From Experience to Skill: Multi-Agent Generative Engine Optimization via Reusable Strategy Learning

Beining Wu[1,*], Fuyou Mao[2,*], Jiong Lin[1,*], Cheng Yang[1,*], Jiaxuan Lu[3], Yifu Guo[4,5], Siyu Zhang[4], Yifan Wu[6], Ying Huang[1], Fu Li[1,†]

[1]Hangzhou Dianzi University, [2]Central South University,
[3]Shanghai Artificial Intelligence Laboratory, [4]Sun Yat-sen University, [5]Ramus,
[6]Hong Kong University of Science and Technology (GuangZhou)

[*]Equal contribution. [†]Corresponding authors.

Correspondence: lifu@hdu.edu.cn

Abstract

Generative engines (GEs) are reshaping information access by replacing ranked links with citation-grounded answers, yet current Generative Engine Optimization (GEO) methods optimize each instance in isolation, unable to accumulate or transfer effective strategies across tasks and engines. We reframe GEO as a strategy learning problem and propose MAGEO, a multi-agent framework in which coordinated planning, editing, and fidelity-aware evaluation serve as the execution layer, while validated editing patterns are progressively distilled into reusable, engine-specific optimization skills. To enable controlled assessment, we introduce a Twin Branch Evaluation Protocol for causal attribution of content edits and DSV-CF, a dual-axis metric that unifies semantic visibility with attribution accuracy. We further release MSME-GEO-Bench, a multi-scenario, multi-engine benchmark grounded in real-world queries. Experiments on three mainstream engines show that MAGEO substantially outperforms heuristic baselines in both visibility and citation fidelity, with ablations confirming that engine-specific preference modeling and strategy reuse are central to these gains, suggesting a scalable learning-driven paradigm for trustworthy GEO. Code is available at https://github.com/Wu-beining/MAGEO.

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 (Shi et al., 2024; Asai et al., 2024). This paradigm improves user efficiency but also reshapes creator-traffic dynamics: web pages increasingly serve 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). Traditional SEO signals (Sun and Yu, 2025), such as keyword density and link structure, are often ineffective under semantically driven generation (Yu et al., 2024; Li et al., 2025a). Optimization must therefore move beyond search ranking toward improving citation accuracy and semantic influence within generated answers, a challenge central to Generative Engine Optimization (GEO). Crucially, effective GEO requires not only per-instance content improvement but also the ability to accumulate reusable optimization strategies that transfer across queries and engines.

Recent work has begun formalizing and evaluating GEO. GEO and GEO-Bench (Aggarwal et al., 2024) quantify exposure via position- and word-count-based measures alongside subjective impression ratings. RAID (Chen et al., 2025b) infers intent through staged planning and rewriting to align content with latent user needs. CC-GSEO-Bench (Chen et al., 2025a) emphasizes the impact on answer quality, proposing dimensions including exposure, faithful credit, and causal impact.

However, several deployment-oriented gaps remain, as shown in Figure 1. First, many metrics treat surface visibility and semantic influence separately without jointly enforcing faithful attribution, allowing exposure gains to coincide with miscitation or hallucination. Second, evaluations often rely on offline or semi-simulated pipelines where retrieval noise and ranking drift confound the effects of content edits (Ru et al., 2024; Jin et al., 2025). Third, and most critically, all existing approaches

Findings of the Association for Computational Linguistics: ACL 2026, pages 43305–43315

