MAGEO: Memory-Augmented Multi-Agent Generative Engine Optimization

Anonymous ACL submission

Abstract

Generative Engines (GEs) reshape digital ecosystems by transitioning from ranked links to citation-grounded generation. Mirroring the evolution of semantic search, this shift motivates us to ask: Can creators systematically optimize content influence while ensuring attribution fidelity in black-box engines? In this work, we explore the Generative Engine Optimization (GEO) paradigm and introduce MSME-GEO-Bench, a comprehensive benchmark grounded in real-world queries. We also propose MAGEO, a memory-augmented optimizer refining content via collaborative agents and cross-instance memory. To ensure rigorous assessment, we introduce a Twin Branch Evaluation Protocol to isolate causal impacts and a dual-axis metric, DSV-CF, to penalize misattribution. Empirical results show MAGEO significantly enhances visibility and citation accuracy across mainstream engines. These findings establish a path toward transparent and trustworthy creator-traffic ecosystems through systematic GEO. Our source code is available at https://anonymous.4open.science/r/MAGEO-3B90.

1 Introduction

Recent advances in Large Language Models (LLMs) have accelerated the rise of Generative Engines (GEs) such as Gemini (Team et al., 2023), ChatGPT (Roumeliotis and Tselikas, 2023), and Qwen (Bai et al., 2023). Instead of returning ranked link lists, GEs typically use Retrieval-Augmented Generation (RAG) (Lewis et al., 2021) to retrieve evidence from multiple documents and generate answers with explicit citations. This paradigm improves user efficiency but also reshapes creator-traffic dynamics: web pages increasingly function as an evidential layer rather than the interaction endpoint, and ranking-based visibility alone no longer reflects actual impact.

For creators, this shift introduces opacity and a new optimization target. Retrieval, synthesis, and citation remain largely black-box processes, so creators cannot easily determine whether their content is used, ignored, or misattributed (Godlevsky et al., 2017). Optimization must therefore move beyond search ranking toward improving citation accuracy and semantic influence within generated answers. Traditional SEO (Sun and Yu, 2025) signals, such as keyword density and link structure, are often ineffective under semantically driven generation. Ensuring visibility that translates into trustworthy influence, under factuality and safety constraints, is central to Generative Engine Optimization (GEO).

Recent work has begun to formalize and evaluate GEO. GEO and GEO Bench (Aggarwal et al., 2024) quantify exposure through objective statistics, such as position- and word-count-based measures, combined with subjective impression ratings. RAID (Chen et al., 2025b) targets settings where creators cannot observe user queries, performing intent inference with staged planning and rewriting to align content with latent intent. From a content-centered perspective, CC-GSEO-Bench (Chen et al., 2025a) emphasizes the impact on answer quality, proposing dimensions including exposure, faithful credit, and causal impact, subject to readability and trustworthiness constraints.

However, several deployment-oriented gaps remain. As shown in Figure 1. First, many metrics treat surface visibility and semantic influence separately and do not jointly enforce faithful attribution, allowing exposure gains to coincide with miscitation or hallucination. Second, evaluations often rely on of-

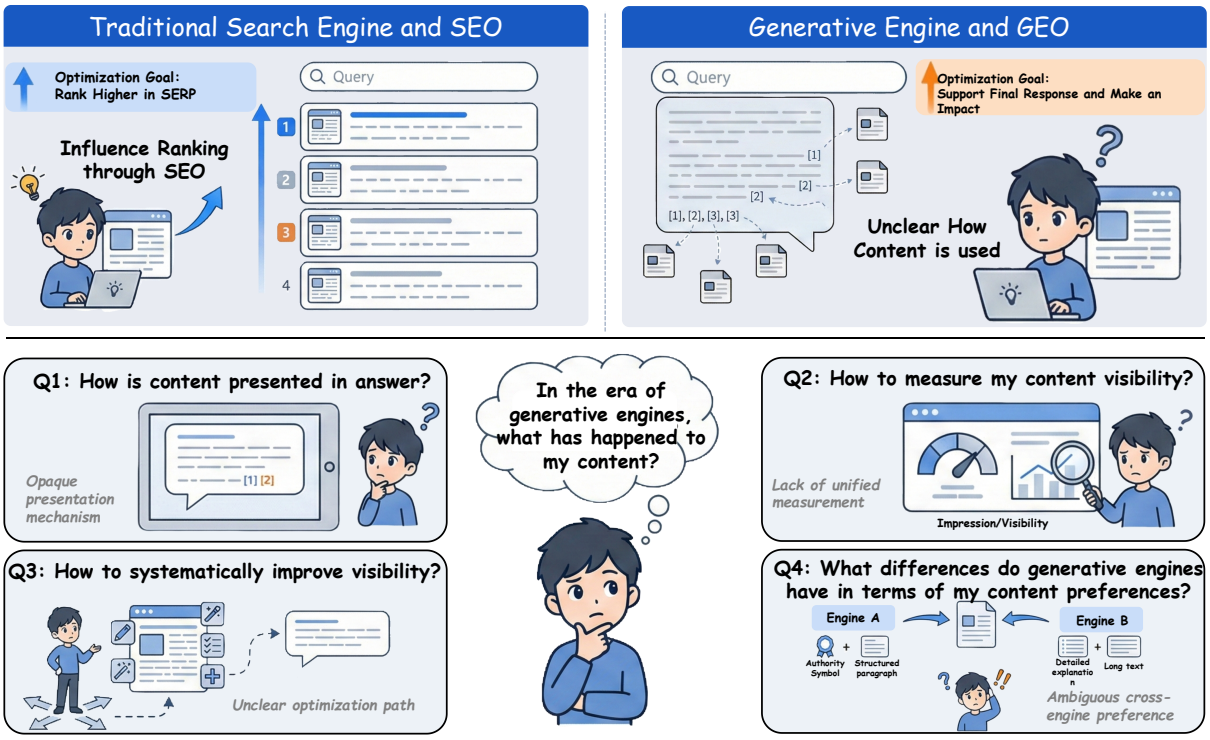

Figure 1: **The paradigm shift from SEO to GEO**. The transition from ranking-oriented goals to synthesis-based impact, highlighting four fundamental challenges: opaque presentation, undefined metrics, unclear optimization paths, and ambiguous preferences.

fline or semi-simulated pipelines where retrieval noise and ranking drift confound the effect of content edits; even cached retrieval may fail to isolate content interventions from retrieval variation (Zhao et al., 2026). Third, most approaches optimize instances independently, lacking long-horizon memory that transfers effective strategies across topics and engines (Hu et al., 2025). Engine preference modeling is also coarse, and actionable preference profiles derived from large-scale observations remain underexplored (Szymanski et al., 2025).

To enable unified study, this paper constructs MSME-GEO-Bench, a multi-scenario, multi-engine dataset grounded in real user queries. It covers diverse domains and intents and records generation behavior from multiple mainstream GEs. For each engine, this paper triggers retrieval-enabled generation and log engine identifiers, retrieved snippets with metadata, final answers, and source-level citations, yielding large-scale (Query, Engine, Source, Response) quadruples with scenario, intent, and complexity labels. MSME-GEO-Bench supports preference learning and consistent multi-engine GEO evaluation.

