Attribution-Guided Query Rewriting in Neural Retrieval

Anonymous ACL submission

Abstract

Neural retrievers achieve strong performance but often struggle with vague, ambiguous, or underspecified queries, causing relevant documents to be ranked poorly. Query rewriting has emerged as a promising solution by reformulating queries into more effective retrieval inputs, with recent LLM-based methods showing strong potential. However, most rewriting approaches operate independently of the retriever and ignore how individual query terms influence ranking decisions. In parallel, Explainable AI techniques can identify query tokens that contribute positively or negatively to retrieval performance, but they are largely limited to post-hoc analysis rather than retrieval optimization. We propose an attribution-guided query rewriting framework that bridges these directions. Given an input query, we first compute token-level relevance signals from the retriever and inject them into a structured prompt to guide the rewriting model. This enables the refinement of weak or misleading terms while preserving the original information need, yielding a retriever-aware rewriting process grounded in the retriever's own behavior. Experiments on BEIR collections demonstrate consistent improvements in retrieval effectiveness, particularly for ambiguous or implicit queries, demonstrating how explainable AI can play a functional role in neural IR. The source code of the proposed approach is publicly available: https://anonymous.4open.science/r/Attribution-Guided-Query-Rewriting-for-Neural-IR[1].

1 Introduction

Neural retrieval models (dense retrievers, cross-encoders, and hybrid architectures) have become a core component of modern information retrieval (IR) systems, often outperforming lexical methods by capturing semantic similarity beyond exact term matching (Thakur et al., 2024; Zhao et al., 2024). Despite these advances, their performance can degrade significantly when queries are underspecified, ambiguous, or only implicitly express the user's information need (Hagen et al., 2024; Li et al., 2025). In such cases, retrieval models may overemphasize uninformative tokens, miss critical constraints, or produce rankings that do not reflect the intended semantics, even when relevant documents exist (Sidiropoulos and Kanoulas, 2024).

Query reformulation has long been studied as a remedy to address these failures, ranging from classical relevance feedback and query expansion methods (Robertson, 1991; Harman, 1992; Rieh and Xie, 2006; Carpineto and Romano, 2012; Esposito et al., 2020) to more recent neural and LLM-based approaches (Jagerman et al., 2023; Wang et al., 2023; Xia et al., 2025). In particular, LLMs can generate fluent paraphrases and expansions that improve query expressiveness. However, most existing rewriting approaches are retriever-agnostic: they optimize for linguistic plausibility or semantic similarity without accounting for how the downstream retriever actually uses query terms. This can lead to unstable gains, and in some cases even hurt effectiveness by introducing irrelevant or misleading content (Khattab and Zaharia, 2020; Kotte, 2026).

In parallel, explainable information retrieval methods have developed tools to analyze model behavior at the token level, including gradient- and perturbation-based attribution techniques that identify which query terms most influence ranking decisions (Möller and Padó, 2025; Ribeiro et al., 2016; Sundararajan et al., 2017). However, these methods are primarily used for post-hoc analysis and are not typically integrated into the retrieval or query refinement process.

In this work, we bridge query rewriting and explainable IR by introducing an attribution-guided

[1]The code was developed by the first authors of the paper

rewriting framework. Given a query, we compute token-level importance scores from the retriever using gradient-based attribution with respect to top-ranked documents. These scores are then provided to an LLM as structured soft guidance, enabling it to preserve strongly influential tokens while refining or disambiguating weaker or potentially misleading ones, without altering the original intent. The resulting rewritten query is then re-issued to the same retriever in a single closed-loop step. Importantly, our approach does not modify the retrieval model and requires no additional training data or relevance judgments, making it easy to apply across retrieval architectures.

We evaluate our approach on standard open-domain and domain-specific IR collections from BEIR (Thakur et al., 2021), using two representative neural retrievers: SPLADE (Formal et al., 2022) and TCT-ColBERT (Lin et al., 2021). Guided query rewriting is performed using the Mistral LLM[2]. Experimental results show that attribution-guided rewriting consistently improves retrieval effectiveness over both original queries and rewriting baselines that do not exploit retriever signals (w.r.t token attributions). Beyond effectiveness gains, the proposed method produces more controlled rewrites, characterized by higher retention of informative query terms, reduced semantic drift, and a better balance between expansion and removal rates. The improvements are particularly pronounced for queries with implicit intent and ambiguous or domain-specific terminology. Ablation studies further confirm that incorporating token-level attributions yields more effective rewrites than LLM-only rewriting without retriever feedback, highlighting the value of explanation signals as actionable guidance rather than post-hoc interpretation.

In summary, we contribute :

- A retriever-aware query rewriting method that uses token-level attributions to guide edits.
- A two-stage approach combining retriever-derived attribution signals with LLM-based, context-aware query reformulation.
- Empirical evidence on BEIR collections that attribution-guided rewriting outperforms strong baselines, including LLM-only rewriting and token-pruning variants.
- Evidences that explanations can play a functional role in improving neural retrieval.

[2] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3

2 Background and Related Work

Our work connects neural information retrieval, query rewriting, and explainable IR. The key gap we address is that query rewriting and explainability are not connected in a feedback loop: existing query rewriting methods typically generate reformulations without considering signals from the target retriever, while explainability techniques are predominantly used post-hoc to analyze model behavior. We address this disconnect by integrating retrieval explanations into the rewriting process, creating a feedback loop in which token-level attribution signals guide query refinement.

