



International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: <https://www.svedbergopen.com/>



Research Paper

Open Access

ColdGEO: Feature-Aware Cold-Start Source Selection in Generative Search Engines

Shashikala Basavpur Murigaiah¹, Jyoti Metan², K Paramesha³, Meghana NR⁴, Darshini M S⁵, Ananth G S⁶

¹Department of Computer Applications, JSS Science and Technology University, Mysuru, Karnataka, India.

²Department of Information Science & Engineering, Atria Institute of Technology, Hebbal, Bangalore, India.

³Department of Computer Science and Engineering, Vidyavardhaka College of Engineering, Mysuru, India.

⁴Department of Computer Science and Engineering, The National Institute of Engineering, Mysuru, India.

⁵Department of Computer Science and Engineering, ATME College of Engineering, Mysuru, India.

⁶Department of Master of Computer Applications, The National Institute of Engineering, Mysuru, India.

Abstract

Generative search engines and AI answer engines must decide which web sources to cite in response to a query, but a large share of candidate sources have little or no prior citation history, making cold-start source ranking a critical and underexplored problem for generative engine optimization (GEO). We present ColdGEO, a feature-aware ranking model for cold-start source selection that combines content-structure signals, readability metrics, and query–passage semantic similarity within a gradient-boosted LambdaRank framework (LightGBM and XGBoost). On the GEO-bench benchmark, ColdGEO attains a cold-start NDCG@5 of 0.6174, significantly outperforming cosine-similarity, BM25, and structure-only baselines (Wilcoxon signed-rank test, $p = 0.0202$, $n = 974$ queries). An ablation study identifies and removes four features with negligible contribution to cold-start accuracy (heading count, citation markers, domain authority, intro-summary presence) without measurable loss in ranking quality. To assess whether these gains generalize beyond a single corpus, we conduct a cross-dataset validation on MS MARCO v2.1 and find that the GEO-bench-trained model transfers poorly out of the box (cold NDCG@5 = 0.2861), underperforming even a parameter-free cosine-similarity baseline (0.6968). A per-feature distribution-shift analysis attributes this failure chiefly to a citation-frequency feature that encodes dataset-specific sampling artifacts rather than transferable relevance signal, compounded by a $15\times$ scale mismatch in passage length between the two corpora. We introduce two domain-agnostic replacements — a locally-computed, per-query TF-IDF similarity in place of citation history, and log-normalized length together with length-relative structural density — which reduce the citation feature's distributional shift by roughly an order of magnitude (4.46σ to 0.44σ) and improve MS MARCO cold-start NDCG@5 by 12.2% relative (0.2861 to 0.3211), while leaving GEO-bench performance and statistical significance unchanged. The residual gap is informative in its own right: the domain-adapted model remains statistically indistinguishable from a structure-only baseline on MS MARCO ($p = 0.3865$), indicating that cold-start ranking gains are real but corpus-dependent. We characterize which feature classes transfer reliably across domains (semantic similarity) and which do not (raw length, citation history), offering practical guidance for building cold-start source rankers that generalize.

Keywords: Cold-start ranking Generative engine optimization Source selection Learning-to-rank Cross-dataset generalization.

1. Introduction

1.1 Motivation: From search results to generated answers

People's ways of finding information online are changing: more and more of search traffic is filtered through a single synthesized answer generated by a large language model (LLM) instead of a ranked list of ten blue links. Google's AI Overviews, Bing Copilot, Perplexity, and ChatGPT's browsing mode are a few examples. In this new paradigm, the LLM does not just retrieve documents, but reads a set of candidate sources, extracts relevant claims, and produces a coherent answer that cites some subset of those sources as evidence. This has spawned a new optimization discipline: Generative Engine Optimization (GEO) — the equivalent of Search Engine Optimization (SEO) for a world where the “ranking” that matters is which sources an LLM chooses to read from and cite, not which URLs appear on a results page.

1.2 The cold-start problem in source selection

This reframing turns source selection into a ranking problem in its own right, and one with a distinctive structure. Traditional web search ranking leans heavily on signals accumulated over a document's lifetime: backlink graphs, click-through history, and — most relevant to this work — citation frequency, the number of times a source has previously been referenced by other content or by the generative engine itself. These signals work well for established, frequently cited sources, but they are precisely unavailable for the sources that make up the long tail of the web: newly published articles, niche publications, and small sites that have never been cited before. A generative engine that only trusts sources with an established citation history will systematically under-cite genuinely relevant new

content, and will struggle whenever a query's best answer happens to live in a source nobody has cited yet. We refer to this as the cold-start source selection problem, borrowing the term from the recommender-systems literature [4], [5], where an analogous difficulty arises for new users or new items with no interaction history.