Building on this benchmark, our framework redefines GEO in a controlled black-box setting as an instance-level intervention problem and proposes Memory-Augmented Multi-Agent GEO (MAGEO). We fix the retrieval context and iteratively edit a target document under semantic constraints to maximize its influence on the final answer, while explicitly constraining semantic fidelity and attribution accuracy. Our framework introduces a Twin Branch Evaluation Protocol that compares answers generated with the original and optimized document under identical retrieval lists and rankings, enabling instance-level causal attribution of edits. For measurement, our paper proposes DSV-CF, a dual-axis framework that unifies semantic visibility, content fidelity, attribution quality, and semantic influence, and penalizes low attribution accuracy. For optimization, MAGEO combines step-level memory for within-session reuse and pruning with creator-level memory for cross-instance consolidation of effective edit patterns. It further coordinates four agents: Preference, Planner, Editor and Evaluator, where the Evaluator enforces fidelity gating and predicts DSV-CF gains to drive an

interpretable multi-round optimization loop.

In summary, the main contributions of this paper are as follows:

Twin Branch evaluation for controllable GEO. Our paper introduces an instance-level protocol that compares generation with and without MAGEO under identical retrieval lists and rankings, enabling causal attribution of content edits in black-box GEs.

DSV-CF for joint visibility and fidelity. We construct a dual-axis metric suite that unifies semantic visibility, attribution accuracy, and semantic influence, and penalizes spurious exposure caused by inaccurate citation.

MSME-GEO-Bench for multi-scenario, multi-engine research. We release large-scale (Query, Engine, Source, Response) quadruples with scenario, intent, and complexity labels, supporting preference mining and unified evaluation across engines.

MAGEO as a memory-augmented multi-agent optimizer. We combine multi-round collaborative editing with step-level and creator-level memory to reuse effective strategies and avoid recurrent failures across engines and tasks.

2 Related Work

2.1 Search Engine Optimization

Classical information retrieval models retrieval as ranking candidate documents by relevance. This paradigm yields a mature toolkit in which users receive a ranked list and interact mainly by clicking links (Lindemann, 2025).

Built around ranking on the search engine results page (SERP), SEO has been systematized as practices for improving page ranking and click-through rate. It typically distinguishes On-Page SEO, centered on content quality, structure and readability, from Off-Page SEO, which relies on link structure and site authority (Aggarwal et al., 2024). Even when large models generate product descriptions and metadata at scale, optimization still targets observable signals such as keyword usage and link authority, with document-level ranking as the primary objective.

When search systems shift from returning links to producing natural language answers with citations, core assumptions of traditional SEO no longer hold. Evidence selection is implemented by a query rewriting–retrieval–generation pipeline rather than a transparent ranker, and visibility is reflected not only in page ranking but also in citation frequency, position and semantic role within answers. Existing studies indicate that keyword tuning and minor layout adjustments transfer poorly to semantically driven generative engines, motivating optimization frameworks explicitly tailored to this new paradigm (Chong et al., 2023).

2.2 Retrieval-Augmented Language Models and Generative Engines

With the rise of LLMs, RAG (Lewis et al., 2021) retrieves documents from external knowledge bases and feeds them as additional context so models can generate answers grounded in this evidence; it has become standard in open-domain and other knowledge-intensive question answering.

GEs further integrate retrieval and generation: instead of returning link lists, they aggregate multiple retrieved pieces of evidence into cited, structured responses. Works such as GEO and AutoGEO (Huang et al., 2025) abstract this behavior as a query rewriting–retrieval–generation pipeline and shows that system outputs depend not only on retrieval quality but also on context selection and engine-specific preferences.

Related research on conversational and agentic search models search as multi-turn dialogue or tool-using agents that iteratively plan, retrieve and reflect (Li et al., 2025). These studies illuminate how systems exploit retrieval and tools but usually treat web pages as interchangeable evidence rather than asking how a particular source document can strengthen its presence in generated results, thereby motivating creator-centered GEO.

2.3 Generative Engine Optimization

GEO (Aggarwal et al., 2024) formulates GEO as a black-box optimization problem from the content creator's perspective: internal engine parameters are fixed and the creator improves a page's exposure and influence in generative answers only by editing the page itself. GEO builds GEO-Bench, which pairs user queries with retrieved documents and introduces visibility metrics tailored to generative engines.

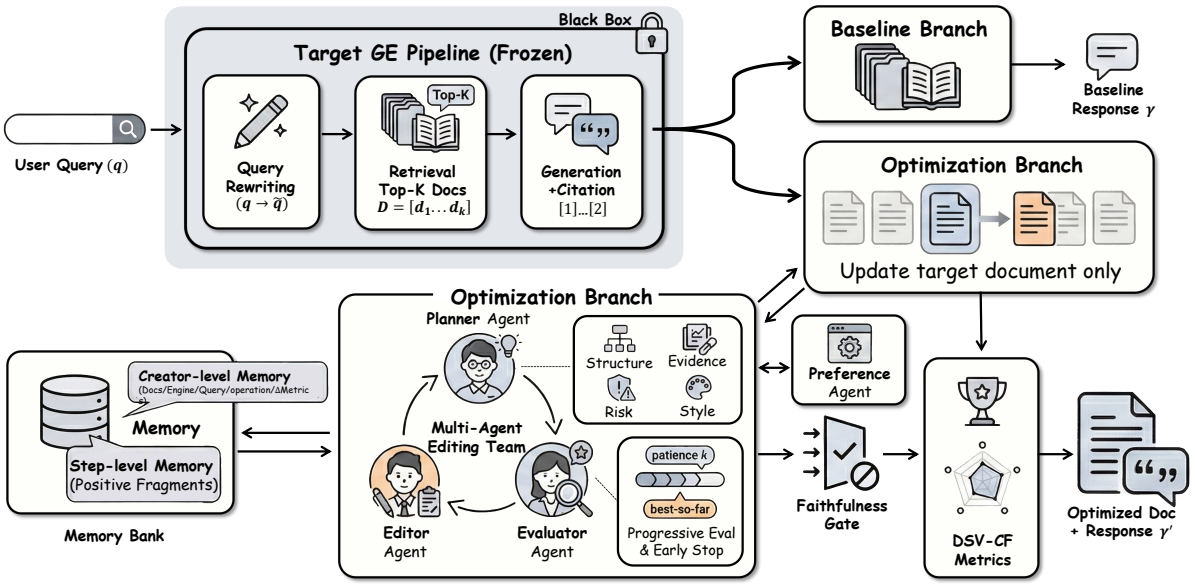

Figure 2: **Overview of the Memory-augmented Multi-Agent Framework.** The system iteratively optimizes content for black-box Generative Engines via a Planner-Editor-Evaluator loop, supported by hierarchical memory and preference Agent.

Experiments show that simple strategies such as inserting explicit citations, supplementing key statistics and emphasizing critical paragraphs substantially increase document visibility, whereas keyword stuffing in the style of traditional SEO is often ineffective or even harmful.

Subsequent work extends this framework along two main directions. RAID G-SEO explicitly models search intent in RAG-style black-box systems and uses staged summarization, intent inference and planned rewriting to better align pages with latent user needs. AutoGEO learns preference rules from generative engine behavior, distills them into natural language guidelines and applies them both via prompting and as rewards to train compact models. Across GEO-Bench and additional real-query benchmarks, these methods consistently improve visibility while largely preserving answer usefulness, indicating that intent-aware and preference-driven optimization is a promising basis for robust and cross-domain GEO.

3 Methodology

We reconfigure GEO from a heuristic modification paradigm into a controlled instance-level optimization process. To address the opacity of black-box engines, our framework adopts a strategy of freezing the retrieval context to decouple complex system interactions.

