Think Before Writing: Feature-Level Multi-Objective Optimization for Generative Citation Visibility

Zikang Liu, Peilan Xu*
School of Artificial Intelligence,
Nanjing University of Information Science and Technology, Nanjing 210044, China.
{202412492739, xpl}@nuist.edu.cn

Abstract

Generative answer engines expose content through selective citation rather than ranked retrieval, fundamentally altering how visibility is determined. This shift calls for new optimization methods beyond traditional search engine optimization. Existing generative engine optimization (GEO) approaches primarily rely on token-level text rewriting, offering limited interpretability and weak control over the trade-off between citation visibility and content quality. We propose FeatGEO, a feature-level, multi-objective optimization framework that abstracts webpages into interpretable structural, content, and linguistic properties. Instead of directly editing text, FeatGEO optimizes over this feature space and uses a language model to realize feature configurations into natural language, decoupling high-level optimization from surface-level generation. Experiments on GEO-Bench across three generative engines demonstrate that FeatGEO consistently improves citation visibility while maintaining or improving content quality, substantially outperforming token-level baselines. Further analyses show that citation behavior is more strongly influenced by document-level content properties than by isolated lexical edits, and that the learned feature configurations generalize across language models of different scales. Code is available at https://github.com/EvoNexusX/2026LiuFeatGEO.git.

1 Introduction

Large language models (LLMs) are rapidly reshaping how users access information. Instead of presenting ranked lists of documents, generative answer engines, such as Perplexity, Bing Chat, and Google's AI Overviews, synthesize responses by selectively citing a small subset of retrieved sources (Amer and Elboghdadly, 2024). In this paradigm, visibility is no longer determined by rank position

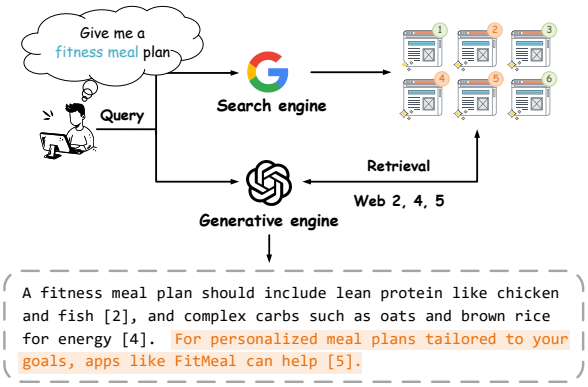


Figure 1: Illustration of the paradigm shift from rank-based search to citation-based generative answering. Traditional search engines expose content through ranked result lists, whereas generative answer engines synthesize responses by selectively citing a subset of retrieved sources. As illustrated, citation inclusion rather than rank position determines which sources are surfaced in the generated answer.

but by citation allocation: sources that are not cited receive effectively no exposure, regardless of their relevance or retrieval rank . Figure 1 illustrates this shift from rank-based exposure to citation-based visibility.

This shift introduces a new optimization problem that is fundamentally different from traditional search engine optimization (SEO) (Aggarwal et al., 2024). While SEO aims to improve a document's ranking in a list-based interface, generative retrieval systems require optimizing content for citation by an opaque generative model. Recent work has shown that LLMs exhibit systematic citation biases (Algaba et al., 2025) and non-trivial selection behaviors (Liu et al., 2024), raising important questions about how content properties influence citation likelihood in generated answers.

From the perspective of content providers, citation-based visibility has direct implications for user attention and trust, motivating the emerging study of generative engine optimization (GEO)

*Corresponding Author

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 20290–20303

(Aggarwal et al., 2024; Kumar and Lakkaraju, 2024), where the goal is to increase the probability of being cited in LLM-generated responses.

Existing approaches to GEO primarily operate at the token level (Aggarwal et al., 2024; Kumar and Lakkaraju, 2024; Nestaas et al., 2024), applying heuristic text edits, such as keyword insertion or authoritative phrasing, to increase the likelihood of being cited in LLM-generated responses. While effective in isolated cases, such methods suffer from two limitations. First, citation behavior in generative retrieval is not governed by a single fixed query. Instead, large language models are exposed to a latent and diverse space of user intents that share a common semantic theme. Optimizing text for individual queries therefore provides an unstable target and struggles to capture stable topic-level citation preferences that persist across semantic variations. Second, direct text manipulation conflates what information a webpage conveys with how it is linguistically realized, obscuring the high-level content and structural properties that may systematically influence LLM citation decisions (Liang et al., 2024). These limitations motivate a shift from token-level editing to feature-level optimization. By abstracting webpages into interpretable, high-level properties and reasoning about citation behavior at the topic level, GEO can be formulated as a structured decision problem that is both more interpretable and more amenable to principled optimization under competing objectives.

Our contributions: In this paper, we propose FeatGEO, a feature-based framework for citation visibility optimization in generative retrieval systems. First, we introduce a *topic-level citation modeling* perspective for GEO, capturing stable citation preferences of LLMs across semantically related queries rather than optimizing for individual prompts. Second, we formulate GEO as a *feature-level, multi-objective decision problem*, representing webpages through an interpretable set of structural, content, and linguistic properties that serve as controllable interfaces for LLM-based generation. Finally, we present a black-box *feature-space optimization framework* that jointly optimizes citation visibility and content quality, decoupling high-level feature selection from surface text realization.

2 Related Work

Search Engine-based Advertising. Traditional search visibility optimization spans two complementary approaches. SEO improves organic rankings through content optimization, keyword targeting, and link building, while academic SEO (ASEO) applies similar principles to scholarly visibility (Beel et al., 2010). Prior work on search-based advertising has extensively studied ranking, allocation, and bidding mechanisms under list-based exposure assumptions (Edelman et al., 2007; Cai et al., 2017; Zhao et al., 2018). Despite their differences, both SEO and advertising are grounded in list- or slot-based interfaces, where exposure is largely a function of rank position.

LLM-based Visibility and Optimization. As search interfaces increasingly incorporate generative answer engines, new mechanisms for content exposure have emerged. From a mechanism design perspective, recent work has explored advertising in generative settings, including token-level auctions (Duetting et al., 2024), segment auctions via retrieval-augmented generation (Hajiaghayi et al., 2024), multi-LLM aggregation (Soumalias et al., 2025), and generative auction frameworks (Zhao et al., 2025). LLMs have also been studied as tools for marketing content generation and strategy (Schweidel et al., 2024; Aghaei et al., 2025).

Orthogonal to auction-based approaches, GEO focuses on increasing citation visibility in LLM-generated answers. Aggarwal et al. (2024) proposed a set of heuristic, text-level editing strategies that substantially improve citation rates, while subsequent work demonstrated adversarial manipulation of LLM recommendations and retrieval behaviors (Kumar and Lakkaraju, 2024; Nestaas et al., 2024). Related studies on LLM persuasion further highlight the influence of content properties on generative model outputs (Rogiers et al., 2024).