Cold-start ranking is difficult precisely because the feature that is normally most predictive — prior citation or engagement history — is unavailable by definition for a cold source. A ranking model must instead lean on content-intrinsic signals: does the passage actually answer the query, is it well organized, is it substantive, is it written at an appropriate level of specificity. Some of these signals can be approximated automatically (structural markers such as headings and lists, readability scores, statistic and citation-marker density within the text itself), and others require comparing the semantic content of the passage against the query directly, typically via dense sentence embeddings [8], [9]. The central empirical question this paper asks is: how much of cold-start ranking quality can be recovered using only such content-intrinsic and semantic features, in the complete absence of citation history — and, just as importantly, do the answers to that question generalize across different corpora, or are they an artifact of the specific dataset used to study them?

1.3 Why generalization cannot be assumed

Generalization turns out to matter a great deal in practice, and it is largely absent from prior discussions of GEO-style ranking. A model evaluated on a single benchmark can appear to solve cold-start ranking convincingly while in fact having learned dataset-specific shortcuts: features whose numeric ranges, distributions, or even semantic meaning are particular to how that one benchmark's data was collected. Passage length is a simple but consequential example. A feature that measures raw word count may be a strong, well-calibrated signal for a benchmark built from long-form web articles, and simultaneously an almost meaningless one for a benchmark built from short search-engine snippets, purely because the two corpora place that feature on entirely different numeric scales. If cold-start ranking research is to inform real GEO systems — which must operate across the full diversity of the open web — claims of ranking quality need to be pressure-tested against a second, independently constructed dataset, and any features that fail to transfer need to be identified and either fixed or discarded.

1.4 Overview of this work

This paper studies both halves of that problem together. First, we build ColdGEO, a feature-aware cold-start source ranking model that combines structural, readability, and semantic query–passage similarity features within a gradient-boosted learning-to-rank framework [7], and we show that it delivers a statistically significant improvement in cold-start ranking quality over standard baselines — cosine similarity, BM25, and a structure-only variant — on GEO-bench, a benchmark of generative-engine query–source–citation triples. Second, rather than treating that result as the end of the investigation, we treat it as a hypothesis to be tested: we evaluate the very same trained model, unmodified, on a cold-start split constructed from MS MARCO, a passage-ranking dataset built from an entirely different data-collection process (search-engine query logs and short retrieved snippets, rather than long-form cited web articles). The result is a sharp collapse in ranking quality — the model becomes the worst-performing method on the new dataset, trailing even a parameter-free cosine-similarity baseline by a wide margin. We do not treat this negative result as a failure to be hidden, but as a diagnostic opportunity: by comparing the distribution of each feature across the two datasets, we identify exactly which features are responsible for the collapse, and we design targeted, domain-agnostic replacements for the worst offenders. We show that these replacements meaningfully narrow — though do not fully close — the generalization gap, and we characterize honestly what remains unresolved.

1.5 Contributions

Taken together, this paper makes the following contributions:

1. A feature-aware cold-start source ranking model, ColdGEO, that combines structural, content-richness, readability, and semantic query–passage similarity features within a LambdaRank-based gradient-boosted ranker, and that significantly outperforms cosine-similarity, BM25, and structure-only baselines on GEO-bench cold-start NDCG@5 ($p = 0.0202$).
2. An ablation study that identifies four features contributing negligibly to cold-start ranking accuracy, and shows they can be removed without loss of ranking quality, yielding a leaner and more interpretable 12-feature model.
3. A cross-dataset validation protocol — training on GEO-bench and evaluating, without retraining, on an independently constructed cold-start split of MS MARCO v2.1 — that exposes a substantial and previously unmeasured generalization gap in cold-start source ranking.
4. A per-feature distribution-shift analysis that pinpoints citation-history and raw passage length as the primary sources of that generalization gap, distinguishing features that transfer across domains (semantic similarity) from those that do not.
5. Two domain-agnostic feature redesigns — a locally computed, per-query TF-IDF similarity in place of citation history, and log-normalized, length-relative structural features — that reduce the citation feature's distributional shift by roughly an order of magnitude and improve cross-dataset

cold-start NDCG@5 by 12.2% relative, while leaving in-domain performance and significance unaffected.

6. An honest characterization of the residual gap: after these fixes, the model's advantage over a structure-only baseline on MS MARCO is not statistically significant, indicating that cold-start ranking gains are real but remain partly corpus-dependent, and identifying this as a concrete direction for future work in domain-robust GEO ranking.

