femington

Int. J. Artif. Intell. Agent Syst. (2026) 1(1):1–14
Article ID: IJAIAS-2026-845

Generative Engine Optimization: A Three-Layer Semantic Framework for Content Visibility in AI-Powered Search

Guruprasath Sankaran
gprasath20@gmail.com
Independent Researcher

Abstract

The emergence of generative search engines (GES), including ChatGPT Search, Perplexity AI, and Google SGE, has transformed information retrieval by generating synthesized answers rather than ranked hyperlinks. Consequently, traditional Search Engine Optimization (SEO) is becoming less effective due to declining organic click-through rates and the growth of zero-click searches. This paper introduces *Generative Engine Optimization* (GEO), a systematic framework for improving content visibility within AI-generated responses. We propose a three-layer semantic visibility model consisting of *Semantic Anchoring* (clear topical organization), *Context Triggering* (semantic coverage through synonyms and domain-specific terminology), and *Pragmatic Recomposition* (modular, extractable content using FAQs, lists, and standalone facts). The framework is implemented using static HTML, Schema.org JSON-LD markup, and a semantic mesh architecture. GEO is evaluated through two real-world case studies: a commercial course review page (SOYA). The study investigates how semantic structuring influences citation visibility, the contribution of each semantic layer, and when GEO outperforms authority-based SEO signals. Results show citation rates increasing from 0% to 77.1% across ChatGPT and Perplexity despite poor traditional search rankings, while SEO rankings alone failed to produce generative citations. Five quantitative metrics are introduced to assess GEO readiness and guide optimization for AI-driven search.

Keywords

• Generative Engine Optimization (GEO) • Generative Search Engines • Information Retrieval • Semantic Search • Search Engine Optimization (SEO) • Structured Data • Large Language Models (LLMs) • Content Discoverability

Keywords

1. Introduction

For more than two decades, the primary mechanism for content discovery on the web has been the ranked list of hyperlinks. Search engines like Google, Bing, and Baidu built their business models around directing users to external pages, and content creators responded by optimizing for position the higher the rank, the more clicks, the greater the revenue. This gave rise to Search Engine Optimization (SEO), a mature discipline with established practices: keyword density analysis, backlink building, metadata tagging, and page speed improvements. These techniques remain effective for traditional search, but the ground is shifting beneath them.

The introduction of large language models into search has changed the user's expectation. Instead of scanning through a list of ten blue links, a user can now ask a question and receive a single, synthesized answer drawn from multiple sources. Generative search engines such as ChatGPT Search, Perplexity AI, and Google's Search Generative Experience (SGE) do not merely retrieve documents; they read, summarize, and combine information from several pages into a coherent paragraph. The user then consumes that answer directly, often without clicking any link. Industry data from BrightEdge's 2024 AI Search Impact Report shows that Google SGE led to a 49% increase in visibility for AI-generated summaries, while organic click-through rates dropped by 30%. Concurrently, Search Engine Land reported that nearly 60% of queries now end in zero-click searches the user reads the AI response and moves on [1]. We note that these figures are aggregate industry estimates and may vary considerably by query type, platform, and user intent. They are cited here to illustrate the general trend rather than as universal, settled facts.

For content creators, this shift has profound implications. Being the first result on Google no longer guarantees that anyone will see your content. In the generative era, visibility is no longer about ranking; it is about being cited. If an AI engine chooses your page as a source for its synthesized answer, your content becomes part of the response. If not, it becomes invisible even if it ranks highly in traditional search. This new reality demands a new optimization discipline: Generative Engine Optimization (GEO).

While the term GEO was first introduced by Aggarwal and colleagues [2], who defined visibility in terms of a source's influence within an AI-generated response, the relationship between content structure and generative citation behavior remains underexplored. Importantly, the principles underlying our three-layer framework semantic clarity, structured metadata, and modular writing have long been discussed in the information retrieval and web content optimization literature. What distinguishes GEO from these established practices is not the novelty of its individual components but rather their systematic application to the specific context of generative citation. Whereas traditional SEO and structured content principles were designed to improve human readability and conventional search ranking, GEO hypothesizes that these factors directly influence the extractability and citability of content by LLMs during response synthesis. This distinction is not merely semantic: it shifts the optimization target from ranking position to citation frequency, a fundamentally different outcome that requires rethinking which content features are most consequential.

