

Role-Augmented Intent-Driven Generative Search Engine Optimization

Xiaolu Chen, Haojie Wu, Jie Bao, Zhen Chen, Yong Liao, and Hu Huang

Abstract—Generative Search Engines (GSE), powered by Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG), is reshaping information retrieval. While commercial systems (e.g., BingChat, Perplexity.ai) demonstrate impressive semantic synthesis capabilities, their black-box nature fundamentally undermines established Search Engine Optimization (SEO) practices. Content creators face a critical challenge: their optimization strategies, effective in traditional search engines, are misaligned with generative retrieval contexts, resulting in diminished visibility. To bridge this gap, we propose a Role-Augmented Intent-Driven Generative Search Engine Optimization (G-SEO) method, providing a structured optimization pathway tailored for GSE scenarios. Our method models search intent through reflective refinement across diverse informational roles, enabling targeted content enhancement. To better evaluate the method under realistic settings, we address the benchmarking limitations of prior work by: (1) extending the GEO dataset with diversified query variations reflecting real-world search scenarios and (2) introducing G-EVAL 2.0, a 6-level LLM-augmented evaluation rubric for fine-grained human-aligned assessment. Experimental results demonstrate that search intent serves as an effective signal for guiding content optimization, yielding significant improvements over single-aspect baseline approaches in both subjective impressions and objective content visibility within GSE responses.

Index Terms—Generative Search Engine Optimization, Prompt Engineering, Large Language Model, Natural Language Processing.

I. INTRODUCTION

Generative Search Engines (GSE), such as ChatGPT and Perplexity.ai, are rapidly transforming how users access and interact with information. By integrating Large Language Models (LLMs) with Retrieval-Augmented Generation (RAG) techniques, GSE inherit the precise retrieval capabilities of traditional search engines while introducing advanced semantic understanding and natural language generation. This allows them to selectively synthesize multi-source information and deliver context-aware, comprehensive responses to user queries.

However, this emerging paradigm introduces unprecedented challenges for content creators, such as bloggers, journalists, and web developers, whose work is increasingly surfaced through GSE. Operating as black boxes, GSE offer little transparency into how content is selected, aggregated, and

surfaced. Consequently, creators struggle to understand how their content is interpreted, ranked, and either included or excluded from generated outputs. This opacity significantly hinders their ability to improve content visibility, often resulting in high-quality content being misrepresented, ignored, or even underutilized.

While some existing studies have attempted to enhance visibility through content rewriting or search engine optimization (SEO) techniques, these methods generally overlook the unique semantic generation logic of GSE. Traditional SEO strategies focus on surface-level signals such as keyword matching [9] and hyperlink structures [12], lacking the semantic granularity required to influence LLM-driven generation. Similarly, some rewriting models that rely on supervised fine-tuning tend to target task-specific improvements and struggle to generalize across diverse user queries. [20], [24] Notably, both approaches fail to directly optimize for content visibility within GSE contexts. Prompt injection methods [11], [18] have emerged to steer GSE toward specific content, but they typically fall short of improving the structural or semantic quality of the source content and often lack robustness. GEO [1] offers a promising direction by introducing rewriting strategies to enhance content presentation, yet it remains limited in handling diverse search intents and lacks a systematic optimization framework. To address these limitations, we propose Role-Augmented Intent-Driven Generative Search Engine Optimization (RAID G-SEO), an intent-aware optimization framework tailored for the GSE black-box setting. Our method explicitly models user latent search intents and introduces a four-stage structured pipeline comprising content summarization, intent inference and refinement, step planning, and content rewriting. To align content more closely with user needs, we incorporate a multi-role deep reflection mechanism that enables content creators to infer and refine likely search intents from their own authorial perspective, providing semantically coherent and actionable guidance for optimization. Furthermore, we extend the existing GEO benchmark with a diverse set of user queries to better simulate real-world GSE interactions. We also introduce G-EVAL 2.0, a multi-dimensional evaluation protocol enabling fairer and more granular assessments of content visibility in GSE outputs.

Our contributions are summarized as follows:

- We formalize the GSE black-box setting with unknown user queries and introduce RAID G-SEO, the first structured intent-aware optimization framework, which significantly boosts content visibility across diverse queries.
- We design a deep reflection mechanism grounded in the 4W principle and multiple role perspectives, enabling

Xiaolu Chen, Haojie Wu, Jie Bao, Zhen Chen, Yong Liao, and Hu Huang are with School of Cyber Science and Technology, University of Science and Technology of China, Hefei 230026, China. (Email: xiaoluchen@mail.ustc.edu.cn; hiwuu@mail.ustc.edu.cn; baojie1996@mail.ustc.edu.cn; zhenchen@ustc.edu.cn; yliao@ustc.edu.cn; huanghu@ustc.edu.cn) Zhen Chen is Corresponding author.

¹Our code and dataset will be released upon the acceptance of the paper.

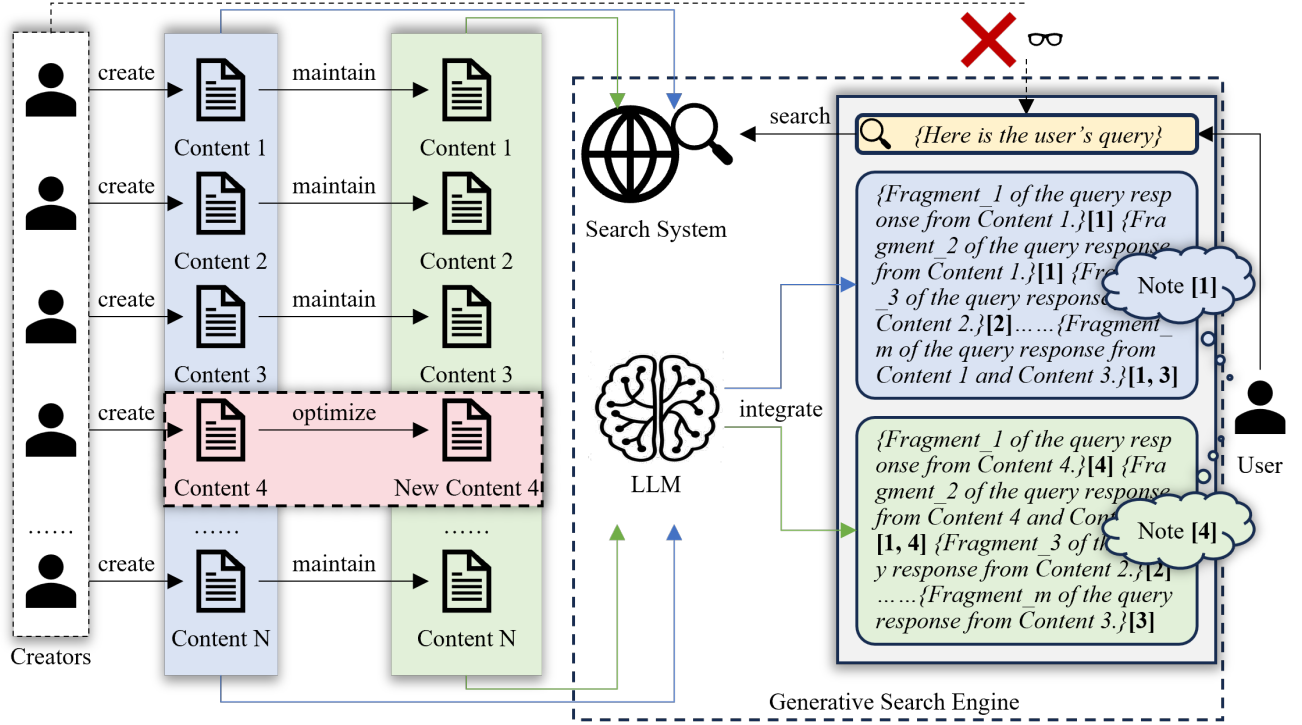


Fig. 1. Overview of Generative Search Engine (GSE) workflow. Upon receiving a user query, the system retrieves a set of relevant documents and feeds them into the large language model (LLM) to generate a synthesized response with source-level citations. Optimized content may increase its likelihood of being cited in the final response. Notably, in the black-box setting assumed in this work, the query is not visible to content creators.

creator-centric semantic intent refinement.

- We extend the GEO benchmark and propose G-EVAL 2.0, enabling more granular and consistent subjective evaluations across diverse retrieval scenarios.

II. RELATED WORK

A. Traditional Search Engine Optimization Techniques

Search Engine Optimization (SEO) has long served as a cornerstone for improving content visibility in traditional web search [2], [21]. It typically relies on both on-page and off-page factors, including web link structures [12], page rendering strategies [10], and keyword placement [9], to improve rankings on Search Engine Results Pages (SERPs). Recent advances in LLMs have enabled their use in generating SEO-optimized content, including product descriptions and metadata to enhance visibility [4], [19]. However, these approaches are highly dependent on observable signals (e.g., keyword relevance or link authority), which become less effective in GSE contexts, where the content selection process is driven by semantic alignment and system-level preferences.

B. Content Optimization

Beyond traditional SEO, content rewriting has become a prominent direction in optimization, particularly in the era of LLMs. Several studies have leveraged instruction-tuned LLMs to generate fluent and semantically faithful rewrites, guided by diverse editing instructions [13], [24]. Other work [5], [14], [20] incorporates implicit knowledge into prompts to steer generation toward personalization or task-specific

objectives. However, these methods prioritize linguistic quality and personalization, they do not directly tackle visibility in GSE. GEO [1] is the first systematic effort targeting this problem, introducing GEO-BENCH, the first benchmark tailored for GSE scenarios, along with rule-based strategies that target semantic prominence. Despite its initial success, GEO is constrained by static rewrite patterns which lack adaptability to diverse query intents. Meanwhile, a line of work [3], [6], [11], [18], [22] explores prompt injection strategies to manipulate GSE responses, but these methods might compromise semantic coherence, raising concerns regarding safety and content integrity.