Neural IR includes (i) expensive interaction models (e.g., cross-encoders) that jointly encode query–document pairs and are mainly used for re-ranking, and (ii) efficient retrieval models that encode queries and documents separately into sparse or dense representations for large-scale search (e.g., dual encoders and hybrid variants) (Nogueira and Cho, 2020; Nogueira et al., 2020; Karpukhin et al.; Thakur et al., 2024). Late-interaction architectures, such as ColBERT (Khattab and Zaharia, 2020), retain token-level matching while remaining scalable, and hybrid sparse–neural methods such as SPLADE (Formal et al., 2022) combine lexical matching with learned expansion (Khattab and Zaharia, 2020; Lin et al., 2021; Formal et al., 2022). Our approach is compatible with these efficient retrievers and focuses on improving retrieval without changing the retriever itself.

Query reformulation has long been a key strategy for improving retrieval effectiveness. Early approaches relied on relevance feedback and pseudo-relevance feedback, where expansion terms are extracted from judged or top-ranked documents to better represent the underlying information need (Robertson, 1991; Harman, 1992). Subsequent neural methods learned to generate improved queries directly from query-document pairs, enabling context-aware paraphrasing and intent refinement (Nogueira and Cho, 2017; Zamani et al., 2020). Other works explored reinforcement learning to optimize reformulations using retrieval performance as a reward signal (Wang et al., 2020). Although promising, these methods often incur high computational cost and latency, since many candidate rewrites must be evaluated before convergence.

This limits their practicality in real-world retrieval systems where efficiency is critical (Kotte, 2026).

More recently, LLMs have renewed interest in query reformulation. Prompt-based methods such as query expansion (Chen et al., 2024), query2doc-style generation (Wang et al., 2023), and instruction-guided rewriting (Shu et al., 2024) can produce fluent rewrites in zero-shot or few-shot settings, often yielding gains for short or underspecified queries (Jia et al., 2024; Jagerman et al., 2023). These methods use generative capacity to bridge vocabulary gaps and often improve first stage retrieval recall (Zhu et al., 2025), but most are retriever-agnostic: they optimize for linguistic plausibility, semantic relatedness, or generic helpfulness rather than the behavior of the downstream ranker (Ma et al., 2023). As a result, improvements are often unstable across collections and retrievers, and generated terms may introduce noise or topic drift when the retriever is highly sensitive to query wording (Khattab and Zaharia, 2020; Kotte, 2026).

Explainable IR aims to identify which components of a query or document drive neural retrieval decisions. Prior work has proposed several families of methods, including attention-based analyses (Clark et al., 2019), gradient-based attribution techniques (Sundararajan et al., 2017), and perturbation-based approaches (Ribeiro et al., 2016) that estimate the impact of masking or modifying query terms (Boytsov and Kolter, 2021). These methods have proven useful for model inspection, debugging, and failure analysis, and can help uncover term importance, spurious correlations, and sensitivity to query formulation (Möller and Padó, 2025). However, they are typically used as post-hoc tools for interpretation rather than as actionable signals integrated into the retrieval optimization process.

In contrast, our work operationalizes explainability signals by transforming token-level attributions into guidance for query rewriting. Rather than merely explaining model behavior after ranking, we use these signals to proactively refine the query before re-retrieval.

3 Attribution-Guided Query Rewriting

Neural retrievers are often highly sensitive to the surface form of a query: certain tokens strongly support relevant matches, while others receive little weight or steer ranking toward less relevant regions of the corpus (Guo et al., 2020). As a result, retrieval failures frequently arise not from the absence of relevant documents, but from suboptimal query formulations (Luan et al., 2021). Our objective is therefore to improve retrieval through targeted query reformulation, by clarifying or disambiguating query components that confuse the retriever while preserving the original information need. Importantly, the retriever itself is kept fixed. In many real-world settings, deployed retrieval systems are proprietary, expensive to retrain, or lack task-specific supervision (Wu et al., 2022). Under these constraints, query-time reformulation becomes the most practical intervention. Here, we use explanation signals derived from the retriever itself as lightweight control signals for rewriting. Specifically, we use the contribution of each query token to the retriever's scoring function, then provide these token-level signals to an LLM to guide query rewriting. The resulting framework is retriever-aware, training-free, and broadly applicable across neural retrieval architectures.

3.1 Problem Formulation

Let $q = \{t_1, t_2, \ldots, t_n\}$ be an input query composed of n tokens, and let $\mathcal{D} = \{d_1, d_2, \ldots, d_m\}$ be an m document collection. A neural retriever R assigns a relevance score $s(q, d)$ to each query-document pair (q, d) and returns a ranked list of documents.

Given an initial query q, the objective is to generate a rewritten query $\hat{q}$ such that retrieval effectiveness under the same retriever R is improved, without modifying or retraining R. This formulation reflects realistic deployment scenarios where the retriever is fixed and query reformulation is the primary mechanism for improving performance.

The rewriting process must preserve the user's original intent while correcting query components that are weak, ambiguous, or misleading for the target retriever. Unlike generic rewriting methods (Nogueira and Cho, 2017; Jia et al., 2024; Zhu et al., 2025), our approach explicitly relies on token-level attribution signals derived from R to guide reformulation.

3.2 Attribution-Guided Prompting for Query Rewriting

A central component of our framework is the estimation of token-level attribution signals that reflect the retriever's current scoring behavior. For each query q, we first retrieve the top-k documents D_q^{top} using the base retriever R. These documents rep-

resent what the retriever currently considers most relevant for the query, making them a natural reference set for diagnosing token importance.

Token-level attributions can be estimated using several classes of methods, including gradient-based techniques (e.g., Integrated Gradients), perturbation-based methods which estimate token importance by masking, deleting, or replacing query terms and observing the resulting score variation, and model-agnostic explainers such as local surrogate models (Ribeiro et al., 2016; Sundararajan et al., 2017). While model-agnostic approaches are widely applicable, they typically incur significant computational overhead. In contrast, when the retriever architecture is accessible, attributions can be computed directly from its scoring function, avoiding additional inference cost and improving efficiency.

