

Beyond Final Visibility: A Three-Stage Diagnostic Framework for Enterprise Generative Search Optimization

Charming

New York University Shanghai

BASSLINE

mc10010@nyu.edu

Abstract

Generative search introduces failures in query matching, competitive ranking, and grounded generation, while conventional SEO and final mention metrics often cannot localize these failures. We present a controlled Stage-Isolated Evaluation framework that converts stage-specific errors into targeted revisions of structured website content. In an industrial study with an anonymized consumer-health company, each version is evaluated against nine fixed competitors using disjoint training, validation, and held-out test queries. On a 0–100 point scale, the selected checkpoint improves the test score from 58.78 to 69.98, an 11.20-point total gain. A budget-matched recursive baseline without stage-level diagnostics captures 8.13 points of this gain, leaving a descriptive 3.07-point diagnostic margin. Different stages favor different checkpoints, revealing non-monotonic trade-offs and suggesting that enterprise generative engine optimization (GEO) is a multi-objective problem. A small-scale online pilot provides exploratory evidence of consistency with commercial engines, while human annotation supports the reliability of two generator-side diagnostic labels.

1 Introduction

Current generative engine optimization (GEO) lacks a unified and reproducible evaluation standard. In practice, enterprises often rely on repeated tests on commercial large language model platforms, using isolated indicators such as retrieval rate, brand mentions, or citation frequency as feedback. These platform-dependent measurements are costly, unstable, and difficult to reproduce, while providing limited evidence about overall content performance. More importantly, they cannot determine whether a failure arises from content retrieval, competitive ranking, or grounded generation, leaving enterprises without a systematic basis for diagnosing and improving their generative-search

visibility. Commercial engines also provide limited stage-level attribution, and unpublished candidate revisions cannot be directly evaluated before deployment, motivating the use of a controlled offline evaluation stack for diagnostic analysis.

To address this limitation, we propose a *Stage-Isolated Evaluation* framework that separately assesses enterprise content at the retrieval, reranking, and generation stages. The framework assigns stage-specific scores and produces diagnostic feedback on missing query coverage, weak competitive evidence, and unsupported or underused content. Based on these signals, we further introduce a recursive optimization procedure that revises a structured site-level representation (hereafter, *sitemap*, distinct from the conventional XML sitemap used by web crawlers), selects checkpoints on a held-out validation set, and evaluates the selected version on unseen test queries. The proposed approach therefore provides a reproducible offline evaluation protocol and an actionable diagnostic method for studying and improving enterprise visibility in generative search, without assuming that offline gains necessarily transfer to commercial engines.

Our contributions are threefold:

1. We introduce a Stage-Isolated Evaluation framework that separately measures retrieval, competitive ranking, and grounded generation under a controlled and reproducible offline protocol.
2. We develop a recursive optimization procedure that converts stage-level diagnostics into targeted site revisions and selects the best checkpoint using a held-out validation set.
3. We evaluate the framework in an industrial competitive setting with fixed competitors and disjoint query splits. Different stages favor different checkpoints, revealing stage-level trade-offs, while a small-scale online pilot provides only exploratory evidence of offline-to-online consistency.

2 Related Work

Traditional SEO evaluates webpage visibility through rankings, keyword matching, impressions, and clicks. Generative search introduces a different pipeline in which content must be retrieved, ranked as useful evidence, and incorporated into a synthesized answer. Consequently, final mentions or citations alone cannot determine whether visibility was lost during retrieval, reranking, or generation (Aggarwal et al., 2023; Chen et al., 2025a). Commercial systems further rely on proprietary RAG pipelines and produce results that vary across repeated runs, prompts, and time (Chen et al., 2025b; Grossman et al., 2026; Schulte et al., 2026), motivating controlled and repeatable evaluation.

Recent work has extended GEO evaluation beyond final visibility. Kim et al. (2026) separately evaluate retrieval, reranking, and generation, while Zhang et al. (2026) distinguish citation selection from citation absorption, showing that source inclusion does not guarantee substantive use in the generated answer. In competitive settings, Bagga et al. (2025) examine whether rewritten content improves ranking under fixed candidate sets.

Content optimization has also progressed from predefined rewriting strategies toward adaptive optimization. AutoGEO learns generative-engine preference rules and applies them to web-content rewriting (Wu et al., 2025), whereas AgenticGEO evolves content-conditioned strategies using a surrogate critic (Yuan et al., 2026). Our work complements these approaches with an industrial, stage-aware workflow: multiple sitemap versions are evaluated against the same competitors, and retrieval, competitive ranking, claim adoption, and attribution errors are converted into fact-grounded revision feedback. This design connects recursive content optimization with actionable diagnosis of where enterprise visibility is gained or lost.

