Annotation-Free Reinforcement Learning Query Rewriting via Verifiable Search Reward

Anonymous ACL submission

Abstract

Optimizing queries for Retrieval-Augmented Generation (RAG) systems poses a significant challenge, particularly across diverse modal indices. We introduce RL-QR, a novel annotation-free reinforcement learning framework for query rewriting that eliminates the need for costly human-annotated data. By leveraging verifiable search rewards derived from index-aligned synthetic queries, RL-QR overcomes human-annotation dependencies, extending its applicability to various modalities and index domains. Experimental results demonstrate the framework's robustness, achieving substantial retrieval performance gains of up to 3.9× on lexical retrievers and 3.5× on semantic retrievers on the MTEB VIDORE V2 benchmark for unstructured visual documents, along with consistent 5% to 10% improvements on MS MARCO v2.1 and internal industrial datasets.

1 Introduction

Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) has proven to be a powerful and widely adopted approach across numerous domains, from natural language processing to multi-modal applications. Its ability to integrate external knowledge into generation tasks has made it a cornerstone of modern retrieval systems. Modern AI assistants (Hurst et al., 2024; Comanici et al., 2025) adopt RAG as core function for correcting factually, delivering out-domain knowledge and beyond.

In practice, when serving RAG systems across various domains and index formats, adapting queries through rewriting proves to be more effective and cost-efficient than rebuilding retrievers. For lexical retrievers, creating domain-specific dictionaries can enhance performance. However, this approach depends on manual annotation, which is not scalable and increases operational costs. For semantic indices, retrievers can be fine-tuned with domain-specific data. Yet, this introduces the burden of maintaining domain-specific retrievers, generating training data, and conducting retraining. Moreover, updating retrievers typically requires reindexing, which adds complexity to RAG system dependencies and further raises operational costs. In contrast, query rewriters transform queries into the representation space of the retrievers, allowing compatibility across different retrievers and index types. From a system maintenance and deployment perspective, developing a query rewriter is generally more cost-effective than enhancing retrievers or re-indexing. It also promotes modularity in RAG architecture by decoupling the query rewriting module from retriever components, avoiding the need for domain-specific retriever development.

Although query rewriting is central to RAG systems, generalized approaches remain largely unexplored due to their reliance on costly human annotation. Recent studies have proposed implicit learning methods that reward the model when the final answer is correct, requiring annotated index-query-answer-verifier sets (Jin et al., 2025). Others use explicit learning, which rewards the model when relevant documents are retrieved, but this approach depends on expensive per-query annotations of both positive and negative document pairs (Wang et al., 2025). While these methods show promise within narrow domains, they face major limitations: they require extensive human effort and are largely restricted to curated, text-only data sources—making them unsuitable for real-world, unstructured document collections.

This study introduces an annotation-free **R**einforcement **L**earning framework for **Q**uery **R**ewriting with verifiable search reward (**RL-QR**). The framework's core novelty lies in synthesizing queries in an index-aligned manner, which permits the proposed verifiable search reward to directly exploit the resultant search score for training. Crucially, this methodology replaces the need for posi-

tive query-corpus human-annotations, thus ensuring off-the-shelf operation and achieving general applicability across different index modalities and domains.

Our experiments demonstrate robust and significant improvements on text-modal and multi-modal unstructured documents with the RAG agent framwork. Especially, upon the general visual document retrieval benchmark MTEB VIDORE V2 (Macé et al., 2025) the conventional text-parsing based RAG system benefits upto $3.9\times$ retrieval recall. For the other benchmarks, the text-modal retrieval benchmark MS MARCO v2.1 (Nguyen et al., 2016) and internal industrial unstructured document benchmark, RL-QR steadily achieves 5% to 10% performance gains. It supports the effectiveness and the adaptability over various index domains and modalities.

In summary, our contributions are

- **Annotation-Free RL Framework with Verifiable Rewards:** We introduce RL-QR, a novel reinforcement learning framework that eliminates the dependency on costly human-annotated data for query rewriting. By leveraging index-aligned synthetic queries to generate verifiable search rewards, our approach explicitly optimizes the rewriter using the resultant search scores.

- **Universal Adaptability and Modular System Integration:** We propose a retriever- and index-agnostic solution that ensures general applicability across diverse modalities, including text-parsed and unstructured visual documents. This modular design decouples the query rewriting module from the retriever, significantly reducing system maintenance overhead by removing the need for domain-specific retriever development or expensive re-indexing processes.