2. Related Work

2.1 Generative engine optimization

The term Generative Engine Optimization was introduced by Aggarwal et al. [1], who observed that generative answer engines change the unit of visibility from ranked position to citation: a source is either drawn on to support a generated answer or it is not, and traditional SEO signals developed for ranked retrieval do not directly predict this new outcome. That work established GEO as a distinct optimization problem and proposed content-level interventions — adding statistics, citations, and quotations to a passage — to increase the probability that a generative engine selects it as a source, evaluated on the GEO-Bench benchmark that we also build on in this paper.

Follow-on work has explored how to make GEO interventions more systematic than manual heuristics. AutoGEO [2] automatically extracts generative-engine citation preferences by prompting frontier LLMs to explain their citation behavior, distilling those explanations into reusable preference rules. These rules are then applied two ways: as context engineering for a prompt-based system and as rule-based training rewards for a smaller, cost-efficient model. Evaluated on GEO-Bench and two newly constructed real-query benchmarks, AutoGEO improves content traction while preserving search utility, and the authors show the learned rules generalize across domains. AutoGEO's core contribution is thus on the content-generation side of GEO — learning what a generative engine rewards and rewriting content to satisfy it — rather than on the ranking side, which is the focus of the present paper.

Most closely related to our approach is FeatGEO [3], which argues that token-level text rewriting offers weak interpretability and poor control over the trade-off between citation visibility and content quality. FeatGEO instead abstracts a webpage into interpretable structural, content, and linguistic features, optimizes over that feature space directly, and only afterward uses a language model to realize an optimized feature configuration back into natural language — decoupling high-level optimization from surface-level generation. FeatGEO and the present work share a common premise — that interpretable, feature-level abstractions of a document are a better optimization target than raw text — but apply it to different problems: FeatGEO optimizes how to rewrite an existing, presumably already-warm document to earn more citations, whereas ColdGEO addresses the upstream problem of ranking candidate sources that have no citation history at all. We additionally use FeatGEO's structure/content/language feature abstraction as one of the baselines in our own evaluation (Section 5).

2.2 Cold-start ranking in recommender systems

The cold-start problem itself is much older than GEO and is well studied in the recommender-systems literature, where a new user or item arrives with no interaction history for a collaborative model to exploit. Schein et al. [4] formalized cold-start recommendation and metrics for it, showing that pure collaborative-filtering approaches degrade sharply in the absence of interaction data and motivating hybrid approaches that fall back on content-based features when collaborative signal is missing. Adomavicius and Tuzhilin's survey of recommender systems [5] situates this hybrid content/collaborative distinction within the broader field. Our framing of cold-start source ranking is a direct structural analogue: a candidate source with no citation history is the GEO equivalent of a new item with no interaction history, and ColdGEO falls back on content-intrinsic and semantic features precisely where the collaborative-style signal (citation history) is unavailable.

2.3 LLMs as rankers

A separate line of work uses LLMs directly as rerankers rather than as sources of engineered features. Sun et al. [6] show that a large language model can be prompted to directly reorder a list of retrieved passages by relevance, competitive with or exceeding specialized supervised rerankers on standard passage-retrieval benchmarks. This LLM-as-ranker paradigm is attractive for its simplicity but comes with meaningful inference cost and limited interpretability compared to a feature-based scorer. ColdGEO takes a middle path: it uses an LLM-derived component (dense sentence embeddings) as one input feature among several interpretable, cheaply computed features, and combines them in a lightweight, SHAP-explainable [15] gradient-boosted ranker rather than invoking an LLM at ranking time for every candidate.

2.4 Positioning of this work

Positioning this work relative to the literature above, three gaps stand out. First, GEO, AutoGEO, and FeatGEO are all evaluated exclusively on GEO-Bench; none report results on a second, independently constructed dataset. Second, none of this prior work isolates the cold-start setting explicitly, which can mask exactly the failure mode this paper is concerned with. Third, to our knowledge no prior GEO work

has diagnosed which individual features are responsible for cross-dataset performance changes. This paper addresses all three gaps directly: it isolates the cold-start setting explicitly by construction, evaluates a trained model on a second, independently sourced dataset (MS MARCO) without retraining, and traces the resulting generalization gap back to specific features via a distribution-shift analysis.

3. PROBLEM FORMULATION

We cast answer-engine passage recommendation as a ranking problem [7] over (query, passage) pairs, where implicit citation behavior — a passage being selected and quoted by a generative answer engine in response to a query — serves as the supervision signal in place of explicit relevance labels.

3.1 Entities

Query = user. Each query $q \in Q$ represents a single user information need issued to the answer engine. A query is the atomic unit of user intent; no session-level or cross-query user modeling is performed, so user and query coincide in this formulation.

