OFF-GEO: MITIGATING OFF-TARGET VISIBILITY IN GENERATIVE ENGINE OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative Engine Optimization (GEO) methods use large language models (LLMs) to rewrite source content and improve its visibility in generative engine responses. However, these gains can extend to queries that a content provider does not wish to target. Mitigating these undesired gains is challenging because desired and undesired queries may rely on overlapping source content, so reversing GEO edits can affect both. We propose OFF-GEO, a framework that learns a shared editing rule for GEO-optimized sources. We optimize the rule to maximize desired visibility subject to a minimum average reduction in undesired visibility. Candidate rules are compared using visibility changes measured from generative engine responses, while feedback from edits that fail to reduce undesired visibility guides rule refinement. The learned rule guides edits to new optimized sources without query inputs or generative engine feedback. Experiments on E-commerce and GEO-Bench show that OFF-GEO reduces undesired visibility more than desired visibility across all metrics, while keeping desired visibility above that of the original sources. Despite removing portions of GEO-optimized content, OFF-GEO also maintains competitive answer quality on desired queries.

1 INTRODUCTION

Generative Engine Optimization (GEO) (Aggarwal et al., 2024) aims to improve the visibility of content in responses produced by generative engines. These engines retrieve information from multiple sources and synthesize it into an answer. A source's visibility depends on whether its content is selected, how much it contributes to the response, and how prominently it is attributed.

Nevertheless, the effects of GEO can extend beyond its target queries. Patterns that make content attractive to an engine may operate across different information needs, increasing its visibility even when the query changes. We call these *off-target visibility increases*. These gains become undesirable when they increase visibility for query types the content provider does not wish to target. Our goal is to mitigate these undesired increases while retaining visibility on desired queries.

Existing GEO work typically uses generative engine feedback to guide source rewriting toward higher visibility on target queries. Specifically, some methods rewrite source content to improve visibility for a given query (Wu et al., 2026b;a), while others coordinate edits across multiple queries to improve overall visibility (Zhou et al., 2026). Our setting requires selective control: visibility should decrease on undesired queries while being retained on desired queries. This is difficult because the same passage may support both query types, so removing an off-target expansion can also weaken useful content. Moreover, source editing takes place before future queries are available. A practical method therefore needs reusable editing guidance that transfers to new sources without query inputs or generative engine feedback at inference time.

We propose OFF-GEO, a framework that learns a shared editing rule to mitigate undesired visibility gains while retaining desired visibility (see Fig. 1). We iteratively refine candidate rules using feedback from unsuccessful edits, measuring visibility changes on both desired and undesired queries. Candidates are evaluated on the same training documents, and rule selection maximizes aggregate desired visibility retention subject to a minimum average reduction in undesired visibility. The learned rule guides deletions and restorations in new optimized documents. A verifier checks the referenced spans and excludes edits flagged as affecting content central to the original source.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

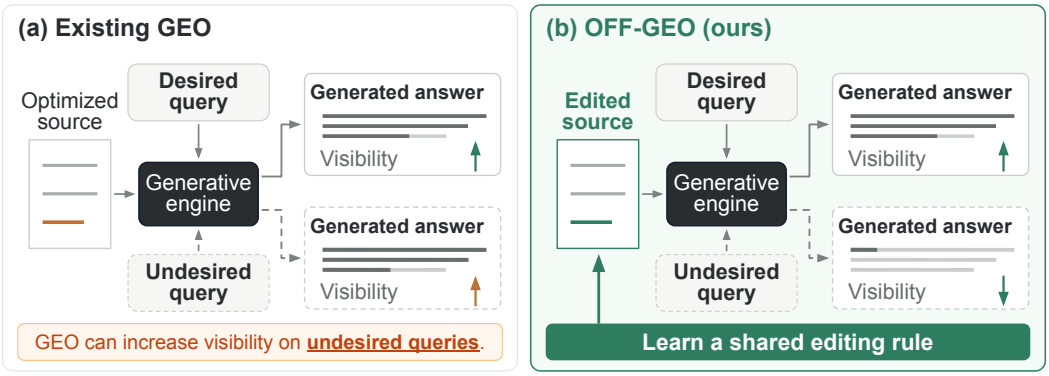


Figure 1: **OFF-GEO mitigates undesired visibility gains introduced by GEO while retaining desired gains.** (a) GEO edits can increase a source's visibility on both desired and undesired queries. (b) OFF-GEO learns a shared editing rule from generative engine feedback and applies it to optimized sources without access to future queries. Solid and dashed paths denote desired and undesired queries, respectively.

During inference, this editing procedure applies the shared rule across documents without access to future queries or generative engine feedback.

We evaluate OFF-GEO on the E-commerce benchmark from AutoGEO (Wu et al., 2026b) and GEO-Bench (Aggarwal et al., 2024). Across both benchmarks, OFF-GEO reduces undesired visibility more than desired visibility on every metric. Desired visibility remains above that of the original sources, preserving part of GEO's benefit on desired queries. A qualitative example illustrates how editing reduces off-target elaboration while retaining the information needed for the desired query. Despite removing portions of the GEO-optimized sources, OFF-GEO also maintains competitive answer quality on desired queries in the E-commerce evaluation, with usefulness and citation support largely preserved.

Our contributions are as follows:

- We identify GEO-induced visibility gains on undesired queries and formulate their selective mitigation. Since the same source content can contribute to visibility under both query types, we maximize desired visibility retention subject to a minimum average reduction in undesired visibility.
- We introduce OFF-GEO, a framework that learns a shared editing rule through generative-engine feedback and constrained candidate selection. The learned rule transfers visibility control across sources, enabling edits to new sources without access to their queries or engine feedback at inference time.
- Experiments on E-commerce and GEO-Bench show that OFF-GEO reduces undesired visibility more than desired visibility while preserving desired visibility gains over the original sources. Qualitative analysis illustrates how editing changes source use, and answer utility evaluation on E-commerce shows that usefulness and citation support are largely preserved on desired queries.