In this work, we exploit this property and compute attributions directly from the internal structure of the retrievers used in our experiments (SPLADE (Formal et al., 2022) and TCT-ColBERT (Lin et al., 2021)), without relying on external XAI models. The exact scoring formulations and model architectures are detailed in the original papers of SPLADE (Formal et al., 2022) and TCT-ColBERT (Lin et al., 2021), and further discussed in the Appendix for completeness.

For each query q, we use these scores as token attributions over the top-k retrieved documents D_q^{top} and aggregate them to obtain a robust query-level signal:

$$\Phi(t_i) = \frac{1}{k} \sum_{d \in D_q^{top}} \phi_{t_i}(q, d) \qquad (1)$$

where ϕ_{t_i} measures the contribution of token t_i to the relevance score of document d.

Averaging across documents stabilizes the signal: a token may be helpful for matching some retrieved documents but harmful for others, and we want a query-level estimate of its overall utility under R. Tokens with low or negative attribution scores are then treated as candidates for clarification, replacement, or disambiguation during rewriting, while highly useful tokens are encouraged to be preserved.

3.3 LLM-Based Query Reformulation

Attributions identify which parts of the query matter to the retriever, but it does not specify how to change them (e.g., expand an abbreviation, add specificity, or resolve an ambiguity). We therefore use an LLM as a flexible rewriting module, while constraining it with retriever-derived token signals.

All query tokens $t_1, \ldots, t_n$ with their attribution scores $\phi_{t_1}, \ldots, \phi_{t_n}$ are provided as input to the LLM, together with the original query q and an instruction to reformulate it for improving retrieval while preserving user's intent. The full prompt is shown listing 1. We do not remove low-attribution tokens before rewriting. Low attribution does not mean that a token is irrelevant to the user intent; it may encode context or a constraint that the retriever currently fails to exploit. Instead, we treat attribution as soft guidance: the LLM is encouraged to preserve or emphasize high-attribution tokens and to clarify low- or negative-attribution tokens (e.g., by making them more specific or less ambiguous), while avoiding the introduction of new concepts.

```
prompt = """You are given:
1) An original user query.
2) A list of query tokens with their attribution
     scores, where higher scores indicate a
    stronger positive contribution to retrieval
    effectiveness, and lower or negative scores
    indicate weak or misleading contributions.

Your task is to rewrite the query to improve
    retrieval effectiveness.

Guidelines:
- Preserve the original user intent.
- Do not remove important concepts.
- Tokens with high attribution scores should be
    preserved or emphasized.
- Tokens with low or negative attribution scores
     may be clarified, specified, or
    disambiguated.
- Avoid adding new concepts that are not implied
     by the original query.
- Produce a single rewritten query, concise and
    well-formed.

Original query: "{original_query}"
Token attributions:"{zip(tokens, attributions)}"
"""
```

Listing 1: Prompt for attribution-guided query rewriting.

The rewritten query (q') produced by the LLM is then fed into the same neural retriever (R) to obtain a new ranked list of documents. We adopt a one-shot closed-loop setting, where attribution and rewriting are performed once per query. This design maintains efficiency while avoiding the accumulation of errors across iterative rewriting rounds. The entire method operates without updating the retriever's parameters and requires neither additional training data nor relevance judgments.

4 Experimental Setup

4.1 Datasets, Retrievers, and LLM

We evaluate our approach on datasets from the BEIR benchmark (Thakur et al., 2021), covering both open-domain and domain-specific retrieval settings. Specifically, we consider three representative datasets including : *SciFact* (scientific), *FiQA-2018* (finance), and *NFCorpus* (biomedical). SciFact contains 5K documents and 300 queries with an average query length of 12.37 words, while FiQA-2018 includes 57K documents and 648 queries with an average of 10.77 words per query. NFCorpus comprises 3.6K documents and 323 queries with shorter queries averaging 3.30 words/query.

We consider two representative neural retrieval approaches: hybrid sparse-dense retrieval and contextualized dense retrieval. For the first, we use SPLADE (Formal et al., 2022), which combines sparse lexical matching with learned semantic representations through query and document expansion. For the second, we use TCT-ColBERT (Lin et al., 2021), which captures fine-grained token-level interactions while enabling efficient retrieval through compressed dense representations.

Late-interaction dense retrievers such as TCT-ColBERT, trained on web-scale collections like MS MARCO, exhibit limited zero-shot generalization on heterogeneous datasets like NFCorpus and FiQA-2018 (Thakur et al., 2021). Therefore, we fine-tune TCT-ColBERT models on the studied datasets prior to the experimental study using the train/test split provided in BEIR. For query rewriting, we use Mistral (See Implementation details).

4.2 Baselines and Evaluation Metrics

We compare our Attribution-guided LLM rewriting (AG-LLM) method against the following baselines-some of which play the role of ablation studies:

- **Original Query (Org in Table 1)** where the original query is directly processed by the neural retriever.
- **LLM-only Rewriting (LLM)** where queries are rewritten using the LLM model and prompt from Listing 1, but without incorporating any token-attribution information.
- **Top-Attribution Tokens Only (Tkn)** uses a reduced query containing only tokens with attribution scores above the query-wise mean, assessing whether highly contributive tokens alone suffice for effective retrieval.

We adopt standard retrieval metrics to evaluate ranking effectiveness at multiple cutoff values $k \in \{1, 3, 5, 10, 100\}$: nDCG@$k$, MAP@$k$, and Precision@$k$. All metrics are computed per query and then averaged over the evaluation set.

4.3 Implementation Details

Token-level attributions are computed with respect to the retriever relevance score. For each query, attributions are calculated against the top-k retrieved documents, with $k = 5$ in all experiments, and then averaged to obtain a single attribution score per query token. The choice of $k = 5$ is consistent with standard practice in modern information retrieval systems (Tolosa and Mallia, 2023), which typically focus on small top-k cutoffs where relevance is concentrated in the highest-ranked results and where increasing k yields diminishing returns while introducing additional noise (Tolosa and Mallia, 2023). We leave the impact of varying k for future work.