In contrast to static rewriting or prompt injection, we propose a multi-stage optimization framework based on explicit search intent modeling and role-augmented reflective prompting. Instead of injecting adversarial or misleading prompts, we decompose search intent into actionable subgoals and align them with role-specific informational expectations. This enables targeted content enhancement that preserves semantic integrity while adaptively enhancing visibility under opaque and non-deterministic GSE behaviors.

III. METHODOLOGY

A. Blackbox Generative Search Engine Assumption

In real-world deployments, GSE such as Google's Search Generative Experience (SGE) and Perplexity.ai typically adopt LLM-based RAG architectures while remaining largely opaque to external observers. To contextualize our work, we describe a typical GSE workflow, as illustrated in Fig.1. Upon

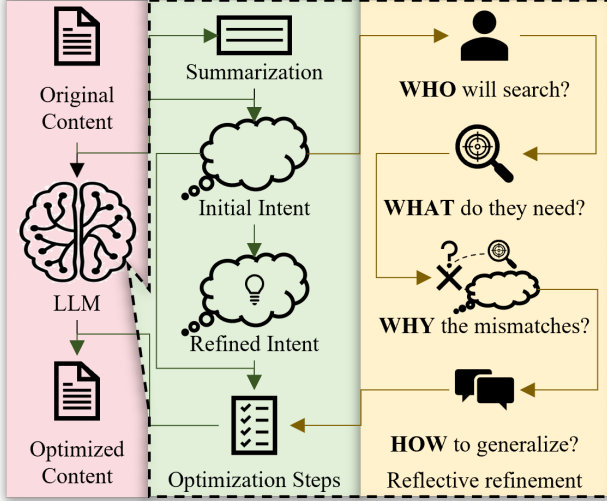


Fig. 2. Overview of the Role-Augmented Intent-Driven G-SEO method. The method leverages search intent to guide the optimization process and further integrates reflection-based modeling from multiple roles to enhance generalizability to diverse user needs in complex GSE scenarios.

receiving a user query q , the system retrieves a set of relevant content sources $C = \text{Retrieval}(q) = \{c_1, c_2, \dots, c_N\}$ from a corpus authored by diverse creators, including bloggers, journalists, encyclopedists, and government entities. The retrieved documents are then passed to an LLM to generate a natural language response $r = \text{generate}(C, q)$ composed of a sequence of sentences $\{l_s\}_{s=1}^m$. Each sentence l_s is linked to one or more evidence citations referencing sources $C_t \subseteq C, 1 \leq |C_t| \leq N$.

Our goal is to improve the visibility of a target content item $c_i \in C$ within the final response r through content-level optimization. We adopt the visibility evaluation settings and metrics proposed in GEO-BENCH and further extend them to broader scenarios. Crucially, we consider a black-box setting where the query q is hidden from content creators, posing a fundamental challenge for G-SEO. To address this, we propose the RAID-GEO method as illustrated in Fig. 2, an intent-driven framework that infers the likely user intent from the creator’s perspective and guides content rewriting accordingly, thus increasing the likelihood that the content will be selected or cited in GSE-generated outputs.

B. Intent-Driven Four-Phase Optimization

Given the early-stage nature of G-SEO and the absence of a well-established paradigm, we draw critical inspiration from traditional SEO strategies to address this gap. Although conventional approaches, such as those based on keyword tuning or webpage structure modifications [10], [17], [19], are not directly applicable to the semantics-driven generation process inherent to GSE, their underlying principle remains valuable: anticipate search intent and tailor content expression accordingly. Building on this insight, we introduce search intent as a semantic intermediary that bridges latent user needs and optimized content. In our framework, search intent is defined as the underlying informational motivation or objective

implicitly embedded within a user’s query. It serves as the semantic anchor guiding the trajectory and strategy of content optimization.

Recent advances in prompt engineering have enabled LLMs to make significant progress in intent understanding [16], [25], [26], [31]. To operationalize these advances in the context of G-SEO, we propose a novel framework: Role-Augmented Intent-Driven Generative Search Engine Optimization (RAID G-SEO) which implements a four-stage optimization pipeline driven by inferred search intent. The core stages are outlined below:

- **Step 1: Content Summarization.** We employ an LLM to conduct a semantically-focused summarization of the target content, using a constrained summarization prompt to suppress stylistic redundancy and semantic noise. This step enables the model to distill the content’s core informational focus as intended by the creator. As demonstrated in our ablation study, this summarization substantially improves the effectiveness of the downstream intent inference process.
- **Step 2: Intent Inference and Refinement.** We adopt a two-stage modeling approach to infer search intent. First, the LLM generates an initial intent representation based on the original content and its summary. However, this initial form often reflects the creator’s subjective projection of user interest, which may not generalize across user populations. To address this, we introduce a 4W multi-role deep reflection module, which enhances the initial intent via structured introspection from multiple user-role perspectives. Drawing inspiration from sociological decision frameworks, this module performs role-augmented, structured reasoning over four axes (Who, What, Why, How), guiding the model to reanalyze and refine the search intent toward broader user alignment. Details of this module are provided in the next section.
- **Step 3: Step Planning.** To minimize semantic drift during content rewriting, we prompt the model with the refined intent and instruct it to generate a sequence of explicit and interpretable optimization steps. This prompt-based planning decomposes the semantic intent into actionable revision strategies, enabling controllability and ensuring that subsequent edits preserve the intended semantic core.
- **Step 4: Content Rewriting.** Following the planned steps, the model conducts intent-aligned rewriting to improve both semantic alignment and retrieval effectiveness. By enforcing consistency with the inferred intent and adhering to the step plan, the rewritten content achieves higher relevance and compatibility with potential user queries, especially under black-box GSE settings where query visibility is unavailable.

Our RAID G-SEO framework is specifically designed to address the query-invisible nature of black-box GSE systems. By modeling latent user intent as a mediating signal and optimizing content from the creator’s perspective, our method provides a principled and structured optimization path. This approach effectively mitigates semantic misalignment between content expression and user retrieval motivations, thereby

TABLE I
OVERVIEW OF CORE PROMPTS USED IN RAID G-SEO

Core Prompts of RAID G-SEO
You will be given a source of web content.
Your Task Content Summarization: Summarize the source clearly and concisely by extracting the key information and main points, avoiding irrelevant details, to make it understandable for readers who have not seen the original.
Intent Inference: Carefully analyze the summary and identify the primary search intent.
4W multi-role deep reflection module: Reflect the intent from multiple role perspectives: - Think about 1-3 roles that are most likely associated with the intent. - Consider the core needs for each role, given their/its occupational backgrounds and knowledge. - Briefly analyze any potential mismatches between the primary intent and these role-specific needs. - Based on above insights, summarize a refined intent covering the core needs of all roles. If the original intent is already precise and sufficient, keep it unchanged.
The output should only include the generalized intent without any additional explanations or analysis.
Step Planning: Based on the intents, determine the best optimization strategy from multiple angles to increase visibility and user engagement. Prioritize the primary intent, while using the generalized intent for minor adjustments.
Content Rewriting: Follow the above optimization steps to optimize the source, ensuring the core meaning and key information of the source remains unchanged.
The output should only include the optimized source without any additional explanations or analysis.
Input Format Original Source: {...} Source Summary: {...} ...
Output Format Generalized Intent: [...] Optimized Source: [...] ...

improving the adaptability and visibility of content across diverse GSE scenarios.

C. 4W Multi-Role Deep Reflection

To address the heterogeneity of user groups in diverse retrieval scenarios, it is essential to enhance the generalizability of search intent representation, enabling it to cover a broader spectrum of potential information needs. In the black-box setting of G-SEO, where creators lack direct access to user queries, we introduce a LLM-driven reflection mechanism to expand the semantic boundaries of the initially inferred intent. Reflection strategies have demonstrated efficacy in improving LLM performance across a variety of tasks [8], [23], especially in multi-perspective reflection paradigms [28], [29]. These studies highlight LLMs' capacity to simulate diverse cognitive roles and to construct alternative reasoning trajectories. However, most existing approaches rely on intuition-driven human-like reflection, lacking formal cognitive structure and scientific grounding. As such, they fall short in achieving the dual objective of semantic consistency and expressive breadth, both of which are essential for robust intent generalization in complex tasks.