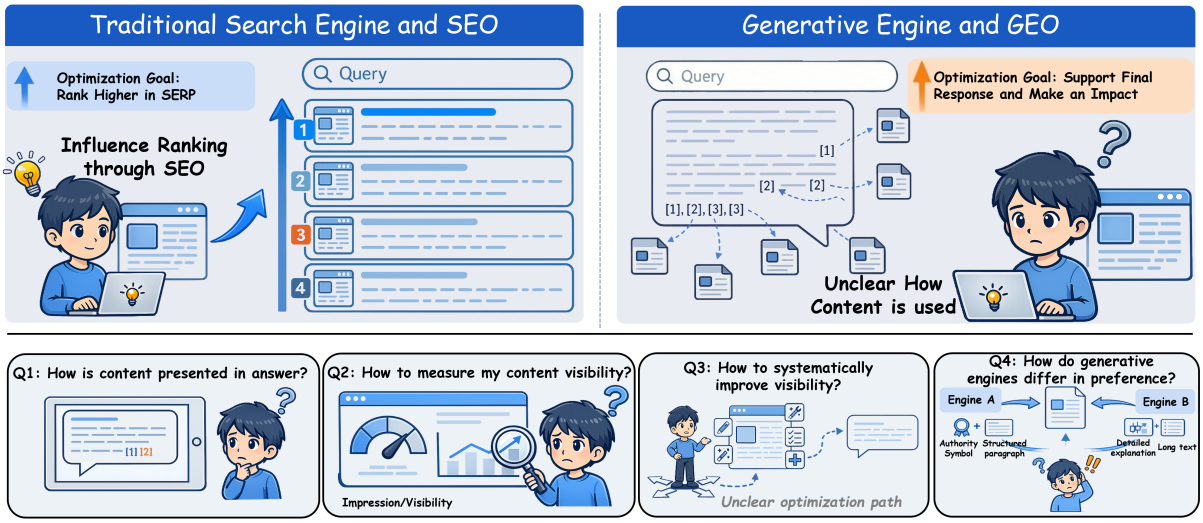

Figure 1: **The paradigm shift from SEO to GEO**. The transition from ranking-oriented goals to synthesis-based impact, highlighting four fundamental challenges: opaque presentation, undefined metrics, unclear optimization paths, and ambiguous preferences.

optimize instances independently, with no mechanism to identify which editing patterns succeeded, abstract them into transferable strategies, or reuse them on subsequent tasks. Engine preference modeling also remains coarse (Szymanski et al., 2025). As a result, current GEO remains trapped in per-instance trial-and-error rather than evolving into a cumulative, skill-building process (Zhong et al., 2024; Gupta et al., 2024).

In this work, we reframe GEO as a strategy learning problem and propose MAGEO, a multi-agent framework that operates on two layers. At the *execution layer*, four coordinated agents (Preference, Planner, Editor, and Evaluator) collaborate through an iterative Generate-Evaluate-Select loop in which the Evaluator enforces a fidelity gate and predicts DSV-CF gains to select the best candidate (Liang et al., 2024; Bo et al., 2024). At the *learning layer*, validated editing patterns are distilled into reusable, engine-specific optimization skills: within a session, effective trajectories and failures are recorded to guide subsequent rounds; across sessions, recurring successful patterns are abstracted into structured strategy skills indexed by engine and scenario for direct reuse (Wang et al., 2025; Zhong et al., 2024). To enable controlled assessment, we introduce a Twin Branch Evaluation Protocol that compares generation with and without optimized content under identical retrieval lists, enabling causal attribution of edits in black-box engines. We further propose DSV-CF, a dual-axis metric that unifies semantic visibility with attribution accuracy while penalizing spurious citation, and construct MSME-GEO-Bench, a multi-scenario, multi-engine benchmark grounded in real-world queries across diverse life domains.

The main contributions are as follows:

A multi-agent framework with reusable strategy learning. We propose a dual-layer architecture in which multi-agent collaboration serves as the execution layer and strategy skill distillation serves as the learning layer. Validated editing patterns are abstracted into engine-specific skills that transfer across tasks, and ablations confirm that both engine-specific preference modeling and strategy reuse contribute measurably to the overall gains.

Twin Branch protocol and DSV-CF for causal, fidelity-aware evaluation. We introduce an instance-level controlled protocol that isolates the effect of content edits from retrieval variation, together with a dual-axis metric suite that jointly measures visibility and attribution quality while penalizing miscitation.

MSME-GEO-Bench for multi-scenario, multi-engine research. We release large-scale (Query, Engine, Source, Response) quadruples with scenario, intent, and complexity labels across diverse life domains and multiple mainstream engines, supporting robust cross-engine strategy evaluation.

