three has been tested most and works least.

The one purpose-built defense with published numbers succeeds and fails for the same reason. SCI-Defense (Yu et al., 2026c) scores a passage on four persuasion signals, namely authority attribution, narrative purposiveness, comparative claims, and temporal urgency, and on Amazon product descriptions it reaches a precision of 1.000 with recall between 0.830 and 1.000 depending on the attack style. Generic defenses do far worse, since perplexity filters, safety classifiers, and paraphrasing recall almost nothing against content that has to stay readable and so does not look like incoherent jailbreak text. But the same persuasion signals are largely absent from ordinary web passages, where recall against review-style attacks falls to near zero. What the detector is picking up is sales language, abundant in product copy and rare elsewhere, so the result does not travel past the setting that produced it.

Spending more on defense does not reliably buy less attack. In the model of Hu (2026) there is a region where added investment leaves attack probability unchanged and can raise it, because a better-defended slot is a more valuable one to capture. Defense effort is therefore not a quantity a platform can report as a safety record; what matters is where that effort sits relative to attacker incentives and to the competitive value of the visibility at stake. That is the argument for moving the intervention off the content and onto the answer, where Wen et al. (2026) locate contestability, disclosure, and deployment-aligned measurement. The move concedes something. Those measures do not stop manipulation; they make it answerable for afterwards, which is what remains once prevention has been shown not to work at the

Table 15: Governance Challenges Across Five GEO Application Scenarios

Scenario	Distinctive risk	Key evidence	Primary research gap
Generative search	Broad attack surface across domains, query types, and source-selection pathways	Production attacks on Bing and Perplexity (Nestaas et al., 2024); offline-to-online transfer (Pfrommer et al., 2024)	Platform incentives that favor cooperative optimization over adversarial competition
Agentic search	Manipulated recommendations may translate into delegated actions or purchasing decisions	Plugin selection is susceptible to adversarial influence (Nestaas et al., 2024)	Safe delegation protocols and agent-level defenses remain underdeveloped
E-commerce	Strong economic incentives to manipulate product ranking and recommendation exposure	GCG and StealthRank can be applied to product descriptions (Kumar and Lakkaraju, 2024; Tang et al., 2025); e-commerce testbed results (Bagga et al., 2025)	Legal status of ranking manipulation and the limits of market self-correction
Multi-agent systems	Responsibility may be diffused across agents, content providers, brokers, and platform operators	Game-theoretic models of adversarial dynamics (Hu, 2026); multi-agent attack automation (Liu et al., 2025); diffuse-responsibility analyses (Ye et al., 2026; Zhao et al., 2026)	N-player equilibrium conditions, multi-agent defense, and accountability allocation
Political and public-interest information	Manipulation may shape civic knowledge, public opinion, and perceived evidentiary authority	Black-box opinion manipulation can reverse generated stances (Chen et al., 2025e); political queries face elevated poisoning risk (Mochizuki et al., 2026)	Provenance safeguards, authenticity signals, and detection of coordinated opinion manipulation

level of the artifact.

8.4.2 Regulatory Status

Advertising and disclosure law regulates a transaction, and GEO involves none. The rules assume that someone paid someone else to place a message, so they attach the disclosure to the payment. A GEO practitioner buys nothing: they rewrite a page or seed an authority signal, and the engine takes it up on whatever terms the engine uses. Wen et al. (2026) put the same point as a gap in the FTC's native advertising rules, which have no counterpart for influence that is infrastructural rather than purchased. The difficulty is not that the conduct is hard to describe, but that the category the law reaches for, paid placement, is the one thing GEO does not involve.

Nothing at the policy level has closed that gap. The International AI Safety Report (Bengio et al., 2026) names manipulation of AI-mediated information systems as a cross-cutting risk without proposing a mechanism for it, so the problem stands recognized in general terms and untranslated into anything a generative search platform could be held to.

8.4.3 Governance Across Application Scenarios

What differs across deployment settings is less the risk than the question of who can be held to account for it. Generative search has one platform to answer for it; in agentic commerce the answer divides between the engine and whoever the agent transacts with; in a multi-agent pipeline it divides far enough that no party is left holding a duty anyone could enforce. Table 15 gives the distinctive risk, the evidence, and the open research question for each of five settings. Its evidence column is uniform in a way that matters: every entry demonstrates that the setting can be attacked, and none evaluates whether a governance measure holds up in it.

One threat does not respect that partition. The jailbreak methods discussed above carry across settings, so a rule written for product ranking leaves the same technique untouched when it is pointed at citation selection or plugin choice (Kumar and Lakkaraju, 2024; Liu et al., 2025). Scenario-specific governance is needed because the harms differ, and it is incomplete because the methods do not.

8.5 Open Problems

Five problems are specific to the security and governance of GEO, and Section 10 places them within the survey's wider agenda. Two are conceptual, asking what a classification could rest on and what the law should call the conduct. The other three are blocked on measurements nobody is taking: of the incentives a defense actually changes, of what a disclosure does to a reader, and of how the corpus degrades over time.

Classification criteria that do not rest on detection. No feature of a page separates the three classes (Tang et al., 2025; Nimase et al., 2026), so a workable classification has to rest on something other than the artifact. The candidates all lie in the process rather than in the artifact: the objective the optimizer was run against, the fidelity check it was constrained by, and what happened to answer quality afterwards. The first two can be attested but not observed; the third can be observed, but not before the answer has been served. That turns the research question from detection into evidence, namely what a provider would have to publish and what a platform would have to log for a claim of cooperative optimization to be checkable at all. GEO-Bench (Nimase et al., 2026), CC-GSEO-Bench (Chen et al., 2025c), and the utility-aware design of AutoGEO (Wu et al., 2025) each measure one piece, but they measure it for a researcher who already knows which condition produced the text, which is exactly the knowledge a governance regime lacks.