2 Related Work

Generative engine optimization (GEO). Aggarwal et al. (2024) introduce GEO as a framework for improving source visibility in generative engine responses, together with visibility metrics and a benchmark for evaluating content revisions. Recent methods learn editing strategies from feedback or coordinate revisions across multiple queries. AutoGEO (Wu et al., 2026b) extracts preference rules from LLM explanations of visibility differences between documents and uses these rules to guide rewriting. MAGEO (Wu et al., 2026a) uses an LLM evaluator to assess candidate revisions for predicted gains and semantic fidelity. It distills effective editing patterns into reusable strategies

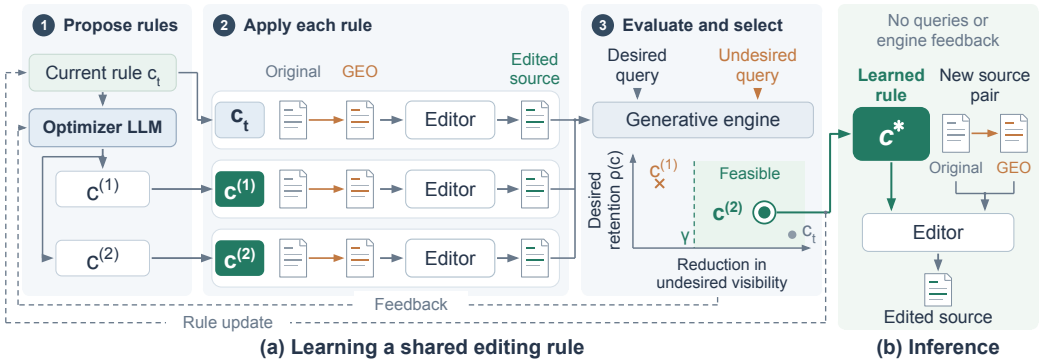


Figure 2: **Learning and applying a shared editing rule in OFF-GEO.** (a) An optimizer LLM proposes candidates from failures of the current rule (c_t). Each candidate and c_t independently edit the same GEO-optimized sources, using the originals as references. Selection maximizes desired visibility retention ($\rho(c)$) subject to a mean undesired visibility reduction of at least (γ). In the schematic, $c^{(1)}$ is infeasible, and $c^{(2)}$ is selected over feasible c_t for its higher retention despite a smaller reduction. (b) The fixed learned rule $c^\star$ edits new source pairs without queries or engine feedback.

for different engines. IF-GEO (Zhou et al., 2026) combines revision requests from multiple queries into a shared rewriting plan, resolving conflicts when the requests target overlapping passages. These methods aim to maximize visibility gains while accounting for content fidelity or cross-query stability. OFF-GEO instead learns a shared editing rule that reduces the off-target effects of GEO on undesired queries while retaining gains on desired queries.

LLMs as optimizers. LLM-based prompt optimization uses evaluation feedback to guide the generation and revision of natural language instructions. APE (Zhou et al., 2023) generates instruction candidates and selects them based on task performance. OPRO (Yang et al., 2024) conditions new proposals on previously evaluated candidates and their scores. ProTeGi (Pryzant et al., 2023) uses critiques of training errors to revise instructions, and PromptAgent (Wang et al., 2023) combines error feedback with Monte Carlo tree search. Evolutionary approaches maintain populations of candidate prompts: Promptbreeder (Fernando et al., 2024) evolves both task prompts and the instructions used to mutate them, while GEPA (Agrawal et al., 2026) combines reflection on execution traces with Pareto-based search. LEAP (Zhang et al., 2024) distills model errors into task-specific principles that guide predictions on unseen examples. For constrained prompt selection, Li et al. (2026) allocate evaluations across a candidate pool to identify prompts that maximize a primary metric while satisfying thresholds on other metrics. OFF-GEO learns a shared editing rule to control the visibility of GEO-optimized sources under desired and undesired queries. The same source content can contribute to visibility under both query types, so removing content to suppress undesired visibility can also diminish desired visibility. We therefore evaluate both effects of each candidate rule on the same sources, reducing undesired visibility subject to a bound on aggregate desired visibility loss. The learned rule is then applied to new sources without access to their queries or generative-engine feedback.

3 OFF-GEO: Mitigating Undesired Visibility Gains in GEO

We introduce OFF-GEO, a framework that learns a shared editing rule c across different sources to reduce GEO-induced visibility gains on undesired queries while retaining visibility on desired queries (Fig. 2). During training, we test candidate rules by editing GEO-optimized documents and measuring their visibility under both query types. Generative engine feedback guides the search for a rule that reduces undesired visibility with limited aggregate loss in desired visibility. The learned rule is then applied to new documents without access to their queries or engine feedback. Algorithm 1 summarizes rule learning and inference.

3.1 Problem Setup

Let d be an original document and $d' = \mathcal{M}(d)$ its GEO-optimized version. Following GEO-bench (Aggarwal et al., 2024), we measure visibility score $v(d, q)$ based on the length and position of passages that cite d in the generative engine's answer to q. Visibility scores are represented on a $[0, 1]$ scale. For example, a document referenced throughout the beginning of an answer receives a higher score than one referenced briefly near the end. The GEO-induced visibility gain for q is $v(d', q) - v(d, q)$.

For each training document d_i, let q_i^D denote the original query targeted by GEO and q_i^U a related query designated as undesired for that source. We represent the training set as $\mathcal{T} = \{(d_i, q_i^D, q_i^U)\}_{i=1}^N$ and keep each query pair fixed throughout rule learning. The query-pair construction procedure is detailed in the experimental setup.

Given $\mathcal{T}$ and the GEO-optimized sources $\{d_i'\}$, we seek a shared editing rule $c^\star$ that revises these sources to retain as much desired visibility as possible while achieving a minimum average reduction in undesired visibility. The learned rule is applied to new source pairs without access to their queries or generative-engine feedback.