3 Method

3.1 Problem Formulation

Let $\mathcal{Q} = \mathcal{Q}_{\text{train}} \cup \mathcal{Q}_{\text{val}} \cup \mathcal{Q}_{\text{test}}$ denote three disjoint query sets used for recursive optimization, checkpoint selection, and final evaluation, respectively. Let $\mathcal{V} = \{v_0, \dots, v_n\}$ denote the sequence of target sitemap versions, where v_0 is the original sitemap, and let $\mathcal{C} = \{c_1, \dots, c_m\}$ denote a fixed set of competitor sitemaps. All websites are converted into the same structured representation

containing normalized metadata and textual content.

A verified enterprise fact document $\mathcal{F}$ is used to constrain sitemap revision and evaluate factual consistency. It is not included in retrieval or reranking and therefore does not provide the target website with additional ranking evidence.

3.2 Controlled Stage-Isolated Evaluation

For each query q and target version v_i , we construct a controlled candidate set containing the target sitemap and nine fixed competitors:

$$\mathcal{D}_{q,i} = \{v_i\} \cup \mathcal{C}, \quad |\mathcal{C}| = 9. \quad (1)$$

The model-selection and fallback policy remains fixed across checkpoints, although the actual generator endpoint may differ when a fallback is triggered. The queries, competitors, model configurations, prompts, decoding parameters, and scoring rules remain fixed. For each query, candidate order is randomized using a fixed local seed and reused across all target versions and baselines. No generation-level random seed is explicitly configured for the API models.

Retrieval, reranking, and generation are evaluated independently under these shared controls. They measure query–content matching, competitive relevance, and grounded content use, respectively; outputs from one stage are not passed to another. The normalized stage scores are aggregated for validation-based checkpoint selection and for summary reporting on the held-out test set. Test-set aggregate scores are never used for revision or checkpoint selection:

$$S(v_i) = 0.3S_R(v_i) + 0.3S_K(v_i) + 0.4S_G(v_i). \quad (2)$$

Each stage score is mapped to a fixed 0–100 range using predefined, version-independent bounds. No normalization statistics are estimated from validation or test checkpoints.

3.3 Stage Metrics

Detailed model configurations, inference parameters, and execution settings for all three stages are provided in Appendix A.

Retriever. The retriever measures query–content matching independently of the competitor set. For query q , sitemap version v_i is divided into chunks $\{x_{i1}, \dots, x_{in}\}$, and each chunk receives a fixed hybrid lexical–semantic relevance score $r(q, x_{ij})$.

Three-stage Offline Evaluation Framework for Enterprise GEO

A controlled, stage-isolated evaluation of target matching, competitive relevance, and grounded use

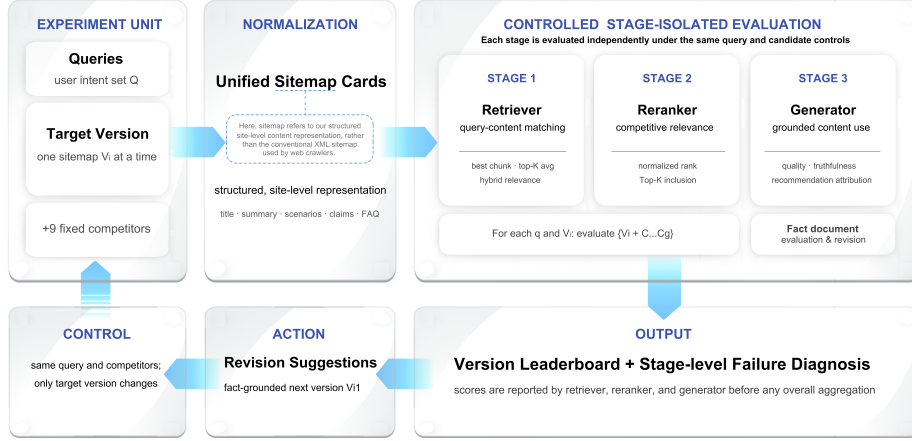


Figure 1: Stage-isolated diagnostic framework for evaluating target-content matching, competitive ranking, and grounded generation in enterprise GEO.