Empirical calibration of the defense paradox. The defense paradox of Hu (2026) is a model without numbers. Its four parameters have not been estimated in a live GEO setting: attack cost, detection probability, expected visibility gain, and discount factor. Without them a platform cannot tell in advance whether a proposed defense will deter manipulation, displace it, or make success worth more. Attack cost and expected visibility gain could in principle be estimated from outside with enough black-box auditing. Detection probability could not, since it is the platform's private information, so the parameter that decides whether more defense deters or instead raises the prize is the one an outside researcher cannot reach.

What disclosure does to a reader. Disclosure is where the governance proposals above tend to land, and none of them has been put in front of a user. The study that is missing is not hard to specify: show readers optimized and unoptimized answers, with and without a provenance label, and measure whether the label moves their confidence in the direction the evidence warrants. Without it the field cannot separate a disclosure that calibrates trust from one that reads as a stamp of approval.

Corpus-level measurement over time. Benchmarks in this survey mostly score a single query or a single document, while the harm that matters most accumulates across many of both. Nobody is tracking what fraction of the retrievable corpus is machine-tailored, how fast that fraction moves, or where it becomes self-reinforcing, even though synthetic sources already hold a measurable share of the material generative engines cite (Allaham and Diakopoulos, 2026) and the two collapse mechanisms above give a reason to expect that share to compound (Yu et al., 2026a; Shumailov et al., 2024). This is a longitudinal measurement problem rather than a modeling one, and it needs a monitored corpus and a time series that no current benchmark provides. It also has to begin before the phenomenon is well understood, because a corpus cannot be sampled retrospectively: whatever baseline goes unrecorded now is not recoverable later.

Legal status. The FTC Act, the EU AI Act, the Digital Services Act, and unfair competition law each reach some conduct that resembles GEO manipulation, and none has been mapped onto it. A mapping would have to settle a three-way question: when optimization is ordinary content strategy, when it triggers a duty to disclose, and when it becomes deceptive conduct. That is doctrinal work the technical literature cannot do for itself.

The five share one obstacle: governance is being asked to certify something the field has not yet made checkable. Until that changes, GEO will keep producing effective techniques faster than it produces ways to tell which of them a platform should allow.

9 GEO in Practice: Industry Perspectives

The preceding sections have analyzed GEO through the lens of academic research. This section addresses RQ6: how has the field developed in practice, and what positions have the major platform operators taken? Three developments define the practitioner landscape in 2025–2026: the emergence of a commercial GEO tooling industry, the first official platform guidance from Google, and the launch of agentic commerce protocols by OpenAI and Google that make AI-mediated product visibility economically

consequential in concrete, measurable terms.

9.1 The Commercial GEO Tooling Industry

The practitioner GEO market emerged rapidly in 2024–2025, driven by the same platform shifts that motivated academic study. By 2026, established SEO platforms had added GEO tracking features alongside dedicated GEO-native tools. Ahrefs added AI Overviews tracking in 2024, including a Brand Radar product that indexes over 100 million prompts across five AI platforms to measure citation rates and identify competitors' citation gaps.[1] Semrush extended its marketing platform with an AI Visibility Score that benchmarks brand mentions across engines and provides optimization recommendations.[2] Newer dedicated platforms—including Profound, OtterlyAI, and TopCited—focus specifically on multi-platform AI citation monitoring rather than traditional SEO metrics. TopCited[3] represents a case in which academic GEO research and practitioner tooling have developed in parallel within the same group: the platform integrates real-time citation monitoring across six AI search platforms (ChatGPT, Gemini, Grok, Perplexity, Google, Bing) with GEO content optimization and SCI-Defense, a detection module for ranking-manipulation attacks of the type studied in Section 6. SCI-Defense achieves Precision = 1.000 and False Positive Rate = 0.000 across three attack strategies and 720 evaluation instances (Yu et al., 2026c), and is integrated into the platform as a production feature at `/try-sci-defense`. TopCited reached organic position 3 on Google for the query "GEO tools AI" in July 2026 and is referenced in field resource lists as a dedicated GEO visibility platform—a demonstration that academic GEO research can directly inform deployed tooling within a compressed development timeline.

The academic literature provides context for evaluating these commercial claims. The low cross-platform citation consistency documented by Grossman et al. (2026) (AIO-vs-SERP Jaccard of 0.178) and by the AI Product Bench (cross-model Jaccard of 0.042) implies that a single "AI visibility score" aggregated across platforms may obscure platform-specific variation that matters for optimization decisions. The temporal instability documented by Schulte et al. (2026)—same-day Jaccard stability of 0.32–0.43—further complicates any visibility metric that relies on periodic sampling. These findings do not invalidate commercial monitoring, but they suggest that methodology transparency (which engines, which query set, how many repetitions, which collection dates) is essential for evaluating competing tool claims.