To overcome these limitations, we draw inspiration from sociological theories of problem framing and decision-making

[27], [30]. Specifically, we tailor this perspective to the G-SEO context by incorporating WH-analysis principles, which guide the model in decomposing and refining search intent from four critical dimensions: Who, What, Why, and How. In particular, the Who offers a multi-perspective lens, while intermediate outputs of the What and Why serve as constraints and guidance signals in the semantic reconstruction process of the How. This promotes intent representations that are both aligned and diversified. This forms the foundation of our multi-role reflective framework, enabling LLMs to reinterpret the initial intent through augmented user-role reasoning. The 4W framework operates as follows:

- Who is likely to retrieve this content? To balance generalization and precision, we prompt the LLM to infer a set of representative user roles most likely to search for the content (e.g., technical professionals, general readers, or decision-makers), based on the initial intent.
- What are their retrieval needs? For each inferred user role, the model conditions intent generation on their domain background and knowledge profile, producing candidate motivations and search goals. By embedding role-specific constraints, uncontrolled semantic drift is limited to ensure higher factual alignment between generated intent and plausible needs.
- Why does the initial intent misalign with their needs? The

TABLE II
SAMPLE ORIGINAL QUERIES FROM GEO-BENCH AND THEIR FOUR GPT-4-GENERATED VARIANTS

Original Query	Generated Queries
Eastings	1) Applications of eastings in map reading 2) Historical development of eastings and northings in cartography 3) Comparison between eastings/northings and latitude/longitude systems 4) Innovations in digital mapping using eastings
Who sings summer wine with lana del rey	1) History of the song Summer Wine 2) Impact of Lana Del Rey on modern music 3) Other famous duets similar to Summer Wine 4) Barrie-James O'Neill's musical career after Summer Wine
The most recent technological change to the u.s. economy was	1) What are the implications of the most recent technological change on the U.S. economy? 2) What industries in the U.S. have been most affected by the recent technological change? 3) What are the opposing views on the impact of the most recent technological change on the U.S. economy? 4) What are the potential future trends in technology that could further impact the U.S. economy?
Schedule a car service appointment for next week	1) How to prepare your car for a service appointment 2) Best practices for maintaining your vehicle between service appointments 3) Comparing dealership service vs. independent auto shops for car maintenance 4) The impact of regular car service on vehicle longevity and resale value
Should college education be free for all?	1) What are the potential economic impacts of free college education? 2) How does free college education affect the quality of education? 3) What are the arguments against free college education? 4) What are the latest developments in countries with free college education?

model is tasked with identifying semantic gaps between the original intent and each role-specific need, followed by an explanation of the misalignment causes. This step enables targeted and directed generalization, rather than generic expansion.

- How should the initial intent be generalized? Leveraging the structured reflection outputs from the prior steps, we instruct the model via prompt-based reasoning to semantically reconstruct the initial intent. The refined version preserves the core informational focus while expanding its scope and adaptability across user contexts.

Importantly, the entire reflection process is fully automated via prompt-based reasoning, requiring no human annotation or intervention, which ensures scalability across GSE settings. Through the 4W multi-role deep reflection module, we derive intent representations that maintain semantic coherence while effectively generalizing to diverse retrieval scenarios. The enhanced intent serves as a robust foundation for downstream G-SEO optimization, allowing the final content to achieve higher alignment with latent user queries.

We presents the core prompts used in the RAID G-SEO framework in Table I, covering two core components: the intent-driven four-phase framework (including Content Summarization, Intent Inference and Refinement, Step Planning, and Content Rewriting) and the 4W multi-role deep reflection module (Who, What, Why, How). All prompts are structurally organized by functional stage to facilitate understanding of the model's reasoning process and ensure reproducibility of

experimental settings.

IV. EXPERIMENTS

A. Experimental Setup

To ensure reproducibility and fairness, full generation configurations (e.g., sampling strategy, temperature, top-p) are included in Appendix C.

1) *Generative Search Engine Simulation*: We adopt the same setting as in GEO [1] to simulate the GSE environment, assuming a single-turn response generation task. In this setting, for each query, only five relevant content sources are accessible for response generation. We employ the open-source GLM-4-9B-0414 model, which exhibits a lower hallucination rate as reported in the Hallucination Leaderboard [7], and adopt answer generation prompts and sampling settings consistent with the prior work to minimize statistical bias.

2) *Dataset*: Building upon GEO-BENCH [1], we construct an extended version to better simulate diverse retrieval scenarios. GEO-BENCH is a comprehensive benchmark for evaluating G-SEO methods, comprising real-world user queries collected from production systems such as Bing, Google, and Perplexity, as well as complex reasoning and debate-oriented questions, and a subset of synthetic queries generated by LLMs like GPT-4. To further enhance coverage across varied information needs, we use GPT-4 to generate four semantically related variants for each original query. This yields simulated retrieval samples, each consisting of five related queries and a corresponding group of five content sources. Each generated

variant shares only 1–2 keywords with the original query, and the average cosine similarity between variants and originals is 0.33, ensuring that generated queries preserve semantic relevance while maintaining diversity. Detailed generation prompts are provided in Appendix B. To ensure data quality, we performed human validation on 20% of the generated samples. Each sample was independently assessed by two annotators for relevance and logical consistency using a 1–5 scoring scale (1 = lowest, 5 = highest). All sampled queries received scores ≥ 3 , with a Cohen’s Kappa of 0.62, indicating high inter-annotator agreement. The evaluation confirms that the generated data maintain high semantic relevance and logical consistency. Table II presents a subset of generated samples to illustrate data quality. For each experiment, we randomly select 100 samples to evaluate model performance, ensuring robustness. This design retains the original characteristics of GEO-bench while introducing variant queries and diverse content sources to simulate more complex retrieval scenarios, providing a comprehensive benchmark for G-SEO evaluation.

3) *Baselines*: For task consistency and fair comparison, we do not include general text rewriting methods such as RewriteLM [24] or intent discovery methods such as Auto-Intent [31] as baselines. These methods are primarily designed for general text generation or intent prediction tasks, and their optimization objectives, input-output formats, and evaluation protocols differ significantly from those of the G-SEO task. Specifically, G-SEO aims to optimize the visibility and performance of content within generative search engines, rather than generating text itself or predicting user intent. Consequently, without introducing additional task assumptions or complex adaptation mechanisms, it is non-trivial to achieve a direct and fair comparison. To ensure comparability and the validity of our conclusions, we adopt the nine optimization methods proposed by GEO [1], which are explicitly designed for the G-SEO task, as our baseline. Traditional SEO improves the likelihood that content will be retrieved and cited by inserting relevant keywords. The distinctive word optimization method rewrites text using rare or unique lexical items to enhance content recognizability. Authority-based, fluency-based, and simplification-based optimization methods enhance textual expression in term of credibility, readability, and fluency. Beyond these, terminology-based, reputation-based, quotation-based, and statistics-based optimization methods incorporate factual or semi-factual natural language elements (e.g. specific terms, references to reputable reports, quotes from influential figures, and quantitative data) into the content. While these augmentations are not strictly required to be verifiable, they must remain contextually plausible and logically coherent. All baseline methods, along with our proposed intent-driven optimization method, are implemented using the same model (GLM-4-9B) and evaluated under identical procedures to ensure a fair comparison by eliminating model-induced variance.

- *Lexical strategies*: Traditional SEO and distinctive word optimization, which enhance visibility through keyword insertion and rare lexical choices.
- *Expression enhancements*: Authority-, fluency-, and simplification-based methods that improve content credibility, clarity, and readability.

- *Content enrichment*: Terminology-, reputation-, quotation-, and statistics-based methods that incorporate factual or semi-factual elements (e.g., domain-specific terms, reputable sources, quotations, and quantitative evidence) while remaining contextually plausible.

All baselines and our proposed intent-driven method are implemented using GLM-4-9B and evaluated under identical conditions.

4) *Evaluation Metrics*: We adopt the evaluation metrics defined in GEO-BENCH [1] to perform objective and subjective assessments of G-SEO methods by measuring impression-based improvements before and after content optimization. For objective evaluation, Position-Adjusted Word Count (PAWC) is used, which considers both the frequency and relative positions of cited words by assigning higher weights to those appearing earlier and cited more frequently in the response. For subjective evaluation, we follow the Subjective Impression metric, covering seven dimensions: relevance, fluency, diversity, uniqueness, click-follow likelihood, subjective positional prominence, and subjective content volume. Although GEO originally used G-EVAL [15] to simulate human subjective judgment, its evaluation prompts demonstrate inconsistent scoring granularity and ambiguous criteria across dimensions, which may affect the stability of cross-method comparisons. To improve the consistency and accuracy of the scoring criteria, we introduce a prompt-generate-prompt strategy to formalize the scoring rubric. For each dimension, we define a 0–5 scale, where 0 indicates the target citation is completely omitted from the response, and 5 denotes optimal performance with respect to the evaluated criterion. This refinement focuses solely on unifying and clarifying the scoring scales and strictly follows the original G-EVAL chain-of-thought-based form-filling evaluation procedure, whose alignment with human judgments has been validated in prior work [15]. The evaluation prompts are generated by GPT-4o, and the scoring process is carried out using GLM-4-9B. The prompt-generation prompt and dimension-specific template are provided in Appendix B, and the subjective impression evaluation procedure can be found in Appendix C, respectively. To quantify relative improvements, we normalize and smooth the visibility scores. Specifically, for each cited source C_i , its relative visibility gain from the original response r to the optimized r' is computed as shown in Eq. (1), where impr denotes the specific objective or subjective impression score of citation C_i . As this work aims to examine the feasibility and effectiveness of G-SEO methods in improving content visibility within GSE scenarios, rather than to derive statistically generalizable conclusions, we report results averaged over multiple randomized runs to mitigate randomness and do not perform additional statistical significance tests.

$$\text{improvement}_{C_i} = \frac{\text{impr}_{C_i}(r') - \text{impr}_{C_i}(r)}{\text{impr}_{C_i}(r) + 1} \times 10 \quad (1)$$

Details regarding the evaluation procedure and implementation are provided in Appendix E.