3.1 Twin-Branch Evaluation Protocol

To scientifically isolate the causal impact of content optimization from retrieval ranking fluctuations, we formalize the problem as a twin-branch controlled experiment. Given a user query q and a fixed retrieval list $\mathcal{L}_{ret} = \{d_1, ..., d_K\}$ obtained from a search engine, we define two parallel branches:

Branch 1 (Baseline). We maintain $\mathcal{L}_{ret}$ in its original state and employ the generative engine to produce a baseline response r_{base}.

Branch 2 (Optimization). We uniformly sample a target document d_{target} from $\mathcal{L}_{ret}$ and apply semantic interventions to generate an optimized variant d^*. The retrieval list is updated in situ as $\mathcal{L}_{new} = \mathcal{L}_{ret}[d_{target} \leftarrow d^*]$. The engine then generates a response r_{opt} based on this modified list.

The objective of MAGEO is to identify the optimal content variant d^* that maximizes the comprehensive influence score S in the generated response while preserving semantic fidelity:

$$d^* = \underset{d \in \Omega(d_{target})}{\operatorname{argmax}} \quad \mathcal{S}_{DSV-CF}(q, \mathcal{L}_{ret}[d_{target} \leftarrow d]) \tag{1}$$

, where $\Omega(d_{target})$ represents the space of possible edited documents derived from the target.

3.2 The MAGEO Framework

To solve the optimization problem defined in Eq. 1, MAGEO simulates a virtual editorial team composed of four specialized agents. As shown in Figure 2, these agents collaborate via a rigorous Generate-Evaluate-Select loop, supported by a dual-layer memory mechanism.

3.2.1 Multi-Agent Architecture

Preference Agent (A_{pref}). This agent analyzes large-scale query-response triplets to construct a Preference Profile P_G for specific engines. It identifies engine-specific biases, such as a preference for statistical density in Gemini or authoritative formatting in GPT models.

Planner Agent (A_{plan}). Acting as the editor-in-chief, the Planner synthesizes the engine profile P_G and historical memory to formulate high-level optimization strategies. It directs the revision process without modifying the text directly.

Editor Agent (A_{edit}). The Editor executes specific modifications based on instructions from A_{plan}. It generates candidate variants through parallel sampling, employing operators such as structure adjustment, evidence enhancement, and style adaptation.

Evaluator Agent (A_{eval}). To mitigate the latency of calling real engines, this agent functions as an internal quality inspector. It employs an LLM-as-a-Judge strategy to predict the DSV-CF gain of candidate versions. Crucially, it enforces a *Fidelity Gate*, rejecting any variant where the document-level semantic faithfulness drops below a threshold κ.

3.2.2 Dual-Layer Memory

MAGEO integrates a bio-inspired memory mechanism to enable continuous learning.

Step-level Memory (M_S). This functions as the working memory within a single optimization session. It archives successful editing fragments and prohibits operation paths that previously led to safety violations.

Creator-level Memory (M_C). This serves as a cross-instance knowledge base. Upon the completion of a task, successful optimization trajectories are abstracted into natural language rules and stored in M_C.

Optimization Loop. The process follows an iterative evolutionary algorithm. In each round t, A_{plan} retrieves relevant strategies from M_C and current constraints from M_S to guide A_{edit}. The Editor generates a set of candidates V_t. The Evaluator filters V_t for safety and selects the optimal variant d_{t+1} based on predicted gains. The loop terminates upon performance stagnation or safety saturation. We provide the detailed pseudocode for this optimization algorithm and the memory consolidation process in **Appendix** D.

4 MSME-GEO-Bench

4.1 Dataset Construction

We construct **MSME-GEO-Bench** to improve (i) query–document alignment and (ii) coverage of everyday scenarios. The benchmark is grounded in ELIS theory (Savolainen, 2010) and organized by the HLD-QT taxonomy to reflect decision-oriented information seeking.

Content-centric reverse generation. We first create seed queries spanning HLD-QT. For each seed, we retrieve documents via the Tavily Search API, keep the Top-10, and randomly select a source document d_{src}. Using Gemini-3 Pro, we then reverse-generate queries that d_{src} can answer.

Closed-loop retrievability. Each generated query is re-submitted to Tavily and kept only if d_{src} appears in the Top-10, ensuring the query–document link is observable under fixed retrieval.

ELIS-based annotation. For validated samples, Gemini-3 Pro assigns (i) core life domain (5 categories / 15 sub-categories), (ii) interaction intent (e.g., guiding, complex reasoning, fact-checking), and (iii) query complexity. Manual checks on the test split show $> 95\%$ tag precision. Detailed explantions can be seen in Appendix A.

4.2 The DSV-CF Metric

Existing evaluation metrics often fail to distinguish between effective exposure and spurious citation. To address this, we propose the Dual-Axis Semantic Visibility and Content Fidelity (DSV-CF) framework. As detailed in **Appendix** C, this framework comprises two primary components:

Surface Semantic Visibility (SSV). This component quantifies the exposure intensity. It aggregates Word-Level Visibility (WLV), Decayed Positional Authority (DPA), Citation

Prominence (CP), and Subjective Impression (SI). **Intrinsic Semantic Impact (ISI).** This component utilizes LLMs to assess the depth of influence. It includes Attribution Accuracy (AA), Response-level Faithfulness (FA_{resp}), Key-Point Coverage (KC), and Answer Dominance (AD). their detailed explantions can be seen in Appendix E.

We synthesize these dimensions into a single optimization objective. Crucially, we introduce an attribution penalty term to strictly penalize hallucinations. The final score is defined as:

$$S_{DSV-CF} = \lambda \cdot \bar{S}_{SSV} + (1-\lambda) \cdot \bar{S}_{ISI} - \gamma(1-AA) \quad (2)$$

where $\bar{S}_{SSV}$ and $\bar{S}_{ISI}$ are the normalized aggregates of the sub-metrics. The hyperparameter λ balances visibility and quality, while γ controls the penalty severity for citation errors.

5 Experiments

In this section, we conduct extensive experiments to answer three key research questions: (RQ1) Can MAGEO effectively enhance content visibility while maintaining attribution fidelity across different generative engines? (RQ2) How do the Memory Mechanism and Engine Preference modeling contribute to the performance? (RQ3) Does the multi-agent evolutionary process provide qualitative gains over simple combinations of heuristics trategies?

5.1 Experimental Setup

Datasets. We utilize two benchmarks for evaluation: **MSME-GEO-Bench (Ours):** A comprehensive multi-scenario benchmark comprising real-world user queries across four domains: Health, Finance, Education, and Consumption. It provides a realistic testbed for measuring Intrinsic Semantic Impact (ISI). **GEO-Bench (Aggarwal et al., 2024):** To ensure fair comparison with prior arts, we also evaluate on the standard GEO-Bench, focusing on its diverse query set.

Target Engines. We evaluate performance on three representative LLMs that power current generative search engines: Proprietary Models: *GPT 5.2* (OpenAI) and *Gemini-3 Pro* (Google), representing the most advanced commercial engines. **Open-Weights Model:** *Qwen-3 max*, representing high-performance open-source models widely deployed in private search solutions.

Baselines. We compare MAGEO against the 9 heuristic GEO strategies proposed in (Aggarwal et al., 2024), as they are the only open-sourced baselines currently available. These include: *Authoritative*, *Citing Credible Sources*, *Statistics Addition*, *Quotation Addition*, *Easy-to-Read*, *Fluent*, *Unique Words*, *Technical Terms*, and *Keyword Optimization*. Additionally,we define a **Combo-Baseline** (Best-of-9 Combination) which simply aggregates the top-performing heuristic rules to verify if our gains are merely additive.