9.2 Google's Official Guidance

The preceding sections draw their evidence from outside the engines: audits and benchmarks run against systems whose retrieval and selection steps are not open to inspection. In 2026 the operators began to account for those steps in their own documentation. Google's account appeared on May 15, 2026: *Optimizing your website for generative AI features on Google Search*, a page for site owners under a new "Generative AI fundamentals" heading in Search Central.[4]

Google answers that the work of making a site more visible in AI search experiences is itself SEO. The guide treats AEO and GEO as names for that work: "From Google Search's perspective, optimizing for generative AI search is optimizing for the search experience, and thus still SEO." The reason it gives is architectural: AI Overviews and AI Mode are rooted in the core Search ranking and quality systems, reaching the Search index through retrieval-augmented generation and query fan-out. Eligibility therefore runs through Search, and a page must be indexed and eligible to be shown with a snippet before it can appear in a generative feature. One condition is specific to the AI features, in that a site must also be included in them in Search Console. Section 2 arrives at the same precondition from published measurements, so the platform's account and ours agree on this point.

The guide also lists what site owners can stop doing. Google states that its systems ignore `llms.txt` and similar files, that content need not be broken into small pieces, and that no particular writing style is needed for generative AI search; it adds that seeking inauthentic mentions helps less than it appears, since the generative features draw on both the core ranking systems and the systems that block spam. Structured data is called unnecessary for these features while still recommending it for the rest of SEO. The guide is equally direct about the monitoring tools described above: no third-party tool has access to

[1] https://ahrefs.com/blog/geo-generative-engine-optimization/

[2] https://www.semrush.com/blog/generative-engine-optimization/

[3] https://topcited.ai

[4] https://developers.google.com/search/docs/fundamentals/ai-optimization-guide, accessed August 14, 2026.

Google's ranking or AI systems, which makes any claim to report "internal" Google metrics false on Google's own account, although it stops short of telling site owners not to use them. All of this covers one operator and its own features and says nothing about ChatGPT, Perplexity, or Claude, so a tactic Google ignores may still matter elsewhere. What it does show is that practitioner advice in 2025–2026 and the platform's account of what it reads have diverged far enough to be worth testing.

Measurement followed. In June 2026, Google added Generative AI performance reports to Search Console, initially for a subset of UK site owners; these give impressions, but not clicks, for content appearing in AI features.[5] Microsoft had moved earlier, putting an AI Performance report into public preview in Bing Webmaster Tools in February 2026, which gives citation counts in Copilot and Bing AI answers together with the grounding queries behind them, and neither impressions nor clicks.[6] Both operators report from inside their own systems, across the full set of responses those systems serve, where an external audit works from samples. What they report is not the same quantity, however, and neither reports clicks, so AI features have no click-through rate and the two platforms cannot be placed on a common scale. In Section 10 we call for a canonical metric that subsumes impressions, citation rate, and rank change, so that findings from different sources can be compared. The platform reports sharpen that call rather than answering it: they are the most complete visibility data available, and they cannot be pooled across operators.

9.3 Agentic Commerce Protocols and Commercial GEO

The commercial stakes of product GEO increased substantially with the launch of agentic commerce protocols. OpenAI launched ChatGPT Instant Checkout in September 2025, initially with Etsy merchants and subsequently rolling out to over one million Shopify merchants including established consumer brands; the checkout flow is powered by Stripe's Shared Payment Token and processes through the open Agentic Commerce Protocol (ACP).[7] Google launched the Universal Commerce Protocol (UCP) on January 11, 2026, with Shopify, Etsy, Wayfair, Target, and Walmart as launch partners, alongside endorsements from American Express, Mastercard, Stripe, and Visa.[8]

These protocols create a direct link between AI citation and commercial transaction: a product that is not cited in a ChatGPT or Gemini response cannot be purchased through the agent interface, regardless of its search index position. This transforms AI citation from a visibility metric into a transactional precondition, with commercial consequences directly analogous to the revenue-per-rank-change figures estimated by Bagga et al. (2025) for e-commerce queries.

OpenAI has stated that product results in ChatGPT are not advertisements and not influenced by commercial partnerships, with ranking based on availability, price, quality, primary-seller status, and Instant Checkout participation.[9] This separates the ChatGPT commerce model from traditional paid-search advertising, but it introduces its own optimization surface: merchants have an incentive to ensure their product feeds are structured, complete, and protocol-compliant so that agent systems can parse and surface them. The academic study of what signals influence agent-mediated product selection—beyond the signals that influence static generative responses—remains largely open, as noted in Section 10.

9.4 The Academic–Practitioner Gap

The comparison reveals not a single divide, but a mismatch in objectives, observability, and evaluation timescales. Academic GEO has primarily asked whether controlled interventions improve visibility, citation, or ranking on static benchmarks and simulated pipelines. Practitioners, by contrast, must determine whether brands and products appear on live, continuously changing engines. Commercial tooling accordingly centers on monitored prompts, brand mentions, citations, share-of-voice, and cross-platform comparisons. These quantities are operationalized differently across vendors: Semrush's AI Visibility Score combines topic coverage and mention consistency, whereas Ahrefs constructs Brand Radar using its own prompt-selection and coverage methodology.[10] Practitioner scope is also expanding

[5] https://developers.google.com/search/blog/2026/06/gen-ai-performance-reports, accessed August 14, 2026.

[6] https://blogs.bing.com/webmaster/February-2026/Introducing-AI-Performance-in-Bing-Webmaster-Tools-Public-Preview, accessed August 14, 2026.