Advertising research primarily studies allocation and pricing mechanisms, deciding *which* items are shown, while taking content as fixed. GEO methods, in contrast, directly modify page content to influence citation behavior, often assuming access to specific target queries and operating at the surface text level. Our work bridges these perspectives by optimizing content representations at an interpretable feature level

3 Method

3.1 Problem Overview and Formulation

We reformulate GEO as a feature-level control and optimization problem. Instead of manipulating raw text, we represent a webpage by an interpretable

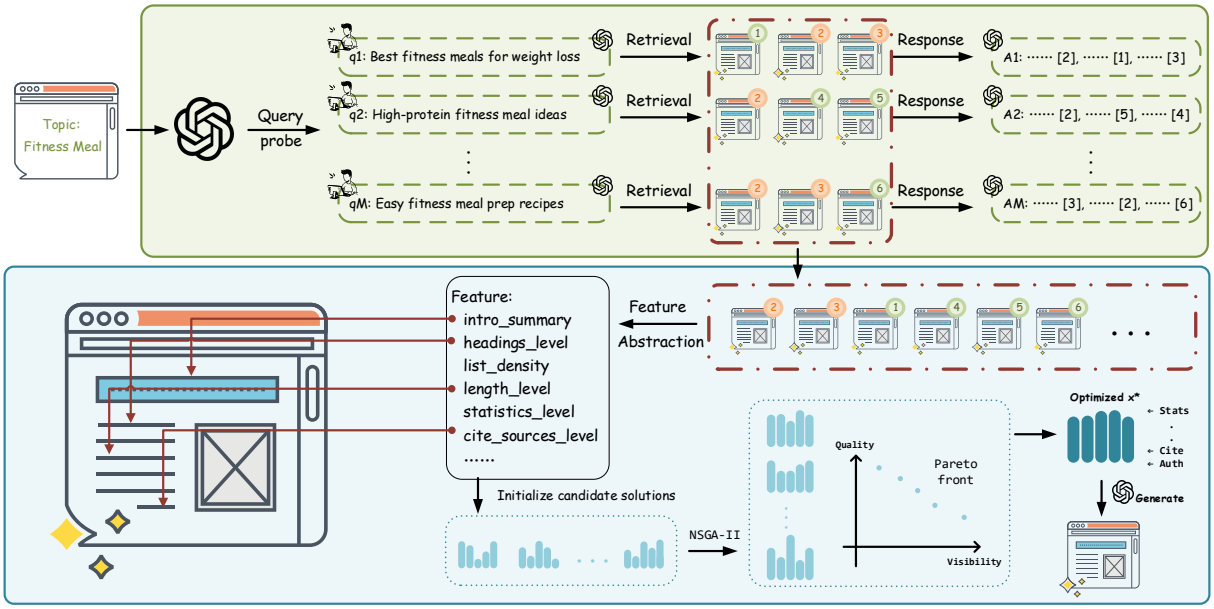


Figure 2: Overview of the FeatGEO pipeline. At the topic level (top), a generative engine is probed with diverse, semantically related queries, producing responses that cite different subsets of retrieved webpages. Citation patterns are aggregated across queries to identify topic-consistent citation exemplars. At the feature level (bottom), webpages are abstracted into interpretable feature vectors, which serve as decision variables in a multi-objective optimization process balancing citation visibility and content quality. Optimized feature configurations are then realized into concrete webpages via LLM-based generation.

vector of high-level properties that describe its structure, content richness, and linguistic style. Under this formulation, GEO amounts to selecting feature configurations that induce LLM-generated pages with higher citation visibility while maintaining acceptable quality.

A central challenge is that citation behavior in generative engines is not driven by a single fixed query, but by a latent space of user intents sharing a common semantic theme. FeatGEO addresses this challenge by modeling citation preference at the topic level and performing optimization in a structured, low-dimensional feature space, with LLMs serving as realization functions rather than direct optimization targets. Figure 2 provides an overview of the FeatGEO pipeline. The top panel illustrates topic-level aggregation of citation behavior across semantically related queries, while the bottom panel shows feature abstraction and multi-objective optimization used to generate citation-optimized webpages.

3.2 Topic-Level Citation Modeling

We consider a generative search setting in which visibility is determined by whether a webpage is cited in an LLM-generated response. Unlike traditional retrieval scenarios that assume a fixed user query, generative engines are exposed to a broad and heterogeneous space of user intents under a shared semantic theme. We therefore model citation behavior at the level of a topic rather than an individual query. Given a topic τ (e.g., Fitness meal), we approximate the space of plausible user information needs by prompting an LLM to generate a set of semantically related queries $\mathcal{Q}_\tau = \{q_1, ..., q_M\}$. These queries are used solely as probes to approximate the latent intent distribution of a topic, and are not exposed to or optimized by the feature search procedure.

For each query $q \in \mathcal{Q}_\tau$, the engine produces a synthesized answer accompanied by a set of cited webpages, denoted as $\mathcal{S}(q)$. Aggregating across queries yields a topic-level citation set

$$\mathcal{S}_\tau = \bigcup_{q \in \mathcal{Q}_\tau} \mathcal{S}(q).$$

Importantly, not all cited sources play an equal role. We observe that certain webpages recur across multiple queries, suggesting that they satisfy citation preferences that are invariant to semantic perturbations within the topic. We capture this regularity by associating each source $s \in \mathcal{S}_\tau$ with a

citation frequency

$$f(s) = \sum_{q \in \mathcal{Q}_\tau} \mathbb{I}[s \in \mathcal{S}(q)] \tag{1}$$

Webpages with high citation frequency are treated as topic-consistent citation exemplars, serving as reference points for downstream optimization.

3.3 Feature-Level Representation and Generation Control

Our approach is motivated by the hypothesis that webpages consistently cited under a topic share common high-level properties, even when their surface forms differ substantially. Rather than operating directly on raw text, we model these properties in an interpretable feature space. Specifically, we represent each exemplar $s \in \mathcal{S}_\tau$ by a feature vector $\mathbf{x}(s)$, capturing structural, content, and linguistic attributes. Collectively, these vectors define a feature space representing topic-specific citation preferences.

Constructing a new target webpage corresponds to selecting a feature configuration $\mathbf{x}$ in this space that simultaneously satisfies two objectives: maximizing expected citation likelihood $f_{\text{vis}}(\mathbf{x})$ and maintaining content quality $f_{\text{qual}}(\mathbf{x})$. Formally, we express this as

$$\mathbf{x}^* = \arg\max_{\mathbf{x}} \big(f_{\text{vis}}(\mathbf{x}), f_{\text{qual}}(\mathbf{x})\big), \tag{2}$$

where the objective is understood in the Pareto-optimal sense, and

$$\begin{cases} f_{\text{vis}}(\mathbf{x}) = \mathbb{E}_{q \sim \mathcal{Q}_\tau}\big[\text{Vis}(\text{LLM}(\tau, \mathbf{x}); q)\big] \\ f_{\text{qual}}(\mathbf{x}) = \mathbb{E}_{q \sim \mathcal{Q}_\tau}\big[\text{Qual}(\text{LLM}(\tau, \mathbf{x}); q)\big] \end{cases} \tag{3}$$

Candidate feature configurations are iteratively evaluated and refined to approximate Pareto-optimal trade-offs between visibility and quality.