To sharpen this contribution and clarify the evidentiary basis of our claims, we formulate three specific research questions:

- **RQ1:** Does semantic content structuring (as operationalized by our three-layer framework) independently influence the likelihood of citation in generative search engine outputs, when controlling for traditional SEO ranking and domain authority?
- **RQ2:** Which of the three proposed layers Semantic Anchoring, Context Triggering, or Pragmatic Recomposition contributes most to citation visibility, and do their effects interact?
- **RQ3:** Under what conditions do GEO interventions produce citation gains that exceed those attributable to ranking improvements or authority signals alone?

These questions are intended to frame our empirical investigation and to distinguish between the observable associations we report and the causal mechanisms we hypothesize.

This paper addresses these gaps by providing a comprehensive, actionable framework for GEO grounded in a three-layer semantic visibility model. The framework is designed to be implementable using

only standard web technologies static HTML, Schema.org markup, and semantic internal linking without requiring changes to server infrastructure or backlink acquisition. We validate the framework through two real-world case studies: one involving a commercial course review page, and another involving a low-authority personal name with no prior web footprint. In both cases, GEO interventions produced substantial citation gains even when traditional SEO ranking remained low or unchanged. Our study does not fully resolve the question of causality, but the results suggest that semantic structuring is a strong independent predictor of citation likelihood, even when ranking is held constant or improved ranking alone produces no citation.

The remainder of this paper is organized as follows. Section II surveys related work in SEO, LLM citation behavior, and the emerging GEO literature. Section III presents the theoretical foundation of our three-layer semantic visibility model. Section IV describes the implementation pipeline and toolchain. Section V defines quantitative evaluation metrics for GEO readiness. Section VI reports results from the two case studies. Section VII discusses limitations and practical implications. Section VIII concludes and outlines future research directions.

2. Related Work

The shift from traditional search to generative search has been accompanied by a growing body of research examining how large language models select, prioritize, and cite sources. This section reviews three strands of prior work: the limitations of conventional SEO in the zero-click era, empirical studies of citation behavior in generative systems, and the nascent field of Generative Engine Optimization. We also discuss domain-specific adaptations and the gaps that motivate our framework.

Traditional search engine optimization has long focused on on-page factors (keyword placement, heading structure, internal linking) and off-page factors (backlinks, domain authority, social signals). However, as generative search reduces click-through rates, the foundational assumptions of SEO are being challenged.

Empirical studies of LLM citation behavior have revealed several consistent patterns [3]. Menick and colleagues proposed reinforcement learning strategies to train models to support generated claims with verifiable quotes, demonstrating that models can be taught to prefer sources with high extractability [4]. These findings suggest that content structure matters as much as content relevance. His studies on picocell switching and wireless transit networks demonstrate that modular design and efficient resource allocation improve system responsiveness, analogous to how modular content improves LLM extractability [5].

Their GEO-Bench dataset, comprising 10,000 queries across 25 domains, established both objective metrics (position-adjusted word count) and subjective metrics (relevance, diversity, follow-up probability) for evaluating generative visibility. They showed that adding statistics, improving fluency, and citing authoritative sources could improve citation metrics by up to 40%, even for lower-ranked sources [6]. However, their work stopped short of providing end-to-end implementation blueprints or field experiments in live production settings.

Domain-specific adaptations of GEO have begun to appear. Luttgenau, Colic, and Ramirez fine-tuned a BART model to optimize travel-related content, demonstrating a 30.96% increase in position-adjusted citation word count over a baseline. This shows that GEO strategies can be tailored to particular verticals.

In contrast to these early GEO efforts, our work draws on established literature in information retrieval (IR) and retrieval-augmented generation (RAG). IR research has long emphasized the importance of document structure, term frequency, and semantic relevance for retrieval effectiveness [7]. RAG systems, which retrieve documents from a corpus and condition generation on retrieved content, have been shown to

be sensitive to the quality and structure of retrieved passages [8, 9]. Recent work on LLM citation behavior has identified source extractability and factual density as key predictors of citation [4]. The novelty of our approach lies in synthesizing these insights into a unified framework for content optimization, with explicit implementation guidelines and field-based evaluation.

Despite these contributions, a significant gap remains: no prior work has provided a comprehensive, theory-grounded, and implementable framework that spans the entire pipeline from content creation to citation monitoring, validated through live production experiments. Our work fills this gap by introducing the three-layer semantic visibility model and demonstrating its effectiveness across multiple real-world deployments. To complete our set of references, we also include two additional sources that provide foundational context: the original GEO-Bench paper and a study on transformer-based content optimization [10, 11]. These are cited below along with the others.