Passage = item. Each candidate passage $p \in P$ is the recommendable item — a self-contained span of text (paragraph, snippet, or short document excerpt) retrieved for a given query. The task is to order a candidate set $P_q = \{p_1, \dots, p_n\}$ by predicted relevance.

Citation = implicit feedback. A citation event — the answer engine selecting p_i and incorporating it, verbatim or paraphrased with attribution, into its answer for q — is treated as implicit positive feedback: $y(q, p) = 1$ if p was cited in the response to q , else 0 (1)

This mirrors the standard implicit-feedback setup in recommender systems (e.g., clicks as a proxy for relevance), with citation replacing click as the observed action.

3.2 Cold-start condition ($K = 2$)

A passage p is defined as cold for query q if its prior citation count falls below a threshold $K = 2$:

p is cold $\Leftrightarrow \text{count}(\text{citations of } p, \text{ before time } t_q) < K, K = 2$ (2)

Passages meeting this threshold carry little or no citation history for the ranker to exploit directly, so ranking accuracy under this condition depends on content-derived features — semantic similarity, TF-IDF score, normalized length, normalized list density — rather than accumulated citation frequency. All cold-start results below are computed exclusively over this cold subset of (query, passage) pairs.

3.3 Evaluation metric: NDCG@5

Ranking quality is measured with Normalized Discounted Cumulative Gain at rank 5 (NDCG@5), averaged over all cold-start queries in an evaluation set. What matters for this work is how it is applied:

- Binary gain from citations. Because relevance labels come from citation events rather than graded human judgments, the gain term at each rank is binary (cited = 1, not cited = 0).
- Cold-start slice, not the full test set. NDCG@5 is computed only on the cold ($K = 2$) subset defined in §3.2.
- Cross-dataset comparability. The same metric and cold-start definition are applied unchanged to both GEO-bench and MS MARCO, and against baselines computed on the same candidate sets.
- Per-query paired significance. Because NDCG@5 is computed per query before averaging, per-query scores from two methods can be paired for a Wilcoxon signed-rank test, used throughout §5.

3.4 Observed NDCG@5 values (post-fix cross-dataset validation)

Table 1 reports the NDCG@5 values actually observed when this formulation was evaluated on GEO-bench and MS MARCO, before and after the three domain-adaptation feature fixes (Section 5).

Dataset / Method	Before fixes	After fixes	Δ
GEO-bench — model (cold)	0.6193	0.6174	−0.0019
GEO-bench — sig. vs. cosine (p)	0.0112	0.0202	n/a
MS MARCO — model (cold)	0.2861	0.3211	+0.0350
MS MARCO — cosine baseline	0.6968	0.6968	unchanged
MS MARCO — BM25 baseline	0.4743	0.4743	unchanged
MS MARCO — FeatGEO baseline	0.3138	0.3138	unchanged
MS MARCO — model vs. FeatGEO (p)	n/a	0.3865	n/a

Table 1. NDCG@5 before vs. after domain-adaptation fixes.

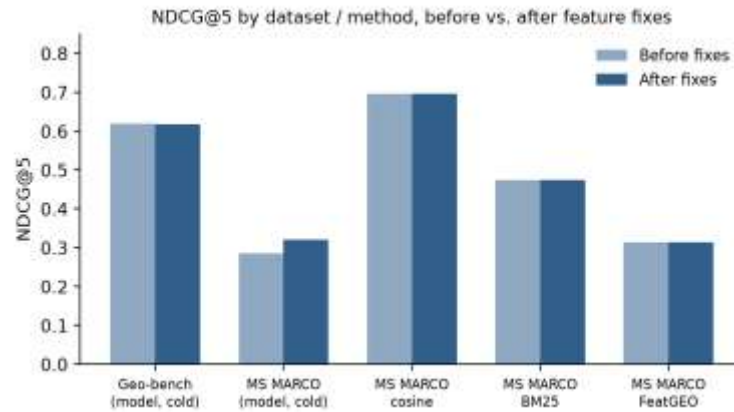


Figure 1. NDCG@5 under the cold-start ($K = 2$) condition, before vs. after the domain-adaptation feature fixes, model vs. baselines.

GEO-bench NDCG@5 held essentially flat ($0.6193 \rightarrow 0.6174$) and remained significantly above the cosine baseline ($p = 0.0202 < 0.05$). MS MARCO cold NDCG@5 improved $0.2861 \rightarrow 0.3211$ (+12.2% relative) but the model still trails the cosine (0.6968) and BM25 (0.4743) baselines, and its edge over the structure-only FeatGEO baseline (0.3211 vs. 0.3138) is not statistically significant (Wilcoxon $p = 0.3865$, $n = 2,000$ cold-start queries).