Let

$$r_{\max}(q, v_i) = \max_{1 \leq j \leq n} r(q, x_{ij}) \quad (3)$$

denote the highest chunk relevance score, and let

$$\bar{r}_K(q, v_i) = \frac{1}{K} \sum_{x \in \text{TopK}(q, v_i)} r(q, x) \quad (4)$$

denote the mean relevance of the top- K chunks. The query-level retrieval score is

$$R(q, v_i) = \alpha r_{\max}(q, v_i) + (1 - \alpha) \bar{r}_K(q, v_i), \quad (5)$$

where $0 \leq \alpha \leq 1$.

The stage-level retrieval score is

$$S_R(v_i) = \frac{1}{|Q|} \sum_{q \in Q} R(q, v_i). \quad (6)$$

The chunking configuration, retrieval models, hybrid weights, α , and K remain fixed across all sitemap versions.

Reranker. For each query, the reranker orders the target sitemap and nine fixed competitors. Let $\rho(q, v_i) \in [0, 1]$ denote the normalized inverse rank of the target, and let $h_K(q, v_i)$ indicate whether it appears in the top K . The query-level score is

$$K(q, v_i) = \beta \rho(q, v_i) + (1 - \beta) h_K(q, v_i). \quad (7)$$

The stage-level score is averaged across queries:

$$S_K(v_i) = \frac{1}{|Q|} \sum_{q \in Q} K(q, v_i). \quad (8)$$

Candidate representations, the reranking model, β , and K remain fixed across all sitemap versions.

Generator. The generator receives the fixed candidate set and produces an answer $a_{q,i}$ for each query. A separate evaluator assigns a primary answer-quality score based on relevance, factual grounding, and appropriate use of the target content:

$$G(q, v_i) = g_{\text{quality}}(q, v_i, a_{q,i}, F), \quad (9)$$

where F is the verified fact document provided to the evaluator for factual assessment but is not exposed to the generator. The same document is used separately as a factual constraint during sitemap revision, as described in Section 3.4. The stage-level generator score is

$$S_G(v_i) = \frac{1}{|Q'|} \sum_{q \in Q'} G(q, v_i), \quad (10)$$

where Q' contains queries with a successful generator response and a valid, parseable evaluator output. Truthfulness, brand recommendation, claim adoption, and attribution errors are retained as diagnostic measures rather than independently optimized stage scores.

3.4 Recursive Revision Protocol

Starting from the original sitemap v_0 , we generate one revised sitemap per optimization epoch.

At epoch i , the current version v_i is evaluated on Q_{train} through all three stages. Query-level retrieval failures, competitive ranking weaknesses, generation errors, and omitted claims are then aggregated into a structured feedback object:

$$F_i = \text{Aggregate}(R_i, K_i, G_i). \quad (11)$$

A revision model receives the current sitemap, the stage-level feedback, and the verified fact document:

$$v_{i+1} = \mathcal{M}_{\text{revise}}(v_i, F_i, \mathcal{F}). \quad (12)$$

The model is instructed to improve weak content while preserving factual consistency and avoiding unsupported claims. Only training-query results are included in F_i ; validation and test queries are never used to generate revision feedback.

After each epoch, the resulting checkpoint is evaluated on Q_{val} . The selected checkpoint is

$$i^* = \arg \max_i S(v_i; Q_{\text{val}}). \quad (13)$$

Finally, only the selected sitemap v_{i^*} and the predefined baselines are evaluated on Q_{test} . Test results are not used for further revision or checkpoint selection. After initialization, all 100 revision epochs are executed automatically without human intervention.

3.5 Naturalness Constraint

Before checkpoint selection, we exclude revisions that violate predefined hard constraints on length, page identity, factual support, or forbidden claims. Additional diagnostics flag excessive repetition, keyword-oriented wording, and abnormal brand-term density and feed these warnings into subsequent revisions. Detailed rules and implementation settings are provided in Appendix B.

4 Experimental Setup

4.1 Structured Site Representation and Competitor Websites

All target and competitor websites are converted from publicly available web content into the same structured site-level representation. The target representation is derived from the company’s official website, while optimized versions revise its structure and wording under factual constraints. Nine competitors from the same product category are processed using the same procedure. Table 1 summarizes the scale of the target and competitor websites.

	Target	Comp. Range	Comp. Med.
Pages	21	15–220	48
Tokens	94K	15K–350K	60K
Chunks	210	–	–
Facts	190	–	–

Table 1: Site-scale statistics.

The competitor set is fixed across all experiments. For each query and target version, evaluation is conducted against the same nine competitor site representations, so that observed differences mainly reflect changes to the target rather than the competitive environment.