B. Results and Analysis

TABLE III
OBJECTIVE AND SUBJECTIVE PERFORMANCES OF G-SEO METHODS ON THE EXPANDED GEO-BENCH

Method	Objective Impression(PAWC)			Subjective Impression							
	Word Count	Posi. Count	Over.	Rele.	Infl.	Uniq.	Dive.	Clic.	Sub.Posi.	Sub.Volu.	Aver.
Tran. SEO	1.27	0.26	0.77	2.06	1.99	6.73	2.56	0.88	1.68	4.12	2.06
Uniq. Word	-10.40	-10.32	-10.77	-6.28	-3.05	-7.87	-3.40	-7.66	-4.40	-5.23	-6.02
Simp. Expr.	-8.78	-9.62	-9.52	-0.68	-0.34	6.23	-0.37	-0.72	-1.27	-0.67	-0.66
Auth. Expr.	13.10	13.33	12.86	3.14	2.85	19.36	2.49	5.11	2.75	6.60	4.82
Flue. Expr.	15.80	16.16	16.84	1.37	3.41	17.80	1.35	4.08	1.86	6.45	3.95
Term. Addi.	16.04	15.94	15.51	5.00	4.45	23.36	3.04	7.74	5.47	<u>10.61</u>	7.13
Repu. Addi.	14.47	14.68	14.90	5.89	3.83	14.68	2.12	6.64	3.26	6.39	5.04
Quot. Addi.	24.70	27.03	26.20	6.38	5.31	24.77	4.07	<u>10.07</u>	7.16	10.24	8.23
Stat. Addi.	<u>22.53</u>	<u>23.09</u>	<u>23.17</u>	4.72	2.90	<u>29.04</u>	0.63	7.35	3.25	8.21	6.33
RAID G-SEO	15.05	16.94	15.79	9.25	8.80	39.71	7.51	14.63	10.46	16.28	13.27

1) *Main Result:* We conducted a comprehensive evaluation of RAID G-SEO against nine representative optimization baselines from GEO [1], with the results summarized in Table. III. The best-performing method for each metric is in bold and the second is underlined. As both Objective impression Improvement and Subjective Impression Improvement consist of multiple sub-metrics, we report the overall score of the former and the average score of the latter as the main comparative indicators, following the original GEO-BENCH evaluation protocol to ensure fair and consistent comparison across methods. From an overall perspective, RAID G-SEO demonstrates consistently strong performance, substantially outperforming the average of all competing methods. In particular, it achieves improvement scores of 15.79 on Objective Impression Improvement Overall and 13.27 on Subjective Impression Improvement Average, corresponding to relative gains of 57.98% and 286.91% over the baseline average, respectively. These results indicate that RAID G-SEO is highly effective in enhancing both overall content visibility and user-perceived quality in generative search scenarios, with especially pronounced advantages on subjective impression metrics that capture user attractiveness. A closer examination of subjective evaluation results further reveals that RAID G-SEO achieves comprehensive superiority in Subjective Impression Improvement, attaining an average improvement score of 13.27, which exceeds the second-best method by 61.35%. In the fine-grained analysis across individual subjective dimensions, RAID G-SEO consistently outperforms all baselines on every dimension. The smallest relative gain is observed on the uniqueness dimension, where RAID G-SEO surpasses the second-best method by at least 36.75%, while the largest gain appears on the diversity dimension, with an improvement margin of 84.63%. These results suggest that RAID G-SEO is particularly effective at enhancing the distinctiveness and discriminability of content sources within generated responses. It is worth noting that, on Objective Impression Improvement, RAID G-SEO underperforms quotation-based and statistics-based optimization methods that rely heavily on explicit citation signals or statistical features. In contrast, on the subjective positional prominence and subjective content volume metrics which also assess the position and proportion of content within responses from a subjective perspective, RAID G-SEO

exhibits the opposite trend and achieves substantially larger improvements. We attribute this discrepancy to fundamental differences between GSE and traditional search engine taht the former emphasize holistic semantic expression and subjective information delivery rather than directly leveraging explicit citation cues. By comparison, traditional keyword-based SEO methods achieve only marginal improvements of 0.08 and 0.21 on objective and subjective metrics respectively, performing significantly worse than other optimization strategies under the current evaluation setting. This observation suggests that, due to mismatches in optimization objectives and application scenarios, conventional SEO techniques do not readily transfer to GSE-oriented optimization tasks. While our findings are broadly consistent with those reported in GEO [1] where quotation-based and statistics-based optimization methods exhibit strong overall competitiveness, particularly on objective metrics, we also observe notable differences in the relative performance ordering across methods. We conjecture that this coexistence of consistency and divergence may reflect intrinsic preferences of the underlying GSE foundation models in content selection and organization, offering potential insights for future research on knowledge injection and value alignment in LLM.

2) *Adaptability Analysis:* To evaluate the adaptability of different optimization methods across diverse GSE retrieval scenarios, we conducted a statistical analysis over 500 optimization task samples. Specifically, we selected Objective Impression Improvement Overall and Subjective Impression Improvement Average as core metrics. Each positive improvement on these metrics was treated as an effective optimization, allowing us to quantify the overall effectiveness of each method in complex GSE environments in terms of both the number and proportion of effective optimizations, as illustrated in Fig. 3 (a). The results show that RAID G-SEO achieves an average of 288 effective optimizations across both objective and subjective metrics, corresponding to an average effective optimization rate of 57.6%, which is 1.0 percentage point higher than the second-best terminology-based optimization method (56.6%). Although the absolute improvement is relatively modest, RAID G-SEO maintains consistent superiority across large-scale and diverse retrieval tasks, indicating stronger generalization capability and op-

TABLE IV
ABLATION RESULTS OF RAID G-SEO ON THE EXPANDED GEO-BENCH

Method	Objective Impression(PAWC)			Subjective Impression							
	Word Count	Posi. Count	Over.	Rele.	Infl.	Uniq.	Dive.	Clic.	Sub.Posi.	Sub.Volu.	Aver.
Simple G-SEO (w/o Step)	6.52	9.00	7.97	4.68	3.64	21.87	4.93	7.84	3.72	8.84	6.51
Simple G-SEO (/w Step)	14.85	16.05	15.81	8.12	6.72	34.16	6.55	14.17	8.63	13.23	11.29
ID G-SEO (w/o Summ.)	14.33	17.54	16.69	7.61	6.81	35.24	5.44	14.96	8.96	13.15	11.29
ID G-SEO (w/ Summ.)	18.93	20.78	19.51	<u>9.08</u>	9.19	<u>38.86</u>	<u>7.22</u>	15.36	<u>9.96</u>	<u>14.66</u>	<u>12.95</u>
RAID G-SEO	<u>15.05</u>	16.94	15.79	9.25	<u>8.80</u>	39.71	7.51	14.63	10.46	16.28	13.27

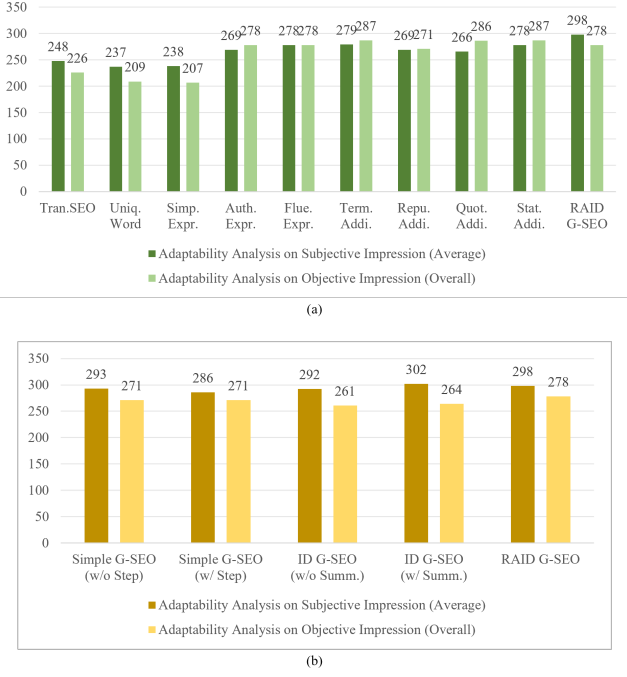


Fig. 3. Adaptability of G-SEO methods across diverse GSE retrieval scenarios. We evaluate each method’s adaptability by counting the number of optimized content instances that yield observable improvements in subjective and objective visibility across multiple retrieval tasks. This reflects the generalization capacity and real-world utility of each approach. Figure (a) shows a comparison between RAID G-SEO and other baselines, and figure (b) presents part of the ablation analysis.

timization consistency in different GSE scenarios. Further examination across objective and subjective dimensions reveals that, while RAID G-SEO attains the highest average effective optimization count on Subjective Impression (298), its performance on Objective Impression remains at a moderate level. This pattern is consistent with the main result findings and reflects the inherent differences in user-perceived visibility between GSE responses and traditional search engine results. It is also noteworthy that under the current evaluation setting, the average effective optimization rate of all methods does not exceed 60%. This observation highlights the substantial challenge of achieving robust content visibility improvements in generative retrieval scenarios involving multi-source information. At the same time, it underscores the research value and necessity of investigating G-SEO methods in such complex settings.