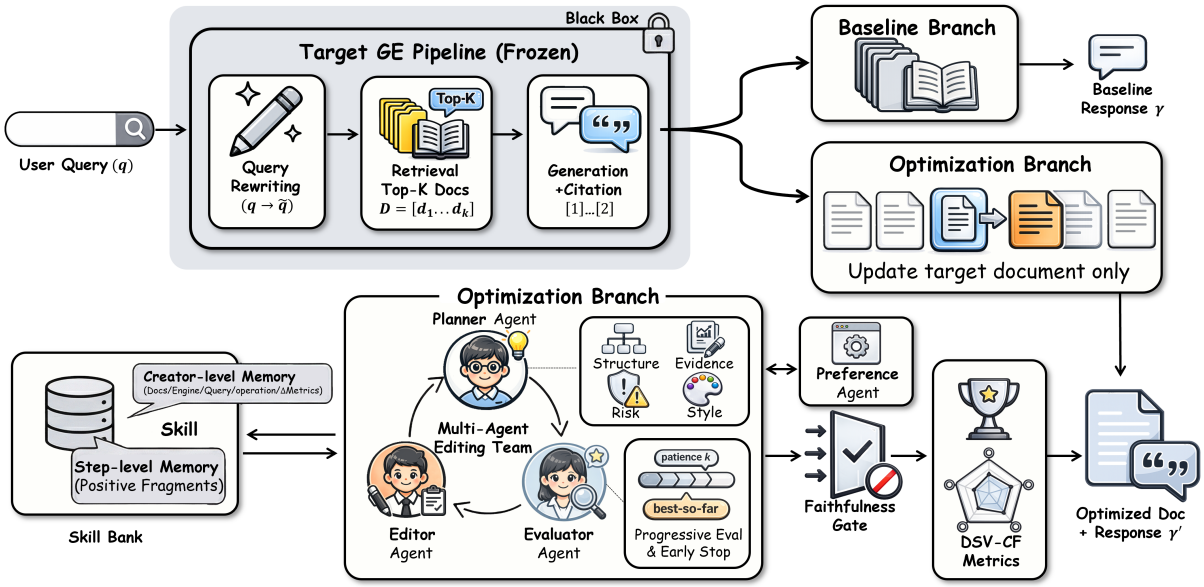

Figure 2: **Overview of MAGEO under the Twin-Branch protocol.** The upper panel compares the baseline branch and the optimization branch under the same frozen retrieval list. The lower panel is a detailed view of the optimization branch, showing how the Preference, Planner, Editor, and Evaluator agents interact with the Skill Bank.

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 retrieved evidence into cited, structured responses. Works such as GEO and AutoGEO abstract this behavior as a query rewriting–retrieval–generation pipeline and show that system outputs depend not only on retrieval quality but also on context selection and engine-specific preferences (Aggarwal et al., 2024; Huang et al., 2025).

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., 2025b). 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.

GEO (Aggarwal et al., 2024) formulates creator-side optimization in black-box generative systems: internal engine parameters are fixed, and creators improve a page's exposure and influence in generative answers only by editing the page itself. GEO-Bench pairs user queries with retrieved documents and introduces visibility metrics tailored to generative engines. Experiments show that simple strategies such as inserting explicit citations, adding 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 reward design. 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:

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

where $\Omega(d_{target})$ denotes the space of candidate edits derived from the target document. This controlled protocol also provides the causal feedback signal for strategy learning: only when the effect of each edit is reliably attributed can the system determine which strategies are worth retaining as reusable skills.

3.2 The MAGEO Framework

To solve the optimization problem defined in Eq. 1, MAGEO operates on two layers, as shown in Figure 2. At the *execution layer*, four specialized agents collaborate through a rigorous Generate-Evaluate-Select loop to iteratively optimize content (Yang et al., 2025, 2026; Yu et al., 2025). At the *learning layer*, validated editing patterns are consolidated into a Skill Bank for reuse on subsequent tasks.

3.2.1 Multi-Agent Architecture

Preference Agent (A_{pref}). This agent analyzes large-scale query-response quadruples to construct a Preference Profile P_G for specific engines. It identifies tendencies such as statistical density, formatting preferences, and rhetorical patterns that are implicitly rewarded by different engines.

Planner Agent (A_{plan}). Acting as the editor-in-chief, the Planner synthesizes the engine profile P_G, the current response state, and relevant strategy skills retrieved from the Skill Bank to formulate high-level revision strategies. It decides *what* should be improved, but does not directly edit the document.

Editor Agent (A_{edit}). The Editor executes the concrete modifications specified by A_{plan}. It generates candidate variants through parallel sampling, including structural adjustment, evidence enhancement, and style adaptation.

Evaluator Agent (A_{eval}). To reduce the latency of repeatedly calling the external engine, this agent functions as an internal quality inspector. It predicts DSV-CF gains using an LLM-as-a-Judge protocol and applies a *Fidelity Gate*, rejecting any vari-