[7] https://openai.com/index/buy-it-in-chatgpt/

[8] https://developers.google.com/merchant/ucp

[9] https://openai.com/index/buy-it-in-chatgpt/

[10] Semrush, "Where does the data in Semrush's AI Visibility Toolkit come from?": https://www.semrush.com/kb/1607-semrush-ai-visibility-data; Ahrefs, "Brand Radar Methodology": https://ahrefs.com/blog/

from answer visibility to agentic-commerce readiness. Google's Universal Commerce Protocol (UCP), for example, requires participating merchants to configure protocol integration and provide product-feed eligibility and compliance data for agent-mediated checkout.[11]

Platform operators occupy a third position. Google exposes impressions for appearances in generative AI features through Search Console, whereas Bing reports total citations, cited pages, citation trends, and sampled grounding queries through Webmaster Tools.[12] Because the two platforms expose different units of observation, their first-party signals are not directly comparable and do not map cleanly onto vendor-defined visibility scores. Google also explicitly cautions that third-party tools do not have access to its internal ranking or AI systems, underscoring the distinction between platform observables and externally estimated visibility.[13]

The resulting gap reflects a trade-off rather than a failure on either side. Static benchmarks provide repeatability and controlled attribution, but gains measured in these environments may not transfer to proprietary, stochastic, and non-stationary engines. Live-platform monitoring provides greater deployment relevance, but its interpretation depends on the query set, sampling frequency, engine versions, and collection dates used by the monitoring system. Agentic commerce widens this mismatch further: success may depend not only on whether a source or product is cited, but also on whether the product is discoverable, eligible, protocol-compliant, and technically actionable within an agent-mediated transaction.[14] Bridging the gap therefore requires evaluating transfer explicitly: pairing reproducible benchmark results with dated and repeated live-engine audits, disclosing query and sampling protocols, and testing whether offline gains predict first-party impressions, citation counts, or downstream commerce outcomes. The central question is not which side measures GEO correctly, but which reproducible measures remain predictive of visibility in deployed systems.

10 Open Problems

We close by consolidating the open problems surfaced throughout this survey, organized by theme.

1. **A unified, engine-aware evaluation protocol and metric standardization.** The field lacks a standard protocol that reports engine versions and dates, mandates repeated measurement with variance (Schulte et al., 2026), and reconciles live-audit realism with static-benchmark reproducibility (Kim et al., 2026; Wen et al., 2026). Disentangling citation, absorption, and exposure (Zhang et al., 2026b; Alipour et al., 2026) into agreed metrics is a prerequisite. At least six partially overlapping metric families (visibility, citation, absorption, exposure, robustness, security) currently coexist without agreed operationalizations, making cross-study comparison unreliable. Specific open sub-problems include: (a) choosing a canonical primary metric that subsumes impression, citation rate, and rank change; (b) standardizing the repeated-measurement protocol across static benchmarks and live audits; (c) bridging the reproducibility gap between offline benchmarks (GEO-Bench, C-SEO Bench) and live-platform audits, whose results diverge in non-trivial ways; and (d) establishing a community benchmark analogous to GLUE/SuperGLUE for systematic leaderboard-style tracking of GEO method progress across engines and domains.

2. **Reproducibility under stochasticity and non-stationarity.** Closing the gap between irreproducible black-box audits of proprietary engines and reproducible offline study—the academic blind spot of Wen et al. (2026)—remains unsolved.

3. **Robust, generalizable defenses.** Current detectors are evaded by stealth-optimized attacks (Tang et al., 2025; Nimase et al., 2026) and miss relevance-based manipulation (Yu et al., 2026c); retrieval

brand-radar-methodology/, accessed August 20, 2026.

[11] Google Universal Commerce Protocol documentation: https://developers.google.com/merchant/ucp; Merchant Center product-data requirements: https://developers.google.com/merchant/ucp/guides/merchant-center, accessed August 20, 2026.

[12] Google, "Introducing Search Generative AI performance reports in Search Console": https://developers.google.com/search/blog/2026/06/gen-ai-performance-reports; Microsoft, "Introducing AI Performance in Bing Webmaster Tools Public Preview": https://blogs.bing.com/webmaster/February-2026/Introducing-AI-Performance-in-Bing-Webmaster-Tools-Public-Preview, accessed August 20, 2026.

[13] Google, "Optimizing your website for generative AI features on Google Search": https://developers.google.com/search/docs/fundamentals/ai-optimization-guide, accessed August 20, 2026.

[14] OpenAI, "Buy it in ChatGPT: Instant Checkout and the Agentic Commerce Protocol": https://openai.com/index/buy-it-in-chatgpt/; Google UCP checkout integration: https://developers.google.com/merchant/ucp/guides/checkout, accessed August 20, 2026.

poisoning largely defeats existing defenses (Ben-Tov and Sharif, 2024; Chen et al., 2025a); and training-time poisoning survives alignment (Zhang et al., 2024). Defenses that separate honest from deceptive persuasion, rather than persuasion from its absence, are needed.

4. **Formalizing the benign–adversarial boundary.** An intent- and effect-aware, operationalizable definition that distinguishes cooperative optimization from manipulation (Bardas et al., 2025; Kumar and Lakkaraju, 2024) would let both platforms and researchers reason about the gray zone (Section 8).

5. **Agentic and multi-agent GEO.** Visibility, attacks, and defenses at the level of multi-step trajectories and evidence ecosystems (Ye et al., 2026; Zhong et al., 2026; Zhao et al., 2026) are nascent; the shift of the manipulation surface to routing/brokerage layers is largely unexplored.

6. **Multimodal and cross-lingual GEO.** Beyond initial multimodal results (Du et al., 2026; Zhang et al., 2026a) and observations of cross-language instability (Chen et al., 2025b), optimization and defense for image, video, and multilingual content are underdeveloped.