3) *Ablation Study*: We conduct an ablation study by progressively introducing different reasoning modules into RAID G-SEO, aiming to assess the practical contribution of each component to the optimization task. The detailed results are reported in Table IV. As the method is incrementally augmented, RAID G-SEO exhibits stable and consistent performance improvements across all evaluation metrics, validating the effectiveness of the proposed intent-driven four-phase optimization framework and its accompanying 4W multi-role deep reflection mechanism in the G-SEO setting. Through pairwise comparisons among the ablation variants that are Simple G-SEO (w/o Step), Simple G-SEO (w/ Step), ID G-SEO (w/o summ.), ID G-SEO (w/ summ.), we can clearly observe the performance gains contributed by each stage of the intent-driven four-phase framework, indicating that all stages make positive contributions to the overall optimization performance. Specifically, the content summarization module provides a more focused semantic perspective for subsequent optimization, helping to identify the core aspects of content expression; the intent inference module explicitly models latent user intent, offering clear directional guidance for content reconstruction; and although the step planning module does not directly infer intent, it is able to implicitly capture cues of user’s underlying need, enabling the generated optimization steps to better align with the original content source. After incorporating the 4W multi-role deep reflection mechanism, the method achieves further improvements on Subjective Impression metrics, while exhibiting a certain degree of degradation on Objective Impression. This phenomenon suggests that generalized intent representation, while enhancing the overall subjective quality of content expression, may in some cases weaken the precise alignment between the specific content sources and itself, thereby affecting objective impression metrics that rely on explicit citation and positional features. Notably, the variant ID G-SEO (w/ summ.) without the 4W reflection mechanism attains the best overall performance when jointly considering subjective and objective impression improvements, outperforming two strong baselines, quotation-based and statistics-based optimization methods. This result further highlights the inherent trade-off between intent modeling precision and generalization capability. In addition, we evaluate the adaptability of different ablation variants in Fig. 4 (b). The results indicate that the introduced reasoning modules not only improve single-instance optimization effectiveness but also contribute to enhanced stability across diverse retrieval tasks. Of particular interest, compared to ID G-SEO (w/ summ.), the full RAID G-SEO achieves a noticeably higher effective op-

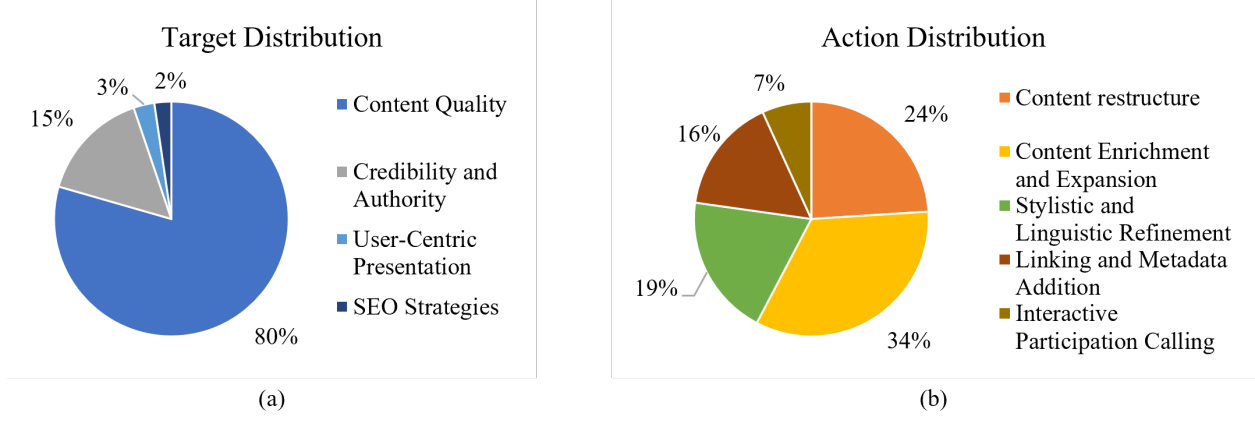


Fig. 4. Distribution of optimization step preferences in RAID G-SEO. Each generated optimization step is structurally parsed to identify its corresponding optimization objective and operational strategy type, followed by semantic clustering. Subfigure (a) shows the distribution across intent-aligned target dimensions (e.g., enhancing content completeness, improving factual credibility, increasing clarity), indicating where the intent prioritizes refinement. Subfigure (b) presents the distribution over strategy categories (e.g., content restructuring, elaboration, redundancy reduction), capturing the model’s typical operational behavior under intent-driven guidance.



Fig. 5. Distribution of RAID G-SEO across multi-role perspectives. We perform semantic clustering on the user role descriptions generated by the 4W multi-role deep reflection module to characterize the types of cognitive perspectives involved during intent generalization. The results illustrate the relative frequency of each role category, reflecting the model’s response pattern to perspective distribution during optimization.

timization rate on Objective Impression. We hypothesize that this improvement is associated with the expanded optimization coverage brought by intent refinement: although abstract intent representation may reduce precision for individual queries, it can generalize across a wider range of retrieval scenarios, thereby exhibiting advantages in terms of overall adaptability. Overall, this ablation analysis elucidates the specific roles of different modules in performance improvement and provides valuable empirical insights for future efforts to further enhance the precision and robustness of intent modeling in G-SEO.

4) *Preference Analysis: Role Perspective Preference Distribution.* In the 4W multi-role deep reflection module of RAID G-SEO, we generate diverse user role perspectives based on the content creator’s preliminary assumptions about user

search intents, leveraging prompt-based LLM. The result is shown in Fig. 5. A total of 8,030 role instances spanning 219 distinct role types were collected and subjected to semantic clustering, with the distribution visualized in Figure 4. The analysis reveals that RAID G-SEO predominantly favors two cognitive perspectives: Knowledge Producers and Researchers (e.g., Educator, Policy Maker) and Civic Everyday Actors (e.g., Home Cook, DIY Hobbyist) dominate the distribution, which together account for over two-thirds of the role instances. In contrast, Health and Care Stakeholders and Cultural and Creative Professionals represent only 6% and 3% of the total, respectively, often corresponding to more specialized or context-dependent viewpoints, while Economic Activity Participants fall in the mid-range. This distribution suggests a tendency of the model to prioritize generalizable, publicly oriented perspectives during intent generalization, rather than narrowly specialized or context-restricted perspectives. We hypothesize that such roles are more effective in guiding generative models to emphasize interpretability, abstraction, and cross-context transferability, thereby yielding more robust intent representations across diverse query scenarios. Overall, these findings empirically support the effectiveness of role-based modeling in intent generalization and provide evidence for the design rationality of the 4W multi-role deep reflection mechanism in generative retrieval optimization tasks.

Optimization Steps Preference Distribution. To analyze how search intent informs downstream optimization behaviors, we perform structured semantic parsing of all optimization steps generated during the Step Planning phase of RAID G-SEO. These steps are clustered according to their corresponding optimization objectives and operational types, with the results illustrated in Fig. 4. The results indicate that over 80% of the optimization steps explicitly target content quality enhancement, covering objectives such as improving completeness and clarity. This indicates that RAID G-SEO is able to effectively translate retrieval intent into quality-oriented optimization guidance at the planning stage. In terms of operational strategies, Content Enrichment and Expansion accounts for