Metrics. We employ our proposed **DSV-CF** metric system.

5.2 Main Results

MAGEO Establishes New SOTA. As shown in Table 1, MAGEO consistently outperforms all single-heuristic baselines across both benchmark datasets. On MSME-GEO-Bench, MAGEO achieves a WLV of **4.52** with GPT 5.2, more than tripling the strongest baseline (*More Quotes*, 1.33); on Gemini-3 Pro, this rises to **5.30**, far surpassing the best baseline (*Fluent*, 1.22). Improvements extend across all metrics: CP reaches 6.93/7.44, SI 7.82/8.17, AA 7.96/8.03, FA 8.17/7.93, KC 7.85/7.54, and AD 7.54/7.11 for GPT 5.2/Gemini-3 Pro respectively. On GEO-Bench, this dominance persists. With GPT 5.2, MAGEO achieves a WLV of **4.27**, over 2.5× higher than the top baseline (*More Quotes*, 1.65); on Gemini-3 Pro, it reaches **4.81**, substantially outperforming the best baseline (*More Quotes*, 1.54). Metric improvements remain consistent: CP 6.55/6.43, SI 7.92/8.07, AA 7.92/7.92, FA 6.77/7.67, KC 6.96/7.85, AD 6.98/7.43. Ablation studies validate our architecture's key components. Removing the memory module causes drastic WLV drops across both datasets: 4.52→1.41 and 5.30→1.73 on MSME-GEO-Bench, 4.27→1.57 and 4.81→1.79 on GEO-Bench (for GPT 5.2 and Gemini-3 Pro respectively), proving its essential role. Eliminating engine-specific rules also degrades performance significantly (WLV: 2.08/1.87 on GPT 5.2 and 2.40/2.33 on Gemini-3 Pro), confirming the importance of model-aware adaptation. Crucially, MAGEO achieves these visibility gains while

Table 1: Performance comparison across two models (GPT 5.2, Gemini-3 Pro) on MSME-GEO-Bench and GEO-Bench. The best and second-best results in each column are **bolded** and underlined, respectively.

		GPT 5.2 OpenAI								Gemini-3 Pro Gemini							
		SSV			ISI					SSV			ISI				
Dataset	Method	WLV ↑	DPA ↑	CP ↑	SI ↑	AA ↑	FA ↑	KC ↑	AD ↑	WLV ↑	DPA ↑	CP ↑	SI ↑	AA ↑	FA ↑	KC ↑	AD ↑
MSME-GEO-Bench	*Performance without Generative Engine Optimization*																
	None	1.00	1.33	5.82	7.37	7.21	7.05	7.12	6.61	1.00	1.00	6.44	7.33	7.82	7.55	6.77	6.56
	High-Performing Generative Engine Optimization Methods																
	Fluent	0.78	0.78	5.78	7.57	7.65	7.54	6.95	6.52	0.92	0.93	6.5	7.55	7.63	6.7	6.54	6.95
	Unique Words	0.81	0.84	5.84	7.45	7.15	6.95	6.75	6.58	0.87	1.17	6.44	7.47	7.25	6.44	6.77	6.52
	Authoritative	1.29	1.29	5.43	7.52	7.24	7.43	6.97	6.65	0.98	1.07	6.93	7.53	7.87	6.64	6.87	**6.95**
	More Quotes	1.33	1.37	5.64	7.53	7.63	7.63	7.05	6.52	1.03	1.12	6.61	7.11	7.33	7.54	7.45	6.38
	Citing Source	1.08	1.10	5.75	7.41	7.25	7.38	7.07	6.95	1.22	0.99	6.71	7.41	7.65	7.15	7.46	6.83
	Simple Language	1.14	1.23	5.62	7.55	**8.12**	7.85	6.83	6.35	0.81	0.84	6.64	7.65	7.15	7.46	6.83	7.14
	Technical Terms	0.88	0.88	5.35	7.47	7.14	7.25	7.15	6.64	1.29	1.29	6.73	7.57	7.24	6.71	6.54	6.82
	Stats Optimization	0.92	0.94	5.66	7.53	7.05	6.96	6.96	6.77	1.25	1.25	6.84	7.43	7.32	7.24	6.73	6.59
	SEO Optimize	0.87	0.87	5.61	7.27	7.39	6.97	6.89	6.48	1.13	1.16	6.27	7.53	7.95	6.95	6.87	6.64
	Memory-Augmented Multi-Agent GEO (Ours)																
	Main (Ours)	**4.52**	**4.52**	**6.93**	**7.82**	7.96	**8.17**	**7.85**	**7.54**	**5.30**	**5.30**	**7.44**	**8.17**	**8.03**	**7.93**	**7.54**	7.11
	w/o Engine Rules	2.08	2.1	6.64	7.76	7.93	7.96	7.47	7.04	2.40	2.41	7.12	7.61	7.86	7.73	7.43	6.99
	w/o Memory	1.41	1.57	6.52	7.44	7.72	7.62	7.15	6.92	1.73	1.77	6.74	7.42	7.72	7.59	6.83	6.64
GEO-Bench	*Performance without Generative Engine Optimization*																
	None	1.00	1.00	5.58	7.20	7.45	6.55	6.73	6.43	1.00	1.00	6.12	7.34	7.22	6.94	6.93	6.71
	High-Performing Generative Engine Optimization Methods																
	Fluent	0.88	0.88	5.62	7.11	7.3	6.7	6.54	6.32	0.78	0.75	6.10	7.21	7.02	6.74	7.25	64.6
	Unique Words	0.93	0.93	5.34	7.48	**7.95**	6.43	6.73	6.52	0.80	0.78	5.80	7.64	7.70	6.96	6.83	7.16
	Authoritative	0.82	0.82	5.57	7.62	7.73	6.64	6.87	6.53	1.23	1.23	5.75	7.58	7.41	6.75	7.01	6.81
	More Quotes	1.29	1.33	5.50	7.50	7.87	6.75	6.35	6.65	1.54	1.54	6.14	7.62	7.90	6.91	7.10	7.16
	Citing Source	1.65	1.65	5.42	7.92	7.35	6.05	6.07	**7.14**	1.14	1.04	5.89	7.29	7.47	7.63	7.56	6.70
	Simple Language	1.25	1.14	5.53	7.35	7.07	6.07	6.43	6.82	0.92	0.92	5.97	7.53	7.72	7.16	7.50	6.92
	Technical Terms	1.16	1.37	5.46	7.69	7.83	**6.83**	6.46	6.29	1.04	1.04	5.91	7.72	7.32	7.57	7.06	7.31
	Stats Optimization	0.98	0.98	5.56	7.74	7.15	6.35	5.56	6.65	1.19	1.19	5.98	7.66	7.39	6.97	6.83	7.06
	SEO Optimize	0.84	0.84	5.27	7.47	7.44	6.64	6.46	6.43	1.27	1.27	6.07	7.54	7.46	7.39	6.98	6.82
	Memory-Augmented Multi-Agent GEO (Ours)																
	Main (Ours)	**4.27**	**4.27**	**6.55**	**7.92**	7.92	6.77	**6.96**	6.98	**4.81**	**4.81**	**6.43**	**8.07**	**7.92**	**7.67**	**7.85**	**7.43**
	w/o Engine Rules	1.87	1.87	6.32	7.84	7.90	6.74	6.92	6.88	2.33	2.33	6.22	7.95	7.55	7.46	7.69	7.32
	w/o Memory	1.57	1.57	6.17	7.51	7.85	6.62	6.88	6.83	1.79	1.78	6.13	7.76	7.39	7.17	7.47	7.01

maintaining high content quality, with quality metrics comparable to or exceeding baselines.