7. **Fairness, concentration, and governance.** Whether GEO can be steered to counteract rather than amplify winner-take-most concentration (Schulte et al., 2026; Alipour et al., 2026), and how to operationalize answer-level disclosure and contestability (Wen et al., 2026), are pressing socio-technical questions.

8. **Reconnecting the feedback loop.** Restoring fine-grained, process-level feedback to generative search (Dai et al., 2025) so that quality-degrading manipulation is surfaced through user signals remains an open systems-design challenge.

10.1 Reproducibility Problem Checklist

Drawing on our systematic audit of 20 open-source GEO projects (Section 5.4), we consolidate the reproducibility barriers encountered into four categories. This checklist is intended as a practical reference for researchers attempting to reproduce or build on existing GEO work.

Category 1: Hardware and compute dependencies.

- **GPU-gated generation stages.** Five projects require a GPU-hosted LLM for their generation or training stages: AutoGEO-Mini (A100×2, 48h training), StealthRank (7B LLM generation), LLM Rank Optimizer (GCG adversarial suffix generation), PoisonArena (Llama-3-8B LLM judge), and AgentGEO (local retrieval model). Workaround: use pre-computed results where available, or substitute API-based LLMs for generation.
- **CUDA hard-coding.** GASLITE contained five hard-coded `.cuda()` calls that prevented CPU/MPS execution. Fix: replace with `.to(model.device)` throughout.
- **Framework version conflicts.** numpy/pandas binary incompatibility (numpy ABI mismatch) blocked several projects on macOS. Fix: `pip install --force-reinstall numpy pandas`.

Category 2: API and third-party service dependencies.

- **Search API requirements.** AgentGEO depends on ChatNoir (academic registration required); GEO/AEO Tracker requires six commercial engine API keys for full operation. Workaround: run in demo/degraded mode; search results absent but LLM optimization pipeline remains functional.
- **LLM API hard-coding.** FeatGEO's `core.py` hard-coded Gemini as the rewrite engine regardless of the `engine_llm` parameter. Fix: one-line patch to route by `engine_llm` parameter value.
- **Proprietary engine access.** Live-platform studies (AIO Benchmark, Kirsten et al.) cannot be reproduced without access to the same engine version used at collection time. No workaround; results are inherently non-reproducible beyond the data release.

Category 3: Data access barriers.

- **Manual download required.** RAG-Convincingness data is hosted on Google Drive and requires manual download; no programmatic access. CONFLICTBANK is 5.79 GB; streaming mode (`streaming=True`) is required to avoid full download.
- **Git LFS dependencies.** E-GEO uses Git LFS for data files; `git clone` without LFS produces empty stubs. Fix: `git lfs install && git lfs pull`.
- **Hard-coded local paths.** AgentGEO's `optimization_config.yaml` contained an absolute path (`/Users/erv1n/...`) for the disk cache directory. Fix: replace with relative `./cache`.

Category 4: Engine version and non-stationarity.

- **Model version drift.** Results obtained on `gpt-3.5-turbo-16k`, `gpt-4o`, or `gemini-2.5-pro` may not replicate on current default API endpoints, which silently update. Best practice: pin model versions via API parameters and report the exact model string.
- **Live engine non-stationarity.** Citation sets from live engines (Perplexity, Google AIO, Bing) change between runs due to index updates, personalization, and A/B testing. Same-day Jaccard stability is as low as 0.32–0.43 (Schulte et al., 2026). Best practice: report collection date, number of repeated queries, and variance.
- **Offline–online gap.** The most ecologically valid results were obtained on live engines under conditions that cannot be exactly replicated. Offline benchmarks (GEO-Bench, C-SEO Bench) provide reproducibility at the cost of ecological validity; live audits provide ecological validity at the cost of reproducibility. Both should be reported together where possible.

11 Conclusion

Generative Engine Optimization has emerged, in a remarkably short time, as the discipline governing visibility in an information ecosystem increasingly mediated by generative and agentic AI. In this systematization we traced the evolution from SEO to GEO across six dimensions of divergence, positioned GEO relative to RAG and generative, conversational, and agentic search, and organized the field's methods, evaluation practices, attacks, and defenses into coherent taxonomies, and set that literature against how GEO is practiced by vendors and platform operators. Two themes recur throughout. First, the object of optimization is steadily moving up the pipeline—from page text toward multi-step agent trajectories—tracking the growing autonomy of the engines themselves. Second, the techniques of benign optimization and adversarial manipulation are mechanistically inseparable, differing only in intent and in their effect on answer quality and fairness; this duality is the source of both the field's central evaluation difficulties and its most consequential governance challenges. We hope the taxonomies, comparisons, and consolidated open problems presented here provide a stable scaffold for a field that is otherwise moving faster than its own foundations.

References

P. Aggarwal, V. Murahari, T. Rajpurohit, A. Kalyan, K. Narasimhan, and A. Deshpande. Geo: Generative engine optimization. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 5–16, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704901. doi: 10.1145/3637528.3671900. URL https://doi.org/10.1145/3637528.3671900.

S. Alipour, M. Kargar, and M. Zihayat. When attention becomes exposure in generative search, 2026. URL https://arxiv.org/abs/2601.01750.

M. Allaham and N. Diakopoulos. Synthetic sources?: Auditing generative search engine citations for evidence of ai-generated sources, 2026. URL https://arxiv.org/abs/2605.23684.