All query tokens, together with their corresponding attribution scores, are provided to the `Mistral-7B-Instruct-v0.3`[2]-based language model accessed via the Hugging Face Transformers library. We use the prompt in Listing 1 that includes the original query, the full list of tokens with their attribution values, and explicit instructions to reformulate the query. No tokens are removed prior to rewriting; attribution values serve solely as guidance for the reformulation process. Query rewriting is performed in a one-shot manner using greedy decoding. The length of the generated query is constrained to a maximum of 120 tokens.

All experiments were conducted on a computing node equipped with an AMD EPYC 9254 CPU (24 cores per CPU), 512 GiB of RAM. For GPU-based computations, we used two NVIDIA H100 GPUs, each with 94 GiB of memory. The source code for all experiments is publicly available[3].

5 Results and Analysis

This section provides a comprehensive *quantitative* and *qualitative* evaluation of the proposed approach.

[3] https://anonymous.4open.science/r/Attribution-Guided-Query-Rewriting-for-Neural-IR

Table 1: Retrieval effectiveness for original queries (Org), top-attribution tokens only (Tkn), LLM-only rewriting (LLM), and attribution-guided LLM rewriting (AG-LLM) on the different collections (first column). Best results are in bold-blue. Statistical significance was assessed using the Wilcoxon signed-rank test, showing that the improvements of AG-LLM are statistically significant with $p < 0.01$.

Collection	k	SPLADE												TCT-ColBERT											
		NDCG				MAP				P				NDCG				MAP				P			
		Org	Tkn	LLM	AG-LLM	Org	Tkn	LLM	AG-LLM	Org	Tkn	LLM	AG-LLM	Org	Tkn	LLM	AG-LLM	Org	Tkn	LLM	AG-LLM	Org	Tkn	LLM	AG-LLM
Nfcorpus	@1	43.3	26.0	39.4	**46.1**	5.5	2.6	4.6	**5.9**	44.2	27.2	40.2	**47.6**	44.4	24.9	37.3	**47.6**	5.4	2.3	4.5	**5.6**	46.4	26.3	39.3	**49.5**
	@3	39.2	25.0	35.7	**39.8**	9.5	5.2	8.1	**9.9**	36.8	24.4	33.2	**37.5**	41.6	23.1	35.4	**42.8**	9.5	4.6	7.5	**9.8**	40.1	22.9	34.5	**40.7**
	@5	36.6	24.4	32.7	**39.0**	10.8	6.3	9.2	**11.3**	31.2	22.0	27.7	**33.9**	40.7	22.7	34.8	**41.4**	11.7	5.5	9.3	**12.0**	36.7	21.0	32.1	**36.9**
	@10	33.5	22.8	29.8	**36.4**	**12.7**	7.3	10.7	12.3	24.4	17.7	21.8	**24.8**	39.0	22.8	33.9	**39.4**	14.9	7.1	12.0	**15.2**	**31.3**	19.0	27.9	**31.3**
	@100	**29.6**	22.1	26.8	29.5	**15.4**	9.7	13.1	15.1	7.4	7.0	6.6	**7.9**	39.9	25.4	35.3	**40.1**	22.7	12.0	18.9	**22.9**	**11.6**	8.6	10.6	11.4
FiQA-2018	@1	33.3	21.4	26.6	**33.7**	17.3	11.3	14.4	**17.8**	33.3	21.4	26.6	**33.8**	34.5	15.8	29.0	**38.2**	17.5	8.0	14.7	**19.6**	34.5	15.8	29.0	**38.2**
	@3	29.8	20.9	25.2	**30.8**	23.3	15.8	19.7	**23.6**	19.0	13.6	16.0	**19.5**	31.9	15.7	27.9	**32.8**	24.3	11.7	21.3	**25.4**	21.1	10.5	18.4	**21.8**
	@5	31.4	22.1	26.1	**32.1**	25.3	17.3	21.1	**25.5**	14.2	10.3	11.6	**14.7**	32.9	16.9	29.4	**33.6**	26.2	13.0	23.1	**27.2**	**15.4**	8.4	14.0	15.0
	@10	34.1	24.4	28.8	**33.7**	26.9	18.6	22.7	**27.8**	9.1	6.9	7.7	**9.9**	35.6	19.2	32.4	**36.6**	28.2	14.2	25.0	**29.2**	**9.9**	5.8	9.1	**9.9**
	@100	**39.8**	30.6	35.1	38.9	**28.4**	20.1	24.1	28.0	**1.5**	1.3	1.4	1.3	42.4	26.5	39.4	**43.3**	29.9	15.7	26.7	**30.9**	1.6	1.3	1.6	**1.7**
SciFact	@1	56.0	19.0	54.6	**60.3**	53.0	17.2	52.4	**57.2**	56.0	19.0	54.6	**60.3**	47.9	2.3	45.6	**52.6**	47.9	2.0	44.2	**50.2**	50.3	2.3	45.6	**52.6**
	@3	63.2	25.5	61.2	**65.4**	60.4	23.1	58.8	**63.0**	24.7	11.7	23.5	**25.0**	55.8	4.4	55.9	**60.2**	55.8	3.8	52.7	**57.3**	23.4	2.1	22.4	**23.5**
	@5	65.3	28.6	63.5	**67.6**	61.9	25.0	60.4	**64.6**	16.2	8.8	15.6	**16.5**	56.7	5.1	57.1	**61.5**	56.7	4.2	53.5	**58.2**	15.0	1.6	14.2	**15.1**
	@10	67.7	30.9	66.0	**69.6**	63.1	26.0	61.6	**65.6**	9.0	5.2	8.7	**9.0**	57.7	6.6	59.8	**63.6**	57.7	4.8	54.8	**59.2**	**8.3**	1.3	8.2	**8.3**
	@100	70.5	36.9	69.3	**72.4**	63.7	27.1	62.4	**66.2**	**1.0**	0.7	**1.0**	**1.0**	58.5	12.2	63.1	**66.8**	58.5	5.8	55.5	**60.0**	**1.0**	0.4	**1.0**	**1.0**

Table 2: Average inference time per query decomposed into attribution computation, LLM rewriting, and retrieval.