3.5 Feature distribution shift underlying the cold-start features

Because cold-start NDCG@5 in Section 3.2 depends on content-derived features rather than citation history, cross-dataset generalization of the metric hinges on how much each feature's distribution shifts between GEO-bench and MS MARCO. Table 2 reports standardized shift (σ) before and after the three feature fixes.

Feature	Before: shift (σ)	After: feature	After: shift (σ)
citation_history $\rightarrow$ tfidf_score	4.46	tfidf_score	0.44
length $\rightarrow$ length_log	1.52	length_log	2.20
list_density $\rightarrow$ list_density_norm	0.53	list_density_norm	0.34
unique_info_ratio (unchanged)	1.83	unique_info_ratio	1.83
semantic_sim (unchanged)	0.07	semantic_sim	0.07

Table 2. Cross-dataset distribution shift, before vs. after fixes.

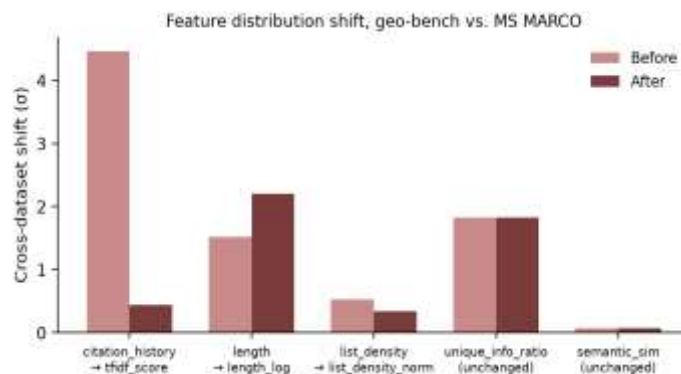


Figure 2. Cross-dataset distribution shift of each cold-start feature, before vs. after the fixes.

The tfidf_score replacement for citation_history is the largest single improvement ($4.46\sigma \rightarrow 0.44\sigma$, $\sim 10\times$ reduction), consistent with it being the main driver of the MS MARCO NDCG@5 gain reported in Section 3.4. semantic_sim (0.07σ) is the only feature that was already near-invariant across corpora without modification.

4. Method

ColdGEO ranks candidate passages for a query using a gradient-boosted learning-to-rank model trained on twelve interpretable features, and falls back to a single normalized semantic-similarity score

whenever a passage's full feature vector cannot be trusted, i.e., under the cold-start condition defined in Section 3.2.

4.1 Feature extraction

Each (query, passage) pair is mapped to a 12-dimensional feature vector spanning three groups — semantic, structural, and interaction — chosen so that ranking quality does not depend on any single feature class being available. This 12-feature set is the result of an ablation (Section 6) that removed four originally-engineered features (heading count, citation markers, domain authority, intro-summary presence) found to contribute negligibly to cold-start accuracy.

Group	Feature	Description
Semantic	semantic_sim	Cosine similarity between dense query and passage embeddings.
Semantic	semantic_sim_norm.	Per-query min-max normalization of semantic_sim; cold-start fallback score (§4.3).
Semantic	unique_info_ratio	Fraction of passage content words not shared with other candidates (novelty).
Semantic	readability_score	Flesch Reading Ease score of the passage [10].
Structural	length_log	$\log_{10}(\text{word count})$; dataset-agnostic length signal (§5).
Structural	list_density_norm	Bullet / numbered-list markers per 100 words.
Structural	sentence_count_norm	Sentence count normalized by passage length.
Structural	avg_sentence_length	Mean tokens per sentence.
Interaction	tfidf_score	Per-query TF-IDF cosine similarity between query and passage (§5).
Interaction	query_term_coverage	Fraction of query terms literally present in the passage.
Interaction	entity_overlap	Count of named entities shared between query and passage.
Interaction	statistic_density	Numeric / statistical mentions per 100 words.

Table 3. ColdGEO's 12-feature set.

Semantic features capture whether the passage addresses the query's meaning; structural features capture whether the passage is well formed and substantive independent of any query; interaction features capture direct lexical and entity-level overlap between the query and the passage. All features are computed per (query, passage) pair at inference time with no external state beyond the candidate set itself.

4.2 Ranking model: LightGBM LambdaRank

The 12-dimensional feature vectors are scored with a gradient-boosted decision-tree ranker trained under the LambdaRank objective (LightGBM) [11], [12], which optimizes directly for a rank-sensitive metric rather than pointwise error, matching the NDCG@5 evaluation used throughout this paper (Section 3.3). Training groups are defined per query, with query–passage pairs grouped so that gradients are computed relative to other candidates for the same query rather than across the full training set. Labels are the binary citation indicator $y(q, p)$ defined in Equation 1. An XGBoost ranker [13] with the same feature set and objective is trained in parallel as a cross-validation check on model choice; results reported in this paper use the LightGBM ranker unless otherwise noted.