P. S. Bagga, V. F. Farias, T. Korkotashvili, T. Peng, and Y. Wu. E-geo: A testbed for generative engine optimization in e-commerce, 2025. URL https://arxiv.org/abs/2511.20867.

N. Bardas, T. Mordo, O. Kurland, M. Tennenholtz, and G. Zur. White hat search engine optimization using large language models, 2025. URL https://arxiv.org/abs/2502.07315.

A. M. Bean, R. E. Payne, G. Parsons, H. R. Kirk, J. Ciro, R. Mosquera-Gómez, S. Hincapié M, A. S. Ekanayaka, L. Tarassenko, L. Rocher, and A. Mahdi. Reliability of llms as medical assistants for the general public: a randomized preregistered study. *Nature Medicine*, 32:609–615, 2026. doi: 10.1038/s41591-025-04074-y.

M. Ben-Tov and M. Sharif. Gasliteing the retrieval: Exploring vulnerabilities in dense embedding-based search, 2024. URL https://arxiv.org/abs/2412.20953. ACM CCS 2025.

Y. Bengio, S. Clare, C. Prunkl, M. Andriushchenko, B. Bucknall, M. Murray, R. Bommasani, S. Casper, T. Davidson, R. Douglas, D. Duvenaud, P. Fox, U. Gohar, R. Hadshar, A. Ho, T. Hu, C. Jones, S. Kapoor, A. Kasirzadeh, S. Manning, N. Maslej, V. Mavroudis, C. McGlynn, R. Moulange, J. Newman, K. Y. Ng, P. Paskov, S. Rismani, G. Sastry, E. Seger, S. Singer, C. Stix, L. Velasco, N. Wheeler, D. Acemoglu, V. Conitzer, T. G. Dietterich, F. Heintz, G. Hinton, N. Jennings, S. Leavy, T. Ludermir, V. Marda, H. Margetts, J. McDermid, J. Munga, A. Narayanan, A. Nelson, C. Neppel, S. D. Ramchurn, S. Russell, M. Schaake, B. Schölkopf, A. Soto, L. Tiedrich, G. Varoquaux, A. Yao, Y.-Q. Zhang, L. A. Aguirre, O. Ajala, F. Albalawi, N. AlMalek, C. Busch, J. Collas, A. C. P. de Leon Ferreira de Carvalho, A. Gill, A. H. Hatip, J. Heikkilä, C. Johnson, G. Jolly, Z. Katzir, M. N. Kerema, H. Kitano, A. Krüger, K. M. Lee, J. R. L. Portillo, A. McLysaght, O. Molchanovskyi, A. Monti, M. Nemer, N. Oliver, R. Pezoa, A. Plonk, B. Ravindran, H. Riza, C. Rugege, H. Sheikh, D. Wong, Y. Zeng, L. Zhu, D. Privitera, and S. Mindermann. International ai safety report 2026, 2026. URL https://arxiv.org/abs/2602.21012.

S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1):107–117, 1998. ISSN 0169-7552. doi: https://doi.org/10.1016/S0169-7552(98)00110-X. URL https://www.sciencedirect.com/science/article/pii/S016975529800110X. Proceedings of the Seventh International World Wide Web Conference.

L. Chen, X. Yang, Y. Lu, J. Zhang, X. Sun, Q. Liu, S. Wu, J. Dong, and L. Wang. Uncovering competing poisoning attacks in retrieval-augmented generation, 2025a. URL https://arxiv.org/abs/2505.12574. KDD 2026; benchmark named PoisonArena.

M. Chen, X. Wang, K. Chen, and N. Koudas. Generative engine optimization: How to dominate ai search, 2025b. URL https://arxiv.org/abs/2509.08919.

M. Chen, X. Wang, K. Chen, and N. Koudas. Navigating the shift: A comparative analysis of web search and generative ai response generation, 2026a. URL https://arxiv.org/abs/2601.16858.

Q. Chen, J. Chen, H. Huang, Q. Shao, J. Chen, R. Hua, H. Xu, R. Wu, R. Chuan, and J. Wu. Cc-gseo-bench: A content-centric benchmark for measuring source influence in generative search engines, 2025c. URL https://arxiv.org/abs/2509.05607.

X. Chen, H. Wu, J. Bao, Z. Chen, Y. Liao, and H. Huang. Role-augmented intent-driven generative search engine optimization, 2025d. URL https://arxiv.org/abs/2508.11158.

X. Chen, H. Zhou, J. Bao, H. Wu, Y. Gao, L. Pan, Y. Liao, and Z. Chen. A survey for generative search engine optimization, 2026b. URL https://www.researchgate.net/publication/404612480_

A_Survey_for_Generative_Search_Engine_Optimization. Preprint, April 2026.

Z. Chen, Y. Gong, J. Liu, M. Chen, H. Liu, Q. Cheng, F. Zhang, W. Lu, and X. Liu. Flippedrag: Black-box opinion manipulation adversarial attacks to retrieval-augmented generation models. In *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security*, CCS '25, page 4109–4123. ACM, Nov. 2025e. doi: 10.1145/3719027.3765023. URL http://dx.doi.org/10.1145/3719027.3765023.

N. Craswell, O. Zoeter, M. Taylor, and B. Ramsey. An experimental comparison of click position-bias models. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, WSDM '08, page 87–94, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781595939272. doi: 10.1145/1341531.1341545. URL https://doi.org/10.1145/1341531.1341545.