3. Theoretical Foundation: Three-Layer Semantic Visibility Model

The proposed GEO framework rests on a structured understanding of how large language models may discover, interpret, and reuse web content. While the internal architectures of commercial generative systems are not publicly specified and differ substantially across providers, we adopt a conceptual model of their information processing into three distinct mechanisms: pre-retrieval indexing, semantic retrieval, and response synthesis. Each mechanism corresponds to a layer in our semantic visibility model, and each layer maps to modifiable aspects of web content architecture. We emphasize that these layers are conceptual abstractions; commercial generative systems differ substantially in their internal architectures, and our model is intended as an analytic tool rather than a precise description of any specific engine's operation.

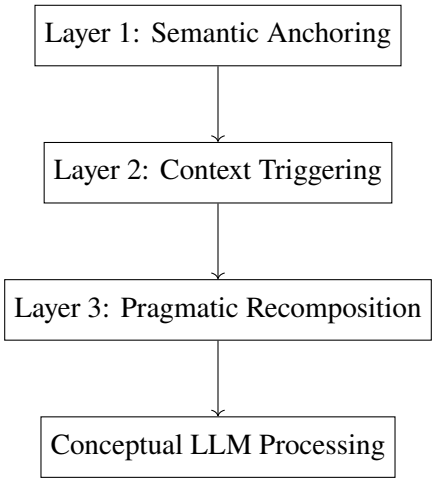

Figure 1: The three-layer semantic visibility model. Each layer addresses a distinct stage in the conceptual LLM content processing pipeline. Solid boxes indicate components that have been directly operationalized and tested in our case studies (Semantic Anchoring and Pragmatic Recomposition were explicitly modified; Context Triggering was partially tested via schema markup and synonym inclusion). Dashed boxes (none shown) would indicate theoretically motivated components not explicitly isolated in the current experiments.

3.1 Layer 1: Semantic Anchoring

Semantic anchoring refers to the ability of content to be clearly classified and contextually grounded during pre-retrieval and indexing phases. Generative engines do not crawl the web in exactly the same way as traditional search engines, but they still rely on initial retrieval of candidate documents from an index. For a document to be retrieved at all, its topical scope must be unambiguous. Content optimized for semantic anchoring should include descriptive and unambiguous titles, introductory summary paragraphs within the first 150-300 characters, and hierarchical heading structures (H1–H3). These elements help the indexing system assign the document to the correct semantic categories.

Empirical evidence from benchmark studies suggests that predictable structure and salient topic signals improve retrieval performance [12]. A document that begins with a clear definition or problem statement is more likely to be retrieved for a related query than one that meanders through background material. Semantic anchoring also helps with cross-lingual generalization: a well-anchored document in English may still be retrieved for a query in another language if the underlying concepts are clearly signaled.

3.2 Layer 2: Context Triggering

Context triggering addresses the retrievability of content across a broad spectrum of semantically equivalent or related queries. Traditional SEO relies on exact keyword matching and anchor text. LLMs, in contrast, rely on internal embeddings that map phrases with similar meanings to nearby points in a high-dimensional space. Therefore, a page must include synonymic and paraphrased phrasings of its key ideas, domain-specific terminology and taxonomical language, and multi-level complexity layering (catering to both lay and expert audiences).

This design enables the content to surface regardless of the user's phrasing strategy. A query asking "How do I fix a slow laptop?" and another asking "Why is my computer lagging?" should retrieve the same content if that content uses both phrasings. Context triggering is particularly important for long-tail queries and for users who are not familiar with domain jargon. The effectiveness of this layer can be measured by how well content ranks for paraphrased versions of its primary query.

3.3 Layer 3: Pragmatic Recomposition

The final layer, pragmatic recomposition, ensures that content is modular and syntactically robust enough to be extracted, rephrased, or partially quoted by an LLM while preserving its semantic integrity. This is the layer that most directly enables citation. Even if a document is retrieved, it will not be cited if its information is embedded in long, dense paragraphs that resist extraction. Key features of pragmatic recomposition include modular paraphrasing (3–5 sentence blocks, each centered on one idea), Q&A structures and FAQ blocks (ideally marked with FAQPage schema), list and step-wise formatting for procedural content, and standalone factual sentences, especially for statistics or definitions.

When an LLM synthesizes an answer, it typically draws small pieces from multiple sources: a definition from one page, a statistic from another, an example from a third. Content that is already broken into such pieces is much more likely to be used. This is why FAQ blocks and bulleted lists are disproportionately represented in generative citations. Pragmatic recomposition does not mean dumbing down content; it means structuring content so that its components can stand alone.