3.2 Learning a Shared Editing Rule

Candidate rule generation. We initialize the current rule c_0 as empty string. At each iteration, we evaluate the current rule on a training minibatch and collect failure cases where editing does not reduce undesired visibility. An optimizer LLM examines the current rule, the original and edited documents, and the changes in both desired and undesired visibility relative to the GEO-optimized documents to produce a critique of these failures. Conditioned on the current rule and this critique, it generates up to K self-contained candidate rules. Each candidate must remain document-agnostic and actionable using only the original and optimized sources, without access to query text. If no failures are found, we retain the current rule and proceed to the next training iteration.

Rule editing and evaluation. Each rule is applied independently to the same GEO-optimized source d_i'. Edits are not accumulated across candidate rules or learning iterations. We evaluate the candidates alongside the current rule on shared training documents. For each rule c, the editing function $\mathcal{E}$ uses an LLM to propose deletions or restorations in the optimized document d_i', with the original document d_i provided as a reference. A verifier checks the quoted spans and excludes edits that the editor flags as affecting core content, yielding the revised document $d_{i,c} = \mathcal{E}(d_i, d_i'; c)$. We then measure the revised document's visibility under both query types.

To evaluate a candidate rule c, we provide the generative engine with q_i^D as the user query, $d_{i,c}$ as the target document, and the other documents available for that query. The engine generates an answer to q_i^D. We repeat the procedure for q_i^U, and then evaluate both queries again with d_i' replacing $d_{i,c}$. For each query, all other documents remain fixed. We define the change in the target document's visibility caused by editing as

$$\Delta v_i^a(c) = v(d_{i,c}, q_i^a) - v(d_i', q_i^a), \qquad a \in \{D, U\}. \tag{1}$$

Negative values indicate reduced visibility relative to the GEO-optimized source.

Constrained rule selection. Edits that reduce undesired visibility may also reduce desired visibility. We therefore seek rules that achieve a specified undesired reduction while preserving as much desired visibility as possible.

We measure desired retention using the ratio of total visibility after editing to that of the GEO-optimized sources:

$$\rho(c) = \frac{\sum_{i=1}^N v(d_{i,c}, q_i^D)}{\sum_{i=1}^N v(d_i', q_i^D)}. \tag{2}$$

The ratio measures retained total desired visibility relative to the optimized sources: $\rho(c) = 1$ denotes equal aggregate visibility, and $\rho(c) > 1$ denotes an increase. The denominator is fixed across rules and is assumed positive.

Algorithm 1 OFF-GEO: Learning and applying a shared editing rule

Require: $\mathcal{T}$, optimized sources $\{d'_i\}$, candidate limit K
Require: Iterations L, selection budget B, threshold γ
Ensure: Shared editing rule $c^\star$

$c \leftarrow \varnothing$
for $t = 1, \ldots, L$ **do**
  Evaluate c on a training minibatch
  Collect failures with $\Delta v_i^U(c) = v(d_{i,c}, q_i^U) - v(d'_i, q_i^U) \geq 0$
  if no failures are found **then**
    Retain c and continue to the next training iteration
  Critique failures and propose up to K rules in $\mathcal{C}_{\text{new}}$
  $\mathcal{C} \leftarrow \{c\} \cup \mathcal{C}_{\text{new}}$
  Reset statistics for all rules in $\mathcal{C}$
  Initialize the Successive Rejects schedule with budget B
  while $|\mathcal{C}| > 1$ **do**
    Evaluate all rules in $\mathcal{C}$ on the next shared batch
    Update cumulative average Δv_i^U and pooled retention ρ for each rule
    if any rule has average $\Delta v_i^U > -\gamma$ **then**
      Remove the rule with largest average Δv_i^U from $\mathcal{C}$
    else
      Remove from $\mathcal{C}$ the rule with the smallest pooled desired retention $\rho(c)$ in Eq. 2
  if the surviving rule is new and feasible **then**
    $c \leftarrow$ the surviving rule
$c^\star \leftarrow c$
return $c^\star$
Application: $\mathcal{E}(d, d'; c^\star)$ for new sources (d, d').

We maximize $\rho(c)$ subject to a minimum average reduction in undesired visibility:

$$\max_{c} \quad \rho(c) \qquad \text{s.t.} \quad \frac{1}{N} \sum_{i=1}^{N} \Delta v_i^U(c) \leq -\gamma, \tag{3}$$

where $\gamma > 0$ specifies the reduction threshold. The constraint excludes rules that leave the optimized sources unchanged. Among feasible rules, the objective favors greater desired retention rather than further undesired reduction. We assess the reduction across training sources because a shared rule can affect individual documents differently. We use $\gamma = 0.02$ in our experiments. This corresponds to a minimum average reduction of two percentage points in undesired visibility.

We compare the current rule and newly proposed candidates using a constrained variant of Successive Rejects (Audibert & Bubeck, 2010). At each phase, all surviving rules are evaluated on the same additional training documents. We update their mean undesired-visibility changes and pooled desired-retention ratios using the evaluations accumulated so far. If any rule fails the reduction constraint, we eliminate the failing rule with the smallest average reduction in undesired visibility. If all surviving rules satisfy the constraint, we eliminate the rule with the lowest pooled desired retention. The current rule follows the same elimination criteria as the new candidates.

We repeat this process until one rule remains. This rule becomes the current rule for the next iteration if it satisfies the reduction constraint. Otherwise, we retain the previous rule. After training, we fix the resulting rule $c^\star$ for editing new sources.

3.3 Inference

For an unseen pair (d, d'), OFF-GEO applies the frozen rule $c^\star$ through the same editing procedure used during training, producing $\mathcal{E}(d, d'; c^\star)$. The original source provides the reference for restorations and core-content assessment, while the optimized source supplies the text to be edited. Deployment uses neither the evaluation queries nor generative engine feedback, and all model parameters remain fixed. The reduction constraint is used during rule selection. Its satisfaction on new sources is assessed empirically.