- **Robust Empirical Effectiveness:** Extensive experiments demonstrate that RL-QR achieves substantial performance gains, proving its efficacy in "unlearning" chat-oriented behaviors to prioritize retrieval intent, including up to 3.9x improvement on lexical retrievers and 3.5x on semantic retrievers within the MTEB VIDORE V2 benchmark. Additionally, the framework consistently delivers 5% to 10% recall enhancements on the MS MARCO v2.1 benchmark and internal industrial datasets, proving its efficacy in "unlearning" chat-oriented behaviors to prioritize retrieval intent.

2 Related Works

Our work focuses on enhancing the query rewriter for RAG systems, with an emphasis on handling multi-modal (unstructured imaged documents) and text-modal (text-parsed documents) indices with real-world unstructured data. In this section, we provide an overview of the research background, covering the evolution of RAG, the integration of various modalities in RAG systems, and the role of query rewriting.

2.1 Retrieval-Augmented Generation (RAG)

RAG is a hybrid approach that integrates retrieval-based and generation-based techniques to improve the performance of language models on knowledge-intensive tasks. By leveraging external knowledge sources, RAG enables models to produce more accurate and contextually relevant responses. The paradigm has gained significant attention due to its ability to combine the strengths of retrieving pertinent documents and generating coherent text. In the real-world, RAG systems are widely adopted with online web search (*e.g.,* OpenAI (Hurst et al., 2024), and Gemini (Comanici et al., 2025)) and industrial domains with credential documents.

Early research on RAG established its effectiveness across various natural language processing tasks (Lewis et al., 2020). Subsequent studies have proposed advancements, such as improved retrieval mechanisms using dense retrieval methods (Karpukhin et al., 2020) and the integration of structured knowledge bases like databases or graphs (Edge et al., 2024). RAG has also shown promise in multi-task and few-shot learning scenarios (Izacard et al., 2023), where the retrieval component compensates for limited training data by accessing external information. However, challenges remain, particularly in optimizing the retrieval process, which depends heavily on the quality of the input query—an issue that motivates the exploration of query rewriting (Ma et al., 2023).

2.2 Modalities in RAG

While RAG was initially designed for text-based applications, recent applications have extended their scope to the real-world unstructured documents including slide decks, web pages, blogs, papers and

so on supported by document parsing approaches. This expansion is critical for tasks where knowledge sources span multiple formats, requiring systems to integrate and reason over heterogeneous inputs.

Multi-modal RAG systems have been explored for image-as-embedding (Faysse et al., 2024) or parsing-documents-to-text (Feng et al., 2025). Image-as-embedding approaches (Faysse et al., 2024) embeds imaged document into document embedding as document semantic embedding (Zhang et al., 2025). Parsing-documents-to-text approaches (Wei et al., 2024; Feng et al., 2025) converts documents into plain text, enabling the present text retrievers (Robertson et al., 2009; Zhang et al., 2025). For text-modal data, such as parsed text from structured documents, the challenge lies in effectively retrieving and utilizing information from long-form or hierarchically organized content (Larson and Truitt, 2024).

2.3 Query Rewriting for RAG

Query rewriting is a pivotal component in RAG systems, as the effectiveness of the retrieval step hinges on how well the query is formulated. A poorly designed query can lead to irrelevant or low-quality retrieved documents, undermining the generation process. Conversely, an optimized query enhances the relevance of retrieved information, directly improving the overall system performance.

Traditional query rewriting techniques, such as query expansion and reformulation, have roots in information retrieval and rely on heuristics or statistical methods to refine queries (Zhu et al., 2016). In the context of RAG, however, query rewriting must align with the needs of the retriever (Ma et al., 2023). Recent efforts have introduced learning-based approaches, including neural network models and reinforcement learning, to dynamically adapt queries based on system feedback (Wang et al., 2025; Chan et al., 2024; Li et al., 2024; Ma et al., 2023; Jin et al., 2025). Despite these advances, existing query rewriting techniques often require extensive annotated data or are constrained to specific domains (Liu et al., 2021). Our work addresses these gaps by developing a generalized reinforcement learning framework for query rewriting, tailored to enhance retrieval across diverse indices without relying on large-scale human annotations.

3 Method

In this section, we describe our proposed framework, annotation-free reinforcement learning query rewriting via verifiable search reward (RL-QR). Illustrated in the Alg. 1, the learning process consists of two-steps: (1) index-algined query synthesis where the answer for the query necessarily requires the source index resulting the index matching queries, (2) and reinforcement learning the query rewriter based on the verifiable search reward.