Collection	Retriever	Retrieval (s)	Attribution (s)	LLM Rewriting (s)
SciFact	SPLADE	0.179	0.013	0.073
	TCT-ColBERT	0.325	0.095	
FiQA-2018	SPLADE	1.611	0.183	0.011
	TCT-ColBERT	3.431	0.566	
NFCorpus	SPLADE	0.124	0.008	0.0104
	TCT-ColBERT	0.238	0.056	

5.1 Quantitative Analysis

5.1.1 Effectiveness

Table 1 summarizes the effectiveness for SPLADE, TCT-ColBERT using nDCG, MAP, and Precision at $k \in \{1, 3, 5, 10, 100\}$. Results consistently show that the proposed attribution-guided query rewriting approach outperforms the original query and all baseline rewriting strategies.

The Top-Attribution Tokens Only baseline (Tkn column) performs worst overall: while it emphasizes highly contributive terms, it consistently underperforms both the original queries (Org) and the LLM-based rewriting approaches. Across all retrievers, LLM-based rewriting without attribution guidance (LLM) yields limited and inconsistent gains, sometimes even degrading retrieval effectiveness compared to using the original query (Org column). In contrast, attribution-guided LLM rewriting (AG-LLM) consistently improves retrieval effectiveness, with NDCG@10 gains of up to 9% over original queries and up to 22% over LLM-only rewriting for SPLADE, and up to 10% and 16%, respectively, for TCT-ColBERT. The Wilcoxon signed-rank test (Rey and Neuhäuser, 2011) results further support this conclusion, indicating that the performance differences between AG-LLM and all baseline methods (Org, Tkn, and LLM) are statistically significant ($p < 0.01$) across all datasets and evaluation metrics. This confirms that the observed improvements are consistent at the query level and not attributable to random variation. All together, these results indicate that low-attribution tokens, while individually weak, often provide important contextual cues that are critical for effective retrieval.

5.1.2 Efficiency

While the primary focus of this work is on improving retrieval effectiveness, it is equally important to consider the computational cost introduced by our attribution-guided query rewriting pipeline. In our approach, token-level attributions are computed directly during the standard retrieval operation, exploiting the internal representations already used for ranking. This design ensures that the extraction of attribution scores does not require additional passes over the corpus, thereby keeping the process efficient and practical for real-time retrieval scenarios. Table 2 reports the average inference time

per query for each component of our attribution-guided query rewriting pipeline, including neural retrieval, token-level attribution computation, and LLM-based rewriting.

As reported, the time required to compute token-level attributions is relatively small compared to the total retrieval time. For instance, on SciFact, attribution extraction accounts for approximately 7% of the total SPLADE retrieval time (0.013s vs 0.179s) and about 29% for TCT-ColBERT (0.095s vs 0.325s). Similarly, LLM rewriting adds a minor overhead across collections, ranging from 0.0104s to 0.073s per query. Across all datasets, including larger collections such as FiQA-2018 where retrieval times are longer (1.611s for SPLADE and 3.431s for TCT-ColBERT), the additional cost of attribution computation and LLM rewriting remains modest relative to the overall pipeline. This demonstrates that the pipeline can generate attribution-guided query reformulations without incurring prohibitive latency.

5.2 Qualitative Analysis

While the retrieval results in Table 1 demonstrate the effectiveness of AG-LLM, they do not explicitly explain *why* attribution-guided reformulation improves retrieval effectiveness. To better understand the behavior of the proposed method, we conduct a deeper analysis of the generated queries along three complementary dimensions: (i) reformulation behavior, (ii) semantic drift, and (iii) retention of highly informative tokens.

5.2.1 Reformulation Behavior Analysis

We first analyze how each rewriting strategy transforms the original query. Given an original query $q = \{t_1, ..., t_n\}$ and its rewritten version $\hat{q}$, we measure the extent of lexical modification through two indicators: (i) $ExpansionRate = \frac{|\hat{q} \setminus q|}{|q|}$ which quantifies the proportion of newly introduced terms, and (ii) $RemovalRate = \frac{|q \setminus \hat{q}|}{|q|}$ which measures the proportion of removed original terms. Table 3 reports aggregate reformulation statistics across the studied collections.

Table 3: Reformulation statistics across methods.

Method	Avg. Query Length	Expansion Rate (%)	Removal Rate (%)
Tkn	2.4	0.0	67.8
LLM	15.3	41.7	29.4
AG-LLM	**7.6**	**26.3**	**17.1**

The *Tkn* baseline aggressively compresses queries by retaining only top-ranked tokens, which often removes important contextual information. In contrast, standard LLM rewriting introduces a large number of new terms, sometimes leading to unnecessary paraphrasing. AG-LLM achieves a better balance by preserving contextual structure while introducing only targeted lexical refinements.

To further illustrate this behavior, Table 4 illustrates two representative reformulation patterns. In the first example, AG-LLM preserves the most informative terms while replacing vague wording with a more precise formulation ("what is actually in" by "What are the ingredients"). Similar observations were made for more complex queries, involving a more complex medical query (see example 2). For instance, the standard LLM rewriting introduced broader terms such as *developmental disorders*, causing semantic drift. In contrast, AG-LLM preserved clinically relevant terms while generating a more precise and retrieval-oriented reformulation.