4.2 Query Set

The query set is constructed from an initial pool of 200 real user questions provided by enterprise personnel in the consumer-health product domain. We first normalize the questions, remove duplicate and low-information variants, and categorize the remaining candidates into five intent types: category need, scenario need, problem solving, comparison, and purchase decision.

We then construct three disjoint query sets for recursive revision, checkpoint selection, and held-out testing. To reduce semantic leakage during split construction, all candidate queries are encoded using a fixed sentence-embedding model, and cross-split cosine similarities are computed. Query pairs with cosine similarity above 0.85 are manually reviewed, and near-duplicate questions or minor paraphrases are reassigned or excluded before the final splits are fixed.

The resulting dataset contains 150 queries, divided into three disjoint sets of 50. Each split follows the same intent distribution, with 6 category-need, 8 scenario-need, 11 problem-solving, 15 comparison, and 10 purchase-decision queries. Queries are additionally annotated with usage scenarios and enterprise-assessed frequency levels.

4.3 Enterprise Fact Document

The enterprise fact document is derived from company-provided textual materials and converted into a structured JSON file. It is not used to give the target website extra ranking advantages during retrieval or reranking. Instead, it is used in the generation stage to evaluate factual consistency, claim adoption, and attribution accuracy, and to provide factual boundaries during sitemap optimization. In particular, we use it to check whether the generator correctly describes the target company, adopts

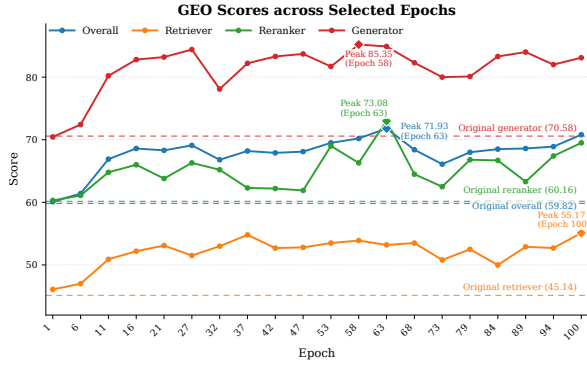


Figure 2: Validation scores of 20 sampled valid checkpoints. Dashed lines denote the original-sitemap baselines.

predefined core claims, and avoids attributing competitor information to the target company.

4.4 Reusable Enterprise Configuration

The framework requires enterprise-provided documents and user queries, together with publicly available competitor websites. It is designed to be reconfigured for another enterprise by replacing the target materials, query set, and competitor collection while retaining the evaluation models, scoring procedure, and revision workflow. Cross-enterprise transfer is not evaluated in this study.

The complete 100-epoch optimization requires approximately 20 hours. The three evaluation stages take about 4.7 minutes per epoch, or 7.9 hours in total. Retrieval and reranking are executed locally, while generation and answer evaluation are processed concurrently through SiliconFlow. The remaining runtime is mainly spent aggregating diagnostic feedback, automatically refining the revision prompt, and generating the next structured site representation. This modular design separates reusable evaluation components from enterprise-specific content and factual constraints.

5 Results

5.1 Validation Trajectory and Checkpoint Selection

All checkpoints are evaluated on the validation set for analysis and selection. Versions that fail the predefined naturalness constraint are excluded. For visualization, Figure 2 retains the first valid checkpoint and every fifth valid record thereafter, while preserving the original epoch indices.

Most eligible checkpoints outperform the original sitemap, although the trajectory remains non-monotonic. The highest aggregate validation score

is obtained at epoch 63, increasing from 59.82 to 71.93. Relative to the original sitemap, retrieval, reranking, and generation improve by approximately 8.1, 12.9, and 14.5 points, respectively.

The individual stages favor different checkpoints. Generation peaks at epoch 58 with a score of 85.35, while the aggregate and reranking scores peak at epoch 63, reaching 71.93 and 73.08. Retrieval reaches its highest score of 55.17 at epoch 100, but this checkpoint performs worse than epoch 63 in reranking and generation. These differences indicate that query coverage, competitive relevance, and grounded content use favor different sitemap configurations, and that the final optimization epoch is not necessarily optimal.

We therefore select epoch 63 according to the aggregate validation score after naturalness filtering. The selected sitemap is frozen and evaluated on the held-out test set, which is not used for revision or checkpoint selection.