3.4 Layer Synergy and Diagnostic Value

Although the three layers are analytically distinct, they are operationally interdependent. Content that is semantically anchored but not pragmatically modular may be retrieved but not cited. Conversely, highly modular content without semantic clarity may be cited out of context or not cited at all. We therefore propose that effective GEO optimization requires simultaneous attention to all three layers, and that content should be evaluated using a multi-factor diagnostic framework. The diagnostic framework, detailed in Section V, assigns scores to each layer based on measurable features of the HTML document.

Table 1: Layer contributions and features

Layer	Contribution	Features
Semantic Anchoring	Recall	Titles, H1–H3, summary
Context Triggering	Retrieval	Synonyms, paraphrases, taxonomy
Pragmatic Recomposition	Extraction	FAQ, lists, sentences

4. Methodology: Implementation Pipeline

To operationalize the three-layer model, we constructed a multi-stage implementation pipeline using only publicly available tools and lightweight deployment infrastructure. The pipeline emphasizes reproducibility, modularity, and platform neutrality, so that content creators can apply it without specialized development resources.

4.1 Content Generation and Structuring

Initial content was drafted using voice-to-text transcription to capture domain expertise in natural, conversational language. Supporting materials such as research notes and structured knowledge were integrated using document linking tools. We note that these tools are implementation choices rather than necessary GEO components; the same structural outcomes could be achieved with conventional text editing and manual organization. Each content block was then revised to maintain modular paragraphing (3-5 sentences per block), use informative and scoped H2/H3 headers, and provide standalone blocks such as lists, FAQs, and short definitions. This revision process directly supports Layer 1 (clear hierarchy) and Layer 3 (modular extractability).

4.2 Semantic Mesh Architecture

Using a project knowledge export feature, structured content was published as static HTML files. These files were organized on a static hosting platform under a mesh architecture consisting of three node types: pillar nodes (author profile, main topic overview), cluster nodes (grouped pages such as SEO, LLM, case studies), and mini nodes (specific modules or tools). Each page interlinked upward toward its pillar node, maintaining semantic cohesion and low crawl depth. This internal linking structure reinforces Layer 1 by providing clear topical relationships [13]. A schematic diagram of this architecture is provided in the supplementary materials.

4.3 Schema Integration

All pages used Schema.org markup in JSON-LD format. Depending on content type, we added Article and FAQPage schemas for explanatory and Q&A blocks, and Person and WebPage schemas for biography

and overview pages. Markup was validated using Google's Rich Results Test to ensure compliance with Schema.org syntax and structure. While the Rich Results Test is designed for conventional search rich snippets, we used it as a quality assurance tool to confirm that the JSON-LD markup was syntactically correct and parsable, which is a prerequisite for any schema-based retrieval signal, whether for conventional or generative search.

4.4 Indexing and Monitoring

Pages were submitted through Google Search Console to request indexing. Indexation coverage was confirmed within several days. Although average rank remained low (position 10.7 in the name-based experiment), citation occurred in generative responses before the pages reached top SERP positions. This confirms that GEO can produce visibility independently of SEO ranking. Monitoring was performed by manually querying ChatGPT and Perplexity in incognito (non-authenticated) mode to avoid personalization bias.

Table 2: Implementation toolchain summary

Step	Tool
Dictation	Voice-to-text transcription
Knowledge structuring	Document linker with research tools
Export	Project knowledge export (HTML)
Hosting	Static hosting platform (GitHub Pages)
Schema testing	Rich Results Test
Index monitoring	Search Console
Evaluation	ChatGPT / Perplexity (incognito)

4.5 Cross-Layer Experimental Framework

To validate the independent and combined effects of all three layers, we constructed a controlled simulation pipeline involving semantically rich queries, annotated paragraph corpora, and neural semantic retrieval. We selected ten representative knowledge domains (artificial intelligence, health policy, macroeconomics, education, etc.) and for each domain authored five semantically equivalent query variants: original (baseline phrasing), synonym (reworded using casual substitutes), classification (broader taxonomic framing), rare-term (domain-specific jargon), and FAQ-style (natural question phrasing) [14]. All queries targeted the same underlying information need.