4 EXPERIMENTAL RESULTS

Our experiments investigate whether OFF-GEO can selectively control the visibility of the same source across desired and undesired queries. We assess its ability to reduce undesired visibility while retaining desired visibility (RQ1), then examine the resulting changes in source use (RQ2) and answer quality (RQ3). To understand the behavior and limitations of the learned rule, we study how performance evolves over training iterations (RQ4), and relates to similarity between desired and undesired queries (RQ5).

4.1 EXPERIMENTAL SETUP

Datasets. We evaluate on the E-commerce benchmark from AutoGEO (Wu et al., 2026b) and GEO-Bench (Aggarwal et al., 2024). For each target source, we use its original query as the desired query q_D and pair it with a semantically related but distinct query as the undesired query q_U. This provides a non-trivial setting for selective unlearning while avoiding near-duplicate queries that represent essentially the same information need. Specifically, we embed queries within the same split using `all-MiniLM-L6-v2` (Reimers & Gurevych, 2019), exclude queries with cosine similarity above 0.6 to q_D, and select the most similar query among the remaining candidates as q_U. All methods are evaluated using the same target sources, (q_D, q_U) pairs, and candidate-document sets. The training pool contains all available GEO-optimized training sources. For visibility evaluation, we exclude sources whose undesired visibility under AutoGEO$_{\text{API}}$ is zero or equals the default score assigned when an answer cites no source. This filtering depends only on AutoGEO$_{\text{API}}$ and is independent of OFF-GEO's outputs.

Metrics. Following GEO-Bench (Aggarwal et al., 2024) and AutoGEO (Wu et al., 2026b), we measure source visibility using three citation-based metrics: Word, the normalized amount of generated text attributed to the target source. Pos, which additionally accounts for where the source-linked content appears in the response, assigning greater weight to earlier positions. Overall, which combines content contribution and position into a unified visibility score. We report all three metrics for both the desired query q_D and the undesired query q_U. We assess whether editing reduces visibility under q^U while retaining as much visibility as possible under q^D.

Baselines. We compare three settings. Original uses the original source without GEO optimization. AutoGEO$_{\text{API}}$ applies the official AutoGEO rewriting procedure to increase source visibility for the desired query. OFF-GEO is applied to the GEO source to mitigate off-target visibility while preserving its desired-query visibility gains.

Implementation details. We generate AutoGEO$_{\text{API}}$ rewrites using Gemini 2.5 Pro (Comanici et al., 2025) with AutoGEO's official rule set (Wu et al., 2026b). Gemini 2.5 Flash-Lite (Comanici et al., 2025) serves as the generative engine for visibility evaluation, while Gemini 2.5 Pro is used as the document editor. We learn the editing rule on the training split over three iterations of candidate generation and constrained Successive Rejects. Once learned, the rule is fixed and applied to test sources without access to test-time queries or generative engine feedback.

4.2 OFF-TARGET VISIBILITY GAINS FROM GEO

RQ1. Can OFF-GEO mitigate undesired visibility gains while retaining desired visibility? We evaluate OFF-GEO on the E-commerce benchmark from AutoGEO (Wu et al., 2026b) and GEO-Bench (Aggarwal et al., 2024). Table 1 compares original sources, GEO-optimized sources, and sources rewritten by OFF-GEO. Relative to GEO-optimized sources, OFF-GEO reduces visibility on *undesired* queries more than on *desired* queries across all metrics and both benchmarks. The learned rule therefore suppresses GEO-induced visibility gains on undesired queries while retaining a substantial portion of GEO's benefit on desired queries.

RQ2. How does reducing undesired visibility change the engine's use of a source? Table 2 compares the engine's answers under AutoGEO$_{\text{API}}$ and OFF-GEO on a GEO-Bench test case. Visibility differences are reported in percentage points (pp). For the desired query, Overall visibility

Table 1: **OFF-GEO reduces undesired visibility more than desired visibility across both benchmarks.** We report mean source visibility (%) on the evaluated test subsets. Original provides the reference before GEO optimization. For each metric, Δ^D and Δ^U denote the desired and undesired visibility changes from AutoGEO$_{\text{API}}$ to OFF-GEO, in percentage points. Mean gap averages $\Delta^D - \Delta^U$ across Word, Pos, and Overall; positive values indicate that the undesired reduction exceeds the desired loss on average. Bold marks the lower undesired visibility and higher desired visibility between AutoGEO$_{\text{API}}$ and OFF-GEO, as well as the positive mean gaps.

	Undesired visibility ↓			Desired visibility ↑			Mean gap
Method	Word	Pos	Overall	Word	Pos	Overall	(pp) ↑
E-commerce							
Original (reference)	22.38	22.47	22.57	19.67	20.12	20.01	—
AutoGEO$_{\text{API}}$	41.51	40.51	41.26	**34.06**	**34.32**	**34.43**	—
OFF-GEO	**30.14**	**30.24**	**30.47**	27.07	27.43	27.52	**+3.88**
GEO-Bench							
Original (reference)	21.79	21.78	21.78	21.70	21.57	21.76	—
AutoGEO$_{\text{API}}$	39.55	39.55	40.09	**38.86**	**39.47**	**39.84**	—
OFF-GEO	**32.41**	**32.46**	**32.87**	32.69	33.33	33.51	**+0.94**

Table 2: **OFF-GEO reduces undesired visibility while preserving the core information needed for the desired query.** This qualitative example is sampled from GEO-bench. Values report Overall visibility (%); parentheses show changes from AutoGEO$_{\text{API}}$ in percentage points.