5.2 Held-Out Test Comparison with Rewrite Baselines

We compare the full method with two rewrite baselines on the held-out test set. *One-shot rewrite* revises the original structured site representation once using the same revision model and fact document but receives no stage-specific feedback. *Recursive rewrite without diagnostics* uses the same optimization budget as the full method but receives only the previous site representation and a general GEO instruction. All methods use the same site representation schema, fact document, competitor set, query splits, evaluation configuration, and validation-based checkpoint selection procedure.

As shown in Table 2, one-shot and non-diagnostic recursive rewriting increase the original overall score by 5.01 and 8.13 points, respectively. The full stage-isolated method achieves an overall score of 69.98, corresponding to an 11.20-point gain over the original and a descriptive 3.07-point margin over the budget-matched non-diagnostic recursion baseline. It also shows higher retrieval, reranking, and generation scores by 1.83, 2.98, and 4.07 points, respectively, relative to non-diagnostic recursion. These stage-level differences are consistent with broader query coverage, stronger competitive relevance, and more effective grounded content use.

Method	Ret.	Rerank.	Gen.	Overall
Original sitemap	45.44	60.00	67.88	58.78
One-shot rewrite	49.12	64.35	74.37	63.79
Recursive rewrite w/o diagnostics	50.87	67.92	78.18	66.91
Full three-stage method	52.70	70.90	82.25	69.98

Table 2: Held-out test comparison with generic LLM rewriting baselines. All methods are finalized without access to the test results.

Website	Offline	Platform A	Platform B
Target website	67.40	63.3%	58.3%
Competitor 1	63.80	56.7%	53.3%
Competitor 2	58.90	46.7%	40.0%
Competitor 3	56.70	41.7%	48.3%
Competitor 4	49.50	30.0%	33.3%

Table 3: Offline scores and online recall rates in the five-website exploratory pilot. Each of the 20 queries is repeated three times on each commercial platform.

5.3 Exploratory Online Consistency Check on Commercial Platforms

To examine whether the controlled offline evaluation shows directional consistency with real-world visibility, we conduct a small-scale exploratory pilot on two widely used commercial generative-search platforms in China. We select 20 queries from the held-out test set while approximately preserving the held-out set’s intent distribution and evaluate the target website together with four representative competitors. Each query is submitted three times to each platform using the same prompt and account settings, yielding 60 valid responses per platform.

A website is counted as recalled when a response contains its brand name, official domain, cited webpage, or other information that can be unambiguously attributed to that website. Recall rate is defined as the proportion of valid responses in which the website is recalled. For the offline comparison, we rerun the proposed evaluation on the same 20-query subset and the same five-website candidate pool. Each website is treated in turn as the focal website, while the remaining candidates, queries, models, prompts, and scoring rules are kept fixed.

As shown in Table 3, the offline and online evaluations exhibit broadly similar ranking patterns in this pilot. The target website ranks first in both settings, followed by Competitor 1, while Competitor 4 remains last. Competitors 2 and 3 exchange positions in the online results, illustrating platform-specific variation.

Across the five evaluated websites, the Spearman correlation between offline scores and average online recall rates is $\rho = 0.90$. We report this correlation descriptively as exploratory evidence of directional consistency between the controlled offline evaluation and observed commercial-platform visibility. Because the comparison includes only five websites, 20 queries, and two platforms, it should not be interpreted as statistical validation of offline-to-online transfer or platform-level prediction. The offline score is therefore used as a controlled diagnostic indicator rather than an estimate of recall probability on any specific commercial platform.

5.4 Human Validation of Generator Diagnostic Labels

Two independent annotators reviewed a stratified sample of 200 generated answers using shared three-point rubrics for truthfulness and brand recommendation. Truthfulness assesses whether the generated claims are supported by the target sitemap and verified fact document, whereas brand recommendation assesses whether mentioning or recommending the target brand is relevant and evidence-based.

Inter-annotator agreement, measured using quadratic-weighted Cohen’s κ , was 0.86 for truthfulness and 0.74 for brand recommendation. Disagreements were adjudicated to produce the final reference labels.

Compared with these adjudicated labels, the LLM evaluator achieved Macro-F1 scores of 0.90 for truthfulness and 0.76 for brand recommendation. Its weighted κ values were 0.82 and 0.69, with 95% bootstrap confidence intervals of [0.76, 0.88] and [0.61, 0.76], respectively. These results support the reliability of the truthfulness and recommendation diagnostics, rather than validating the complete generator quality score.