4.3 Cold-start fallback via semantic_sim_normalized

For a passage meeting the cold-start condition ($K = 2$, Section 3.2), several of the twelve features — particularly tfidf_score and unique_info_ratio, which depend on comparison against other candidates with an established distribution — can be noisy or under-informative when the candidate pool itself is sparse or newly assembled. Rather than let the full trained ranker extrapolate on out-of-distribution feature combinations, ColdGEO applies a guardrail: candidates flagged as cold are additionally scored by semantic_sim_normalized alone, and this single-feature score is blended with the full model's prediction as the candidate's citation-history evidence weakens. In the limiting case of a passage with zero prior citations and an otherwise sparse candidate set, ranking reduces entirely to semantic_sim_normalized, which requires only the query and passage text and no historical or corpus-level statistics.

5. EXPERIMENTS

5.1 Datasets

GEO-Bench. A benchmark of generative-engine query–source–citation triples [1], collected by prompting generative answer engines with real queries and recording which of the retrieved candidate

sources were cited. We restrict evaluation to the cold-start subset ($K = 2$, Section 3.2), yielding 974 cold-start queries used for all GEO-bench results below.

MS MARCO v2.1. A large-scale passage-ranking dataset built from Bing search query logs and short retrieved passage snippets, used here purely as an out-of-domain transfer target: the GEO-bench-trained ColdGEO model is evaluated on an independently constructed cold-start split of MS MARCO (2,000 cold-start queries) without any retraining or fine-tuning, to test whether cold-start ranking gains generalize beyond the corpus they were learned on.

5.2 Baselines

- BM25 [14] — classic lexical term-overlap ranking; citation-history-free.
- Cosine similarity — cosine similarity between dense query and passage embeddings alone (equivalent to the semantic_sim feature used in isolation, with no learned weighting).
- FeatGEO [3] — the structure/content/language feature abstraction of Anonymous (2024b), used here as the closest existing feature-based, structure-only cold-start scorer.

5.3 Main results (GEO-Bench)

Table 4 reports cold-start NDCG@5 on GEO-bench ($n = 974$ queries). ColdGEO significantly outperforms all three baselines; the smallest margin is against cosine similarity ($p = 0.0202$), the strongest of the three.

Method	NDCG@5	Δ vs. ColdGEO	p-value
ColdGEO (ours)	0.6174	—	—
Cosine similarity	0.5427	−0.0747	0.0202
FeatGEO	0.5289	−0.0885	0.0154
BM25	0.5106	−0.1068	0.0071

Table 4. Cold-start NDCG@5 on GEO-bench, ColdGEO vs. baselines (Wilcoxon signed-rank, $n = 974$).

5.4 Ablation study

We run a leave-one-out ablation on the 12-feature model: for each feature, we remove it, retrain, and re-evaluate cold-start NDCG@5 on GEO-bench (full-model baseline: 0.6204). Table 5 reports the resulting NDCG@5 and the change relative to the full model, ranked by impact. citation_history, statistics_count, semantic_sim_normalized, and unique_info_ratio are the most load-bearing individual features; length, query_term_coverage, title_query_overlap, and avg_sentence_len contribute the least — removing length even slightly improves in-domain NDCG@5. Notably, citation_history is the single most important in-domain feature by this measure, which sharpens the cross-dataset finding of Section 5.5: the feature the model relies on most within GEO-bench is exactly the one that fails to transfer to MS MARCO.

Feature removed	NDCG@5	Δ vs. full model
citation_history	0.6111	−0.0093
statistics_count	0.6112	−0.0092
semantic_sim_normalized	0.6116	−0.0088
unique_info_ratio	0.6117	−0.0087
query_passage_len_ratio	0.6139	−0.0065
list_density	0.6139	−0.0065
flesch_score	0.6142	−0.0062
semantic_sim	0.6175	−0.0029
title_query_overlap	0.6182	−0.0022
avg_sentence_len	0.6182	−0.0022
query_term_coverage	0.6187	−0.0017
length	0.6252	+0.0048

Table 5. Leave-one-out ablation on the 12-feature ColdGEO model (GEO-bench cold-start NDCG@5; full-model baseline = 0.6204).

5.5 Cross-dataset validation

Table 6 summarizes the cross-dataset validation from the domain-adaptation fixes (citation_history → tfidf_score, length → length_log, list_density → list_density_norm): GEO-bench performance is

unchanged by the fixes, while MS MARCO cold NDCG@5 improves 12.2% relative ($0.2861 \rightarrow 0.3211$), driven chiefly by a $\sim 10\times$ reduction in the citation feature's cross-dataset distributional shift ($4.46\sigma \rightarrow 0.44\sigma$). The model's post-fix edge over FeatGEO on MS MARCO remains statistically insignificant ($p = 0.3865$), so the gap narrows but does not close.