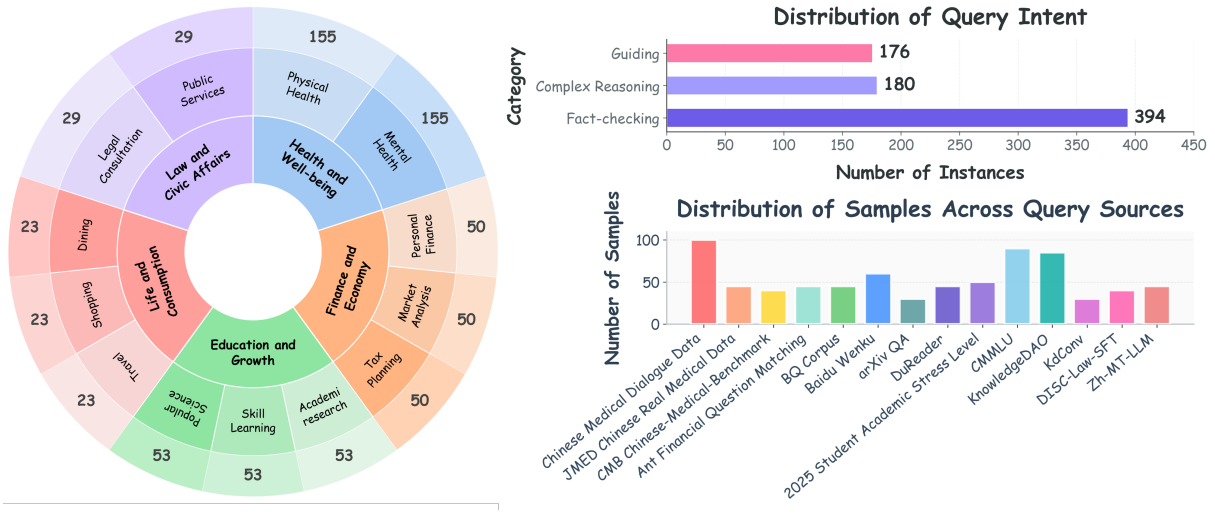

Figure 3: **Statistics analysis of MSME-GEO-Bench.** (left) Distribution of query scenarios. Our benchmark covers 5 major domains and 15 sub-category query types. (right) Distributions of query intent and sample sources. MSME-GEO-Bench incorporates a diverse array of user intents and data sources, enabling a comprehensive and multi-faceted evaluation of Generative Engine Optimization.

ant whose document-level semantic faithfulness falls below a threshold κ.

3.2.2 Skill Bank

The Skill Bank serves as the learning layer of MA-GEO, transforming optimization experience into reusable strategy skills through a three-stage lifecycle: discovery, consolidation, and retrieval (Yu et al., 2026).

Step-level Memory (M_S). This memory supports skill discovery within a single optimization session by recording the outcome of each editing attempt. Variants that produce positive DSV-CF gains are retained as positive fragments, while strategies that trigger fidelity or safety failures are flagged. These records constitute the raw material from which candidate skills are identified.

Creator-level Memory (M_C). This memory handles skill consolidation across optimization sessions. After a successful session, recurring effective patterns in M_S are abstracted into structured strategy skills. Each skill is characterized by its applicability conditions (engine type, scenario), the editing operations it prescribes, and its observed effectiveness (ΔMetrics). Skills are indexed by engine and scenario and stored in the Skill Bank for cross-instance reuse. To keep the Skill Bank scalable, we enforce a capacity limit for each engine-scenario combination and evict entries by recency or usage frequency once the limit is exceeded.

Skill Retrieval. When a new optimization task arrives, A_{plan} queries the Skill Bank with the current engine type and scenario to retrieve matching strategy skills. Retrieved skills serve as prior guidance for revision planning, narrowing the search space and reducing the exploration rounds required.

Optimization Loop. In each round t, A_{plan} retrieves relevant skills from the Skill Bank and current constraints from M_S to guide A_{edit}. The Editor produces a candidate pool V_t. The Evaluator filters V_t using the fidelity gate and selects the best surviving variant d_{t+1} based on predicted gains. The loop terminates when the score plateaus or the edit budget is exhausted. Upon termination, M_S is consolidated into the Skill Bank via skill consolidation, completing the learning cycle.

4 MSME-GEO-Bench

4.1 Dataset Construction

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

Our construction pipeline contains four stages. First, we create seed queries spanning the HLD-QT space, retrieve candidate documents with the Tavily Search API, keep the Top-10 results, and randomly lock one source document d_{src}. We then use Gemini-3 Pro to reverse-generate user queries that d_{src} can answer, enforcing strong semantic align-

ment between the query and the selected source. Second, we perform strict closed-loop retrieval validation by re-submitting each generated query to Tavily and retaining it only if d_{src} still appears in the Top-10. This step ensures that the query–document link is observable under fixed retrieval and makes subsequent optimization effects measurable in a realistic black-box setting. Third, for each validated sample, Gemini-3 Pro assigns three labels: core life domain, interaction intent, and query complexity. As summarized in Figure 3, the resulting benchmark covers 5 major domains and 15 sub-categories together with diverse intent types and source distributions. Finally, to reduce model-specific construction bias, we apply structured prompting, lightweight rule-based filtering, and sampled human quality checks. Manual inspection on the test split shows over 95% tag precision.