Query	**AutoGEO$_{\text{API}}$**	**OFF-GEO**
Desired query q^D "what items does habitat for humanity take?"	Retained: Explains that Habitat for Humanity ReStores accept donations such as furniture, appliances, and building materials, together with examples of accepted items. Overall: 41.38%	**Retained:** The information about items accepted by Habitat for Humanity ReStores, including furniture, appliances, and building materials. Overall: 42.38% (+1.00 pp)
Undesired query q^U "Should more effort be made to protect endangered species?"	**Retained:** Habitat for Humanity ReStores accept donations such as furniture and appliances. **Off-target:** Elaborates on accepted items and states that donating items "diverts them from landfills, promoting sustainability and reducing waste," even though the query asks about protecting endangered species. Overall: 43.30%	**Retained:** A brief description of Habitat for Humanity ReStores and their accepted donations, including furniture and appliances. **Omitted:** The sustainability claim about reducing landfill waste. The answer instead explicitly acknowledges that the provided results do not contain information about protecting endangered species. Overall: 20.00% (−23.30 pp)

changes only slightly, from 41.38% to 42.38% (+1.00 pp), while all essential information is retained. For the undesired query, Overall visibility falls from 43.30% to 20.00% (−23.30 pp). This larger change is accompanied by reduced use of query-irrelevant content. Especially, the answer removes an off-target claim and explicitly acknowledges that the retrieved results do not address the query. Although the source remains cited, its off-target contribution is reduced, while the essential content of the desired answer is preserved.

RQ3. Can OFF-GEO reduce undesired visibility while maintaining competitive answer quality? We evaluate answers to desired queries on E-commerce using AutoGEO's Generative Engine Utility (GEU) evaluation, with `gpt-4o-mini` as the judge. Table 3 shows that OFF-GEO remains competitive with AutoGEO$_{\text{API}}$ across the four quality metrics. Recall and Insight remain close to

Table 3: **OFF-GEO maintains competitive answer quality on desired queries.** We report AutoGEO's official GEU scores for answers to desired queries on E-commerce, with `gpt-4o-mini` as the judge. Bold marks the best value in each column.

Method	Precision ↑	Recall ↑	Clarity ↑	Insight ↑
Original	88.83	94.85	53.11	41.48
$\text{AutoGEO}_{\text{API}}$	**91.79**	96.31	**55.19**	**43.30**
OFF-GEO	90.25	**96.63**	53.59	42.85

Table 4: **Rule refinement narrows background removal while preserving essential context.** We show the rules selected after the first iteration and at the end of training on GEO-Bench. The final rule explicitly retains definitions and background needed to understand the source's subject. Visibility changes (points, relative to $\text{AutoGEO}_{\text{API}}$) are measured on the training minibatch.

Stage	Rule	Δ^U	Δ^D
Iteration 1	Remove any general introduction or background for a broader field; cut all definitions, historical summaries, and fundamental principles not unique to the document's subject.	-7.1	-10.7
Final	Remove general introductions to a much broader field, but retain definitions, historical context, and fundamental principles essential for a non-expert to understand the document's subject.	-4.1	-2.3

$\text{AutoGEO}_{\text{API}}$, while Precision and Clarity decrease modestly. Compared with Original, OFF-GEO scores higher on all four metrics. Together with the visibility results, these findings suggest that OFF-GEO can reduce off-target source use while preserving answer quality on desired queries.

4.3 Analysis of OFF-GEO

We analyze how rule-learning iterations and query similarity relate to the effects of OFF-GEO. We report changes in both undesired and desired visibility relative to the GEO-optimized sources, using the same held-out sources and query pairs within each benchmark.

RQ4. How does the learned rule change over iterations? Table 4 shows how the GEO-Bench rule changes during training. After the first iteration, the selected rule calls for removing definitions, historical summaries, and fundamental principles whenever they are not unique to the source's subject. On the training minibatch where it is next evaluated, this rule reduces undesired visibility but also substantially reduces desired visibility. The revised rule, retained through the final iteration, removes general introductions to fields much broader than the source's topic. It preserves definitions, historical context, and fundamental principles when they are needed to understand that topic. The revision thus shifts the criterion from whether background is unique to the subject to whether it is necessary for understanding it. This example illustrates how the shared editing rule develops a more explicit distinction between removable background and essential explanatory context.

RQ5. Does higher query similarity weaken undesired visibility reduction? We examine whether more similar desired and undesired queries lead to smaller reductions in undesired visibility or larger losses in desired visibility. We measure query similarity using cosine similarity between `all-MiniLM-L6-v2` embeddings. Our pairing protocol excludes candidates above the predefined similarity threshold. In our query pool, candidates above this threshold were typically near-duplicates of the desired query, asking for essentially the same information. We select the most similar remaining query within the same split as the undesired query.

Fig. 3 relates query similarity to the visibility changes produced by the learned rule $c^\star$ on held-out GEO-Bench sources. We report Spearman's rank correlation coefficient ρ between query similarity and each visibility change, with p denoting the corresponding p-value for testing the null hypothesis of zero correlation. Across all three metrics, the correlations are close to zero and are not statistically significant. Higher similarity is therefore not consistently associated with smaller undesired visibility reductions or larger desired visibility losses. The has-room subset shows the same pattern.