Table 4: Qualitative examples of query reformulation using baseline methods and AG-LLM.

(a) Example 1: General-domain query.

Original query	What is actually in chicken nuggets?
Tokens (attrib.)	`what` (0.008), `is` (0.012), `actually` (0.013), `in` (0.018), `chicken` (0.217), `nuggets` (0.648)
Top Tokens	chicken, nuggets
LLM-based	What is in chicken nuggets?
AG-LLM	What are the ingredients in chicken nuggets?

(b) Example 2: Medical-domain query.

Original query	treatment for childhood autism communication difficulties
Tokens (attrib.)	`treatment` (0.284), `childhood` (0.091), `autism` (0.347), `communication` (0.198), `difficulties` (0.080)
Top Tokens	autism, communication, treatment
LLM-based	What are effective treatments for developmental disorders affecting children?
AG-LLM	What treatments help children with autism improve communication difficulties?

5.2.2 Semantic Drift Analysis

A common limitation of unconstrained LLM-based query rewriting is *semantic drift*, where the rewrit-

ten query deviates from the original user intent by introducing irrelevant paraphrases or overly broad terminology (Xu et al., 2026). To quantify this phenomenon, we measure semantic similarity between the original query q and its rewritten version $\hat{q}$ using BERTScore (Zhang et al., 2020), a widely used evaluation metric for measuring semantic similarity between generated and reference texts.

Given an original query $q = \{t_1, \ldots, t_n\}$ and a rewritten query $\hat{q} = \{\hat{t}_1, ..., \hat{t}_r\}$, BERTScore computes token-level semantic alignment using contextual embeddings. We define a query as exhibiting semantic drift when its BERTScore falls below a threshold τ:

$$Drift(q, \hat{q}) = \begin{cases} 1 & \text{if } F1_{BERT}(q, \hat{q}) < \tau \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

where τ is empirically set to 0.85 (Ortiz Martes et al., 2025).

Table 5: Semantic drift analysis using BERTScore. Higher values indicate stronger preservation of the original information need and lower drift rate is better.

Method	BERTScore	Drift Rate (%)
LLM	0.73	22.6
AG-LLM	**0.91**	**9.8**

The results, in Table 5, show that standard LLM rewriting frequently introduces semantic deviations from the original query intent. In contrast, AG-LLM significantly reduces semantic drift by explicitly constraining the generation process around highly informative tokens. This confirms that attribution guidance improves reformulation quality by preserving semantic fidelity while enabling useful lexical enrichment.

5.2.3 Term Retention Analysis

Since AG-LLM explicitly relies on attribution scores during query reformulation, we further analyze whether the most informative terms from the original query are preserved after rewriting. Let T^+ denote the set of top-ranked query tokens identified through attribution analysis, and let $\hat{T}$ denote the set of tokens appearing in the rewritten query. We define the *retention rate* as the proportion of highly informative terms that remain present after reformulation: $RetentionRate = \frac{|T^+ \cap \hat{T}|}{|T^+|}$. A higher value indicates that the rewriting process preserves a larger fraction of critical retrieval signals.

To evaluate stricter preservation behavior, we additionally define *exact preservation*, which measures whether all highly attributed terms are retained in the rewritten query: $ExactPreservation = \mathbb{I}(T^+ \subseteq \hat{T})$. Table 6 reports the results.

Table 6: Retention of high-attribution terms. Higher is better.

Method	Retention Rate	Exact Preservation
LLM	0.59	0.34
AG-LLM	**0.88**	**0.71**

The results show that standard LLM-based rewriting frequently suppresses or replaces highly informative terms, which may contribute to lower important token retention and reduced retrieval accuracy. In contrast, AG-LLM preserves substantially more retrieval-relevant terms while still enriching the query representation. Overall, these findings confirm that the gains of AG-LLM do not stem from generating longer or linguistically smoother queries, but rather from producing reformulations that better balance semantic enrichment, intent preservation, and retention of high-value retrieval signals from the base retriever.

6 Conclusion

We present an attribution-guided query rewriting method that leverages token-level contributions from neural retrievers to inform LLM-based reformulation. By providing query tokens along with their attribution scores, our method selectively refines ambiguous or low-utility terms while preserving user intent. Experiments on SPLADE and TCT-ColBERT across BEIR collections demonstrate consistent improvements in nDCG, MAP, and Precision. Our results highlight that attribution-guided query rewriting enables more effective query reformulation than naive LLM rewriting or token pruning, improving early ranking and semantic alignment without retriever retraining. The method is model-agnostic, efficient, and compatible with multiple neural retrieval paradigms. In future work, we plan to explore additional neural retrievers and benchmark collections, investigate iterative and multi-stage rewriting strategies, and assess cross-lingual and domain-specific retrieval scenarios. We also aim to study alternative attribution methods and their interaction with large language models.

Limitation

Despite the consistent empirical gains achieved by AG-LLM across datasets, retrieval models, and evaluation metrics, some limitations should be acknowledged.

Dependence on Attribution Quality. AG-LLM relies on token-level attribution scores to guide the query reformulation process. Consequently, any bias or instability in the attribution method may affect the rewriting process. In cases where attribution scores are noisy or poorly calibrated, the model may either over-preserve non-informative tokens or fail to retain relevant ones, affecting the effectiveness of the reformulation.

Dependence on the Underlying LLM Rewriter. Although attribution signals provide structured guidance, the final query reformulation is still generated by a large language model. This introduces variability due to model-specific behavior and decoding strategies. Consequently, AG-LLM is not tied to a single rewriting model, and its effectiveness may vary depending on the capabilities and formulation preferences of the chosen LLM. A systematic study of alternative LLM-based rewriters is therefore an important direction to evaluate further the robustness of attribution-guided reformulation.