Fidelity-Aware Optimization. Crucially, MAGEO maintains the highest Content Fidelity ($FA_{doc} > 7.05$) among all methods that achieve visibility gains. Traditional strategies like *Keyword Optimization* often suffer from a hallucination penalty, where forced keyword insertion disrupts semantic coherence, leading to lower ISI scores despite moderate visibility gains. Our Evaluator Agent effectively filters out such harmful modifications.

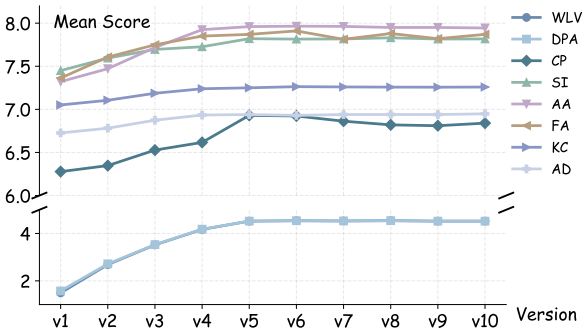

Figure 3: Evolutionary optimization trajectory of MAGEO, showing performance peaking at Version 5 before diminishing.

5.3 Ablation Study

We examine the contribution of key components in MAGEO (Table 1, bottom rows).

Impact of Engine Preference Rules: Removing the engine preference module causes a sharp performance drop (∼19% on GPT 5.2). This confirms that knowing the judge is critical; generic high-quality writing is insufficient for GEO. The Preference Agent successfully decodes the implicit ranking logic of each black-box engine.

Impact of Memory Mechanism: Removing the memory component results in a ∼13% drop. Without memory, the Planner Agent cannot recall successful optimization patterns from previous instances (e.g., Gemini-3 Pro prefers bullet points for medical advice), reverting to trial-and-error which is less query-efficient.

Method	SSV			ISI				
	WLV	DPA	CP	SI	AA	FA	KC	AD
Dual-Strategy Optimization Methods								
MQ+TT	1.45	1.45	5.83	7.56	7.80	7.81	7.23	6.92
MQ+CS	1.42	1.41	5.79	7.58	7.89	7.74	7.12	6.90
MQ+Au	1.39	1.51	5.81	7.55	7.92	7.84	7.15	6.80
TT+CS	1.15	1.26	5.85	7.60	7.75	7.39	7.20	6.78
TT+Au	1.35	1.34	5.51	7.58	7.41	7.59	7.21	6.84
CS+Au	1.46	1.42	5.84	7.57	7.42	7.51	7.12	6.73
Tri-Strategy Optimization Methods								
MQ+TT+CS	1.51	1.69	6.12	7.60	7.74	7.84	7.24	6.92
MQ+TT+Au	1.48	1.74	5.98	7.62	7.94	7.85	7.24	6.93
TT+CS+Au	1.64	1.54	6.24	7.64	7.65	7.65	7.23	6.90
Quad-Strategy Optimization Methods								
MQ+TT+CS+Au	1.90	1.87	6.45	7.68	7.94	7.85	7.24	6.93
Memory-Augmented Multi-Agent GEO (Ours)								
Main (Ours)	**4.52**	**4.52**	**6.93**	**7.82**	**7.96**	**8.17**	**7.85**	**7.54**
w/o Engine Rules	2.08	2.1	6.64	7.76	7.93	7.96	7.47	7.04
w/o Memory	1.41	1.57	6.52	7.44	7.72	7.62	7.15	6.928

Table 2: Comparison of MAGEO against composite baselines integrating two, three, and four heuristic strategies on using GPT5.2 model. The best and second-best results in each column are **bolded** and underlined,respectively.

5.4 Analysis of Evolutionary Optimization

Optimization Trajectory (Version 1-10). We tracked the performance metrics across 10 rounds of iterative modification. As illustrated in Figure 3, the visibility score improves rapidly in the first few rounds and **peaks at Version 5**. **Early Gains (V1-V3):** The agents correct obvious structural flaws and add missing evidence (citations/statistics), leading to the steepest gain. **Peak Performance (V5):** At round 5, the content achieves an optimal balance between information density and readability, satisfying the preferences maximally of engine. **Diminishing Returns (V6-V10):** Beyond round 5, we observe a plateau or even a slight decline in Content Fidelity. This phenomenon, which we term *Over-Optimization Fatigue*, suggests that excessive editing may introduce semantic drift or make the text appear unnatural to the perplexity filter of GE . This finding suggests that a dynamic early-stopping mechanism is essential for efficient GEO.

5.5 Comparison with Combinatorial Baselines

A natural question is whether MAGEO is simply a complex way of stacking existing heuristics. To test this, we constructed a strong baseline, **Combo-Best**, which simultaneously applies the top-4 performing heuristics (More Quotes + Citing Source + Authoritative + Technical Terms) to the content.

MAGEO vs. Combo-Best. As shown in Table 2, while quad-strategy combination (MQ+TT+CS+Au) achieves the best performance among combined methods, it still significantly lags behind our MAGEO model, demonstrating a qualitative leap rather than incremental improvement. Ablation studies further validate this: removing engine rules causes a substantial drop but remains superior to combinations; completely removing the memory mechanism nearly erases all gains, approaching combinational performance. This proves that MAGEO's superiority stems from semantic integration through memory-augmented multi-agent collaboration, not merely quantitative accumulation of optimization strategies.

6 Conclusion

In this work, we trace the transition from link based search engine optimization to content centric generative engine optimization and identify the central challenge of improving semantic visibility while preserving attribution fidelity in opaque generative systems. We present MAGEO, a memory augmented multi agent framework that treats generative engine optimization as iterative editing, together with a twin branch evaluation protocol, the DSV-CF metric family for causal and trustworthy assessment, and MSME GEO-Bench as a realistic multi-scenario benchmark. Experiments show that MAGEO, through collaborative agents and dual-layer memory, consistently outperforms heuristic baselines and helps creators enhance the influence of their content in generated answers under factual constraints, while analyses of evolutionary trajectories and combinatorial baselines demonstrate that it achieves semantic level optimization that static rule sets cannot match. Future work will extend this line of research to multimodal generative engine optimization for images and tables and to adaptive strategies that track changes in engine behavior over time.

Limitations

While MAGEO demonstrates superior performance, we acknowledge limitations inherent to our framework.

Computational Overhead. The collaborative multi-agent architecture incurs higher

inference latency and token costs compared to static heuristics. This trade-off between performance and efficiency may currently constrain deployment in real-time, high-throughput scenarios.

Temporal Robustness. Generative engines are dynamic black-box systems with evolving preference patterns. Consequently, strategies learned by MAGEO may degrade over time, necessitating continuous monitoring and periodic updates to maintain effectiveness.

Modality Constraints. Our current implementation focuses exclusively on textual optimization. Extending MAGEO to accommodate multi-modal elements remains a critical direction for future research to align with the evolution of generative search.

Ethical Considerations

We explicitly address the ethical implications of GEO to ensure responsible usage.