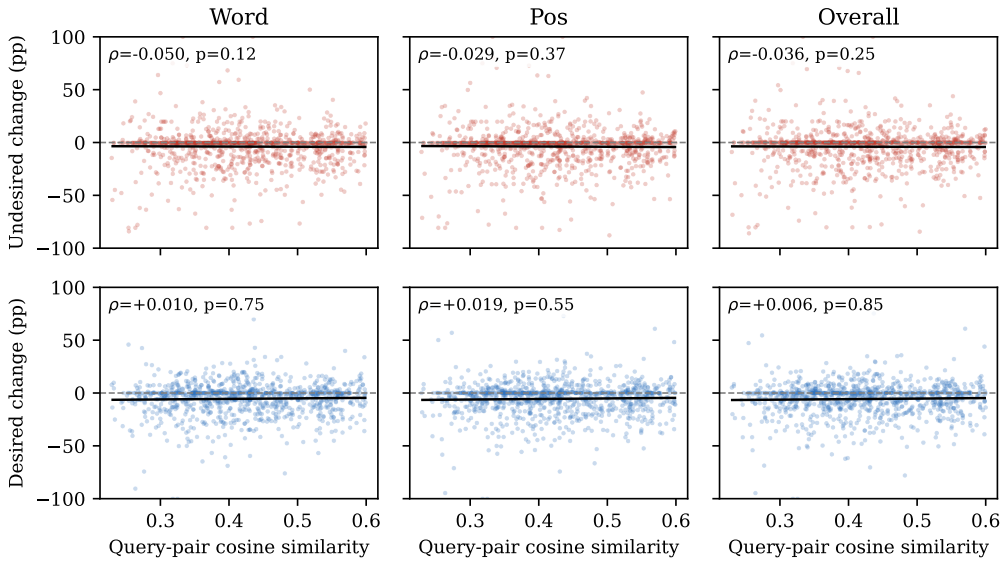


Figure 3: **Visibility changes show no systematic relationship with query similarity on GEO-Bench.** Each point represents a held-out source and its fixed desired–undesired query pair selected by our pairing protocol. The horizontal axis shows cosine similarity between `all-MiniLM-L6-v2` query embeddings. The vertical axis shows undesired visibility changes in the top row and desired visibility changes in the bottom row, relative to the GEO-optimized sources. Columns correspond to visibility metrics. Spearman correlations and p-values are shown in each panel.

Among its pairs with similarity of at least 0.5, undesired visibility decreases more on average than desired visibility, although desired visibility also declines.

These results show that the observed reductions extend to the more similar query pairs admitted by our pairing protocol. Within this setting, we find no evidence that increasing query similarity systematically weakens undesired visibility reduction or increases desired visibility loss.

5 Conclusion

We presented OFF-GEO, a framework for mitigating GEO-induced visibility gains on undesired queries while preserving desired visibility. Our approach optimizes a shared editing rule using generative-engine feedback, maximizing desired visibility subject to a minimum average reduction in undesired visibility. The learned rule guides edits to optimized sources without query inputs or generative-engine feedback. Experiments on E-commerce and GEO-Bench demonstrate reduced undesired visibility with limited loss in desired visibility, while answer-utility evaluation suggests that these edits need not compromise the quality of generated answers. These findings support optimizing source content according to which visibility gains a provider wishes to preserve and which it wishes to reduce.

AI Use Statement

We used generative AI tools to assist with implementation and debugging. We also used generative AI tools for editing portions of the manuscript, including improvements to organization, readability. All AI-assisted work was reviewed and verified by the authors, who take responsibility for the final content, claims, and results of this work.

Reproducibility Statement

We provide the experimental details needed to reproduce OFF-GEO in the appendix. Appendix A describes query construction and filtering, training and evaluation protocols, model configurations, optimization settings, and implementation details. Appendix B provides the complete prompts used for rule learning and document editing. We also report the fixed hyperparameters, candidate-selection procedure, and evaluation protocol used in our experiments.

Ethics Statement

OFF-GEO studies how content providers can reduce unintended off-target visibility while retaining visibility on desired queries. Selective control of source visibility may have dual-use implications, since similar techniques could be used to strategically influence when content is surfaced by generative engines. Our experiments focus on mitigating visibility gains introduced by GEO rather than promoting deceptive or misleading content. We encourage future work to consider safeguards against manipulative uses of content optimization.

References

Pranjal Aggarwal, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, Karthik Narasimhan, and Ameet Deshpande. Geo: Generative engine optimization. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 5–16, 2024.

Lakshya A Agrawal, Shangyin Tan, Dilara Soylu, Noah Ziems, Rishi Khare, Krista Opsahl-Ong, Arnav Singhvi, Herumb Shandilya, Michael J Ryan, Meng Jiang, Christopher Potts, Koushik Sen, Alexandros G. Dimakis, Ion Stoica, Dan Klein, Matei Zaharia, and Omar Khattab. Gepa: Reflective prompt evolution can outperform reinforcement learning, 2026. URL https://arxiv.org/abs/2507.19457.

Jean-Yves Audibert and Sébastien Bubeck. Best arm identification in multi-armed bandits. In *COLT-23th Conference on learning theory-2010*, pp. 13–p, 2010.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025. URL https://arxiv.org/abs/2507.06261.

Chrisantha Fernando, Dylan Sunil Banarse, Henryk Michalewski, Simon Osindero, and Tim Rocktäschel. Promptbreeder: Self-referential self-improvement via prompt evolution. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 13481–13544. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/fernando24a.html.

Donghao Li, Chengshuai Shi, Weijuan Ou, Cong Shen, and Jing Yang. Efficient multi-objective prompt optimization via pure-exploration bandits. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 45274–45299, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/4b0713b4f15946acaecb0676389715ed-Paper-Conference.pdf.

Reid Pryzant, Dan Iter, Jerry Li, Yin Lee, Chenguang Zhu, and Michael Zeng. Automatic prompt optimization with "gradient descent" and beam search. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7957–7968, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.494. URL https://aclanthology.org/2023.emnlp-main.494/.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of*

the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410/.

Xinyuan Wang, Chenxi Li, Zhen Wang, Fan Bai, Haotian Luo, Jiayou Zhang, Nebojsa Jojic, Eric P. Xing, and Zhiting Hu. Promptagent: Strategic planning with language models enables expert-level prompt optimization, 2023. URL https://arxiv.org/abs/2310.16427.

Beining Wu, Fuyou Mao, Jiong Lin, Cheng Yang, Jiaxuan Lu, Yifu Guo, Siyu Zhang, Yifan Wu, Ying Huang, and Fu Li. From experience to skill: Multi-agent generative engine optimization via reusable strategy learning. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 43305–43315, San Diego, California, United States, July 2026a. Association for Computational Linguistics. ISBN 979-8-89176-395-1. URL https://aclanthology.org/2026.findings-acl.2149/.