4.2 The DSV-CF Metric

Existing evaluation metrics often fail to distinguish effective exposure from spurious citation. To address this, we propose the Dual-Axis Semantic Visibility and Content Fidelity (DSV-CF) framework, which contains 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).

We synthesize these dimensions into a single optimization objective:

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

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. In our experiments, we set $\lambda = 0.5$ to impose a symmetric prior between exposure gain and fidelity preservation: larger values tend to over-reward visibility-oriented edits, whereas smaller values reduce the task to conservative rewriting with limited competitive gain. We use $\gamma = 0.5$ as the default attribution penalty because it provides the best overall DSV-CF among representative settings on the test set.

5 Experiments

In this section, we answer five questions: (RQ1) Can MAGEO improve content visibility while preserving attribution fidelity across different generative engines? (RQ2) How well does the LLM-based DSV-CF judge align with human assessment? (RQ3) What is the cost–effectiveness trade-off of multi-agent GEO? (RQ4) How much do the Skill Bank and engine-specific preference modeling contribute? (RQ5) Does the multi-agent evolutionary process provide gains beyond simple combinations of heuristic strategies?

5.1 Experimental Setup

Datasets. We evaluate on two benchmarks. **MSME-GEO-Bench (ours)** is a multi-scenario benchmark comprising real-world user queries across health, finance, education, consumption, and related daily-life domains. **GEO-Bench** (Aggarwal et al., 2024) is the standard benchmark from prior GEO work and is included for direct comparison with previously published baselines.

Target Engines. We evaluate three representative engines. **Proprietary models:** GPT-5.2 (OpenAI) and Gemini-3 Pro (Google). **Open-weights model:** Qwen-3 max, which represents strong open deployments in private search solutions.

Baselines. We compare MAGEO against the nine heuristic GEO strategies released in the official GEO repository (Aggarwal et al., 2024): *Authoritative*, *Citing Credible Sources*, *Statistics Addition*, *Quotation Addition*, *Easy-to-Read*, *Fluent*, *Unique Words*, *Technical Terms*, and *Keyword Optimization*. These are the only publicly released GEO baselines with fully reproducible implementations at the time of our experiments.

Metrics. We use DSV-CF as the optimization and evaluation metric. We also report the constituent sub-metrics to expose the trade-off among visibility, semantic transfer, and attribution fidelity.

5.2 Human Validation of the LLM-as-a-Judge

Because DSV-CF partially relies on LLM judgments, we validate it against human experts. We stratify and randomly sample 100 quadruple from MSME-GEO-Bench across different scenarios, intents, and complexity levels, including both original responses and MAGEO-optimized responses. Three annotators with NLP backgrounds independently read the query, a source-document summary,

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
	Multi-Agent GEO with Strategy Learning (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 Preference	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 Skill Bank	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	6.46
	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
	Multi-Agent GEO with Strategy Learning (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 Preference	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 Skill Bank	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

Table 2: Human validation of the LLM-based DSV-CF judge on 100 sampled triplets.

Metrics	ρ	95% CI	p-value
DSV-CF	0.81	[0.76, 0.85]	$< 1e{-10}$
WLV	0.79	[0.73, 0.84]	$< 1e{-10}$
CF	0.74	[0.67, 0.80]	$< 1e{-9}$

and the response, then score *visibility*, *semantic influence*, and *attribution faithfulness* on a 1-10 scale. We linearly combine these three human scores using the same weights as DSV-CF and average over annotators to obtain a human counterpart of the metric.

As shown in Table 2, the LLM-based judge exhibits strong agreement with human evaluation on the overall metric and key sub-dimensions, with Spearman correlations of 0.81 for DSV-CF, 0.79 for WLV, and 0.74 for CF. In an additional pairwise comparison on 50 sampled response pairs, the agreement rate between the LLM judge and human experts reaches 81.5%, significantly above random choice. These results suggest that the LLM judge is a reliable proxy for scalable GEO evaluation, although it remains an approximation and should be supplemented with sampled human audits in high-risk settings.

5.3 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**, substantially above the best single heuristic. Improvements are also consistent across CP, SI, AA, FA, KC, and AD. On GEO-Bench, the same pattern persists. These gains indicate that MAGEO is not merely increasing superficial exposure: it improves both semantic transfer and citation faithfulness.