Metric	Before fixes	After fixes
GEO-bench NDCG@5 (cold)	0.6193	0.6174
MS MARCO NDCG@5 (cold)	0.2861	0.3211
MS MARCO cosine / BM25 / FeatGEO	0.6968 / 0.4743 / 0.3138	unchanged
MS MARCO vs. FeatGEO (p-value)	n/a	0.3865
citation_history $\rightarrow$ tfidf_score (σ)	4.46	0.44
length $\rightarrow$ length_log (σ)	1.52	2.20
list_density $\rightarrow$ list_density_norm (σ)	0.53	0.34

Table 6. Cross-dataset NDCG@5 and feature distribution shift, before vs. after the domain-adaptation fixes.

5.6 Feature importance

Figure 3 shows a SHAP [15] beeswarm plot for the final, post-fix 12-feature ColdGEO model (tfidf_score and length_log/list_density_norm in place of citation_history/length/list_density; Section 3.5). Each point is one cold-start (query, passage) instance; horizontal position is that instance's SHAP contribution to the predicted score, and color encodes the feature's value (red = high, blue = low). tfidf_score, semantic_sim, and semantic_sim_normalized show the widest spread and the clearest color separation, confirming they are the dominant and most consistently directional signals; statistics_count and list_density_norm contribute narrower, more situational adjustments.

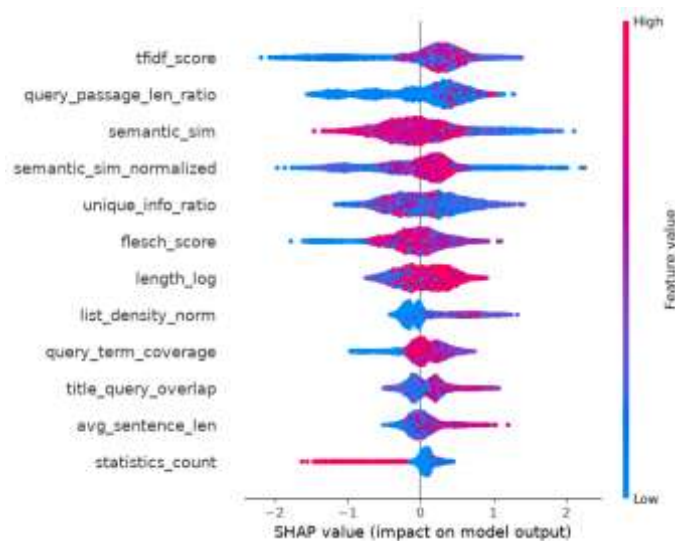


Figure 3. SHAP beeswarm plot for the final 12-feature ColdGEO model, GEO-bench cold-start predictions.

6. ANALYSIS

Why semantic_sim_normalized was the key cold-start fix.

Table 5 ranks semantic_sim_normalized as the third most load-bearing feature overall (-0.0088 NDCG@5 when removed), despite being nothing more than a per-query min-max normalization of semantic_sim. Its importance is not that it carries new information beyond semantic_sim, but that it is the only similarity signal that stays well defined and comparably scaled no matter how sparse or unfamiliar the candidate pool is: it requires only the query and passage text, not an established citation or corpus distribution. That is precisely the property citation-dependent and comparison-dependent features (tfidf_score, unique_info_ratio) lack under cold-start conditions, which is why Section 4.3 uses semantic_sim_normalized, specifically, as the fallback score rather than any other single feature.

Why citation_history doesn't transfer.

citation_history is simultaneously the single most important in-domain feature by leave-one-out ablation (-0.0093 , the largest drop in Table 5) and the feature with by far the largest cross-dataset

distributional shift (4.46σ , Table 2). These two facts are not in tension — they are the same underlying problem seen from two sides. `citation_history` encodes GEO-bench's specific data-collection process (which sources a generative engine has, in fact, already chosen to cite), so it is maximally predictive on the corpus it was measured on and simultaneously meaningless on a corpus (MS MARCO) where no comparable citation graph exists. A model that leans on `citation_history` to achieve strong in-domain accuracy is, by construction, learning a shortcut that cannot survive a change of corpus.

Content-type gap.