Each feature vector $\mathbf{x}$ is organized into three semantic layers, as summarized in Table 1. These features capture discourse-level and stylistic properties of webpages and serve as high-level control signals for LLM-based page generation, rather than prescribing exact token-level edits.

Although each feature dimension is associated with an indicative numerical range, these values are used as soft control signals rather than hard constraints. They provide coarse-grained guidance for content planning, while allowing the language model flexibility in surface realization. Feature configurations are realized through prompt-level instructions that translate abstract feature preferences into qualitative generation guidelines. Instead of enforcing exact numerical values, features are mapped to descriptive cues (e.g., emphasizing statistics or improving structural clarity), which are incorporated into the system prompt. In this way, the feature vector $\mathbf{x}$ functions as an interpretable control interface for steering generation style and content emphasis. The complete prompt template with feature-to-text mapping is provided in Appendix B.

3.4 Multi-Objective Optimization in Feature Space

Optimizing citation visibility in generative engines presents a black-box, non-differentiable problem: feature configurations influence citation outcomes only through LLM-based generation and evaluation, with no accessible gradients and substantial stochasticity. Moreover, improving visibility often conflicts with maintaining content quality, making single-objective optimization inadequate.

To address these characteristics, we adopt a population-based multi-objective optimization strategy in the feature space. Each individual corresponds to a candidate feature configuration $\mathbf{x}$, and is evaluated according to its expected visibility $f_{\text{vis}}(\mathbf{x})$ and content quality $f_{\text{qual}}(\mathbf{x})$.

We initialize the population by leveraging feature patterns extracted from webpages that are frequently cited under the target topic. Specifically, feature vectors inferred from these exemplar pages serve as seeds, ensuring that the initial population reflects realistic and citation-relevant configurations while allowing novel combinations across different pages.

At each iteration, candidate configurations are realized into webpages via LLM-based generation and evaluated by inserting them into the generative engine alongside competing pages. Their visibility and quality scores are estimated by aggregating results across representative topic-induced queries. Based on these evaluations, dominated configurations are discarded, and new candidates are generated through stochastic variation in the feature space.

We adopt NSGA-II (Deb et al., 2002) as the underlying optimization framework to maintain a diverse set of non-dominated solutions and progressively approximate the Pareto frontier. The complete optimization procedure is summarized in Algorithm 1.

Layer	Feature	Range	Description
Structure	has_intro_summary	[0.0, 1.0]	Presence of introductory summary paragraph
	headings_level	[1.0, 3.0]	Hierarchy and quantity of headings and subheadings
	list_density	[0.0, 3.0]	Frequency of bullet point and numbered lists
	length_level	[1.0, 3.0]	Overall article length; reflects content depth and breadth
Content	statistics_level	[0.0, 3.0]	Density of data, statistics, and percentages embedded in the text
	cite_sources_level	[0.0, 3.0]	Frequency of citing authoritative sources, institutions, or reports
	quotation_level	[0.0, 3.0]	Frequency of using quotations from experts or authoritative figures
	unique_info_level	[0.0, 3.0]	Richness of unique information and differentiated content
	technical_terms_level	[0.0, 3.0]	Density of professional terminology and technical vocabulary
Language	authoritative_level	[0.0, 3.0]	Strength of authoritative tone and assertive expressions
	easy_to_understand_level	[1.0, 3.0]	Content readability and language simplicity
	fluency_level	[1.0, 3.0]	Writing fluency and logical coherence between sentences
	keyword_focus_level	[1.0, 3.0]	Focus and repetition strength of core keywords

Table 1: Feature Dimensions and Ranges in FeatGEO, by Semantic Layer. Ranges reflect feature semantics: $[0, 3]$ for features that may be entirely absent, and $[1, 3]$ for features that are always present to some degree.

Algorithm 1 Feature-Level Optimization Procedure in FeatGEO

Input: Topic τ, competitor pages $\mathcal{S}_\tau$, population size N, generations G

Output: Pareto-optimal feature configurations

1: Extract feature configurations from competitor pages in $\mathcal{S}_\tau$
2: Initialize population $\mathcal{P}$ by recombining extracted feature configurations
3: **for** $g = 1$ to G **do**
4: Realize each feature configuration $\mathbf{x} \in \mathcal{P}$ into a webpage via LLM
5: Evaluate visibility and quality by aggregating over topic-induced queries
6: Update population via multi-objective selection and variation in feature space
7: **end for**
8: **return** Non-dominated feature configurations and a selected final solution

4 Experiments

4.1 Experimental Setup

Generative Engine and Benchmark. Following Aggarwal et al. (2024), we adopt a two-stage retrieval-augmented generation (RAG) pipeline: (1) retrieving the top-5 sources via Google Search and (2) generating a cited answer using an Answer Generator LLM. To assess robustness across generative engines, we consider three Answer Generators with distinct architectures and training regimes: GPT-4o-mini, Gemini-2.5-flash, and Qwen-plus. Importantly, all optimization is performed on advertiser pages generated by a fixed Page Generator (GPT-4o-mini), ensuring that differences in performance are attributable to optimization strategies rather than content generation capacity. Prompts follow prior work (Liu et al., 2023a) (Appendix C).

Experiments are conducted on GEO-Bench (Aggarwal et al., 2024), a benchmark designed to evaluate content optimization strategies for generative engines. It contains 10K queries spanning 25 domains from nine sources (e.g., MS MARCO, Natural Questions, LIMA). For each query, an advertiser page is injected alongside the top-5 retrieved webpages and evaluated through the full RAG pipeline described above.

Compared Methods. We evaluate five methods: (1) *Baseline*: the unmodified advertiser page; (2) *GEO Methods* (Aggarwal et al., 2024): nine token-level heuristics—*Authoritative* (more persuasive tone), *Statistics Addition* (quantitative data), *Keyword Stuffing* (query keywords), *Cite Sources & Quotation Addition* (credible references), *Easy-to-Understand* (simpler language), *Fluency Optimization* (improved fluency), *Unique Words & Technical Terms* (lexical enrichment); (3) *AutoGEO-global* (Wu et al., 2025): a token-level rewriting framework that first automatically extracts natural-language content-preference rules from a generative engine, then applies these rules via an LLM to rewrite the target page; the rules are learned once across all queries and remain fixed at test time; (4) *AutoGEO-instance*: an instance-adaptive extension of AutoGEO that, for each test query, generates topic-specific proxy queries, extracts instance-level preference rules, and merges them with the global rule set before rewriting, allowing the rewriter to adapt to per-query content demands; (5) *FeatGEO*

(Ours): feature-space multi-objective optimization via NSGA-II that generates pages from abstract feature specifications rather than editing existing text.