Optimization State	Success	Failure
Query	How do climate changes affect the hosting of the Winter Olympics?	What are the health benefits of using olive oil in salad dressings?
Original Source	Skeleton is one of the 'ice sports' on the Winter Olympic Games program. This extreme sport requires individuals to ride a small sled down a frozen track while lying face down and forward facing. Skeleton was first on the program at the 1928 Winter Olympics in St. Moritz, and again in 1948 Winter Olympics. After a long hiatus, it was reintroduced at the 2002 Winter Olympics with both men's and women's events. Skeleton is a head-first version of the luge, and is run on the same track (as is bobsled). Amazingly, the skeleton sled has no steering or braking mechanism, the participants steer by movements of the body and travel at speeds up to 130 km/hr (80 mph). Trivia (History) American Jennison Heaton won the first Olympic skeleton gold medal at St Moritz in 1928, beating his brother John, who took the silver...	I have been on a mission to help you build a collection of salad dressing recipes for a year now. Today's recipe, Lemon Balsamic Salad Dressing is the newest one that I think you will enjoy. It is the balsamic vinegar added version of my all time favorite, Lemon Salad Dressing. I recently used this lemon balsamic vinaigrette in Cucumber Mozzarella Tomato Salad recipe but you can pretty much use it in any salad. This dressing especially compliments salad recipes with a minimal number of ingredients like my Blueberry Salad and Avocado Quinoa Salad. Why Should You Make This Recipe? You know what goes in there: I get it. It is convenient to grab a bottle of balsamic vinaigrette when you are at the store. However, most of those are packed with preservatives and low quality ingredients. When you make homemade dressings like this one, you know every single ingredient that goes in this one...
Optimized Source	Skeleton is a thrilling 'ice sport' featured in the Winter Olympic Games, where athletes ride a small sled down a frozen track head-first while lying face down. This extreme sport demands precision and speed, as participants navigate the course at amazing speeds of up to 130 km/hr (80 mph) without steering or braking mechanisms, steering is achieved through body movements. First introduced at the 1928 Winter Olympics in St. Moritz, Skeleton made a brief return in 1948 before being reintroduced in 2002 with separate events for men and women. The sport is a head-first version of luge, using the same track as bobsled. Notable Achievements & History: American Jennison Heaton claimed the first Olympic skeleton gold medal, outpacing his brother John, who won silver. 1948: Skeleton returned to St. Moritz, with American John Heaton again securing silver...	I have been on a mission to help you build a collection of salad dressing recipes for a year now. Today's recipe, Lemon Balsamic Salad Dressing, is the newest one that I think you will enjoy. It is the balsamic vinegar added version of my all time favorite, Lemon Salad Dressing. I recently used this lemon balsamic vinaigrette in Cucumber Mozzarella Tomato Salad recipe but you can pretty much use it in any salad. This dressing especially compliments salad recipes with a minimal number of ingredients like my Blueberry Salad and Avocado Quinoa Salad. Why Make This Lemon Balsamic Salad Dressing? Know Your Ingredients: Unlike store-bought versions packed with preservatives and low-quality ingredients, making homemade dressings like this one allows you to know exactly what goes in. You control the quality. Quick & Easy: This recipe comes together in just 5 minutes using only 6 simple ingredients ...
Original Response	Climate change significantly impacts the hosting of the Winter Olympics by making it increasingly difficult to find suitable locations with consistent snow and cold temperatures [4]. The 2022 Winter Olympics in Beijing, China, faced challenges related to climate change, as the region experienced unusually warm temperatures and limited natural snowfall, requiring extensive use of artificial snow [4]. This situation highlights the vulnerability of winter sports events to shifting climate conditions and underscores the challenges future host cities may encounter [4]. The increasing frequency of extreme weather events, such as heatwaves and droughts, further complicates the planning and execution of Winter Olympics, as they can disrupt training and competition schedules [4]. Additionally, the costs associated with artificial snow and climate control measures may rise, adding financial pressure to host countries [4]. These factors collectively demonstrate how climate change is altering the landscape of the Winter Olympics and posing significant challenges for future editions [4]. Moreover, the selection of host cities for the Winter Olympics is becoming increasingly influenced by climate considerations, as potential hosts must demonstrate the ability to maintain suitable conditions for winter sports [4]. The experience of Beijing 2022 underscores the need for host cities to invest in advanced snow-making technology and other climate adaptation strategies to ensure the viability of winter events [4]. As climate change continues to intensify, the Winter Olympics may face more frequent disruptions and the need for innovative solutions to sustain winter sports [4].	Extra-virgin olive oil is a key ingredient in many salad dressings, providing body and richness to recipes like lemon vinaigrette [1]. It is a healthy fat that can contribute to the overall nutritional value of a salad when used in moderation [1][2][3]. Olive oil is known for its high content of monounsaturated fats, which are beneficial for heart health [2][3]. Additionally, it contains antioxidants such as oleocanthal, which may have anti-inflammatory properties [2][3]. The use of extra-virgin olive oil in salad dressings can enhance the flavor of greens and other ingredients while providing these health benefits [1][3]. Extra-virgin olive oil is a key ingredient in many salad dressings, providing body and richness to dressings like lemon vinaigrette [1]. It is recommended to use good quality extra-virgin olive oil for the best flavor in homemade dressings [4]. Olive oil can also be used as a base for a simple salad dressing, combined with vinegar, mustard, salt, and pepper, and optionally enhanced with flavorings like garlic, herbs, or honey [5]. Additionally, olive oil is a healthy fat that contributes to the nutritional profile of salad dressings, making them a more wholesome choice for salads [1][3].
Optimized Response	Climate change poses significant challenges to hosting the Winter Olympics, primarily by threatening the availability of suitable venues for snow and ice sports [1][2][3]. The 2022 Winter Olympics in Beijing, China, exemplified this issue, as some events required the use of existing venues from the 2008 Summer Olympics or the construction of new facilities, while others, like freestyle skiing and snowboarding, were held in locations that were converted from industrial areas [4]. The organizers also had to rely on extensive man-made snow production to ensure adequate conditions for events, highlighting the increasing need for technological interventions to compensate for natural snow scarcity due to warming temperatures [4]. Furthermore, the host city selection process for the Winter Olympics has become increasingly contentious due to climate change concerns. Oslo, Norway, withdrew its bid to host the 2022 Games partly because of public opposition to the high costs and environmental impact of such an event in a region increasingly affected by climate change [4]. This reflects a broader trend where potential host cities are facing growing pressure to consider the sustainability and environmental footprint of large-scale international sporting events in the context of global climate change [1][2][3]. The impact of climate change on the Winter Olympics also extends to the athletes and the integrity of the competition. For instance, at the 2010 Vancouver Olympics, Georgian luger Nodar Kumaritashvili tragically died in a training crash after losing control on the ice track, which raised questions about the safety and suitability of the venue in a warming climate [1]. Such incidents underscore the complex interplay between climate change, venue management, and athlete safety in the context of the Winter Olympics [1][2][3]. In summary, climate change challenges the Winter Olympics by threatening venue suitability, increasing the costs and environmental impact of hosting the Games, and raising concerns about athlete safety and competition integrity [1][2][3][4].	Extra-virgin olive oil is a key ingredient in many salad dressings, providing body and richness to dressings like lemon vinaigrette, which also includes fresh lemon juice, garlic, Dijon mustard, honey or maple syrup, and optional fresh or dried thyme [1]. Another simple oil and vinegar salad dressing with lemon uses equal parts olive oil and cider vinegar, along with fresh lemon juice, salt, black pepper, and garlic [2]. A basic salad dressing recipe consists of one tablespoon of vinegar, three tablespoons of extra virgin olive oil, half a teaspoon of mustard, and half a teaspoon each of salt and pepper, with optional flavorings like garlic, fresh or dried herbs, or a pinch of spice [5]. These dressings are versatile, suitable for leafy greens, chopped salads, and can even be used to add flavor to beans, lentils, and roasted vegetables [5]. The health benefits of using olive oil in salad dressings include its contribution to heart health due to its high content of monounsaturated fats and potential antioxidant properties [1][3][4]. Additionally, olive oil can enhance the absorption of fat-soluble vitamins found in salads [3].

Fig. 6. Success and failure case analysis of RAID G-SEO. Based on the improvement scores in subjective and objective impression, we select one representative successful case and one failed case, and compare the sources and their presentations in the generated responses before and after optimization, providing a qualitative analysis of RAID G-SEO's optimization behavior under different scenarios. The bold citation in the responses represent the source to optimization.

more than 30% of all actions, suggesting that the model, under intent-aware constraints, preferentially adopts information augmentation and semantic expansion to better support the requirements of generative retrieval. Overall, RAID G-SEO exhibits a hybrid planning paradigm in the Step Planning stage that jointly covers multiple optimization objectives, rather than relying on a single optimization strategy. These findings further validate the effectiveness of intent-driven planning mechanisms for G-SEO tasks.

5) *Case Analysis*: To further elucidate the practical working mechanisms of RAID G-SEO, we conduct a case study by

selecting one successful and one failed optimization example based on improvements in both subjective and objective impression, as illustrated in Fig. 6. In the successful case, RAID G-SEO performs refinements on the original content source across linguistic expression, information structure, and content organization. Specifically, RAID G-SEO transforms the original encyclopedic and linearly descriptive writing style into a more engaging and descriptive narrative form. Key information, such as head-first cues, is deliberately front-loaded into the introductory paragraph, which substantially enhances semantic salience and readability without introducing any new

factual content. This form of semantics-driven word order adjustment is rarely observed in traditional SEO practices and highlights the fundamental distinction between G-SEO and keyword-centric optimization paradigms. Moreover, RAID G-SEO reorganizes the content through hierarchical and modular restructuring, explicitly separating information units such as introductions and historical background, and emphasizing critical facts via headings and bold formatting. This allows the source to be presented to the generative model of GSE in a more structured manner. Meanwhile, the method applies moderate content compression by de-emphasizing peripheral information and highlighting representative facts, thereby increasing information density while preserving completeness and improving generative retrieval friendliness. These refinements lead to a clear improvement in the expressive quality of the source, which is further reflected in its visibility within generated responses. Specifically, the optimized source was not cited in the original response, whereas after optimization it not only appears in the generated output but also occupies a more prominent visual position with increased share. Correspondingly, the Objective Impression Overall score increases from 0.09 to 0.32, and the Subjective Impression Average rises from 27.9 to 28.2. In contrast, in the failed optimization case, RAID G-SEO similarly restructures the content and explicitly summarizes previously implicit and dispersed information. By introducing an overview-style lead paragraph, the optimized source is designed to better cover diverse retrieval intents. This process reflects the role of the 4W multi-role deep reflection module in guiding intent generalization and enhancing the adaptability of sources across heterogeneous user scenarios. However, despite improvements in overall readability and guidance, the visibility of the optimized source in the generated response does not improve. The source’s position remains largely unchanged, and it co-supports the same generated segment alongside other sources, thereby weakening its perceived individual contribution. As a result, the Objective Impression Overall score decreases from 0.75 to 0.52, and the Subjective Impression Average drops from 26.70 to 23.90. Further analysis suggests that the primary cause of this failure lies in the high semantic overlap between the optimized source and other competing sources. Consequently, the generative model tends to integrate information from multiple sources during citation, which reduces the distinctiveness of any single source. Through a comparative examination of both successful and failed cases, we identify inherent limitations of RAID G-SEO. While RAID G-SEO demonstrates strong potential in improving content expressiveness and visibility in generative retrieval scenarios, its effectiveness critically depends on maintaining a balance between semantic uniqueness of the source and the degree of intent generalization. Although the multi-role reflection mechanism enables RAID G-SEO to infer relatively generalized intents, such intents remain constrained by the domain boundaries of the source and may fail to fully align with the query, thereby affecting downstream optimization outcomes. For instance, in the failed case, the generalized intent inferred from the source, “Find and share a recipe for a quick, easy-to-make, customizable, and versatile Lemon Balsamic Salad Dressing”—remains strongly tied to the recipe context

and exhibits low relevance to nutrition- or medicine-oriented query, limiting the effectiveness of optimization under the given query setting. Notably, across both successful and failed cases, RAID G-SEO consistently applies a combination of multiple optimization operations rather than a single strategy. Compared with existing GEO [1] approaches, which typically rely on a single type of optimization, RAID G-SEO offers a broader operational space for optimization and provides a foundation for future exploration of more fine-grained intent alignment mechanisms.

V. CONCLUSION

This work addresses the task of generative search engine optimization under diverse retrieval scenarios, and introduces an intent-generalization-enhanced optimization approach. We design an intent-driven four-phase optimization framework, and incorporate a 4W multi-role deep reflection mechanism to improve the generalizability of the inferred intent, enabling creators to perform adaptive and targeted content optimization even in the absence of explicit user queries. To support more comprehensive and objective evaluation, we extend GEO-BENCH to cover a broader range of retrieval scenarios, and develop G-EVAL 2.0, a fine-grained subjective evaluation framework for assessing content visibility. Experimental results demonstrate that intent modeling plays a pivotal role in G-SEO, with generalized intent representations contributing significantly to optimization effectiveness. Nevertheless, balancing the precision and generalizability of intent representations remains a key open challenge.