Fidelity-Aware Optimization. MAGEO's visibility gains do not come from indiscriminate ci-

Method	Qwen-3 max							
	SSV				ISI			
	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
Multi-Agent GEO with Strategy Learning (Ours)								
Main (Ours)	**3.84**	**3.84**	**5.89**	**6.65**	**6.77**	**6.94**	**6.67**	**6.41**
w/o Engine Preference	1.77	1.79	5.64	6.60	6.74	6.77	6.35	5.98
w/o Skill Bank	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 underlined, respectively.

tation amplification. Methods such as *Keyword Optimization* can trigger hallucination penalties by forcing lexical patterns that disrupt semantic coherence. In contrast, MAGEO maintains high fidelity ($FA_{doc} > 7.05$) while increasing visibility, indicating that the Evaluator Agent and fidelity gate filter harmful edits.

Generalization to an open-weights engine. We observe the same trend on Qwen-3 Max. Under the same evaluation pipeline, MAGEO increases WLV/DPA from 1.00/1.00 without GEO to 3.84/3.84, while maintaining strong fidelity-related scores (CP 5.89, SI 6.65, AA 6.77, FA 6.94, KC 6.67, and AD 6.41). In contrast, heuristic baselines on Qwen-3 Max remain around WLV $\leq$ 1.33, indicating that MAGEO is not limited to proprietary engines and also transfers to strong open deployments.

5.4 Ablation Study

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

Impact of Engine-Specific Preference Modeling: Removing the engine preference module causes a sharp performance drop ($\sim$19% on GPT 5.2). This confirms that knowing the judge is critical; generic high-quality writing is insufficient for GEO. Successful optimization also requires alignment with engine-specific preferences learned by the Preference Agent.

Impact of the Skill Bank: Removing the Skill Bank results in a $\sim$13% drop. Without accumulated strategy skills, the Planner Agent cannot leverage 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.

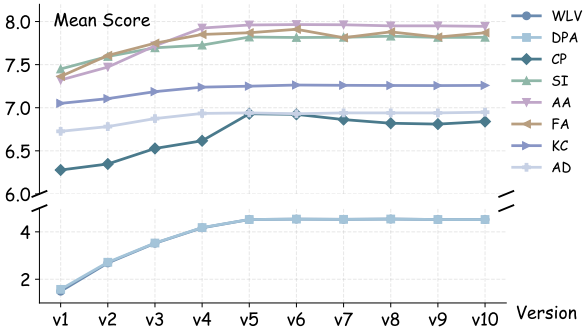

Figure 4: Evolutionary optimization trajectory of MAGEO, showing performance peaking at Version 5 before diminishing.

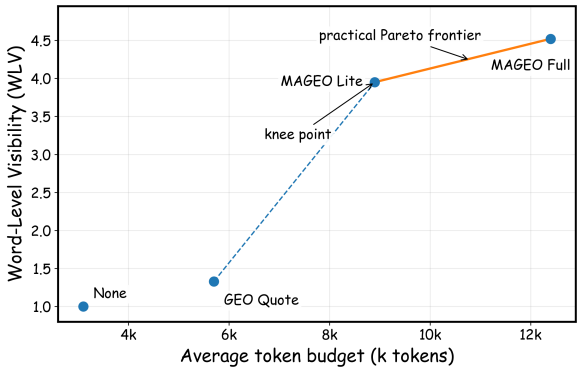

Figure 5: Cost-effectiveness trade-off of MAGEO on MSME-GEO-Bench with GPT-5.2. MAGEO Lite and MAGEO Full form the practical Pareto frontier, with Lite serving as the knee point and the more cost-effective default.

These trends are also qualitatively consistent with cross-engine preference differences: Gemini-3 Pro tends to favor compact and highly structured evidence presentation, GPT-5.2 more often adopts an authority-seeking style with heavier citation formatting, and Qwen-3 Max prefers didactic, safety-aware organization.

5.5 Analysis of Evolutionary Optimization

As shown in Figure 4, the visibility score improves rapidly in the first few rounds and peaks around Version 5. Early gains mainly come from structural repair and missing-evidence completion, with the Skill Bank helping reduce early-round exploration cost by providing validated starting strategies. The peak corresponds to a good balance between information density and readability. Beyond this point, additional edits bring diminishing returns and can

Table 4: Cost-effectiveness comparison on MSME-GEO-Bench with GPT-5.2.