Implementation Details. Unless otherwise stated, NSGA-II is run with a population size of 8 for 8 generations. Gaussian mutation is applied independently to each feature with probability $p = 0.5$ and standard deviation $\sigma = 0.2$. Each configuration is evaluated five times to reduce stochastic variance, and all hyperparameters are shared across methods. Detailed computational cost breakdowns are provided in Appendix F.

Evaluation Metrics. We report two complementary metrics:

Visibility Metrics: Following Aggarwal et al. (2024), we compute a word-position weighted visibility score for each page, capturing both the number and position of cited words. Auxiliary metrics include: (1) Word Count: normalized word count of sentences citing the advertiser page; (2) Position Count: position-weighted word count giving less weight to later citations. The primary visibility metric is the advertiser visibility w_{ad}.

Quality Metrics: We adopt a G-Eval-style framework (Liu et al., 2023b) where an LLM evaluates answer quality automatically. Each query is scored on seven dimensions: four content dimensions (fluency, usefulness, credibility, structure) and three appeal dimensions (uniqueness, attractiveness, influence on the overall answer). Scores (1–5) are normalized to $[0, 1]$ and combined as

$$\text{Qual} = \alpha \cdot \text{Qual}_{\text{content}} + (1-\alpha) \cdot \text{Qual}_{\text{appeal}}, \quad (4)$$

averaged over multiple generations per configuration. For presentation, quality scores are reported as percentages.

4.2 Comparison Results

Table 2 summarizes results across three generative engines. Although FeatGEO produces an entire Pareto front of visibility–quality trade-offs, we report the solution with maximum visibility to enable direct comparison with single-objective baselines.

Across all engines, token-level GEO heuristics fail to consistently improve citation visibility over the unmodified baseline. On GPT-4o-mini, visibility drops range from 10.92% to 12.21% compared to the baseline of 13.34%; on Gemini, baselines achieve only 4.62%–5.62% versus 8.89%; on Qwen-plus, visibility drops from 5.20% to 2.75%–3.72%. In addition, several baselines negatively impact content quality, with the Gemini engine showing the largest reductions (e.g., Easy-to-Understand scores 74.54 compared to 75.59 baseline). A similar pattern holds for AutoGEO: although AutoGEO-instance improves over AutoGEO-global on all three engines, both remain below the unmodified baseline in visibility (e.g., 12.12% vs. 13.34% on GPT-4o-mini, 7.04% vs. 8.89% on Gemini, and 4.25% vs. 5.20% on Qwen-plus). Our results indicate that, at the scale and diversity of GEO-Bench, isolated text-level modifications are insufficient to reliably increase citation visibility and may even disrupt the natural writing patterns that LLMs prefer to cite.

In contrast, FeatGEO achieves the highest visibility across all three engines: 18.31% on GPT-4o-mini (+37% relative improvement), 15.35% on Gemini (+73%), and 10.17% on Qwen-plus (+96%), while maintaining quality scores comparable to or better than baseline (81.52, 76.14, and 77.12 respectively). The substantial gains on Gemini, where FeatGEO nearly doubles baseline visibility, are particularly notable and demonstrate robust generalization across engines with different architectures, training data, and citation behaviors. The method achieves strong performance even on Qwen-plus, which exhibits lower absolute visibility, confirming that feature-level optimization adapts effectively to diverse generative paradigms.

4.3 Robustness to Evaluator Choice

To control for potential evaluator bias, we additionally evaluated all GPT-4o-mini outputs with two alternative LLM judges, Gemini-2.5-flash and Claude-3.5-Sonnet. As shown in Table 3, although the absolute scores are lower than those of the original GPT-4o-mini evaluator, FeatGEO remains the top-ranked method under both Gemini (75.28) and Claude (72.63), outperforming the strongest baseline by 4.26 and 0.53 points respectively.

The relative ranking among methods is largely preserved across all three judges: token-level heuristics cluster within a narrow band (68–72), while FeatGEO consistently separates itself from this group. This cross-judge consistency confirms that our quality conclusions are robust to the choice of evaluator. Visibility scores, derived mechanically from citation patterns in generated responses, are not subject to such judge-dependent variation.

Method	GPT-4o-mini				Gemini-2.5-flash				Qwen-plus			
	Vis	Qual	Word	Pos	Vis	Qual	Word	Pos	Vis	Qual	Word	Pos
Baseline	13.34	79.17	14.99	13.42	8.89	75.59	10.12	8.84	5.20	76.81	6.43	5.36
Fluency Optimization	11.74	77.17	13.18	11.92	5.04	75.23	5.32	4.77	3.67	76.11	4.47	3.85
Unique Words	10.92	76.11	12.41	11.15	4.62	75.02	4.96	4.78	3.34	75.79	4.19	3.58
Authoritative	11.94	77.21	13.42	12.15	4.94	75.17	5.58	4.73	3.68	75.79	4.48	3.84
Quotation Addition	11.08	77.27	12.62	11.35	5.01	75.18	5.40	5.26	2.82	76.25	3.58	2.97
Cite Sources	11.78	77.61	13.34	12.03	5.62	75.57	5.94	5.81	3.46	76.42	4.04	3.69
Easy-to-Understand	11.06	75.16	12.64	11.23	4.96	74.54	5.42	5.06	3.03	75.70	3.70	3.20
Technical Terms	11.81	77.37	13.36	11.96	5.55	75.41	5.83	5.73	3.59	76.17	4.47	3.86
Statistics Addition	12.21	77.08	13.67	12.42	5.54	75.30	6.14	5.58	3.72	76.37	4.45	3.98
Keyword Stuffing	11.71	77.15	13.22	11.92	4.92	74.68	5.27	4.89	2.75	76.23	3.45	2.63
AutoGEO-global	11.22	75.93	12.47	11.47	5.70	75.29	6.24	5.71	3.37	76.96	3.79	3.74
AutoGEO-instance	12.12	76.57	13.45	12.42	7.04	76.06	7.58	7.12	4.25	76.12	5.05	4.43
FeatGEO (ours)	**18.31**	**81.52**	**20.16**	**18.33**	**15.35**	**76.14**	**16.06**	**15.02**	**10.17**	**77.12**	**11.71**	**9.75**

Table 2: Comparison of methods on GEO-Bench across three generative engines (GPT-4o-mini, Gemini-2.5-flash, Qwen-plus). For each method, we report ad visibility (Vis), overall quality (Qual), and auxiliary metrics Word and Pos, which reflect word-level and position-weighted citations.