Yujiang Wu, Shanshan Zhong, Yubin Kim, and Chenyan Xiong. What generative search engines like and how to optimize web content cooperatively. In *International Conference on Learning Representations*, volume 2026, pp. 136858–136888, 2026b.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V Le, Denny Zhou, and Xinyun Chen. Large language models as optimizers. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 12028–12068, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/3339f19c5fcee3ad74502947a32be9e6-Paper-Conference.pdf.

Tianjun Zhang, Aman Madaan, Luyu Gao, Steven Zheng, Swaroop Mishra, Yiming Yang, Niket Tandon, and Uri Alon. In-context principle learning from mistakes. *arXiv preprint arXiv:2402.05403*, 2024.

Heyang Zhou, Jiajia Chen, Xiaolu Chen, Jie Bao, Zhen Chen, and Yong Liao. IF-GEO: Conflict-aware instruction fusion for multi-query generative engine optimization. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 27576–27590, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.1373. URL https://aclanthology.org/2026.findings-acl.1373/.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. Large language models are human-level prompt engineers, 2023. URL https://arxiv.org/abs/2211.01910.

A Additional Experimental Details

Query construction. For each source, we use its original query as the desired query q^D. We encode queries using `all-MiniLM-L6-v2` and compute cosine similarity between their embeddings. To construct the undesired query q^U, we exclude the desired query itself and candidates with similarity above 0.6, then select the most similar remaining query within the same split. This protocol selects related queries while limiting overlap with the desired information need. Query assignments remain fixed throughout rule learning and evaluation.

Rule learning. We learn a shared editing rule over three iterations, starting from an empty rule. At each iteration, we sample a minibatch of 64 training sources and collect failure cases for which the current rule does not reduce undesired visibility. An optimizer LLM critiques these cases using changes in both desired and undesired visibility and proposes up to four candidate rules. If no failures are found, we retain the current rule and proceed to the next iteration.

We compare the candidate rules and the incumbent using constrained Successive Rejects with a selection budget of $B = 200$. Each rule is applied independently to the same GEO-optimized sources. Edits are not accumulated across candidates or learning iterations. At each selection phase, all surviving rules are evaluated on the same additional training sources. If any rule has a mean undesired visibility change greater than $-\gamma$, we eliminate the rule with the largest mean change. Otherwise, we eliminate the rule with the lowest pooled desired visibility retention. The surviving rule replaces the incumbent only if it is new and satisfies the reduction constraint. To reduce sensitivity to candidate-document order during learning, each visibility score is averaged over three orderings.

B Full Prompts and Training Details

This appendix gives the complete, verbatim text of every prompt used to train and deploy OFF-GEO. All prompt placeholders ({ . . . } or `%s`) are kept as-is; they are filled in programmatically at call time.

B.1 Fixed base and initial criterion

Training starts from an empty criterion. The criterion is learned as an addendum to a fixed base, which is AutoGEO's official rule set for each benchmark, the same rules used to produce the $\text{AutoGEO}_{\text{API}}$ rewrites. The base appears only as context in the critique and candidate-generation prompts and is never edited. The edit-application prompt (§B.4) receives only the learned criterion. The fixed bases are:

```
GEO-Bench:
- Ensure all information is factually accurate and verifiable, citing credible
sources.
- Ensure information is current and up-to-date, especially for time-sensitive
topics.
- Ensure the document is self-contained and comprehensive, providing all
necessary context and sub-topic information.
- Explain the underlying mechanisms and principles (the 'why' and 'how'), not
just surface-level facts.
- Maintain a singular focus on the core topic, excluding tangential information,
promotional content, and document 'noise' (e.g., navigation, ads).
- Organize content with a clear, logical hierarchy, using elements like headings,
lists, and tables.
- Present a balanced and objective view on debatable topics, including multiple
significant perspectives.
- Provide specific, actionable guidance, such as step-by-step instructions, for
procedural topics.
- State the primary conclusion directly at the beginning of the document.
- Use clear and unambiguous language, defining technical terms, acronyms, and
jargon upon first use.
- Use specific, concrete details and examples instead of abstract
generalizations.
```

```
- Write concisely, eliminating verbose language, redundancy, and filler content.

E-commerce:
- Ensure all information is factually accurate, verifiable, and current for the
topic.
- Establish credibility by citing authoritative sources, providing evidence, or
demonstrating clear expertise.
- Justify recommendations and claims with clear reasoning, context, or
comparative analysis like pros and cons.
- Organize content with a clear, logical structure using elements like headings,
lists, and tables to facilitate scanning and parsing.
- Present information objectively, avoiding promotional bias and including
balanced perspectives where applicable.
- Provide actionable information, such as step-by-step instructions or clear
recommendations.
- Provide specific, verifiable details such as names, model numbers, technical
specifications, and quantifiable data.
- Structure content into modular, self-contained units, such as distinct
paragraphs or list items for each concept.
- Use clear, simple, and unambiguous language, defining any necessary technical
terms or jargon.
- Write concisely, eliminating verbose language, filler content, and unnecessary
repetition.
```

B.2 GRADIENT (CRITIQUE) PROMPT

At every iteration, this prompt diagnoses why the current criterion fails on the minibatch's failure cases, i.e., sources whose undesired visibility did not decrease, before any fix is proposed. Each failure case also shows its desired-visibility change, so the diagnosis can flag directions that would risk desired visibility.

```
You are improving an instruction that rewrites a document to {verb} its
{metric_name} in a generative-engine RAG system. Below is the CURRENT instruction
and a set of FAILURE examples: documents whose rewrite under this instruction did
NOT {verb} {metric_key} ({delta_cond}). Each example gives 'original' (the
document's true, un-rewritten source), 'rewrite' (the document after the current
instruction), and 'delta' (rewrite score minus the score of the exact text the
instruction was applied to -- which may already differ from 'original').
Diagnose,
in general and reusable terms, why the current instruction fails to {verb}
{metric_key} on these documents. Do not propose a fix yet, only diagnose. Return
<=250 words of plain-text critique, grounded only in the supplied examples.

The instruction is always applied to a document BEFORE any query is chosen or
known -- whichever model later scores {metric_key} never shows the rewriter the
query text, not even the document's own. Diagnose only from the document's own
text
and the instruction's wording; never attribute a failure to "the rewrite doesn't
address the new/specific query," since no query is ever available to act on.
{d_note}

CURRENT INSTRUCTION:
{current}

FAILURE EXAMPLES:
{examples_text}
```