Method	Avg. Tokens	WLV	Avg. Latency
None	3.1k (1.0×)	1.00	–
GEO Quote	5.7k (1.8×)	1.33	12.3s
MAGEO Lite	8.9k (2.9×)	3.95	18.1s
MAGEO Full	12.4k (4.0×)	4.52	38.7s

even slightly reduce faithfulness—a phenomenon we call *over-optimization fatigue*. This observation motivates dynamic early stopping in MAGEO.

5.6 Cost-Effectiveness Analysis

A multi-agent GEO framework is only practically meaningful if its visibility gains are cost-effective relative to additional inference overhead. We therefore measure the total number of input/output tokens across all LLM calls for one query as a unified inference budget and compare four settings on MSME-GEO-Bench with GPT-5.2: *None*, *GEO Quote*, *MAGEO Lite*, and *MAGEO Full*.

Table 4 shows that both MAGEO variants dominate heuristic GEO in the visibility–cost space. **MAGEO Lite** uses $\sim$2.9$\times$ the tokens of GEO Quote but achieves nearly 3$\times$ the WLV, while scaling to Full yields a smaller marginal gain (3.95$\rightarrow$4.52). We therefore recommend **Lite** for cost-sensitive and **Full** for peak-performance applications. Notably, the false-citation ratio (CF) falls from 0.058 (GEO Quote) to 0.047 (Lite) and 0.043 (Full), confirming gains are not driven by hallucinated citations. Paired t-tests confirm both variants significantly outperform GEO Quote ($p < 1e{-}8$), and Full outperforms Lite ($p < 0.01$), though the additional gain is modest. Figure 5 further illustrates that Lite already captures most of the achievable visibility gain (Wu et al., 2025).

5.7 Comparison with Combinatorial Baselines

A natural question is whether MAGEO is simply a more complex way to stack existing heuristics. To test this, we construct **Combo-Best**, which combines the strongest heuristic rules (More Quotes + Citing Source + Authoritative + Technical Terms) in a single pipeline.

As shown in Table 5, Combo-Best is stronger than any individual heuristic but still remains substantially below MAGEO. This indicates that MAGEO's gains are not reducible to additive rule composition. The advantage comes from iterative, engine-aware, and fidelity-aware coordination across agents, combined with reusable strategy skills that guide optimization beyond what static rule composition can achieve.

	SSV			ISI				
Method	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
Multi-Agent GEO with Strategy Learning (Ours)								
Main (Ours)	**4.52**	**4.52**	**6.93**	**7.82**	**7.96**	**8.17**	**7.85**	**7.54**
w/o Engine Preference	2.08	2.1	6.64	7.76	7.93	7.96	7.47	7.04
w/o Skill Bank	1.41	1.57	6.52	7.44	7.72	7.62	7.15	6.92

Table 5: Comparison of MAGEO against composite baselines integrating two, three, and four heuristic strategies on using GPT-5.2. The best and second-best results in each column are **bolded** and underlined, respectively.

6 Conclusion

We reframe GEO as a strategy learning problem and propose MAGEO, a multi-agent framework coupling iterative optimization with reusable skill distillation. Together with Twin Branch, DSV-CF, and MSME-GEO-Bench, it provides a unified pipeline from causal assessment to skill accumulation. Experiments on three mainstream engines confirm substantial gains over heuristic baselines in both visibility and citation fidelity, while ablations validate the contribution of engine-specific preference modeling and strategy reuse. These results suggest GEO is best approached not as ad hoc rule engineering but as a structured learning process in which optimization experience is consolidated into transferable skills. Future work will address multimodal GEO and adaptive skill maintenance that tracks engine distribution drift.

7 Acknowledgements

This paper is part of the 2025 Special Social Science Research Project of the Ministry of Education (Higher Education Counselors Research) entitled "Research on Mechanism Innovation of Systematic Transformation of Ideological and Political Education in Universities under the Background of New AI Technology Application" (Project No. 25JDSZ3109)

Limitations

Despite its strong empirical performance, MA-GEO still has several limitations. First, its multi-agent optimization loop introduces higher token cost and latency than lightweight heuristic methods, which may limit deployment in real-time or high-throughput settings. Second, although MSME-GEO-Bench provides structured annotations and realistic generation logs, its current scale and category distribution still constrain fine-grained subgroup analysis. Third, because Gemini-3 Pro is involved in reverse query generation and annotation, the benchmark may retain some model-specific bias despite our retrieval validation, rule-based filtering, and sampled human checks. Fourth, while the Skill Bank demonstrates measurable gains in ablation, we do not yet provide a formal analysis of skill generalization across unseen scenarios or a learning curve showing how optimization efficiency scales with accumulated experience. Finally, as generative engines evolve over time, learned skills may gradually lose effectiveness, and our current framework is limited to text-only GEO rather than multimodal settings.