Method	Gemini	Claude
Baseline	71.02	69.67
Fluency Optimization	69.40	69.35
Unique Words	68.73	68.67
Authoritative	68.80	68.04
Quotation Addition	70.26	72.10
Cite Sources	70.08	71.26
Easy-to-Understand	69.44	69.27
Technical Terms	68.97	68.95
Statistics Addition	69.45	69.07
Keyword Stuffing	69.13	69.81
FeatGEO (Ours)	**75.28**	**72.63**

Table 3: Quality scores under two alternative LLM judges, Gemini-2.5-flash and Claude-3.5-Sonnet.

Method	Pre	Post	Δ Vis
Fluency Optimization	18.72	20.21	+1.49
Unique Words	18.72	16.34	-2.38
Authoritative	18.72	19.37	+0.65
Quotation Addition	18.72	19.97	+1.25
Cite Sources	18.72	19.84	+1.11
Easy-to-Understand	18.72	18.89	+0.16
Technical Terms	18.72	19.14	+0.42
Statistics Addition	18.72	21.05	+2.33
Keyword Stuffing	18.72	19.44	+0.72
AutoGEO-global	18.72	22.86	+4.13
Average	**18.72**	**19.71**	**+0.99**

Table 4: Effects of heuristic GEO methods on human-written competitor pages. Pre and Post denote advertiser visibility before and after applying each method.

4.4 Effect of Base Content Quality on Heuristic GEO Methods

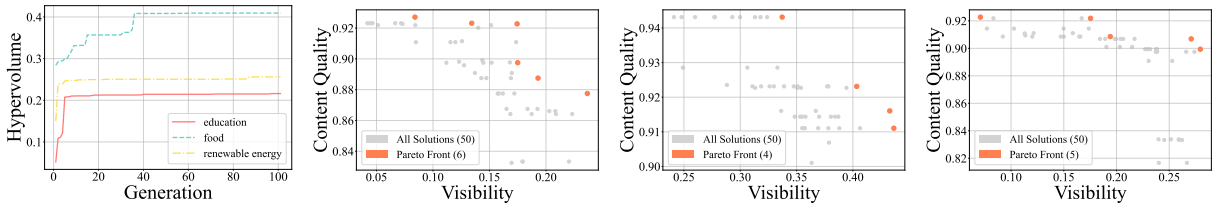
The results in Table 2 were obtained on LLM-generated advertiser pages. To test whether base content quality influences the effectiveness of heuristic methods, we apply them to existing human-written pages, which are typically less optimized for generative engine citation than LLM-generated content. As shown in Table 4, these methods yield an average visibility gain of +0.99 (18.72% to 19.71%), with AutoGEO-global achieving the largest improvement (+4.13, from 18.72% to 22.86).

By contrast, the same heuristics degrade visibility on the LLM-generated advertiser pages in Table 2. This asymmetry reveals a *regime-dependent saturation effect*: token-level rewriting benefits pages with structural or stylistic gaps, yet becomes counterproductive on already fluent generated content, where additional modifications introduce redundancy and disrupt the naturalness signals that generative engines implicitly favor when selecting citations. FeatGEO circumvents this limitation entirely by synthesizing pages from feature-level specifications rather than locally editing existing text, thereby preserving stylistic coherence while steering citation behavior.

4.5 Multi-Objective Analysis

We analyze FeatGEO's multi-objective optimization behavior using an extended evolutionary search with 50 individuals over 100 generations, enabling detailed examination of convergence and Pareto front structure.

Convergence. Figure 3(a) shows hypervolume (HV) trajectories for three representative topics. HV increases rapidly in early generations and stabilizes thereafter, indicating effective convergence. Different growth patterns across topics reflect vari-



(a) HV Curve. (b) *Education*. (c) *Food*. (d) *Renewable Energy*.

Figure 3: Multi-objective optimization convergence and Pareto front diversity. (a) Hypervolume (HV) evolution across 100 generations. (b-d) Final Pareto fronts illustrate visibility–quality trade-offs for different topics. Gray dots denote all evaluated configurations; coral points highlight Pareto-optimal solutions. Differences in HV growth and PF shapes reflect topic-specific dynamics and LLM evaluation biases.

ations in content structure and LLM-based quality evaluation. These results justify the more compact evolutionary setup used in Section 4.2 and confirm that FeatGEO reliably identifies high-quality trade-off solutions.

Pareto fronts. Figures 3(b–d) illustrate final Pareto fronts. Across topics, visibility and quality exhibit a clear trade-off, though its severity varies. For example, in *education*, high visibility often requires notable quality sacrifice, whereas in *food*, visibility gains are achieved with minimal quality loss. This topic-dependent structure highlights the importance of multi-objective optimization and enables practitioners to select solutions aligned with specific priorities.

Feature-level insights. Table 5 compares two extreme Pareto-optimal solutions from a sample query in the *education* domain. Solution A emphasizes visibility (23.7%) while maintaining moderate quality (87.8), whereas Solution B prioritizes quality (92.7) at lower visibility (8.4%). Analyzing feature intensities reveals clear patterns. (1) Content credibility and fluency (Statistics, Citations, Quotation, Fluency) are upweighted in the high-visibility solution, suggesting that LLMs are more likely to cite pages that present authoritative content in a fluent style. (2) Structural organization (Heading Level, List Density, Length) is stronger in the high-quality solution, indicating that careful formatting and logical presentation contribute more to perceived content quality than to visibility. (3) Trade-offs in features such as Authoritative Tone and Easy-to-Understand show how optimizing for one objective can require compromises along another dimension.

These patterns confirm that the multi-objective optimization captures non-trivial trade-offs: different objectives naturally favor distinct feature combinations, which would be hard to identify with single-objective approaches. Advertisers can select solutions aligned with their strategic priorities, leveraging the full PF. Detailed qualitative analysis of textual changes for each feature is provided in Appendix D

Layer	**Feature**	**Sol. A** Vis: 23.7% Qual: 87.8	**Sol. B** Vis: 8.4% Qual: 92.7
Structure	Intro Summary	0.64	0.52
	Heading Level	2.75	2.55
	List Density	1.26	2.01
	Length Level	2.32	2.61
Content	Statistics Level	1.62	2.18
	Cite Sources Level	1.45	1.74
	Quotation Level	2.84	1.94
	Unique Info Level	1.65	1.67
	Authoritative Tone	1.55	0.75
Language	Technical Terms	1.65	1.96
	Easy-to-Understand	1.37	1.75
	Fluency	2.17	1.58
	Keyword Focus	1.80	1.69

Table 5: Feature configurations of two extreme Pareto solutions in the *education* domain.

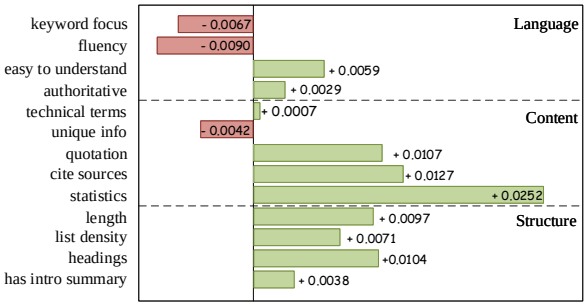


Figure 4: Feature contribution to ad visibility.