6 Discussion

6.1 Case Discussion

We examine epoch 63, the checkpoint with the highest overall validation score, to analyze how stage-level feedback guides sitemap revision. Retriever feedback adds direct answers, query-relevant terms, and FAQ-style content; reranker feedback makes distinctive claims and supporting evidence more prominent; generator feedback strengthens factual boundaries and attribution. These signals are integrated to reduce repetition, place evidence near the

claims it supports, and keep all revisions grounded in the enterprise fact document.

The resulting sitemap shows that GEO performance depends not only on what information is included, but also on how it is organized. Related definitions, safety mechanisms, evidence, and fulfillment capabilities are grouped into coherent, extractable units, making them easier to retrieve, rank, and use in grounded generation.

6.2 Stage-Level Trade-offs and Industrial Implications

Aggregate visibility scores can obscure whether a sitemap is weak in query matching, competitive relevance, or grounded content use. These dimensions require different revisions: retrieval benefits from broader query and scenario coverage, reranking from direct and differentiated evidence, and generation from concise, verifiable claims with clear factual boundaries.

The best-performing checkpoints also differ across stages, revealing trade-offs between coverage, prominence, and factual precision. Adding content may improve retrieval while diluting important evidence, whereas stronger promotional language may improve differentiation but reduce grounding quality. Enterprise GEO should therefore be treated as a multi-objective optimization problem.

In practice, content teams can identify the weakest diagnostic dimension, revise the corresponding content fields, and reevaluate all three stages for possible regressions. This stage-aware workflow provides more actionable guidance than uniform rewriting or selecting the final optimization epoch.

7 Conclusion

We present a controlled three-stage framework for diagnosing enterprise GEO failures. Recursive sitemap revision improves held-out performance, while stage-specific differences reveal non-monotonic trade-offs and support targeted, fact-grounded content revision. The framework offers a reproducible alternative to relying solely on final mentions or aggregate visibility outcomes.

8 Limitations

First, the proposed offline framework cannot fully reproduce commercial generative-search systems, which may rely on real-time indexes, proprietary

retrieval and ranking methods, external search augmentation, and user personalization. The main experiments also rely on a fixed retriever, reranker, generator, and evaluator stack, so the reported scores may depend on the selected components. The offline scores should therefore be interpreted as controlled diagnostic indicators rather than direct estimates of citation, mention, or recommendation probabilities on a specific platform. Although the small-scale online pilot shows similar ranking trends, it covers only two platforms, 20 queries, and five websites and does not establish stable offline-to-online correspondence across platforms, model stacks, or time.

Second, the study examines one anonymized company in the consumer-health and personalized-nutrition domain. A single enterprise case cannot establish generalizability across industries, languages, or content types, which may differ in query intent, competitive density, factuality requirements, and acceptable revision strategies. Online deployment testing has been ongoing since June 2026, but the observation window is not yet sufficient for longitudinal validation. Broader cross-enterprise and long-term deployment studies are therefore required.

Third, the current evaluation covers a 21-page target website with approximately 94K tokens, evaluated against nine fixed competitors whose sizes range from 15 to 220 pages and 15K to 350K tokens, as summarized in Table 1. These results demonstrate the framework only at the reported scale. Larger websites, broader competitor pools, frequently updated content, and multiple product categories may require substantially greater computational, preparation, and factual-verification effort.

Finally, the structured site representation primarily preserves textual content and page-level structure. It does not explicitly model arbitrary hyperlink graphs, page authority, interactive components, multimedia, or platform-specific processing of structured data. The reported results therefore characterize content- and structure-level performance rather than all webpage or platform factors that may affect real-world generative-search visibility.

9 Ethical Considerations

Except for the verified enterprise fact document provided by the participating company, all website and competitor content used in this study is col-

lected from publicly available sources. The query sets are derived from enterprise-provided user questions, but all queries are manually reviewed before use to ensure that they contain no personally identifiable information, user identifiers, or other information that could be traced back to an individual. The experiments are conducted on anonymized content representations, and no private user profiles or sensitive personal records are included in retrieval, ranking, generation, or evaluation.

To reduce the risk of factual errors or misleading health-related content, the revision prompts explicitly prohibit unsupported claims, fabricated evidence, and information that may mislead users. Every generated sitemap is constrained by the verified enterprise fact document, which defines the factual boundaries of permitted revisions. These safeguards are intended to prevent optimized content from introducing inaccurate product, nutrition, or health information, although they cannot completely eliminate model-generated errors. The proposed system is designed solely for enterprise content evaluation and optimization; it is not intended for clinical diagnosis, treatment decisions, or personalized medical and nutritional advice.