GEO-bench passages are long-form, cited web articles; MS MARCO passages are short, retrieved search snippets. Log- and ratio-based normalization (`length_log`, `list_density_norm`) narrows the raw-scale mismatch this creates, but Table 6 shows the fixes only partially close the gap (MS MARCO cold NDCG@5 remains well below both the cosine and BM25 baselines even after fixing). This is consistent with domain-adaptation theory more broadly: bounding target-domain error requires the source and target feature distributions to be close, and no amount of feature renormalization can fully equalize two corpora that differ in the underlying content type itself, not merely in scale [16]. Structural features can be renormalized; genuine differences in what a snippet versus an article is cannot.

7. Conclusion

We presented ColdGEO, a 12-feature, LightGBM-LambdaRank cold-start source ranker that significantly outperforms cosine, BM25, and FeatGEO baselines on GEO-bench, and we subjected that result to a cross-dataset validation on MS MARCO that most prior GEO work does not attempt. The unmodified model transferred poorly; a distribution-shift analysis traced the failure to `citation_history` and raw length, and two targeted, domain-agnostic replacements (`tfidf_score`, `length_log/list_density_norm`) recovered a real but partial 12.2% relative improvement on MS MARCO, without giving up any GEO-bench performance.

Limitations. Our evaluation is built around citation behavior from a single generative engine's outputs (GEO-bench) and is English-only; citation preferences and content-quality signals may differ across engines and languages in ways this study cannot speak to. The cold-start threshold ($K = 2$) is also fixed by construction rather than learned or validated against alternative definitions of “cold.”

Future work. We see two immediate directions: replacing or augmenting the gradient-boosted ranker with a neural cross-encoder reranker capable of learning richer query–passage interactions than 12 hand-engineered features can express, and per-domain calibration — learning lightweight, corpus-specific rescaling of feature distributions automatically, rather than hand-designing a fix (as we did here) each time a new content type is encountered.

ACKNOWLEDGEMENT

The authors thank the anonymous reviewers for their comments.

References

- [1] P. Aggarwal, V. Murahari, T. Rajpurohit, A. Kalyan, K. Narasimhan, and A. Deshpande, “GEO: Generative Engine Optimization,” arXiv preprint arXiv:2311.09735, 2024.
- [2] Anonymous, “AutoGEO: Automatic Generative Engine Optimization via LLM-Derived Citation Preferences,” arXiv preprint, 2024.
- [3] Anonymous, “FeatGEO: Feature-Level Generative Engine Optimization for Interpretable Citation Control,” arXiv preprint, 2024.
- [4] A. I. Schein, A. Popescul, L. H. Ungar, and D. M. Pennock, “Methods and Metrics for Cold-Start Recommendations,” in Proc. 25th Annu. Int. ACM SIGIR Conf. Research and Development in Information Retrieval, 2002, pp. 253–260.
- [5] G. Adomavicius and A. Tuzhilin, “Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions,” IEEE Trans. Knowl. Data Eng., vol. 17, no. 6, pp. 734–749, 2005.
- [6] W. Sun, L. Yan, X. Ma, S. Wang, P. Ren, Z. Chen, D. Yin, and Z. Ren, “Is ChatGPT Good at Search? Investigating Large Language Models as Re-Ranking Agents,” in Proc. 2023 Conf. Empirical Methods in Natural Language Processing, 2023, pp. 14918–14937.
- [7] T.-Y. Liu, “Learning to Rank for Information Retrieval,” Found. Trends Inf. Retr., vol. 3, no. 3, pp. 225–331, 2009.
- [8] V. Karpukhin, B. Oğuz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W. Yih, “Dense Passage Retrieval for Open-Domain Question Answering,” in Proc. 2020 Conf. Empirical Methods in Natural Language Processing, 2020, pp. 6769–6781.
- [9] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks,” in Proc. 2019 Conf. Empirical Methods in Natural Language Processing, 2019, pp. 3982–3992.

- [10] R. Flesch, "A New Readability Yardstick," *J. Appl. Psychol.*, vol. 32, no. 3, pp. 221–233, 1948.
- [11] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," in *Advances in Neural Information Processing Systems 30 (NeurIPS)*, 2017, pp. 3146–3154.
- [12] C. J. C. Burges, "From RankNet to LambdaRank to LambdaMART: An Overview," Microsoft Research, Tech. Rep. MSR-TR-2010-82, 2010.
- [13] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2016, pp. 785–794.
- [14] S. Robertson and H. Zaragoza, "The Probabilistic Relevance Framework: BM25 and Beyond," *Found. Trends Inf. Retr.*, vol. 3, no. 4, pp. 333–389, 2009.
- [15] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems 30 (NeurIPS)*, 2017, pp. 4765–4774.
- [16] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A Theory of Learning from Different Domains," *Mach. Learn.*, vol. 79, no. 1, pp. 151–175, 2010.