S. Dai, W. Wang, L. Pang, J. Xu, S.-K. Ng, J.-R. Wen, and T.-S. Chua. Next-search: Rebuilding user feedback ecosystem for generative ai search, 2025. URL https://arxiv.org/abs/2505.14680. SIGIR 2025 Perspective Paper.

K. Dritsa, T. Sotiropoulos, H. Skarpetis, and P. Louridas. *Search Engine Similarity Analysis: A Combined Content and Rankings Approach*, page 21–37. Springer International Publishing, 2020. ISBN 9783030620080. doi: 10.1007/978-3-030-62008-0_2. URL http://dx.doi.org/10.1007/978-3-030-62008-0_2.

Y. Du, C. Yu, H. Xu, Z. Wang, Y. Zhao, and X. Hu. Multimodal generative engine optimization: Rank manipulation for vision-language model rankers, 2026. URL https://arxiv.org/abs/2601.12263.

R. Grossman, S. Liu, M. K. Chen, M. Smith, C. Borcea, and Y. Chen. How generative ai disrupts search: An empirical study of google search, gemini, and ai overviews, 2026. URL https://arxiv.org/abs/2604.27790. ACM SIGIR 2026.

X. Hu. Dynamics of adversarial attacks on large language model-based search engines, 2026. URL https://arxiv.org/abs/2501.00745.

H. Jin, R. Chen, P. Zhang, Y. Luo, H. Zeng, M. Luo, and H. Wang. Controlling output rankings in generative engines for llm-based search, 2026. URL https://arxiv.org/abs/2602.03608.

S. Kim, W. Jeong, S. Kim, S. Lee, and D. Lee. Sageo arena: A realistic environment for evaluating search-augmented generative engine optimization, 2026. URL https://arxiv.org/abs/2602.12187.

E. Kirsten, J. G. Perdekamp, Q. Wu, M. Upadhyay, K. P. Gummadi, and M. B. Zafar. Characterizing web search in the age of generative ai, 2026. URL https://arxiv.org/abs/2510.11560.

A. Kumar and H. Lakkaraju. Manipulating large language models to increase product visibility, 2024. URL https://arxiv.org/abs/2404.07981.

A. Kumar and L. Palkhouski. Ai answer engine citation behavior an empirical analysis of the geo16 framework, 2025. URL https://arxiv.org/abs/2509.10762.

D. Lewandowski, S. Sünkler, and N. Yagci. The influence of search engine optimization on google's results: A multi-dimensional approach for detecting seo. In *Proceedings of the 13th ACM Web Science Conference 2021*, WebSci '21, page 12–20, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383301. doi: 10.1145/3447535.3462479. URL https://doi.org/10.1145/3447535.3462479.

A. Li and L. Sinnamon. Generative ai search engines as arbiters of public knowledge: An audit of bias and authority, 2024. URL https://arxiv.org/abs/2405.14034.

N. F. Liu, T. Zhang, and P. Liang. Evaluating verifiability in generative search engines, 2023. URL https://arxiv.org/abs/2304.09848.

X. Liu, P. Li, E. Suh, Y. Vorobeychik, Z. Mao, S. Jha, P. McDaniel, H. Sun, B. Li, and C. Xiao. Autodan-turbo: A lifelong agent for strategy self-exploration to jailbreak llms, 2025. URL https://arxiv.org/abs/2410.05295.

Z. Liu and P. Xu. Think before writing: Feature-level multi-objective optimization for generative citation visibility, 2026. URL https://arxiv.org/abs/2604.19113.

F. Lüttgenau, I. Colic, and G. Ramirez. Beyond seo: A transformer-based approach for reinventing web

content optimisation, 2025. URL https://arxiv.org/abs/2507.03169.

O. Martinez. Optimizing visibility in generative engines: A critical survey of generative engine optimization (2023-2026), 2026. URL https://arxiv.org/abs/2607.14035.

R. Mochizuki, S. Komatsu, S. Noguchi, and K. Ataka. Exposing citation vulnerabilities in generative engines, 2026. URL https://arxiv.org/abs/2510.06823.

F. Nestaas, E. Debenedetti, and F. Tramèr. Adversarial search engine optimization for large language models, 2024. URL https://arxiv.org/abs/2406.18382.

O. Nimase, Z. Chen, G. Qi, Y. Zhao, and X. Hu. Geo-bench: Benchmarking ranking manipulation in generative engine optimization, 2026. URL https://arxiv.org/abs/2605.29107.

S. Pfrommer, Y. Bai, T. Gautam, and S. Sojoudi. Ranking manipulation for conversational search engines, 2024. URL https://arxiv.org/abs/2406.03589.

H. Puerto, M. Gubri, T. Green, S. J. Oh, and S. Yun. C-seo bench: Does conversational seo work?, 2025. URL https://arxiv.org/abs/2506.11097. NeurIPS 2025 Datasets and Benchmarks Track.

J. Schulte, M. Bleeker, and P. Kaufmann. Don't measure once: Measuring visibility in ai search (geo), 2026. URL https://arxiv.org/abs/2604.07585.

I. Shumailov, Z. Shumaylov, Y. Zhao, Y. Gal, N. Papernot, and R. Anderson. The curse of recursion: Training on generated data makes models forget, 2024. URL https://arxiv.org/abs/2305.17493.

K. Singh and W. Ngu. Bias-aware agent: Enhancing fairness in ai-driven knowledge retrieval. In *Companion Proceedings of the ACM on Web Conference 2025*, WWW '25, page 1705–1712. ACM, May 2025. doi: 10.1145/3701716.3716885. URL http://dx.doi.org/10.1145/3701716.3716885.