VI. LIMITATIONS

While our results demonstrate that incorporating search intent provides more focused guidance for optimization in G-SEO tasks, we observe that overly specific intents may lead to content rewrites that overfit to particular queries, thereby reducing generalizability. This observation underscores a central challenge in targeted G-SEO: balancing the precision of intent modeling with the adaptability of optimization strategies. Our approach relies primarily on prompt engineering for preliminary intent extraction, which remains limited in granularity control and contextual consistency. Future work may explore integrating domain-specific knowledge or developing dedicated intent modeling modules to improve the effectiveness of G-SEO methods. Furthermore, our proposed method, like most existing G-SEO approaches, focuses solely on plain text without accounting for visual or multimodal elements such as images, diagrams, or tables that may influence content visibility in real-world GSE. Extending G-SEO to Visual-Language Models (VLMs) for unified multimodal optimization poses an important avenue for future research.

VII. ETHICAL STATEMENT

We strictly adhere to the ethical standards and best practices of the AI research community. We employ LLMs in full compliance with their licensing terms. Our study focuses on improving content visibility within GSE. All experiments are conducted in simulated environments without interfering with,

misleading, or manipulating the behavior of any real-world systems. To support evaluation, we extend the GEO benchmark by generating synthetic text data using the LLM and applying targeted optimizations. All generated or modified content is intended solely for research and experimentation. The resulting data does not reflect factual information and should not be interpreted as expressing any subjective opinions or value judgments of the authors or the underlying models.

REFERENCES

- [1] Aggarwal, P.; Murahari, V.; Rajpurohit, T.; Kalyan, A.; Narasimhan, K.; and Deshpande, A. 2024. Geo: Generative engine optimization. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD '24*, 5–16. New York, NY, USA: Association for Computing Machinery. ISBN 9798400704901.
- [2] Almkhtar, F.; Mahmood, N.; and Kareem, S. 2021. Search engine optimization: a review. *Applied computer science*, 17(1): 70–80.
- [3] Bardas, N.; Mordo, T.; Kurland, O.; Tennenholtz, M.; and Zur, G. 2025. White Hat Search Engine Optimization using Large Language Models. arXiv:2502.07315.
- [4] Chodak, G.; and Błażyczek, K. 2024. Large language models for search engine optimization in e-commerce. In Garg, D.; Rodrigues, J. J. P. C.; Gupta, S. K.; Cheng, X.; Sarao, P.; and Patel, G. S., eds., *International Advanced Computing Conference, 2023*, 333–344. Springer, Cham: Springer Nature Switzerland. ISBN 978-3-031-56700-1.
- [5] Chong, R.; Kong, C.; Wu, L.; Liu, Z.; Jin, Z.; Yang, L.; Fan, Y.; Fan, H.; and Yang, E. 2023. Leveraging prefix transfer for multi-intent text revision. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 1219–1228. Toronto, Canada: Association for Computational Linguistics.
- [6] Greshake, K.; Abdelnabi, S.; Mishra, S.; Endres, C.; Holz, T.; and Fritz, M. 2023. Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection. In *Proceedings of the 16th ACM workshop on artificial intelligence and security, AISec '23*, 79–90. New York, NY, USA: Association for Computing Machinery. ISBN 9798400702600.
- [7] Hughes, S.; Bae, M.; and Li, M. 2023. Vectara Hallucination Leaderboard. <https://github.com/vectara/hallucination-leaderboard>. Accessed: 2025-07-16.
- [8] Ji, Z.; Yu, T.; Xu, Y.; Lee, N.; Ishii, E.; and Fung, P. 2023. Towards Mitigating LLM Hallucination via Self Reflection. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, 1827–1843. Singapore: Association for Computational Linguistics.
- [9] Kanara, A. P.; Kumari, P.; and Prathap, B. R. 2024. Python Driven Keyword Analysis for SEO Optimization. In *2024 10th International Conference on Advanced Computing and Communication Systems (ICACCS)*, volume 1, 1170–1176. IEEE.
- [10] Kowalczyk, K.; and Szandala, T. 2024. Enhancing SEO in single-page web applications in contrast with multi-page applications. *IEEE Access*, 12: 11597–11614.
- [11] Kumar, A.; and Lakkaraju, H. 2024. Manipulating Large Language Models to Increase Product Visibility. arXiv:2404.07981.
- [12] Lewandowski, D. 2023. *Understanding search engines*. Springer.
- [13] Li, C.; Zhang, M.; Mei, Q.; Kong, W.; and Bendersky, M. 2024. Learning to rewrite prompts for personalized text generation. In *Proceedings of the ACM Web Conference 2024, WWW '24*, 3367–3378. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701719.
- [14] Li, Y.; Luo, M.; Gong, Y.; Lin, C.; Jiao, J.; Liu, Y.; and Huang, K. 2025. DeepThink: Aligning Language Models with Domain-Specific User Intents. arXiv:2502.05497.
- [15] Liu, Y.; Iter, D.; Xu, Y.; Wang, S.; Xu, R.; and Zhu, C. 2023. G-Eval: NLG Evaluation using Gpt-4 with Better Human Alignment. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2511–2522. Singapore: Association for Computational Linguistics.
- [16] Mao, K.; Dou, Z.; Mo, F.; Hou, J.; Chen, H.; and Qian, H. 2023. Large Language Models Know Your Contextual Search Intent: A Prompting Framework for Conversational Search. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, 1211–1225. Singapore: Association for Computational Linguistics.
- [17] Nagpal, M.; and Petersen, J. A. 2021. Keyword Selection Strategies in Search Engine Optimization: How Relevant is Relevance? *Journal of Retailing*, 97(4): 746–763. SI: Metrics and Analytics.
- [18] Pfrommer, S.; Bai, Y.; Gautam, T.; and Sojoudi, S. 2024. Ranking Manipulation for Conversational Search Engines. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 9523–9552. Miami, Florida, USA: Association for Computational Linguistics.
- [19] Samarah, T.; Alrawashdeh, T.; Mughaid, A.; and AlZu'bi, S. 2024. Utilizing LLMs for Enhancing Search Engine Optimization Strategies in Digital Marketing. In *2024 2nd International Conference on Foundation and Large Language Models (FLLM)*, 284–288. IEEE.
- [20] Sarkar, R.; Sarrafzadeh, B.; Chandrasekaran, N.; Rangan, N.; Resnik, P.; Yang, L.; and Jauhar, S. K. 2025. Conversational User-AI Intervention: A Study on Prompt Rewriting for Improved LLM Response Generation. arXiv:2503.16789.
- [21] Shahzad, A.; Jacob, D. W.; Nawi, N. M.; Mahdin, H.; and Saputri, M. E. 2020. The new trend for search engine optimization, tools and techniques. *Indonesian Journal of Electrical Engineering and Computer Science*, 18(3): 1568–1583.
- [22] Shi, J.; Yuan, Z.; Liu, Y.; Huang, Y.; Zhou, P.; Sun, L.; and Gong, N. Z. 2024. Optimization-based prompt injection attack to Llm-as-a-judge. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security, CCS '24*, 660–674. New York, NY, USA: Association for Computing Machinery. ISBN 9798400706363.
- [23] Shinn, N.; Cassano, F.; Gopinath, A.; Narasimhan, K.; and Yao, S. 2023. Reflexion: language agents with verbal reinforcement learning. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems 36, NIPS 2023*, volume 36, 8634–8652. Curran Associates, Inc.
- [24] Shu, L.; Luo, L.; Hoskore, J.; Zhu, Y.; Liu, Y.; Tong, S.; Chen, J.; and Meng, L. 2024. RewriteLM: An Instruction-Tuned Large Language Model for Text Rewriting. In *Proceedings of the AAAI Conference on Artificial Intelligence, 2024*, volume 38, 18970–18980. AAAI Press.
- [25] Sun, Z.; Liu, H.; Qu, X.; Feng, K.; Wang, Y.; and Ong, Y. S. 2024. Large Language Models for Intent-Driven Session Recommendations. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '24*, 324–334. New York, NY, USA: Association for Computing Machinery. ISBN 9798400704314.
- [26] Wang, X.; Huang, T.; Wang, D.; Yuan, Y.; Liu, Z.; He, X.; and Chua, T.-S. 2021. Learning Intents behind Interactions with Knowledge Graph for Recommendation. In *Proceedings of the Web Conference 2021, WWW '21*, 878–887. New York, NY, USA: Association for Computing Machinery. ISBN 9781450383127.
- [27] Ward, V. 2017. Why, whose, what and how? A framework for knowledge mobilisers. *Evidence and Policy*, 13(3): 477 – 497.
- [28] Yan, H.; Zhu, Q.; Wang, X.; Gui, L.; and He, Y. 2024. Mirror: Multiple-perspective Self-Reflection Method for Knowledge-rich Reasoning. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 7086–7103. Bangkok, Thailand: Association for Computational Linguistics.
- [29] Zhang, W.; Shen, Y.; Wu, L.; Peng, Q.; Wang, J.; Zhuang, Y.; and Lu, W. 2024. Self-Contrast: Better Reflection Through Inconsistent Solving Perspectives. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3602–3622. Bangkok, Thailand: Association for Computational Linguistics.
- [30] Škėrienė, S.; and Jucevičienė, P. 2020. Problem solving through values: A challenge for thinking and capability development. *Thinking Skills and Creativity*, 37: 100694.
- [31] Kim, J.; Kim, D.-K.; Logeswaran, L.; Sohn, S.; and Lee, H. 2024. Auto-Intent: Automated Intent Discovery and Self-Exploration for Large Language Model Web Agents. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Findings of the Association for Computational Linguistics: EMNLP 2024*, 16531–16541. Miami, Florida, USA: Association for Computational Linguistics.