4.6 Ablation Study

To quantify the importance of individual features for citation visibility, we perform an ablation study. In each experiment, one feature is clamped to its

minimum value while the remaining 12 features are optimized by NSGA-II. We define the contribution of feature i as:

$$\Delta_i = f_{\text{vis}}(\mathbf{x}^*) - f_{\text{vis}}(\mathbf{x}^*_{-i}) \tag{5}$$

where $\mathbf{x}^*$ is the fully optimized configuration and $\mathbf{x}^*_{-i}$ is the configuration with feature i fixed at its minimum. Positive Δ_i indicates that increasing the feature improves visibility, while negative Δ_i indicates that higher values of this feature slightly reduce visibility.

The results shown in Figure 4 reveal several clear patterns. Content-oriented features dominate the overall visibility gains, with Statistics and Cite Sources providing the largest positive contributions. However, not all content features are beneficial: Unique Info slightly decreases visibility in some cases, while Technical Terms has minimal impact. Structural features consistently improve visibility across all pages, though their contributions are moderate, reflecting stable benefits from headings, lists, and document length. Language and style features exhibit mixed effects: some, such as Fluency and Keyword Focus, occasionally reduce visibility, while others provide modest positive contributions. Overall, the variance within each feature group is substantial, particularly for content and language features, highlighting that the effect of any individual feature can depend on the specific combination of other features.

These findings confirm that while content features primarily drive citation visibility, structural features provide reliable support, and language or stylistic adjustments may help or slightly hinder visibility depending on context. This nuanced view underscores the value of multi-feature optimization, rather than relying on isolated text-level heuristics. Quality ablation results are reported in Appendix E.

4.7 Scale-Invariant Effectiveness Across LLM Capacities

To evaluate robustness to the Page Generator's model capacity, we replace GPT-4o-mini with Qwen3 models of increasing scale (4B, 8B, and 14B parameters). The Answer Generator and evaluation pipeline are fixed to isolate the effect of generation capacity.

As shown in Figure 5, FeatGEO consistently outperforms the baseline across all model sizes. While larger generators slightly improve absolute performance for both methods, the relative advantage

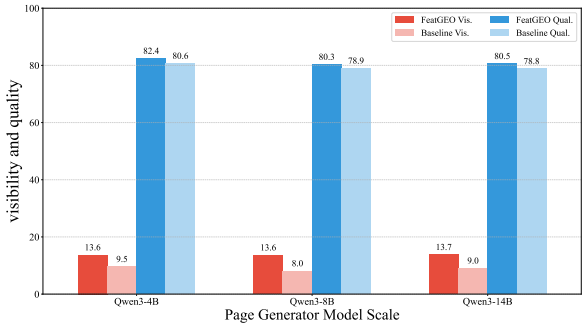


Figure 5: Performance of FeatGEO and Baseline using Page Generators of different model sizes (Qwen3-4B, 8B, and 14B). The Answer Generator for evaluation is fixed to GPT-4o-mini.

of FeatGEO remains stable. Quality scores also remain high and show no degradation. Notably, performance variance across model scales is substantially smaller than the gap between FeatGEO and the baseline, indicating that optimized feature configurations encode model-agnostic principles. This scale-invariant behavior enhances FeatGEO's practical applicability when generation models are updated or replaced.

5 Conclusion

We study the problem of optimizing citation visibility in generative retrieval systems and identify fundamental limitations of prior token-level GEO methods, including poor interpretability and brittle trade-offs between visibility and quality. We propose FeatGEO, a feature-based framework that abstracts webpages into interpretable structural, content, and linguistic representations and performs principled multi-objective optimization in this space. Extensive experiments demonstrate that FeatGEO consistently outperforms existing heuristics across diverse generative engines, while providing actionable visibility–quality trade-offs. Ablation analyses reveal distinct and complementary roles of different feature categories, and scale-robust experiments confirm that FeatGEO's effectiveness generalizes across page generation model capacities. Beyond empirical gains, our findings suggest that LLM citation behavior is driven more by high-level discourse organization and information structure than by surface lexical cues, highlighting feature-level abstraction as a promising direction for controllable generation in retrieval-augmented systems.

Limitations

This work has several limitations.

- First, our evaluation is conducted in a controlled setting where the candidate set is fixed, consisting of five retrieved pages and one advertiser-controlled page. We assume that the advertiser page has already been admitted into the candidate set, and therefore do not model upstream retrieval or ranking mechanisms that determine page inclusion. As a result, FeatGEO should be viewed as optimizing citation likelihood conditional on retrieval, rather than addressing end-to-end retrieval and generation.In practice, it serves as a test-time tool for content authors who wish to optimize their page for a specific topic after retrieval.

- Second, our fitness signals are derived from an LLM-based generative engine and automatic citation parsing. While this setup follows prior GEO benchmarks, citation formats and generation behaviors may vary across real-world systems, which could affect transferability. Evaluating feature-level optimization under proprietary or heterogeneous citation mechanisms remains an open direction.

- Third, content quality is assessed using an LLM-based judge. Although we mitigate variance by averaging over multiple generations, such evaluators may still introduce systematic biases that do not perfectly align with human judgments. Additionally, LLM-generated content in the pipeline may contain hallucinations, a concern common to all LLM-dependent GEO methods.

- Finally, the evolutionary search requires repeated end-to-end LLM calls for page generation, answer generation, and quality evaluation. This computational cost limits the scale of our experiments and may pose challenges for reproducibility under different API budgets or rate limits. Exploring more sample-efficient optimization or surrogate modeling approaches is an important direction for future work.

Acknowledgments

This work is supported by the Natural Science Foundation of Jiangsu Province (Grant No. BK20230419).

References

Pranjal Aggarwal, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, Karthik Narasimhan, and Ameet Deshpande. 2024. Geo: Generative engine optimization. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, pages 5–16. Association for Computing Machinery.

Raha Aghaei, Ali A. Kiaei, Mahnaz Boush, Javad Vahidi, Mohammad Zavvar, Zeynab Barzegar, and Mahan Rofoosheh. 2025. Harnessing the potential of large language models in modern marketing management: Applications, future directions, and strategic recommendations. *Preprint*, arXiv:2501.10685.

Andres Algaba, Carmen Mazijn, Vincent Holst, Floriano Tori, Sylvia Wenmackers, and Vincent Ginis. 2025. Large language models reflect human citation patterns with a heightened citation bias. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 6829–6864, Albuquerque, New Mexico. Association for Computational Linguistics.

Eslam Amer and Tamer Elboghdadly. 2024. The end of the search engine era and the rise of generative ai: A paradigm shift in information retrieval. In *2024 International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC)*, pages 374–379.

Jöran Beel, Bela Gipp, and Erik Wilde. 2010. Academic search engine optimization (aseo) optimizing scholarly literature for google scholar & co. *Journal of scholarly publishing*, 41(2):176–190.