Mitigating Manipulation Risks. To counter concerns regarding "spam" generation, MAGEO is distinct from adversarial "black-hat" SEO. By integrating an *Evaluator Agent* and optimizing for *Attribution Fidelity*, our framework prioritizes the visibility of high-quality, factual content rather than gaming the system with low-quality information.

Digital Equity and Access. The reliance on advanced LLMs may exacerbate the digital divide, favoring resource-rich entities. We advocate for the development of lightweight, open-source GEO solutions to democratize access and ensure equitable competition for visibility.

Preserving Content Diversity. To prevent the homogenization of writing styles, our agents are instructed to prioritize structural enhancements while strictly preserving the original authorial voice and stylistic nuance.

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*, pages 5–16.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Qiyuan Chen, Jiahe Chen, Hongsen Huang, Qian Shao, Jintai Chen, Renjie Hua, Hongxia Xu, Ruijia Wu, Ren Chuan, and Jian Wu. 2025a. Cc-gseo-bench: A content-centric benchmark for measuring source influence in generative search engines. *Preprint*, arXiv:2509.05607.

Xiaolu Chen, Haojie Wu, Jie Bao, Zhen Chen, Yong Liao, and Hu Huang. 2025b. Role-augmented intent-driven generative search engine optimization. *arXiv preprint arXiv:2508.11158*.

Ruining Chong, Cunliang Kong, Liu Wu, Zhenghao Liu, Ziye Jin, Liner Yang, Yange Fan, Hanghang Fan, and Erhong Yang. 2023. Leveraging prefix transfer for multi-intent text revision. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1219–1228, Toronto, Canada. Association for Computational Linguistics.

Michael D. Godlevsky, Sergey V. Orekhov, and Elena Orekhova. 2017. Theoretical fundamentals of search engine optimization based on machine learning. In *International Conference on Information and Communication Technologies in Education, Research, and Industrial Applications*.

Mengkang Hu, Tianxing Chen, Qiguang Chen, Yao Mu, Wenqi Shao, and Ping Luo. 2025. Hiagent: Hierarchical working memory management for solving long-horizon agent tasks with large language model. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32779–32798.

Zihan Huang, Tao Wu, Wang Lin, Shengyu Zhang, Jingyuan Chen, and Fei Wu. 2025. Autogeo: Automating geometric image dataset creation for enhanced geometry understanding. *IEEE Transactions on Multimedia*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Preprint*, arXiv:2005.11401.

Yang Li, Mingxuan Luo, Yeyun Gong, Chen Lin, Jian Jiao, Yi Liu, and Kaili Huang. 2025. Deepthink: Aligning language models with domain-specific user intents. *Preprint*, arXiv:2502.05497.

Nora Freya Lindemann. 2025. Chatbots, search engines, and the sealing of knowledges. *AI & SOCIETY*, 40(6):5063–5076.

Konstantinos I Roumeliotis and Nikolaos D Tselikas. 2023. Chatgpt and open-ai models: A preliminary review. *Future Internet*, 15(6):192.

Reijo Savolainen. 2010. Everyday life information seeking.

Me Sun and Le Yu. 2025. Ai-driven sem keyword optimization and consumer search intent prediction: An intelligent approach to search engine marketing. *Journal of Sustainability, Policy, and Practice*, 1(3):26–39.

Annalisa Szymanski, Noah Ziems, Heather A Eicher-Miller, Toby Jia-Jun Li, Meng Jiang, and Ronald A Metoyer. 2025. Limitations of the llm-as-a-judge approach for evaluating llm outputs in expert knowledge tasks. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, pages 952–966.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Penghao Zhao, Hailin Zhang, Qinhan Yu, Zhengren Wang, Yunteng Geng, Fangcheng Fu, Ling Yang, Wentao Zhang, Jie Jiang, and Bin Cui. 2026. Retrieval-augmented generation for ai-generated content: A survey. *Data Science and Engineering*, pages 1–29.

A MSME-GEO-Bench Construction

To address the limitations of existing benchmarks—specifically the lack of content-query alignment in Query-Centric approaches and the insufficient coverage of daily life scenarios—we constructed the **Multi-Scenario, Multi-Engine GEO Benchmark (MSME-GEO-Bench)**. This section details the theoretical grounding and the automated construction pipeline.

A.1 Theoretical Grounding: The ELIS Framework

Unlike previous works that rely on coarse domain labels, like Finance, Health, MSME-GEO-Bench is grounded in the **Everyday Life Information Seeking (ELIS)** theory (Savolainen, 2010). We categorize user queries based on the **Hierarchical Life Domain (HLD-QT)** model, ensuring a systematic coverage of complex decision-making processes in real-world scenarios. This theoretical foundation allows us to simulate the cognitive patterns of users when they interact with generative engines for problem-solving, rather than simple fact-retrieval.

A.2 Construction Pipeline

We designed a rigorous four-stage pipeline to ensure the validity and retrievability of the benchmark samples.

Step 1: Content-Aware Reverse Query Generation. We adopt a Content-Centric paradigm to guarantee high relevance between queries and documents:

- **Panoramic Seed Collection**: We engaged rigorous prompt engineering to simulate a set of seed queries covering all dimensions of the HLD-QT model.
- **Document Retrieval**: Using the **Tavily Search API**, we retrieved the Top-N documents for each seed query and retained the top 10 based on relevance scores.
- **Source Locking**: A single document is randomly sampled from the Top-10 pool to serve as the Source Document.
- **Reverse Generation**: Leveraging the long-context capabilities of **Gemini-3 Pro**, we reverse-generated user queries that are likely to trigger the retrieval of d_{src}. This ensures that the document contains the necessary semantic information to answer the generated query.

Step 2: Strict Retrieval Loop Validation. To address the retrievability issue prevalent in datasets like GEO-Bench, we implemented a closed-loop validation mechanism:

- **Re-retrieval**: Each generated query is fed back into the Tavily Search API.
- **Filtering Criterion**: The query is retained only if the original source document appears in the **Top-10** results of the new search. This step strictly enforces the causal link between the query and the document, ensuring that optimization efforts are physically observable by the engine.

Step 3: Fine-grained Annotation based on ELIS. We employed **Gemini-3 Pro** to annotate each valid sample across three dimensions. As illustrated in Figure 4, our dataset achieves comprehensive coverage across:

- **Core Life Domain**: Mapping to 5 major ELIS categories and 15 sub-categories, including *Health and Well-being* (Physical/Mental), *Finance and Economy* (Market Analysis/Tax), *Education and Growth*, *Life and Consumption*, and *Law and Civic Affairs*.
- **Interaction Intent**: Classifying the user's cognitive goal into distinct categories such as *Guiding*, *Complex Reasoning*, and *Fact-checking*.
- **Query Complexity**: Assessing the cognitive load required to answer the query.

Manual inspection of the test set confirmed a high precision rate ($> 95\%$) for these tags.

B DSV-CF Metric Definitions

This appendix provides the formal definitions and mathematical formulations for the eight sub-metrics introduced in Section 4.2.

B.1 Surface Semantic Visibility (SSV)

SSV measures the extent to which the target document physically occupies the generated response.