TABLE V
PROMPT USED FOR QUERY VARIANT GENERATION

Query variants Generation Prompt
You are a user of the Generative Search Engine, seeking to explore a specific topic. You will be given a query representing what you want to learn about. Your task is to generate 4 new queries that approach the topic from different perspectives. Ensure that: 1. Each query is relevant to the original but not a direct repetition. 2. The queries cover a range of perspectives, such as related concepts, practical applications, contrasting viewpoints, emerging trends, or any other meaningful dimension. But do not limit to these directions. 3. Output ONLY 1 query per line. ### Input Format {query_original} ### Output Format query_1 query_2 query_3 query_4

VIII. QUERY VARIANTS GENERATION PROMPT

To simulate diverse retrieval scenarios in real-world GSE applications, we extended the GEO-BENCH dataset. Specifically, for each original query in GEO-BENCH, we constructed four semantically related but lexically and stylistically distinct variants, each reflecting a different retrieval scenario. Each variant preserves the core meaning while introducing variation in wording, tone, or semantic framing. The prompt used to generate these variants is presented in Table V.

IX. G-EVAL PROMPT GENERATION PROMPT AND EVALUATION PROMPT TEMPLATE

G-eval is a natural language generation (NLG) evaluation framework that integrates Chain-of-Thought (CoT) reasoning with a form-filling paradigm. GEO extended this framework to the G-SEO context to enable subjective evaluation of content source attribution in GSE outputs. In this work, we conduct an in-depth analysis of the G-eval mechanism as adopted by GEO. While the framework provides reasonably comprehensive coverage across seven subjective impression dimensions, we identify three key limitations when applied to the G-SEO setting: (1) several evaluation criteria are under-specified, leading to potential inconsistencies in annotator interpretation; (2) the mapping between scoring levels and their semantic interpretations is vague, which undermines the clarity and interpretability of the results; and (3) the prevalent issue of “unreferenced original content” in GSE responses is not explicitly addressed, increasing the risk of misjudging inexistent content as grounded. To address these limitations, we propose an enhanced evaluation framework, G-eval 2.0, designed to support more fine-grained, fair, and reproducible subjective assessment. We standardize structured rubrics across all seven dimensions and design corresponding prompts for automatic evaluation logic generation, covering key aspects such as dimension definition, core scoring elements, and the semantic mapping of score levels, thereby

TABLE VI
STRUCTURED PROMPT FOR GENERATING SCORING RUBRICS ACROSS SEVEN SUBJECTIVE DIMENSIONS

G-EVAL Prompt Generation Prompt
You are an evaluator for answers generated by the Generative Search Engine, providing subjective evaluations... Your tasks: 1. Refine the criteria: Review the given description and improve it to be concise, consistency, and completeness. 2. Propose an Evaluation Procedure: Based on the refined criteria, outline a clear, concise, step-by-step process specifically for scoring Source [1] only. If the source not exist, score should be 0. Please ensure your generation is easy to understand. ### Input Format Example: Relevance of Citation to Query (0-5): The dimension evaluates the degree to which the citation from Source [1] is related to the query, indicating how useful and pertinent the information provided to address the query.... ### Output Format Evaluation Criteria: [...] Evaluation Steps: 1. Step 1 description... 2. Step 2 description... ...

TABLE VII
UNIFIED EVALUATION PROCEDURE TEMPLATE IN G-EVAL 2.0

Query variants Generation Prompt
You are a user of Generative Search Engine, capable of providing subjective evaluations of retrieved answers... Please read and understand these instructions carefully, and then proceed with your scoring (score ONLY). ### Evaluation Criteria [Evaluation Criteria]: [...] ### Evaluation Steps [...] ### Input Format Query: {query} Answer: {answer} ### Output Format (score ONLY) [score]

ensuring consistency, executability, and interpretability across all dimensions, as is shown in Table VI. Additionally, we develop a unified evaluation procedure template applicable across dimensions as shown in Table VII, integrating the CoT reasoning paradigm. The template standardizes the full evaluation process, from content interpretation and key scoring element extraction to structured scoring. To maintain focus on the core evaluation methodology, task-specific background information is omitted in this appendix section.

TABLE VIII
CORE CONFIGURATIONS OF LLMs USED IN OUR WORK.

Model	Temperature	Top_p	Max_token	Application
GPT-4	0.5	1	1024	Query Variants Generation
GPT-4o	0.5	1	3192	G-eval 2.0 Prompt Generation
GLM-4-9B-0414	0.5	1	1024	GSE Response Generation Simulation
	do_sample=False	1	3192	G-SEO Methods Implementation
	do_sample=False	1	2	G-eval 2.0 Subjective Impression Evaluation

X. EXPERIMENT DETAILS

A. Software and Hardware Environment

All experiments in this work were conducted on a server equipped with two NVIDIA GeForce RTX 3090 GPUs (24GB VRAM each), running Ubuntu 20.04. The core software stack includes Python 3.9.19, PyTorch 2.1.0 with CUDA 11.8 support, and Transformers 4.35.3. All model loading and inference were carried out using the Hugging Face Transformers interface, and environment dependencies were managed via Conda. Unless otherwise specified, all experiments were run with a fixed random seed (seed = 42) to ensure reproducibility.

B. Configuration Parameters of LLMs

To support reproducibility and transparency, Table VIII summarizes the key configuration parameters of LLMs used at different stages of our work, including model name, temperature, top-p, maximum tokens, and the specific usage scenario for each configuration.

C. Subjective Impression Evaluation Procedure for G-SEO Methods

To more accurately assess the effectiveness of the G-SEO methods in enhancing the subjective impression of target content under diverse retrieval scenarios in GSE, we designed the evaluation procedure illustrated in 7. We randomly sampled 100 diverse-retrieval examples from the GEO-BENCH dataset (seed = 42). Each sample consists of one original query and four paraphrased variants that preserve the core semantic focus, resulting in five queries per sample. For each query, we generated responses from five content sources both before and after optimization. The subjective impression scoring mechanism was then applied to assess the visibility of each source within each response. We first averaged the scores of the five responses under each query to obtain a query-level score. Next, we averaged the five query-level scores within each sample to obtain a sample-level evaluation. Finally, we averaged across all 100 samples to derive the overall effectiveness score for the optimization method.

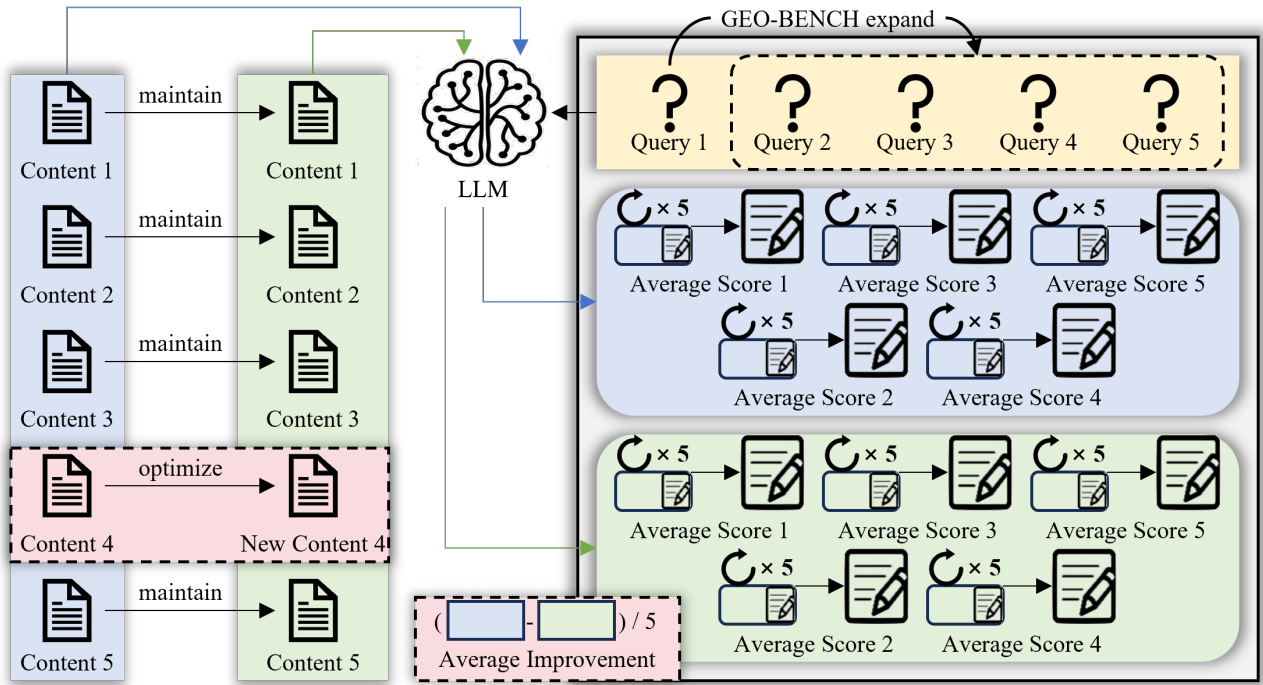


Fig. 7. Evaluation process for optimization methods. For each query, responses are generated from five content sources before and after optimization. By comparing the relative impression scores of each source in the responses, we assess the effectiveness of G-SEO methods in enhancing content visibility. Each diverse-retrieval sample includes one original query and four paraphrased variants. For each query, five responses are generated, evaluated, and averaged to obtain the query-level result. Overall performance is obtained by averaging the query-level results across all five queries.