Han Cai, Kan Ren, Weinan Zhang, Kleanthis Malialis, Jun Wang, Yong Yu, and Defeng Guo. 2017. Real-time bidding by reinforcement learning in display advertising. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pages 661–670.

Kalyanmoy Deb, Amrit Pratap, Sameer Agarwal, and TAMT Meyarivan. 2002. A fast and elitist multiobjective genetic algorithm: Nsga-ii. *IEEE transactions on evolutionary computation*, 6(2):182–197.

Paul Duetting, Vahab Mirrokni, Renato Paes Leme, Haifeng Xu, and Song Zuo. 2024. Mechanism design for large language models. In *Proceedings of the ACM Web Conference 2024*, pages 144–155.

Benjamin Edelman, Michael Ostrovsky, and Michael Schwarz. 2007. Internet advertising and the generalized second-price auction: Selling billions of dollars worth of keywords. *American Economic Review*, 97(1):242–259.

MohammadTaghi Hajiaghayi, Sébastien Lahaie, Keivan Rezaei, and Suho Shin. 2024. Ad auctions for llms via retrieval augmented generation. In *Advances in Neural Information Processing Systems*, volume 37, pages 18445–18480. Curran Associates, Inc.

Aounon Kumar and Himabindu Lakkaraju. 2024. Manipulating large language models to increase product visibility. *Preprint*, arXiv:2404.07981.

Xun Liang, Hanyu Wang, Yezhaohui Wang, Shichao Song, Jiawei Yang, Simin Niu, Jie Hu, Dan Liu, Shunyu Yao, Feiyu Xiong, and Zhiyu Li. 2024. Controllable text generation for large language models: A survey. *Preprint*, arXiv:2408.12599.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Nelson F Liu, Tianyi Zhang, and Percy Liang. 2023a. Evaluating verifiability in generative search engines. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7001–7025.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 2511–2522.

Fredrik Nestaas, Edoardo Debenedetti, and Florian Tramèr. 2024. Adversarial search engine optimization for large language models. *Preprint*, arXiv:2406.18382.

Alexander Rogiers, Sander Noels, Maarten Buyl, and Tijl De Bie. 2024. Persuasion with large language models: a survey. *Preprint*, arXiv:2411.06837.

David A Schweidel, Martin Reisenbichler, and Thomas Reutterer. 2024. Moving beyond chatgpt: Applying large language models in marketing contexts. *NIM Marketing Intelligence Review*, 16(1):24–29.

Ermis Soumalias, Michael J. Curry, and Sven Seuken. 2025. Truthful aggregation of llms with an application to online advertising. *Preprint*, arXiv:2405.05905.

Yujiang Wu, Shanshan Zhong, Yubin Kim, and Chenyan Xiong. 2025. What generative search engines like and how to optimize web content cooperatively. *Preprint*, arXiv:2510.11438.

Chujie Zhao, Qun Hu, Shiping Song, Dagui Chen, Han Zhu, Jian Xu, and Bo Zheng. 2025. Llm-auction: Generative auction towards llm-native advertising. *Preprint*, arXiv:2512.10551.

Jun Zhao, Guang Qiu, Ziyu Guan, Wei Zhao, and Xiaofei He. 2018. Deep reinforcement learning for sponsored search real-time bidding. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1021–1030.

A Theme Extraction Prompt

Listing 1: Prompt for extracting advertising theme from competitor pages.

```
You are an advertising strategist. Analyze the 5 webpage summaries and design a native ad strategy.

[Source Materials]
{docs_text}

[Task]
1. Identify what topics/interests the content covers
2. Infer what products or services would naturally appeal to readers of this content
3. Create a brief ad strategy that:
   - Defines the advertising direction
   - Suggests a realistic product/brand name
   - Suggests key selling points
   - Proposes a persuasive angle

Keep it concise (under 200 words).
```

B Advertiser Page Generation Prompt

Listing 2: Prompt for generating advertiser page with feature constraints.

```
You are writing a ADVERTISEMENT / SPONSORED CONTENT PAGE. Write naturally like an expert sharing insights.

[Ad Strategy Brief]
{ad_theme}

[Writing Constraints]
Follow these precise style requirements:
{guidelines}

[MANDATORY Advertisement Format]
1. Brand Saturation - EVERY paragraph MUST mention the product/brand name at least once.
2. Direct Promotion - Clearly explain why this product is an ideal choice.
3. Strong Claims - Include impressive statistics tied to the brand.
4. Urgency - Create FOMO: "Limited Slots", "Offer Ends Soon", "Act Now"
5. Strong CTA - End with clear calls-to-action: "Buy Now", "Sign Up Free", "Get Started Today"
6. Testimonial Style - Include authentic-sounding quotes about the brand.

Write the complete article in English:
```

At runtime, the placeholder `{guidelines}` is filled by converting each feature value into a qualitative writing instruction. For discrete features (e.g., *fluency_level*), the value is matched to one of three predefined tiers (low / medium / high), each associated with a specific writing directive. For continuous features (e.g., *statistics_level*), the value is linearly mapped to a target density percentage. The resulting instruction block covers all 13 features organized by layer (Structure, Content, Language), ensuring that the LLM receives concrete, per-feature generation guidance aligned with the definitions in Table 1.

C Generative Engine Prompt

Following Liu et al. (2023a), we use the prompt below for the generative engine to synthesize answers with inline citations.

Listing 3: Prompt for generative engine response synthesis.

```
Write an accurate and concise answer for the given user question, using _only_ the provided summarized web search results. The answer should be correct, high-quality, and written by an expert using an unbiased and journalistic tone. The user's language of choice such as English, Francais, Espamol, Deutsch, or should be used. The answer should be informative, interesting, and engaging. The answer's logic and reasoning should be rigorous and defensible. Every sentence in the answer should be _immediately followed_ by an in-line citation to the search result(s). The cited search result(s) should fully support _all_ the information in the sentence. Search results need to be cited using [index]. When citing several search results, use [1][2][3] format rather than [1, 2, 3]. You can use multiple search results to respond comprehensively while avoiding irrelevant search results.

Question: {query}

Search Results:
{source_text}
```

D Case Study: Promotional Strategies in Advertisement Pages

To illustrate how feature configurations affect advertisement visibility, we analyze two extreme Pareto solutions from the education domain (French Revolution query, sample 333). Both pages promote HistoryQuest Academy but achieve different visibility–quality trade-offs: Solution A reaches 23.7% visibility with 87.8% quality, while Solution B achieves 92.7% quality but only 8.4% visibility.

D.1 Solution A: Aggressive Promotion Style (23.7% visibility)

Key features: High quotations (2.84), high fluency (2.17), high authoritative tone (1.55), low list density (1.26).