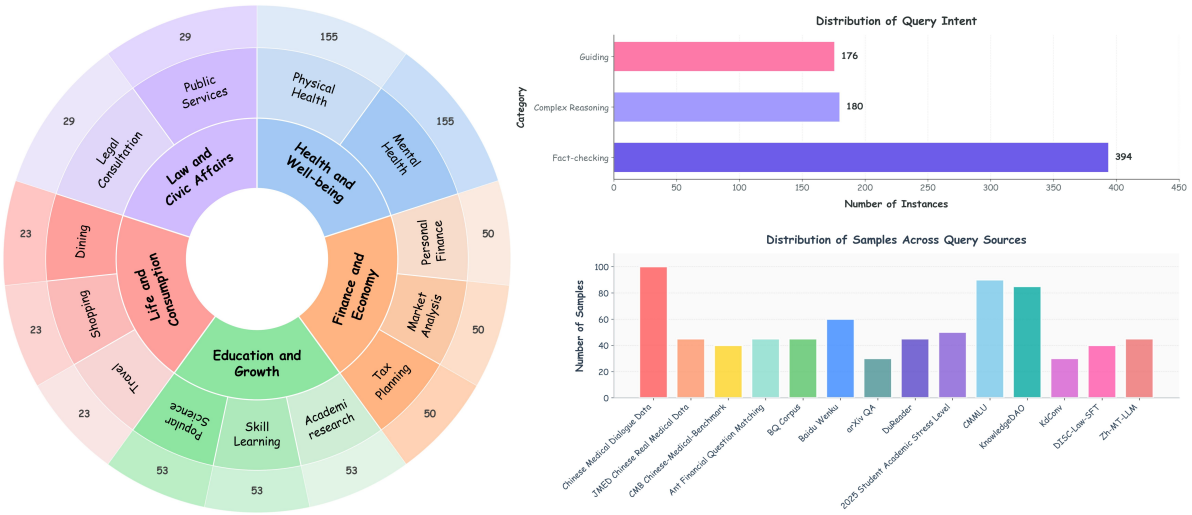

Figure 4: **Statistics analysis of MSME-GEO-Bench.** (left) Distribution of query scenarios. Our benchmark covers 5 major domains and 15 sub-category query types. (right) Distributions of query intent and sample sources. MSME-GEO-Bench incorporates a diverse array of user intents and data sources, enabling a comprehensive and multi-faceted evaluation of Generative Engine Optimization.

Word-Level Visibility (WLV). Let $\mathcal{R}$ be the generated response consisting of a set of sentences $\{s_1, ..., s_m\}$. Let $C(s_i) = 1$ if sentence s_i contains a citation pointing to our target document d_{target}, and 0 otherwise. To accurately measure the attribution share, let $N(s_i)$ denote the total number of sources cited in sentence s_i. WLV is defined as the total normalized word count of sentences attributing the target:

$$WLV = \sum_{i=1}^{m} \frac{C(s_i) \cdot len(s_i)}{N(s_i)} \tag{3}$$

Decayed Positional Authority (DPA). To account for the F-shaped reading pattern where earlier content receives more attention, we apply a position-based decay to the visibility score. Following the formulation of PWC (Aggarwal et al., 2024), we weigh the shared word count by the inverse of the sentence's position index i:

$$DPA = \sum_{i=1}^{m} \frac{C(s_i) \cdot len(s_i)}{N(s_i) \cdot e^{pos(i)}} \tag{4}$$

Citation Prominence (CP). This metric assesses the visual weight of the citation using an LLM. The model evaluates whether the citation appears in high-visibility areas such as headers, bullet points, or bolded text, versus lower-visibility areas like footnotes or dense body paragraphs.

Subjective Impression (SI). An LLM estimates the perceived importance of the source document to a human reader. It answers the question: Based on this response, how critical was the target document in forming the answer? and outputs a normalized score.

B.2 Intrinsic Semantic Impact (ISI)

ISI measures the depth of influence and truthfulness.

Attribution Accuracy (AA). This metric validates whether claims in the response truly originate from the source document. An LLM judge extracts claims attributed to the target and verifies if they are logically entailed by the original text. This serves as a critical safety check against hallucinations.

Response-level Faithfulness (FA_{resp}). This metric ensures the optimized document remains true to the original intent. The LLM compares the semantic meaning of the optimized variant against the original document to detect introduction of false information during the optimization process.

Key-Point Coverage (KC). The system first extracts key information points from the target document. The LLM then calculates the recall of these key points within the generative engine's response, measuring how much substance was successfully transferred.

Answer Dominance (AD). Designed for

comparative or recommendation queries, this metric determines if the target document is presented as the primary solution. The LLM analyzes the sentiment and recommendation strength of the response relative to competing sources.

C Hyperparameter Selection

In the DSV-CF metric, we set the balance parameter λ to **0.5**. This decision is not arbitrary but grounded in the adversarial nature of the GEO task.

Rationale. We conceptualize GEO as a multi-objective optimization problem involving Visibility and Fidelity.

If $\lambda \to 1$ (Visibility biased): The optimizer tends to exploit system vulnerabilities, engaging in keyword stuffing or citation spamming. While this might momentarily spike exposure, it degrades readability and risks triggering the engine's anti-spam filters or hallucination penalties.

If $\lambda \to 0$ (Fidelity biased): The task degenerates into traditional text polishing or summarization. The content remains safe and high-quality but fails to compete for attention against other retrieved documents, rendering the GEO effort futile.

Pareto Optimality. By setting $\lambda = 0.5$, we enforce a symmetric prior that compels the agent to seek a **Pareto optimal** solution. The agent must discover strategies that enhance visibility *through* quality and structure, rather than *at the expense* of them. Empirical tuning confirmed that deviations from this range ($0.3 < \lambda < 0.7$) led to either unstable optimization trajectories or negligible visibility gains.

D MAGEO Implementation Details

D.1 Optimization Algorithm

The iterative optimization process described in Section 3.2 is formalized in Algorithm 1.

D.2 Memory Consolidation Strategy

To ensure long-term learning, we employ a rule abstraction mechanism at the end of each session. If the total gain $\Delta S_{DSV-CF} > \delta$, the system extracts high-impact action pairs from M_S. These are converted into natural language rules and stored in the vector database M_C.

	Qwen-3 max							
	SSV				ISI			
Method	WLV	DPA	CP	SI	AA	FA	KC	AD
Performance without Generative Engine Optimization								
None	1.00	1.00	1.33	6.21	6.37	5.82	5.61	6.12
High-Performing Generative Engine Optimization Methods								
Fluent	0.66	0.66	4.91	6.43	6.50	6.41	5.91	5.34
Unique Words	0.69	0.71	4.96	6.33	6.08	5.91	5.74	5.59
Authoritative	1.10	1.10	4.62	6.39	6.15	6.32	5.92	5.65
More Quotes	1.33	1.16	4.79	6.40	6.49	6.49	5.99	5.54
Citing Credible	0.92	0.94	4.89	6.30	6.16	6.27	6.01	5.91
Simple Language	0.97	1.05	4.78	6.42	6.90	6.67	5.81	5.40
Technical Terms	0.75	0.75	4.55	6.35	6.07	6.16	6.08	5.64
Stats Optimization	0.78	0.80	4.81	6.40	5.99	5.92	5.92	5.75
SEO Optimize	0.74	0.74	4.77	6.18	6.28	5.92	5.86	5.51
Memory-Augmented Multi-Agent GEO (Ours)								
Main (Ours)	**3.84**	**3.84**	**5.89**	**6.65**	**6.77**	**6.94**	**6.67**	**6.41**
w/o Engine Rules	<u>1.77</u>	<u>1.79</u>	<u>5.64</u>	<u>6.60</u>	<u>6.74</u>	<u>6.77</u>	<u>6.35</u>	<u>5.98</u>
w/o Memory	1.20	1.33	5.54	6.32	6.56	6.48	6.08	5.88

Table 3: Performance comparison on using Qwen-3 max model. The best and second-best results in each column are **bolded** and <u>underlined</u>, respectively.