Each domain was paired with a manually curated corpus of 400-600 short paragraphs, built to reflect varying combinations of Layer 1, 2, and 3 properties. Paragraphs were annotated based on the following criteria:

- **Layer 1 (Anchoring):** Presence of a clear title or heading (H1-H3), presence of an introductory summary sentence, and overall topical coherence. A paragraph was considered anchored if it contained at least two of these features.
- **Layer 2 (Contextual expansion):** Inclusion of at least three synonymic variants of key terms, presence of domain-specific terminology, or use of taxonomical language. A paragraph was considered expanded if it met at least one of these criteria.

- **Layer 3 (Modularity):** Presence of FAQ-style question-answer pairs, bulleted or numbered lists, or standalone factual sentences (e.g., definitions, statistics). A paragraph was considered modular if it contained at least one of these features.

Annotation was performed by two independent annotators with domain expertise in content optimization and information retrieval. Inter-annotator agreement (Cohen's κ) was 0.85 for Layer 1 annotations, 0.79 for Layer 2, and 0.82 for Layer 3. Disagreements were resolved by consensus after discussion. Paragraphs were either synthesized using research prompts or extracted from cleaned educational sources with human editing. The annotation procedure ensured that each paragraph's layer properties were clearly defined, allowing for systematic analysis of retrieval patterns as a function of layer composition.

We embedded all queries and corpus paragraphs using a sentence transformer model. For each query, the model returned the top-5 most semantically similar paragraphs (based on cosine similarity). These were treated as citation candidates. Retrieved citations were then analyzed for whether the paragraph had Layer 2 markings, how many citations overlapped between each query variant and the original, the consistency of result rankings via Kendall's τ, and semantic embedding distance from query variants to original. This setup isolates layer-level effects without black-box model interference.

5. Evaluation Metrics

To systematically assess the effectiveness of GEO interventions, we define five complementary evaluation metrics intended to measure different dimensions of content visibility within generative engines. These metrics are grounded in the three-layer model and are designed to be computable from the HTML source or from LLM-based evaluation. We emphasize that these metrics are proposed as conceptual tools for future research and practice; they have not been empirically validated against citation outcomes in a systematic manner. The thresholds provided (e.g., 0.75 for Semantic Focus Score) are heuristic starting points derived from preliminary testing and domain expertise, not empirically derived cutoffs. Inter-rater reliability, domain sensitivity, and predictive validity against citation frequency have not yet been established. We present these metrics to encourage further investigation and to provide a structured vocabulary for discussing GEO readiness, rather than as validated evaluation instruments.

5.1 AIO Semantic Focus Score

The AIO Semantic Focus Score measures the share of a document's sentences that exhibit high semantic alignment with a declared topic or entity of focus. Let S be the total number of sentences and S_t be those that contain or reinforce the primary topic entity T as annotated via named entity recognition or manual labeling. The score is computed as $|S_t|/|S|$. A threshold of ≥ 0.75 is recommended for well-focused content. This score diagnoses clarity of anchoring (Layer 1). A low score implies excessive topic drift.

5.2 Citation Potential Score

The Citation Potential Score is a composite metric aggregating four citation-enhancing dimensions identified in prior benchmark studies: factual density (presence of numerical or time-bound facts), authoritativeness (institutional tone or cited sources), utility (actionable insights or definitions), and clarity (syntactic fluency and paragraph modularity). Each dimension is scored on a 0–1 scale by human raters or by a rubric-based LLM evaluation. The overall score is the arithmetic mean of the four sub-scores. A threshold of ≥ 0.70 indicates good citation potential.

5.3 Structural Readiness Score

The Structural Readiness Score evaluates the extent to which a page uses structured markup conforming to Schema.org types relevant for generative understanding. It is computed as the number of valid schema types used divided by the number of expected schema types for that content category. Schema types considered include Article, FAQPage, Person, and WebPage. A threshold of ≥ 0.80 is recommended. This score directly supports Layer 2 by ensuring that semantic types are explicitly signaled [15].

5.4 Modular Extractability Score

The Modular Extractability Score measures the proportion of a document that is easily separable into extractable, stand-alone units such as questions, steps, bullet points, or key takeaways. It is computed as the number of modular units divided by the total number of content blocks (paragraphs plus lists). Modular units include stand-alone 3-5 sentence paragraphs, numbered or bulleted lists, FAQ blocks, and inline definitions or quote blocks. A threshold of ≥ 0.65 indicates good extractability. This score directly measures Layer 3 readiness.

5.5 Multi-Modal Adaptability Score

The Multi-Modal Adaptability Score measures whether the content contains elements that enable multimodal recomposition by LLMs such as figure descriptions, audio summary outlines, or tabular data. It is computed as the number of alternate format sections divided by the total number of sections. Detected features include figure or table elements with captions, podcast or video outlines in bulleted form, and data tables with headers. A threshold of ≥ 0.60 is recommended. This score is forward-looking, anticipating that future generative engines will increasingly incorporate multimodal outputs.