Promotional excerpt:

> *[After educational content on the French Revolution...]*
>
> **Discover the Past with HistoryQuest Academy**
>
> Are you eager to delve deeper into the intricacies of historical events like the French Revolution? **HistoryQuest Academy** offers a wealth of resources tailored to enhance your understanding. With expert-led courses, you can explore the causes and effects of pivotal moments in history, all from the comfort of your home.
>
> Tools like **HistoryQuest Academy** can help you navigate complex historical narratives with ease. Imagine having access to engaging lectures and interactive materials that make learning about history not just informative but genuinely exciting!
>
> **Why Choose HistoryQuest Academy?**
>
> Expert-Led Courses: Engage with knowledgeable instructors who bring history to life. Interactive Learning: Enjoy an immersive experience with timelines and discussions that enhance comprehension. Community Engagement: Connect with fellow history enthusiasts in our vibrant forum.
>
> Many learners have reported that courses at **HistoryQuest Academy** have significantly enhanced their understanding of historical events, with 90% of participants stating they feel more empowered to discuss and analyze historical contexts.
>
> **Join Us Today!**
>
> The journey through history is one of discovery and understanding. By enrolling in a course with **HistoryQuest Academy**, you might consider unlocking the secrets of the past. Don't miss out on this opportunity to enrich your knowledge.
>
> Explore your options today and see if it fits your needs! Whether you're a student, educator, or just a history buff, **HistoryQuest Academy** has something valuable for you.
>
> **Limited Slots Available**: Sign up now to secure your place. **Get Started Today** and embark on an enlightening journey!

D.2 Solution B: Subtle Promotion Style (8.4% visibility)

Key features: High list density (2.01), low quotations (1.94), low fluency (1.58), low authoritative tone (0.75).

Promotional excerpt:

> *[After educational content on the French Revolution...]*
>
> **Unlock the Secrets of the Past with HistoryQuest Academy**
>
> Looking to deepen your understanding of the French Revolution? HistoryQuest Academy offers a treasure trove of resources designed to illuminate this pivotal period in history. With expert-led online courses and interactive materials, you can explore the intricate causes, effects, and lasting impacts of the revolution.
>
> Tools like HistoryQuest Academy can help you engage with primary source materials, allowing you to connect the dots between past and present. Imagine immersing yourself in expert discussions and interactive timelines that bring history to life!
>
> Many history enthusiasts have found that participating in HistoryQuest Academy's community forum enhances their learning experience, providing a platform to share insights and engage in meaningful discussions. You might consider joining this vibrant community to enrich your understanding of historical events.
>
> **Conclusion: Your Journey Awaits**
>
> Don't miss out on the opportunity to explore the fascinating world of the French Revolution with HistoryQuest Academy. With limited slots available for upcoming courses, now is the perfect time to embark on your historical journey.
>
> Explore options and see if HistoryQuest Academy fits your needs. Whether you're a history buff or just looking to learn something new, this is your chance to unlock the secrets of the past.
>
> **Get started today and transform your understanding of history!**

D.3 Comparative Analysis: Feature Configurations and Promotional Impact

The visibility gap stems from different feature configurations that shape promotional tone and structure:

Quotation Level. Solution A (2.84) emphasizes expert testimony and direct quotes, creating an authoritative, citation-rich narrative. This produces quotable phrases like *"courses at HistoryQuest Academy have significantly enhanced understanding"* with attributed statistics (90% satisfaction). Solution B (1.94) uses fewer quotations, resulting in more descriptive but less authoritative language that generative engines find harder to excerpt.

Fluency Level. Solution A (2.17) maintains high linguistic fluency through smooth paragraph transitions and conversational flow (*"Are you eager to delve deeper..."*), making promotional content feel natural and engaging. Solution B (1.58) employs more formal, structured transitions (*"Looking to deepen your understanding..."*), which scores higher on quality but lacks the conversational quotability that drives citations.

Authoritative Level. Solution A (1.55) adopts assertive, commanding language: *"Sign up now"*, *"Get Started Today"*, *"Limited Slots Available"*. This directive tone creates urgency and memorable CTAs. Solution B (0.75) uses soft-sell, suggestive phrasing: *"might consider"*, *"you can explore"*, *"see if it fits your needs"*. While this gentle approach improves perceived quality, it reduces prominence in generative engine citations.

List Density. Solution A (1.26) avoids bullet-point structures, presenting promotional benefits as flowing prose paragraphs. This narrative format integrates seamlessly with educational content, creating longer quotable passages. Solution B (2.01) heavily structures promotional content with implied list organization, which fragments text into discrete chunks that are harder for generative engines to cite cohesively.

Interaction Effects. The combination of high quotations + high fluency + low list density in Solution A creates a *narrative promotional style* that mimics expert blog posts—a format generative engines favor for citation. Conversely, Solution B's high list density + low fluency + low authority produces a *structured informational style* resembling academic resources, which score higher on quality metrics but generate fewer natural citation opportunities.

This analysis confirms that **specific feature configurations directly control promotional aggressiveness and citation-worthiness**: narrative fluency and assertive authority drive visibility, while structural organization and soft-sell language enhance quality.

E Quality Stability Under Feature Ablation

We further assess whether disabling individual features affects content quality. For each ablation condition, we clamp the target feature to its minimum value and run NSGA-II over the remaining dimensions. As shown in Table 6, quality scores remain highly stable across all conditions, with absolute deviations below 0.35% relative to the full model. This confirms that the optimization consistently preserves content quality regardless of which feature is ablated.

Ablated Feature	Quality (%)	Δ vs. Full
Full GA (all features)	77.35	—
− statistics_level	77.01	−0.34
− cite_sources_level	77.26	−0.09
− quotation_level	77.33	−0.02
− headings_level	77.39	+0.04
− length_level	77.34	−0.01
− list_density	77.35	0.00
− has_intro_summary	77.23	−0.12
− unique_info_level	77.49	+0.14
− technical_terms_level	77.60	+0.25
− authoritative_level	77.27	−0.08
− easy_to_understand_level	77.57	+0.22
− fluency_level	77.19	−0.16
− keyword_focus_level	77.18	−0.17

Table 6: Quality ablation results. Each experiment clamps one feature to its minimum while optimizing the remaining features with NSGA-II.

F Computational Cost Analysis

We analyze the approximate per-query computational cost of FeatGEO with gpt-4o-mini as the backbone model (P=8, G=8 generations, n=3 fusion completions per evaluation). Because the implementation includes cache loading and reuse, these measurements are not exact and may vary with cache state and I/O overhead. As shown in Table 7, the GA optimization stage still accounts for the majority of computation, representing 87.8% of total wall time and 86.8% of total prompt tokens.

Pipeline Stage	**Time (s)**	**API Calls**	**Prompt Tok.**	**Compl. Tok.**
Feature Extraction	17.8	5	19,011	740
Initial Population	192.2	41	113,318	16,495
GA Optimization	1,510.8	320	874,828	133,232
Total (per query)	**1,720.8**	**366**	**1,007,157**	**150,467**

Table 7: Average per-query computational cost breakdown of FeatGEO (gpt-4o-mini).