Z. Su, J. Zhang, X. Qu, T. Zhu, Y. Li, J. Sun, J. Li, M. Zhang, and Y. Cheng. Conflictbank: A benchmark for evaluating the influence of knowledge conflicts in llm, 2024. URL https://arxiv.org/abs/2408.12076.

Y. Tang, Y. Fan, C. Yu, T. Yang, Y. Zhao, and X. Hu. Stealthrank: Llm ranking manipulation via stealthy prompt optimization, 2025. URL https://arxiv.org/abs/2504.05804.

Z. Tian, Y. Chen, Y. Tang, J. Liu, and R. Jia. Diagnosing and repairing citation failures in generative engine optimization, 2026. URL https://arxiv.org/abs/2603.09296.

A. Wan, E. Wallace, and D. Klein. What evidence do language models find convincing?, 2024. URL https://arxiv.org/abs/2402.11782. ACL 2024.

Y. Wen, N. Zhang, H. Yuan, X. Chen, H. Zhang, and H. Guo. Position: Generative engine optimization creates underexamined risks, governance must target concentration, disclosure, and academic blind spots, 2026. URL https://arxiv.org/abs/2606.12439.

B. Wu, F. Mao, J. Lin, C. Yang, J. Lu, Y. Guo, S. Zhang, Y. Wu, Y. Huang, and F. Li. From experience to skill: Multi-agent generative engine optimization via reusable strategy learning, 2026. URL https://arxiv.org/abs/2604.19516. ACL 2026 Findings.

Y. Wu, S. Zhong, Y. Kim, and C. Xiong. What generative search engines like and how to optimize web content cooperatively, 2025. URL https://arxiv.org/abs/2510.11438.

H. Xiong, J. Bian, Y. Li, X. Li, M. Du, S. Wang, D. Yin, and S. Helal. When search engine services meet large language models: Visions and challenges, 2024. URL https://arxiv.org/abs/2407.00128.

H. Xu, U. Iqbal, and J. M. Montgomery. Measuring google ai overviews: Activation, source quality, claim fidelity, and publisher impact, 2026a. URL https://arxiv.org/abs/2605.14021.

Y. Xu, M. Zhang, Z. Ge, H. Li, N. Hu, Y. Zhang, Z. Wen, J. C. Zhang, Q. Li, and L. Chen. Securing retrieval-augmented generation: A taxonomy of attacks, defenses, and future directions, 2026b. URL https://arxiv.org/abs/2604.08304.

N. Yagci, S. Sünkler, H. Häußler, and D. Lewandowski. A comparison of source distribution and result overlap in web search engines, 2022. URL https://arxiv.org/abs/2207.07330.

K.-C. Yang. News source citing patterns in ai search systems, 2025. URL https://arxiv.org/abs/2507.05301.

H. Ye, J. Mao, Z. Guan, and Z. Tian. Ecogeo: Trajectory-aware evidence ecosystems for web-enabled llm search agents, 2026. URL https://arxiv.org/abs/2605.12887.

H. Yu, D. Kim, and Y.-B. Kim. Retrieval collapses when ai pollutes the web. In *Proceedings of the ACM Web Conference 2026*, page 8745–8748. ACM, Apr. 2026a. doi: 10.1145/3774904.3792955. URL http://dx.doi.org/10.1145/3774904.3792955.

J. Yu, M. Yang, Y. Ding, and H. Sato. Structural feature engineering for generative engine optimization: How content structure shapes citation behavior, 2026b. URL https://arxiv.org/abs/2603.29979.

X. Yu, H. Jin, H. Zeng, and H. Wang. Sci-defense: Defending manipulation attacks from generative engine optimization, 2026c. URL https://arxiv.org/abs/2605.21948.

J. Yuan, J. Wang, Z. Wang, Q. Sun, R. Wang, and J. Li. Agenticgeo: A self-evolving agentic system for generative engine optimization, 2026. URL https://arxiv.org/abs/2603.20213.

F. Zhang, Q. Cheng, J. Wan, V. Singh, J. Rao, and K. Boakye. Generative engine optimization: A vlm and agent framework for pinterest acquisition growth, 2026a. URL https://arxiv.org/abs/2602.02961.

K. Zhang, X. He, and J. Yao. From citation selection to citation absorption: A measurement framework for generative engine optimization across ai search platforms, 2026b. URL https://arxiv.org/abs/2604.25707.

Y. Zhang, J. Rando, I. Evtimov, J. Chi, E. M. Smith, N. Carlini, F. Tramèr, and D. Ippolito. Persistent pre-training poisoning of llms, 2024. URL https://arxiv.org/abs/2410.13722.

X. Zhao, C. Li, X. Meng, K. Zhang, and X. Liu. Beyond retrieval: Modeling confidence decay and deterministic agentic platforms in generative engine optimization, 2026. URL https://arxiv.org/abs/2604.03656.

S. Zhong, K. Shen, and C. Xiong. Agentwebbench: Benchmarking multi-agent coordination in agentic web, 2026. URL https://arxiv.org/abs/2604.10938.

H. Zhou, J. Chen, X. Chen, J. Bao, Z. Chen, and Y. Liao. If-geo: Conflict-aware instruction fusion for multi-query generative engine optimization, 2026. URL https://arxiv.org/abs/2601.13938.

W. Zou, R. Geng, B. Wang, and J. Jia. Poisonedrag: Knowledge corruption attacks to retrieval-augmented generation of large language models, 2024. URL https://arxiv.org/abs/2402.07867.