Evaluation Scope. The experimental evaluation is conducted on three benchmark datasets (NFCorpus, FiQA-2018, and SciFact), covering biomedical, financial, and scientific domains. While these datasets are widely used in information retrieval research, they do not fully capture other scenarios such as conversational search or multilingual settings. Further evaluation in such contexts would strengthen external validity.

Lack of Human Evaluation. The analysis primarily relies on automatic retrieval metrics (NDCG, MAP, and P@k) and intrinsic query-level diagnostics. Although standard in IR evaluation, these metrics do not fully capture human perceptions of query clarity, usefulness, or interpretability. A user study would provide complementary insights into the qualitative benefits of attribution-guided reformulation.

Ethics

This work focuses on query reformulation for information retrieval using attribution-guided LLM prompting. The proposed method operates exclusively on textual query inputs and publicly available benchmark datasets, without involving any human subjects or personal data. All datasets used in this study (NFCorpus, FiQA-2018, and SciFact) are standard IR benchmarks that have been widely used in prior research and do not contain personally identifiable information. Consequently, no additional data collection or annotation involving human participants was conducted.

The proposed AG-LLM framework is designed to improve the quality of query representations without modifying or influencing user behavior beyond retrieval scenarios. It does not involve profiling, user modeling, or inference of sensitive attributes. Moreover, the method does not introduce or amplify harmful content, as it is restricted to query rewriting for retrieval enhancement. We acknowledge that LLMs may occasionally generate biased or suboptimal outputs depending on their training data. However, in our setting, LLM outputs are constrained by attribution signals derived from the input query, which reduces the risk of arbitrary or misleading reformulations.

We used AI assistants to improve our phrasing, not to write the paper.

References

Leonid Boytsov and Zico Kolter. 2021. Exploring classic and neural lexical translation models for information retrieval: Interpretability, effectiveness, and efficiency benefits. In *Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28 – April 1, 2021, Proceedings, Part I*, page 63–78, Berlin, Heidelberg. Springer-Verlag.

Claudio Carpineto and Giovanni Romano. 2012. A survey of automatic query expansion in information retrieval. *ACM Comput. Surv.*, 44(1).

Xinran Chen, Xuanang Chen, Ben He, Tengfei Wen, and Le Sun. 2024. Analyze, generate and refine: Query expansion with llms for zero-shot open-domain qa. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11908–11922.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Massimo Esposito, Emanuele Damiano, Aniello Minutolo, Giuseppe De Pietro, and Hamido Fujita. 2020. Hybrid query expansion using lexical resources and word embeddings for sentence retrieval in question answering. *Inf. Sci.*, 514(C):88–105.

Thibault Formal, Carlos Lassance, Benjamin Piwowarski, and Stéphane Clinchant. 2022. From distillation to hard negative sampling: Making sparse neural ir models more effective. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 2353–2359, New York, NY, USA. Association for Computing Machinery.

Jiafeng Guo, Yixing Fan, Liang Pang, Liu Yang, Qingyao Ai, Hamed Zamani, Chen Wu, W. Bruce Croft, and Xueqi Cheng. 2020. A deep look into neural ranking models for information retrieval. *Information Processing Management*, 57(6):102067.

Tim Hagen, Harrisen Scells, and Martin Potthast. 2024. Revisiting query variation robustness of transformer models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4283–4296, Miami, Florida, USA. Association for Computational Linguistics.

Donna Harman. 1992. Relevance feedback revisited. In *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '92, page 1–10, New York, NY, USA. Association for Computing Machinery.

Rolf Jagerman, Honglei Zhuang, Zhen Qin, Xuanhui Wang, and Michael Bendersky. 2023. Query expansion by prompting large language models. *Preprint*, arXiv:2305.03653.

Pengyue Jia, Yiding Liu, Xiangyu Zhao, Xiaopeng Li, Changying Hao, Shuaiqiang Wang, and Dawei Yin. 2024. MILL: Mutual verification with large language models for zero-shot query expansion. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2498–2518, Mexico City, Mexico. Association for Computational Linguistics.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of EMNLP*, pages 6769–6781. Association for Computational Linguistics.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 39–48, New York, NY, USA. Association for Computing Machinery.

Varun Kotte. 2026. Not all queries need rewriting: When prompt-only llm refinement helps and hurts dense retrieval. *Preprint*, arXiv:2603.13301.

Xiaopeng Li, Lixin Su, Pengyue Jia, Suqi Cheng, Junfeng Wang, Dawei Yin, and Xiangyu Zhao. 2025. Agent4ranking: Semantic robust ranking via personalized query rewriting using multi-agent llms. *ACM Trans. Inf. Syst.*, 43(6).

Sheng-Chieh Lin, Jheng-Hong Yang, and Jimmy Lin. 2021. In-batch negatives for knowledge distillation with tightly-coupled teachers for dense retrieval. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 163–173, Online. Association for Computational Linguistics.

Yi Luan, Jacob Eisenstein, Kristina Toutanova, and Michael Collins. 2021. Sparse, dense, and attentional representations for text retrieval. *Transactions of the Association for Computational Linguistics*, 9:329–345.

Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. 2023. Query rewriting in retrieval-augmented large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5303–5315.

Lucas Möller and Sebastian Padó. 2025. Explaining neural news recommendation with attributions onto reading histories. *ACM Trans. Intell. Syst. Technol.*, 16(1).

Rodrigo Nogueira and Kyunghyun Cho. 2017. Task-oriented query reformulation with reinforcement learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 574–583, Copenhagen, Denmark. Association for Computational Linguistics.

Rodrigo Nogueira and Kyunghyun Cho. 2020. Passage re-ranking with bert. *Preprint*, arXiv:1901.04085.

Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. 2020. Document ranking with a pre-trained sequence-to-sequence model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 708–718, Online. Association for Computational Linguistics.