{d_note} adds the desired-visibility change of each failure case:

```
Each example also gives 'd_delta': the change in the document's own citation
visibility (D, on its own true query -- distinct from the CROSS-query metric
above) under the current instruction on that document. Flag whether the direction
```

```
needed to fix these failures risks an unacceptable cost to D, so a later fix can
avoid it.
```

B.3 CANDIDATE-GENERATION PROMPT

Given the critique, this prompt proposes four candidate criteria at temperature 0.95. The model proposes only the criterion; the fixed base is shown as context and cannot be repeated, rewritten, or contradicted.

```
You are writing a CRITERION that identifies which part of a document's content
should be cut or reverted to {verb} its {metric_name} -- the criterion never adds
content, it only names what to remove. It is appended after a FIXED BASE
instruction (shown below, which you must not repeat, rewrite, remove, or
contradict) to form the full rewriting instruction; your job is only to design
the
CRITERION. Given the CURRENT CRITERION (may be empty on the first round) and a
CRITIQUE of why the combined instruction (base + criterion) fails on some
documents, write {num_candidates} different candidate criteria, each a
self-contained alternative (not the whole instruction, not a diff or patch) that
addresses the critique.{step_size_note} Each candidate must stay generic and
document-agnostic -- it will be applied to unseen documents with no case-specific
information, so it must not reference the failure examples' specific content. It
is always applied to a document BEFORE any query is chosen or known -- the model
applying it never sees any query text, not the document's own and not any other.
Do NOT write anything conditional on "the query" (e.g. "if the query is about X,
do Y" or "prioritize the new query's requirements") -- these cannot be followed
and must not appear. Every candidate must be actionable using only the document's
own text. Return ONLY a JSON array of {num_candidates} strings, each string one
candidate criterion by itself (do not repeat the FIXED BASE inside it).

FIXED BASE (do not repeat, rewrite, or contradict this):
{anchor}

CURRENT CRITERION:
{current_addendum}

CRITIQUE:
{gradient}
```

Once the current criterion is non-empty, {step_size_note} is appended to bias candidates toward incremental refinement:

```
Each candidate should be an INCREMENTAL refinement of the CURRENT CRITERION --
prefer small, targeted wording changes (e.g. narrowing scope, softening or
sharpening a single qualifier, adding one clarifying condition) over a sweeping
rewrite of its overall approach, unless the critique shows the current approach
is
fundamentally insufficient rather than merely imprecise.
```

B.4 EDIT-APPLICATION PROMPT AND THE ORTHOGONALITY GUARD

The learned criterion is not applied as a free-form rewriting instruction. Instead, it is used to propose individual edit operations on the AutoGEO$_{\text{API}}$ rewrite, via the following prompt (temperature 0):

```
Below are two versions of a document: ORIGINAL (before any AI rewriting) and
REWRITE (after). REWRITE has been found to leak into unrelated topics -- a
generative search engine cites it even for queries the document is not really
about.

Compare REWRITE against ORIGINAL to find content in REWRITE that was ADDED or
SUBSTANTIALLY EXPANDED relative to ORIGINAL, using this criterion for what counts
```

```
as leaky:
%s

For each piece of leaky content, output ONE edit operation with three fields:
- "search": an EXACT, VERBATIM quote copied character-for-character from REWRITE.
  Do not paraphrase or summarize it.
- "replace": if ORIGINAL has a corresponding, more minimal version of that
  content, quote it EXACTLY (word-for-word) here -- this reverts REWRITE's
  expansion back to what ORIGINAL actually said. If ORIGINAL has no corresponding
  content, use an empty string "" to delete it outright.
- "also_core_to_own_topic": true or false. A document's own core topic is
  whatever ORIGINAL and REWRITE are clearly, primarily about -- infer this from
  their content alone; you are never told what query will later be asked of this
  document, so do not guess at one, judge only from the document's own apparent
  subject. Set this true if the flagged content, despite being leaky toward other
  topics, ALSO does real work for the document's own core topic (e.g. a fact,
  spec, or detail central to what it's actually about) -- removing it would hurt
  the document on its own topic too, not just on unrelated ones. Set it false if
  the content serves no purpose for the document's own topic and exists only to
  connect to other topics, comparisons, or broader context.

Only include operations for content you are confident is leaky under the
criterion above; leave everything else in REWRITE untouched by simply not
mentioning it.

Return ONLY a JSON array of operations, each shaped {"search": "<exact quote from
REWRITE>", "replace": "<exact quote from ORIGINAL, or empty string to delete>",
"also_core_to_own_topic": true/false}. Return an empty array [] if nothing in
REWRITE qualifies as leaky.

ORIGINAL:
%s

REWRITE:
%s
```

Each proposed operation is applied only if it passes mechanical verification: the `search` span must occur verbatim in the current rewrite, and a non-empty `replace` must occur verbatim in the original source. Operations that the model flags `also_core_to_own_topic` are discarded, however confidently they were proposed as "leaky." After editing, markdown headings left with no content beneath them are removed. This per-operation guard complements criterion selection (§3), which, among criteria that meet the undesired-reduction constraint, prefers those that retain the most desired visibility.

B.5 AUTOGEO'S GENERATIVE-ENGINE JUDGE PROMPT

Desired and undesired visibility are both scored using AutoGEO's own generative-engine prompt, loaded unmodified from the official AutoGEO source rather than reimplemented:

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

Question: {query}
```

```
Search Results:
{source_text}
```