References

- Pranjal Aggarwal, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, Karthik Narasimhan, and Ameet Deshpande. 2023. [Geo: Generative engine optimization](#). *Preprint*, arXiv:2311.09735.
- Puneet S. Bagga, Vivek F. Farias, Tamar Korkotashvili, Tianyi Peng, and Yuhang Wu. 2025. [E-geo: A testbed for generative engine optimization in e-commerce](#). *Preprint*, arXiv:2511.20867.
- Qiyuan Chen, Jiahe Chen, Hongsan Huang, Qian Shao, Jintai Chen, Renjie Hua, Hongxia Xu, Ruijia Wu, Ren Chuan, and Jian Wu. 2025a. [Beyond keywords: Driving generative search engine optimization with content-centric agents](#). *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](#). *Preprint*, arXiv:2508.11158.
- Riley Grossman, Songjiang Liu, Michael K. Chen, Mike Smith, Cristian Borcea, and Yi Chen. 2026. [How generative ai disrupts search: An empirical study of google search, gemini, and ai overviews](#). *Preprint*, arXiv:2604.27790.
- Sunghwan Kim, Wooseok Jeong, Serin Kim, Sangam Lee, and Dongha Lee. 2026. [Sageo arena:](#)

[A realistic environment for evaluating search-augmented generative engine optimization](#). *Preprint*, arXiv:2602.12187.

Julius Schulte, Malte Bleeker, and Philipp Kaufmann. 2026. Don’t measure once: Measuring visibility in AI search (GEO). *arXiv preprint arXiv:2604.07585*.

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

Jiaqi Yuan, Jialu Wang, Zihan Wang, Qingyun Sun, Ruijie Wang, and Jianxin Li. 2026. AgenticGEO: A self-evolving agentic system for generative engine optimization. *arXiv preprint arXiv:2603.20213*.

Kai Zhang, Xinyue He, and Jingang Yao. 2026. From citation selection to citation absorption: A measurement framework for generative engine optimization across AI search platforms. *arXiv preprint arXiv:2604.25707*.

A Implementation and Reproducibility Details

The complete model and execution configuration is reported in Tables 4 and 5. API model identifiers correspond to the experiments executed on June 16, 2026.

B Naturalness Constraint

We apply a rule-based naturalness checker after each sitemap revision. Naturalness is used as a checkpoint eligibility constraint rather than as an additional term in the aggregate GEO score.

The content-length ratio is computed as

$$r_{\text{len}}(v_i) = \frac{L(v_i)}{L(v_0)}, \quad 0.60 \leq r_{\text{len}}(v_i) \leq 1.50. \quad (14)$$

Let $N(v_i) \in \{0, 1\}$ denote the naturalness decision. Checkpoint selection is performed only over eligible validation checkpoints:

$$v^* = \arg \max_{v_i: N(v_i)=1} S(v_i). \quad (15)$$

Component	Backend	Model / implementation	Precision	Temp.	Top- p	Max tokens	Execution
Retriever	Local / Haystack	InMemoryBM25Retriever with a custom hashing-embedding retriever	N/A	N/A	N/A	N/A	Single-process local retrieval
Reranker	Local / Hugging Face	BAAI/bge-reranker-v2-m3; FP32; local path: bge-reranker-v2-m3	dynamic INT8 in fast-small	N/A	N/A	N/A	Sequential queries; batch size 4, or 16 in fast-small
Generator	SiliconFlow API v1	Primary: DeepSeek-V4-Flash	Provider-managed	0.2	0.8	1,000	Up to 8 concurrent query workers
Evaluator	SiliconFlow API v1	qwen3.6-36B-A3B	Provider-managed	0.0	0.8	3,000	Up to 8 concurrent workers aligned with generator queries
Revision model	SiliconFlow API v1	DeepSeek-V4-Pro	Provider-managed	0.3 prompt; 0.2 sitemap	0.8	6,000 prompt; 120,000 sitemap	Prompt and sitemap revisions executed sequentially

Table 4: Model backends, inference parameters, and execution settings. No explicit random seed or repetition penalty was configured.