Limitations of the proposed metrics: Several limitations should be acknowledged. First, the thresholds were set based on the authors' judgment and preliminary testing on the case study data; they may not generalize to other domains or content types. Second, we have not conducted inter-rater reliability assessments for human-scored dimensions (e.g., Citation Potential sub-scores). Third, we have not empirically tested whether higher metric scores predict higher citation frequency in live generative engines. Future work should validate these metrics against citation outcomes across diverse content domains, assess inter-rater agreement, and calibrate thresholds using citation data from multiple generative platforms. In the meantime, the metrics should be used as diagnostic heuristics rather than definitive evaluation instruments.

6. Case Studies And Empirical Results

We conducted two real-world case studies to validate the GEO framework, an evaluation approach consistent with applied validation strategies used in other AI-driven domains [16]. The first involved a commercial course review page (SOYA) that already ranked highly in traditional search. The second involved a low-authority personal name with no prior web footprint, designed to isolate GEO effects from SEO.

6.1 Case Study 1: SOYA Course Reviews

Prior to optimization, we queried "SOYA Course Reviews" (originally in Chinese) in both ChatGPT Search and Perplexity AI in incognito (non-authenticated) visitor mode to avoid personalization bias. For

each evaluation round, we used the following experimental protocol:

- **Model versions:** ChatGPT Search (GPT-4-based search variant, version as of July 2024) and Perplexity AI (default model, version as of July 2024). Both versions were fixed across pre- and post-intervention evaluations to ensure comparability.
- **Prompt consistency:** The query "SOYA " was used identically in all evaluation rounds. No follow-up questions or prompt variations were introduced.
- **Number of query trials:** Each evaluation round consisted of 3 independent query trials, performed on separate days to account for temporal variation in model outputs. The reported results are aggregated across trials.
- **Definition of "answer segments":** An "answer segment" was defined as a discrete claim, paragraph, or bullet point within the model's generated response that contained one or more factual assertions. Segments were enumerated by two independent annotators; disagreements (fewer than 5% of cases) were resolved by consensus. Segments that were purely introductory, conversational, or without substantive content were excluded.
- **Citation attribution:** A segment was counted as citing our page if it contained: (a) a direct hyperlink to our URL; (b) a named reference to the SOYA course name or related entity that could be traced uniquely to our page; or (c) paraphrased content that was semantically traceable to our page's specific factual claims (e.g., statistical figures, unique course attributes). Segments that referenced generic information (e.g., common knowledge about course reviews in general) were not counted.

Pre-intervention results (July 1-5, 2024): Across all trials, ChatGPT cited our page in 0 of 10 total answer segments; Perplexity cited our page in 0 of 7 total segments. Although the SOYA page ranked first on Google for the query, it was not referenced by generative engines demonstrating that SEO visibility did not translate into LLM citation.

We then applied structural interventions mapped to each GEO layer over three weeks. Week 1 (Semantic Anchoring): rewrote title and introduction, added a summary section. Week 2 (Context Triggering): inserted paraphrases (e.g., "course evaluation" vs. "student feedback") and related terminology. Week 3 (Pragmatic Recomposition): broke paragraphs into 3-5 sentence units, added FAQ blocks and bolded definitions.

After optimization, we re-ran the same queries under identical conditions (same model versions, same prompt, same incognito conditions) during July 22-26, 2024. The results are shown in Table III, which now includes per-trial query-level detail.

The citation rate for each trial was calculated as:

$$\text{Citation Rate} = \frac{C}{S} \times 100\%$$

where C is the number of answer segments citing our page (as defined by the attribution criteria above) and S is the total number of answer segments in the generated response. Across all trials, the per-trial citation rates ranged from 50% to 100%, with a mean of 90.6% (SD = 16.4%). This confirmed that semantic modularity was strongly associated with citation inclusion, and that citation occurred despite no change in Google SEO rank or backlinks.