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.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

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*.

Xiaohe Bo, Zeyu Zhang, Quanyu Dai, Xueyang Feng, Lei Wang, Rui Li, Xu Chen, and Ji-Rong Wen. 2024. Reflective multi-agent collaboration based on large language models. In *Advances in Neural Information Processing Systems*, volume 37.

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*.

Priyanshu Gupta, Shashank Kirtania, Ananya Singha, Sumit Gulwani, Arjun Radhakrishna, Gustavo Soares, and Sherry Shi. 2024. Metareflection: Learning instructions for language agents using past reflections. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8369–8385, Miami, Florida, USA.

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*.

Jiajie Jin, Xiaoxi Li, Guanting Dong, Yuyao Zhang, Yutao Zhu, Yongkang Wu, Zhonghua Li, Ye Qi, and Zhicheng Dou. 2025. Hierarchical document refinement for long-context retrieval-augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3502–3520, Vienna, Austria.

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.

Xiaoxi Li, Guanting Dong, Jiajie Jin, Yuyao Zhang, Yujia Zhou, Yutao Zhu, Peitian Zhang, and Zhicheng Dou. 2025a. Search-o1: Agentic search-enhanced large reasoning models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 5420–5438, Suzhou, China.

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

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. Encouraging divergent thinking

in large language models through multi-agent debate. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17889–17904, Miami, Florida, USA.

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.

Dongyu Ru, Lin Qiu, Xiangkun Hu, Tianhang Zhang, Peng Shi, Shuaichen Chang, Cheng Jiayang, Cunxiang Wang, Shichao Sun, Huanyu Li, Zizhao Zhang, Binjie Wang, Jiarong Jiang, Tong He, Zhiguo Wang, Pengfei Liu, Yue Zhang, and Zheng Zhang. 2024. Ragchecker: A fine-grained framework for diagnosing retrieval-augmented generation. In *Advances in Neural Information Processing Systems*, volume 37. Datasets and Benchmarks Track.

Reijo Savolainen. 2010. Everyday life information seeking.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Richard James, Mike Lewis, Luke Zettlemoyer, and Wen tau Yih. 2024. Replug: Retrieval-augmented black-box language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8371–8384, Mexico City, Mexico.

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*.

Xiaoqiang Wang, Suyuchen Wang, Yun Zhu, and Bang Liu. 2025. R^{3}mem: Bridging memory retention and retrieval via reversible compression. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 4541–4557, Vienna, Austria.

Yifan Wu, Jingze Shi, Bingheng Wu, Jiayi Zhang, Xiaotian Lin, Nan Tang, and Yuyu Luo. 2025. Concise reasoning, big gains: Pruning long reasoning trace with difficulty-aware prompting. *arXiv preprint arXiv:2505.19716*.

Cheng Yang, Hui Jin, Xinlei Yu, Zhipeng Wang, Yaoqun Liu, Fenglei Fan, Dajiang Lei, Gangyong Jia, Changmiao Wang, and Ruiquan Ge. 2026. Lungnoduleagent: A collaborative multi-agent system for precision diagnosis of lung nodules. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 29793–29801.

Cheng Yang, Jiaxuan Lu, Haiyuan Wan, Junchi Yu, and Feiwei Qin. 2025. From what to why: A multi-agent system for evidence-based chemical reaction condition reasoning. *arXiv preprint arXiv:2509.23768*.

Xinlei Yu, Chengming Xu, Zhangquan Chen, Bo Yin, Cheng Yang, Yongbo He, Yihao Hu, Jiangning Zhang, Cheng Tan, Xiaobin Hu, and 1 others. 2026. Dual latent memory for visual multi-agent system. *arXiv preprint arXiv:2602.00471*.

Xinlei Yu, Chengming Xu, Zhangquan Chen, Yudong Zhang, Shilin Lu, Cheng Yang, Jiangning Zhang, Shuicheng Yan, and Xiaobin Hu. 2025. Visual document understanding and reasoning: A multi-agent collaboration framework with agent-wise adaptive test-time scaling. *arXiv preprint arXiv:2508.03404*.

Yue Yu, Wei Ping, Zihan Liu, Boxin Wang, Jiaxuan You, Chao Zhang, Mohammad Shoeybi, and Bryan Catanzaro. 2024. Rankrag: Unifying context ranking with retrieval-augmented generation in llms. In *Advances in Neural Information Processing Systems*, volume 37.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. 2024. Memorybank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19724–19731.