Algorithm 1 Annotation-free RL-QR

Require: Retriever R, Search Index $\mathcal{I}$, Query synthesis assistant $Assistant$
Ensure: Optimized Policy π_θ

Initialize π_θ
for index I in $\mathcal{I}$ **do**
  $q \leftarrow Assistant(I)$
  $rewards \leftarrow EmptyList$
  for for i in $N_{rollout}$ **do**
    $q'_i \leftarrow \pi_\theta(q)$
    $SearchedIndices_i \leftarrow R(q'_i)$
    $r_i \leftarrow NDCG(I, SearchedIndices_i)$
    $rewards$.append(r_i)
  end for
  $A \leftarrow GroupComputation(rewards)$
  Update policy π_θ using advantages A
end for
return π_θ

3.1 Index-algined Query Synthesis

Query synthesis have already been explored widely (Xu et al., 2024; Cha et al., 2024), and the modern language models are capable enough to conduct the task. In this work, we generate queries for training in index-aligned manner where the answer for the query necessarily requires the corpus (see the prompt provided in Table1).

Specifically, for the conventional text-parsing retrieval system, given the raw data DB_{raw}, the document parsers P and the search engine E, the source index $\mathcal{I}_{text}$ becomes

$$\mathcal{I}_{text} \leftarrow \bigcup_{d \in DB_{raw}} E.index(P.parse(d)) \quad (1)$$

which contains both the search engine index and the parsed corpus. On the other hand, the multi-modal retrieval system does not adopt the parser.

 # Generating Document-Requiring Question and Answer

Read the document carefully, then perform the following three steps:

1. Think of a **scenario** that necessitates the information contained within the document.
2. Create a **question** that logically fits the identified scenario.
3. Provide an **answer** that accurately matches the created question based on the document's content.

Note: If the document's information is insufficient to identify a situation requiring the document, output blank spaces.

Table 1: Prompt template for index-algined query synthesis. Appending the resulting scenario and question is viable to make longer queries. Training the rewriter does not utilize the generated answers, but those are useful for post-training the answer model and evaluating end-to-end RAG.

Consequently, the multimodal index $\mathcal{I}_{multimodal}$ becomes

$$\mathcal{I}_{multimodal} \leftarrow \bigcup_{d \in DB_{raw}} E.index(d) \quad (2)$$

where E utilizes multimodal embedding model (e.g., (Faysse et al., 2024)) for raw document indexing. Along with the query synthesis assistant (e.g., (Bai et al., 2025)) and the instruction (Table 1), it generates index-aligned queries per index as illustrated in Alg. 1, which can be done online and offline train.

3.2 Reinforcement Learning Query Rewriter

It is important to individualize query rewriter with respect to the indices, because each retriever has distinct characteristics. For example, lexical retrievers such as BM25 (Robertson et al., 2009) count on the number of the words, in which simply repeating important word can augment the performance. Whereas, (multi-modal) semantic retrievers that embed (text-parsed-)documents into embedding (Faysse et al., 2024; Zhang et al., 2025) work better if the query-document resembles their trained data, which is hard to manage. The reinforcement learning aims to align user query into the index representation space by the query rewriter per retriever R and its index $\mathcal{I}$. In other words, for N online RAG systems consisting of the data source DB_i and the retriever R_i for $i \in N$, we suggest to have N rewriters respectively, rather than a single universal rewriter.

The precedent RL approaches (Ma et al., 2023; Jin et al., 2025; Nguyen et al., 2025) implicitly train the rewriter by optimizing

$$\max_{\pi_\theta, \pi_{LLM}} \mathbb{E}_{x \sim D, y \sim \pi_\theta(\cdot|x;R), z \sim \pi_{LLM}(\cdot|x;R(y))} [r_\phi(x, z)] \quad (3)$$

where x refers to the sample from the training data D, y denotes the rewritten query by the rewriter, and z represent the final response. π_θ and π_{LLM} are the target rewriter and the final-responding language model. r_ϕ is the reward function but inherently requires human-annotation in query-response or query-retrieval levels.

In contrast, ours optimizes the rewriter explicitly, which down-scales the objective and boosts the training process. Some (Wang et al., 2025) tried explicit rewarding with massive document-wise positive and negative pair annotation, which limits in scaling covering domain and indices. On the other hand, leveraging the synthesized index-aligned queries, we formulate the RL objective function as follows:

$$\max_{\pi_\theta} \mathbb{E}_{x \sim D, y \sim \pi_\theta(\cdot|x;R)} [r_\phi(x, y)] \quad (4)$$

We adopt two function rewards, one for the query rewriting reward and the other for the formatting and redundant penalty. The verifiable search reward $r_{\text{retrieval}}$ uses NDCG (Järvelin and Kekäläinen, 2002) score directly that measures if the target document is retrieved considering ranks.