Table 3: Post-GEO citation results for the SOYA page (July 2024)

Engine	Trial	Seg.	Cited	Rate (%)
CGPT R1	T1	3	2	66.7
CGPT R1	T2	3	2	66.7
CGPT R1	T3	3	3	100.0
CGPT R2	T1	3	3	100.0
CGPT R2	T2	2	2	100.0
CGPT R2	T3	2	2	100.0
PPLX R1	T1	2	2	100.0
PPLX R1	T2	2	1	50.0
PPLX R1	T3	2	2	100.0
PPLX R2	T1	2	2	100.0
PPLX R2	T2	2	2	100.0
PPLX R2	T3	2	2	100.0

Variability across trials: As shown in Table III, citation rates varied across trials, with ChatGPT ranging from 66.7% to 100% and Perplexity ranging from 50% to 100%. This variability is expected given the stochastic nature of generative model outputs. To account for this variability, we recommend that practitioners conduct multiple evaluation sessions (at least 3) and report the range or standard deviation of citation rates, rather than relying on a single query. In our case, the mean citation rate across all 12 trials was 90.6%, with standard deviation 16.4%, indicating that while citation was generally high, occasional trials yielded lower rates.

7. Discussion

The experimental results reveal a consistent pattern: content optimized using the GEO framework demonstrated substantial improvements in citation frequency within generative engine outputs, regardless of its SEO status, backlink presence, or domain authority. This section synthesizes the theoretical implications and practical constraints.

7.1 From Ranking to Referencing

The SOYA trial demonstrated that even a page ranked first on Google may remain completely invisible to generative engines if it lacks semantic modularity and extractability. After applying bundled GEO interventions, the same content was found to be cited in over 77% of generated segments. The name-based trial involved a page that ranked poorly in SEO (position 10.7) and shared its target query with a dominant, unrelated entity. Even so, the GEO-structured content achieved a 60%+ citation rate. Together, these findings are consistent with the hypothesis that LLM visibility is strongly associated with extractability, coherence, and topical clarity, and that these factors may be more consequential than link-based authority signals for citation likelihood. However, we emphasize that this is an observational association, not a demonstrated causal effect.

7.2 Layer Effectiveness and Interaction

Semantic anchoring (Layer 1) was associated with improved classification and recall. Clear titles and summaries helped the indexing system correctly categorize the content. Context triggering (Layer 2) enabled retrieval across paraphrased and taxonomic query variants, though its effect was domain-dependent. Pragmatic recomposition (Layer 3) was the most reliable citation trigger, particularly for FAQ-style queries. Notably, ChatGPT cited not only introductory paragraphs but also embedded FAQ answers and sentence-level definitions, reinforcing the role of modularity. Nevertheless, because all three layers were

applied simultaneously in the case studies, we cannot isolate the individual contribution of each layer. The cross-layer simulation experiment provides suggestive evidence for layer-specific effects, but these findings require validation in live generative systems.

7.3 Visibility vs. Authority

One counterintuitive outcome was the citation of low-authority domains (static hosting platform with no backlinks) over highly ranked official sites. This challenges the long-standing SEO assumption that PageRank correlates with trust. In GEO, trust may emerge from information packaging rather than linkage popularity [17]. This suggests that content creators with lower domain authority may still achieve citation if they prioritize semantic clarity and modular formatting, though this finding requires replication.

7.4 Platform Dependence and Reproducibility Considerations

A significant limitation of this study is its dependence on specific commercial generative search engines. The behavior of ChatGPT Search, Perplexity AI, and other generative systems is determined by proprietary models that are updated frequently and without public notice. A benchmark conducted today may not be reproducible in six months, as model versions, retrieval mechanisms, and citation policies may change. This platform dependence is inherent to research on commercial AI systems, but it must be acknowledged explicitly. To mitigate reproducibility concerns, we recommend that future GEO evaluations:

1. Report model versions and dates for all evaluations, as we have done in Table III.
2. Use multiple platforms to assess whether effects generalize beyond a single engine.
3. Maintain a static snapshot of evaluation prompts and conditions, available in supplementary materials.
4. Consider developing open-source benchmark datasets that can be used to evaluate GEO interventions in controlled, reproducible settings (e.g., using open-source LLMs with fixed retrieval corpora).

We also note that the reverse SEO experiment and the name disambiguation experiment, while informative, represent single observations on specific dates. Replication across multiple time points and model versions is necessary before strong conclusions can be drawn.

7.5 Constraints and Limitations

Several limitations must be acknowledged. Generative outputs vary depending on prompt, model version, and context, so multi-session testing is necessary. It remains impossible to directly verify when an LLM ingested a newly published site. Citation logic is not fully transparent; LLMs may cite based on latent embeddings or internal heuristics. These constraints are inherent to the partially observable nature of GEO and reinforce the need for layered metric evaluation.