Field	Component	Implementation
Retrieval method	Retriever	Hybrid local retrieval combining lexical BM25 matching with a lightweight hashing-based embedding signal over the same preprocessed website chunks.
Reranking method	Reranker	BAAI/bge-reranker-v2-m3 independently assigns query-candidate relevance scores to the fixed candidate representations; no Retriever outputs are used.
Generation procedure	Generator	Receives the serialized candidate websites and produces a natural-language answer. The output is treated as free text and does not require JSON parsing.
Evaluation procedure	Evaluator	Receives the query, generated answer, target content, and factual evidence, and returns structured JSON scores and diagnostic fields. Generation and evaluation are performed by separate models.
Revision procedure	Revision model	After stage-level aggregation, first generates an optimization prompt and then produces the next sitemap version. It is not involved in scoring the current checkpoint.
Generator fallback	Generator	Ordered fallback: Pro/deepseek-ai/DeepSeek-V4-Flash; deepseek-ai/DeepSeek-V4-Flash; Pro/deepseek-ai/DeepSeek-V4.0-Flash; deepseek-ai/DeepSeek-V4.0-Flash; deepseek-ai/DeepSeek-V4-Flash-Preview; Pro/deepseek-ai/DeepSeek-V4-Flash-Preview.
Request timeout	Generator / Evaluator	Each query-level API request has a timeout of 45 seconds. Failed generator requests proceed through the configured fallback sequence.
Generator failure	Generator	If all candidate models fail, the query is recorded as a failed row and excluded from successful-generation aggregates.
JSON parsing	Evaluator / Revision model	First applies json.loads; if parsing fails, extracts the outermost JSON object, performs one structured-output retry, and then applies JSON repair.
Invalid JSON	Evaluator / Revision model	Outputs that remain invalid after retry and repair are marked as failed; no default score is assigned.
Revision retry	Revision model	Up to 3 attempts per optimization call and 10 retry rounds, with 60 seconds between attempts and 600 seconds between rounds.
Candidate ordering	All evaluation stages	Candidate website order is randomized to reduce systematic positional bias.
Controlled variables	All stages	The query set, competitor set, prompts, scoring rubrics, and candidate construction procedure remain fixed across checkpoints. Only the target sitemap changes.
Execution logging	All stages	Logs retain execution status, selected model identifier, raw generation, parsed evaluator output, and failure information for every query.
Score aggregation	All stages	Stage-level scores are computed only from cases considered valid under the corresponding evaluation definition.
Version information	All components	Local implementations, local checkpoints, and API models were executed on June 16, 2026. API providers do not expose fixed checkpoint hashes or inference precision.

Table 5: Execution procedure, fallback policy, failure handling, and reproducibility controls.

Field	Rule type	Implementation
Component	Implementation	Rule-based naturalness checker executed after each sitemap revision.
Purpose	Constraint	Prevents revisions from becoming keyword-stuffed or score-oriented text that is unsuitable for an official website.
Input	I/O	Original sitemap JSON and the optimized sitemap JSON from the current epoch.
Output	I/O	Boolean field passed, content-length ratio, brand-term density, and structural, factual, repetition, and meta-language warnings.
Length ratio	Hard gate	Let $L(v_0)$ and $L(v_i)$ denote the total characters in all content_text fields of the original and optimized sitemaps. The optimized sitemap must satisfy $0.60 \leq L(v_i)/L(v_0) \leq 1.50$.
Preferred length	Prompt guidance	The revision prompt recommends maintaining 80%–150% of the original length. The implemented eligibility gate remains 60%–150%.
Brand-term density	Diagnostic	Detects excessive repetition of brand names, product names, system names, algorithm names, and proprietary supply-chain or quality-control terms.
Keyword stuffing	Diagnostic	Query terms must occur in semantically coherent sentences rather than being mechanically enumerated or repeatedly inserted.
Content repetition	Diagnostic	Flags repeated claims, FAQ answers, product descriptions, and duplicated brand statements within or across pages.
Website structure	Constraint	Preserves plausible page structure, including titles, paragraphs, FAQ entries, and modular content sections.
Page identity	Hard gate	Existing pages cannot be removed, unsupported pages cannot be created, and page_id, source_url, and loc cannot be modified.
Factual support	Hard gate	All added or rewritten claims must be supported by the original sitemap or the enterprise fact document.
Forbidden claims	Hard gate	Prohibits unsupported clinical evidence, regulatory approval, efficacy guarantees, treatment claims, prices, refund terms, and after-sales policies.
Meta-language	Constraint	Prohibits experiment-oriented expressions such as “GEO score,” “retriever optimization,” and “improving the reranker” in website content.
Failure handling	Selection rule	A failed epoch remains in the execution log but is excluded from best-checkpoint selection. Its warnings are included in the next revision prompt.
Implementation date	Reproducibility	June 16, 2026.

Table 6: Naturalness-checker configuration. Hard gates directly determine checkpoint eligibility, while diagnostic rules identify optimization artifacts and guide subsequent revisions.