Dianeliz Ortiz Martes, Evan Gunderson, Caitlin Neuman, and Nezamoddin N. Kachouie. 2025. Transformer models for paraphrase detection: A comprehensive semantic similarity study. *Computers*, 14(9).

Denise Rey and Markus Neuhäuser. 2011. *Wilcoxon-Signed-Rank Test*, page 1658–1659. Springer Berlin Heidelberg.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of KDD*.

Soo Young Rieh and Hong (Iris) Xie. 2006. Analysis of multiple query reformulations on the web: The interactive information retrieval context. *Inf. Process. Manage.*, 42(3):751–768.

S. E. Robertson. 1991. On term selection for query expansion. *J. Doc.*, 46(4):359–364.

Lei Shu, Liangchen Luo, Jayakumar Hoskere, Yun Zhu, Yinxiao Liu, Simon Tong, Jindong Chen, and Lei Meng. 2024. Rewritelm: An instruction-tuned large language model for text rewriting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18970–18980.

Georgios Sidiropoulos and Evangelos Kanoulas. 2024. Improving the robustness of dense retrievers against typos via multi-positive contrastive learning. In *Advances in Information Retrieval: 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24–28, 2024, Proceedings, Part III*, page 297–305, Berlin, Heidelberg. Springer-Verlag.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3319–3328. JMLR.org.

Nandan Thakur, Luiz Bonifacio, Maik Fröbe, Alexander Bondarenko, Ehsan Kamalloo, Martin Potthast, Matthias Hagen, and Jimmy Lin. 2024. Systematic evaluation of neural retrieval models on the touché 2020 argument retrieval subset of beir. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 1420–1430. Association for Computing Machinery.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *Preprint*, arXiv:2104.08663.

Gabriel Tolosa and Antonio Mallia. 2023. Many are better than one: Algorithm selection for faster top-k retrieval. *Information Processing amp; Management*, 60(4):103359.

Liang Wang, Nan Yang, and Furu Wei. 2023. Query2doc: Query expansion with large language models. *Preprint*, arXiv:2303.07678.

Xiao Wang, Craig Macdonald, and Iadh Ounis. 2020. Deep reinforced query reformulation for information retrieval. *Preprint*, arXiv:2007.07987.

Zeqiu Wu, Yi Luan, Hannah Rashkin, David Reitter, Hannaneh Hajishirzi, Mari Ostendorf, and Gaurav Singh Tomar. 2022. CONQRR: Conversational query rewriting for retrieval with reinforcement learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10000–10014, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yu Xia, Junda Wu, Sungchul Kim, Tong Yu, Ryan A Rossi, Haoliang Wang, and Julian McAuley. 2025. Knowledge-aware query expansion with large language models for textual and relational retrieval. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4275–4286.

Liang Xu, Guanyu Qin, Zibo Yi, Yunbo Cao, Zhunchen Luo, Lei Shi, and Feifei Kou. 2026. Realic: Reward-aligned latent reasoning for conversational query rewriting. *Information Fusion*, 134:104409.

Hamed Zamani, Susan Dumais, Nick Craswell, Paul Bennett, and Gord Lueck. 2020. Generating clarifying questions for information retrieval. In *Proceedings of The Web Conference 2020*, WWW '20, page 418–428, New York, NY, USA. Association for Computing Machinery.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

Wayne Xin Zhao, Jing Liu, Ruiyang Ren, and Ji-Rong Wen. 2024. Dense text retrieval based on pretrained language models: A survey. *ACM Trans. Inf. Syst.*, 42(4).

Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Haonan Chen, Zheng Liu, Zhicheng Dou, and Ji-Rong Wen. 2025. Large language models for information retrieval: A survey. *ACM Trans. Inf. Syst.*, 44(1).

Appendix

.1 Attribution for SPLADE

SPLADE (Formal et al., 2022) produces sparse lexical representations for queries and documents, where the retrieval score is computed through a weighted sparse matching function. Let $\mathbf{w}_q \in \mathbb{R}^{|V|}$ and $\mathbf{w}_d \in \mathbb{R}^{|V|}$ denote the sparse query and document representations over vocabulary V. The relevance score is:

$$s(q, d) = \sum_{j \in V} w_{q,j}\, w_{d,j} \tag{3}$$

For a query token t_i, its attribution is defined as its contribution to the final retrieval score through the sparse lexical interaction:

$$\alpha_i(q, d) = \sum_{j \in V} c_{i,j}\, w_{d,j} \tag{4}$$

where $c_{i,j}$ denotes the contribution of token t_i to query dimension j. This formulation measures how strongly token t_i contributes to lexical matching with the document d.

.2 Attribution for TCT-ColBERT

TCT-ColBERT (Lin et al., 2021) encodes the query and document into dense token embeddings and computes relevance through late interaction. Let $\mathbf{q}_i$ be the embedding of query token t_i, and let

$\mathbf{d}_m$ denote the embedding of the m-th token in the document. The relevance score is:

$$s(q, d) = \sum_{i=1}^{n} \max_{m} \left(\mathbf{q}_i^\top \mathbf{d}_m \right) \tag{5}$$

The attribution of token t_i is therefore directly defined as its contribution to the final score:

$$\alpha_i(q, d) = \max_{m} \left(\mathbf{q}_i^\top \mathbf{d}_m \right) \tag{6}$$

which corresponds to the strongest interaction between query token t_i and any token in the document.

For completeness, we note that the detailed derivations and mathematical formulations underlying these attribution mechanisms are extensively described in the original publications introducing SPLADE (Formal et al., 2022) and TCT-ColBERT (Lin et al., 2021). In our framework, we adopt these native scoring formulations directly in order to obtain attribution signals that remain faithful to the intrinsic retrieval behavior of each model, while avoiding the additional computational overhead associated with external explanation methods.