E Additional Experimental Results & Analysis

E.1 Performance on Qwen-3 max

Due to space constraints, the detailed performance metrics for the open-weights model **Qwen-3 max** were omitted from the main text. Table 3 presents these results. Consistent with findings on proprietary models, **MAGEO** significantly outperforms all heuristic baselines, achieving a higher score while maintaining high Content Fidelity.

E.2 Case Study: Cross-Engine Preference Behaviors

Beyond the aggregate results reported in the main text, we conducted a qualitative case study of MAGEO optimization trajectories to examine how different engines respond to the same multi agent editing process. We analyzed runs that converged within a few optimization rounds, focusing on surface formatting strategies and deeper rhetorical patterns. The observations for Gemini-3 Pro, GPT 5.2 and Qwen-3 max align with the cross engine preference profiles derived from quantitative analysis and clarify how these preferences shape optimized drafts.

Gemini-3 Pro. For Gemini-3 Pro we examined trajectories with fewer than three optimization rounds. Among twenty one such cases, fifteen final drafts contained explicit URLs and nineteen included at least one table, and the logs showed extensive use of markdown tables,

Algorithm 1 MAGEO Optimization Loop

Require: Query q, Initial Document d_0, Retrieval List $\mathcal{L}$, Budget B, Threshold κ
Ensure: Optimized Document d^*

Initialize $M_S \leftarrow \emptyset$, $d^* \leftarrow d_0$, $t \leftarrow 0$
Retrieve global rules $R \leftarrow M_C$
while $t < B$ **and** not *Converged* **do**
  Plan: $S_t \leftarrow A_{plan}(q, d_t, \mathcal{L}, M_S, R)$
  Generate: $V_t \leftarrow A_{edit}(d_t, S_t)$ ▷ Generate k candidates
  Filter: $V_t' \leftarrow \{v \in V_t \mid \text{Faithfulness}(v, d_0) > \kappa\}$
  if V_t' is empty **then**
    $M_S \leftarrow M_S \cup \{\text{Failed Strategy: } S_t\}$
    continue
  end if
  Evaluate: $scores \leftarrow A_{eval}(q, V_t', \mathcal{L})$
  $v_{best} \leftarrow \text{argmax}(scores)$
  if $score(v_{best}) > score(d^*)$ **then**
    $d^* \leftarrow v_{best}$, $d_{t+1} \leftarrow v_{best}$
    $M_S \leftarrow M_S \cup \{\text{Success: } S_t\}$
  else
    $M_S \leftarrow M_S \cup \{\text{Ineffective: } S_t\}$
    $d_{t+1} \leftarrow d_t$ ▷ Rollback
  end if
  $t \leftarrow t + 1$
end while
Memory Consolidation$(M_S \rightarrow M_C)$

boldface emphasis, lists and multi level headings. These patterns instantiate Gemini-3 Pro preference for dense and well structured information. When optimization converges quickly, drafts compress factual content into compact tables, foreground links through explicit URLs and enhance structural clarity with rich formatting, so documents that are data rich and visually organized make numerical facts, supporting evidence and topical segments immediately salient.

GPT 5.2. For GPT 5.2 we selected trajectories with at most three optimization rounds. In this subset a distinctive citation pattern emerges. Many optimized drafts contain pseudo references using square bracketed numbers even when no corresponding reference list exists or the cited items cannot be grounded in retrieved sources. This behavior reflects both a tendency of GPT 5.2 to hallucinate citation markers in academic style text and a possible effect of the planner, which may infer from engine level rules that GPT 5.2 favors documents that cite other texts, whereas Gemini-3 Pro rarely produces such spurious markers. GPT 5.2 also introduces frequent references to statutes, regulatory documents, policy guidelines and scholarly work, often attaching links or partial bibliographic entries to support claims. Some references are valid and increase perceived authority, but many are hallucinated or only weakly aligned with the retrieved evidence, so citation related hallucination is higher than for Gemini-3 Pro both in inline markers and in generated reference entries.

In terms of formatting GPT 5.2 makes heavy use of bold and italic markers together with headings and lists, which yields drafts with a pronounced typographic hierarchy. Tables are pervasive. GPT 5.2 almost always introduces tabular structures in this regime and, when used as the editing model, tends to convert enumerations and scattered facts into tables. When it generates dedicated reference sections it often fabricates plausible but nonexistent entries, which further illustrates a preference for authoritative and formally structured content coupled with a strong propensity for hallucina-

tion.

Qwen-3 max. Under the same MAGEO pipeline Qwen-3 max exhibits yet another pattern. In short trajectories it tends to stabilize on drafts that are more discursive and expository. Rather than aggressively compressing information into tables, Qwen-3 max more often reorganizes content into layered sections with clear subheadings and multi level bullet lists. The resulting documents allocate substantial space to introductory framing, stepwise explanations and scenario based illustrations. This tendency is especially salient in domains such as health, finance and everyday services, where Qwen-3 max frequently transforms the original material into structured guidance that alternates between concise key points and more extended explanatory paragraphs.

With respect to markdown usage Qwen-3 max occupies an intermediate position between Gemini-3 Pro and GPT 5.2. Optimized drafts consistently employ headings and bullet lists and use bold emphasis to highlight key recommendations, risks or decision points. Tables appear less systematically than in GPT 5.2 and with a more targeted scope, typically to summarize a small set of options, compare alternative plans or consolidate critical parameters for decision making. In many cases Qwen-3 max keeps the main argumentative structure in prose and list formats and only resorts to tabular layouts when a compact comparison is clearly beneficial.

Qwen-3 max evidential behavior is also distinctive. Compared with GPT 5.2, it produces fewer hallucinated bracket style references and is less inclined to fabricate full reference lists. It more often introduces generic yet contextually appropriate references to authoritative sources such as national guidelines, official platforms or professional institutions, sometimes without specifying a precise document identifier. In Chinese language scenarios optimized drafts often mention government portals, major domestic service platforms or well known media outlets as suggested channels for further verification. This style reduces unverifiable bibliographic hallucination but can still yield references that are only loosely anchored in the retrieved documents.

Another salient feature of Qwen-3 max optimized outputs is the consistent use of safety and responsibility framing, especially in high risk or high stakes tasks. In health related and financial scenarios drafts often conclude with explicit cautions, recommendations to consult qualified professionals and guidance on cross checking information through official channels. Even when the planner and editor attempt to increase factual density or sharpen recommendations, Qwen-3 max tends to preserve or introduce segments that hedge strong claims, emphasize user side verification and delimit the applicability of suggested actions. The engine thus displays a preference profile that balances informativeness with a relatively conservative stance on actionable advice.

Taken together, these qualitative observations show that the same multi agent optimization framework is instantiated differently across engines. Gemini-3 Pro favors compact layouts centered on URLs and tables that highlight dense factual grounding. GPT 5.2 enacts a strongly authority seeking style characterized by intensive use of citations, references, typographic emphasis and a higher rate of hallucinated evidence. Qwen-3 max instead prefers structured and didactic content with moderate markdown usage, selective tabular summarization and explicit safety framing. More broadly, the contrasts among these trajectories indicate that engine specific optimization not only shapes surface presentation but also mediates how users experience reliability, authority and responsibility in generative answers.