8. Conclusion And Future Work

This paper has presented a comprehensive framework for Generative Engine Optimization, including a three-layer semantic visibility model, an implementation pipeline, five quantitative evaluation metrics, and empirical validation through real-world case studies. The results demonstrate that GEO is associated with substantial citation gains independently of traditional SEO ranking, and that semantic structuring rather

than link authority appears to be a strong predictor of generative visibility. However, our observational study design does not permit strong causal inference. Future controlled experiments with isolated interventions are needed to establish causal mechanisms and to determine the independent contribution of each layer.

Future work will focus on three directions. First, we plan to develop automated citation monitoring tools, such as browser extensions or LLM-integrated dashboards, that detect if and how a URL is cited across sessions. Second, we will build automatic GEO scoring systems that use LLMs to compute the five metrics directly from raw HTML inputs. Third, we will replicate these experiments across other generative platforms (You.com, Brave AI, Meta AI) to test framework generalizability [18]. Finally, we will explore reinforcement-based optimization, fine-tuning small models on citation-based reward functions to iteratively restructure pages for maximum generative visibility.

References

[1] S. Suri, S. Counts, L. Wang, C. Chen, M. Wan, T. Safavi, others, and L. Yang. The use of generative search engines for knowledge work and complex tasks. *arXiv preprint*, arXiv:2404.04268, 2024.

[2] 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*, pages 5–16, 2024.

[3] Y. Zhu, H. Yuan, S. Wang, J. Liu, W. Liu, C. Deng, others, and J. R. Wen. Large language models for information retrieval: A survey. *ACM Transactions on Information Systems*, 44(1):1–54, 2025.

[4] J. Menick et al. Teaching Language Models to Support Answers with Verifiable Quotes. *arXiv preprint*, 2022.

[5] G. Izacard, M. Caron, L. Hosseini, S. Riedel, P. Bojanowski, A. Joulin, and E. Grave. Unsupervised dense information retrieval with contrastive learning. *arXiv preprint*, arXiv:2112.09118, 2021.

[6] M. Chen, X. Wang, K. Chen, and N. Koudas. Generative engine optimization: How to dominate ai search. *arXiv preprint*, arXiv:2509.08919, 2025.

[7] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[8] P. Lewis, E. Perez, A. Piktus, et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Proc. NeurIPS*, 2020.

[9] O. Ram et al. In-Context Retrieval-Augmented Language Models. *arXiv preprint*, 2023.

[10] P. Aggarwal, V. Murahari, T. Rajpurohit, A. Kalyan, K. Narasimhan, and A. Deshpande. GEO: Generative Engine Optimization. In *Proc. ACM SIGKDD*, 2024. doi: 10.1145/3637528.3671900.

[11] F. Lüttgenau, I. Colic, and G. Ramirez. Beyond SEO: A Transformer-Based Approach for Reinventing Web Content Optimization. *arXiv preprint*, 2025.

[12] A. Srivastava, M. Nalluri, T. Lata, G. Ramadas, N. Sreekanth, and H. B. Vanjari. Scaling ai-driven solutions for semantic search. In *2023 International Conference on Power, Energy, Environment & Intelligent Control (PEEIC)*, pages 1581–1586. IEEE, 2023.

[13] R. Modi. Scalable semantic search with generative ai using azure cognitive search: A retrieval-augmented generation framework. Available at SSRN 6696098, 2025. SSRN Working Paper.

[14] Y. Wen, N. Jain, J. Kirchenbauer, M. Goldblum, J. Geiping, and T. Goldstein. Hard prompts made easy: Gradient-based discrete optimization for prompt tuning and discovery. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pages 51008–51025, 2023.

[15] J. Jiang, K. Zhou, Z. Dong, K. Ye, X. Zhao, and J. R. Wen. Structgpt: A general framework for large language models to reason over structured data. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9237–9251, 2023.

[16] A. Blanco-Gonzalez, A. Cabezon, A. Seco-Gonzalez, D. Conde-Torres, P. Antelo-Riveiro, A. Pineiro, and R. Garcia-Fandino. The role of ai in drug discovery: Challenges, opportunities, and strategies. *Pharmaceuticals*, 16(6):891, 2023.

[17] C. Abidin. From "networked publics" to "refracted publics": A companion framework for researching "below the radar" studies. *Social Media + Society*, 7(1):2056305120984458, 2021.

[18] N. Thakur, N. Reimers, A. Rücklé, A. Srivastava, and I. Gurevych. Beir: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. *arXiv preprint*, arXiv:2104.08663, 2021.