$$r_{\text{retrieval}}(x, y) = NDCG(\text{index}_x, R(y)) \quad (5)$$

The penalty $r_{penalty}$ targets to match the format, placing the rewritten query inside `<answer>...</answer>`, and reduce redundant generations outside the format. It is similar to and replacable with the well known length penalties. Further, we normalized the penalty r_{penalty} group-wise by ranging `[0.5, 1]` for the non-zero values. The reward function becomes

$$r_\phi(x, y) = \lambda_1 r_{\text{retrieval}(x,y)} + \lambda_2 r_{\text{penalty}(y)} \quad (6)$$

where the lambdas are the hyper-parameters. More specifically, for each sample x, we optimize the

rewriter by maximizing the following object:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}\left[x \sim D, \{o_i\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}(O|q)\right] \left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\left\{\min\left(r_{i,t}(\theta)\hat{A}_{i,t}, \text{clip}\left(r_{i,t}(\theta), 1-\epsilon, 1+\epsilon\right)\hat{A}_{i,t}\right)\right\}\right] \tag{7}$$

where $r_{i,t}(\theta) = \frac{\pi_\theta(o_{i,t}|x,o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}|x,o_{i,<t})}$ be the probability ratio, ϵ and β are hyper-parameters, and $\hat{A}_{i,t}$ is the advantage based on the relative rewards of the outputs inside each group.

4 Experiment

4.1 RAG System Implementation

In this experiment, we adopt three in-house RAG frameworks named *Semantic*, *Lexical* and *Multimodal*. *Semantic* and *Lexical* are the conventional RAG framework containing (1) the document parser that ingests unstructured documents and chunks into corpus (2) the embedding models and (3) the search engine. *Multimodal* is the modern RAG framework without document parsers.

Semantic utlizes an embedding model that converts the user query and the corpus into vector embeddings and searches by its search engine indexed with the corpus vector embeddings. This approach allows the system to understand the user intent and retrieve relevant information even if the phrasing is different and there is no direct keyword overlap.

Lexical is a traditional information retrieval system built upon BM25 scoring algorithm which conducts retireval by matching the exact token between the user query and the corpus. It excels at precision when the query contains specific terms, acronyms, or identifiers that are also present in the source documents, as its ranking is based on term frequency (TF) and inverse document frequency (IDF).

Multimodal is conceptually identical to which of *Semantic*, but has a difference in the raw document compatibility. As it does not have external parser which used to have information leaks, it shows stronger retrieval performance when it comes to unstructured documents (e.g., slide decks, pdf).

4.2 Experiment Data

Experimental setup has two major categories: (1) text-only retrieval task and (2) unstructured document retrieval task. For the text-only retrieval task, we use MS MARCO v2.1 testset 1% which has 10,078 corpus and 1,011 queries. The unstructured visual document retrieval task consists of MTEB VIDORE V2 benchmark with 4,544 visual documents and 327 queries; and in-house real-world industrial data composed of 2,791 documents and 4,398 queries.

The virtue of retrieval task is to maximize recall, which NDCG represents it well with ordinal scoring. Therefore, we deploy NDCG@3 for the target evaluation metric and the reward score.

4.3 Index-Aligned Query Synthesis

We generated queries per index using Qwen3-VL-235B-A22B (Bai et al., 2025) and assumed the generated query has single retrieval target, the source corpus. In other words, the reward becomes if the query retrives the source corpus from the retriever.

4.4 Reinforcement Learning Query Rewriter

We initialize the rewriter model with Qwen3 in `no_think` mode. We train them by the objective single epoch on eight 80GB H100 GPUs without any supervised finetuning. Each training took 1 to 2 hours and the most bottlenecks are the retrieval overhead and the conversation between the search engine and the rewriter. For the GRPO RL training, we adopt `TRL` library of `huggingface` with `deepspeed` and its default training settings: learning rates, the number of rollouts and more.

5 Results

The experiments are designed to systematically compare the retrieval performance between (1) the raw query, (2) the vanilla base model Qwen3, and (3) the RL-QR trained models. Our primary goal is to isolate the performance gains directly attributable to the proposed RL-QR training strategy.

5.1 Overall Performance Comparison

As shown in Table 2, the vanilla Qwen3 base model exhibits inconsistent performance. It could achieve retrieval gains only on the MTEB-*Semantic* and MTEB-*Lexical* tasks, showing poorer performance than the raw queries in all other evaluated settings.

On the other hand, the proposed **RL-QR** method attains robust improvements across *all* index-retriever cases. RL-QR demonstrates substantial gains, obtaining improvements up to $3.9\times$ and from at least 5%. This consistent superiority across diverse retrieval tasks strongly indicates that the performance improvements are a direct result of

Retriever	Rewriter	Text	Unstructured					
		MARCO	MTEB (Vidore V2)					Internal
			ESG	Eco	ESGHL	BioMed	Avg	
Semantic	Raw query	87.62	9.99	9.75	9.69	15.55	11.25	76.52
	Qwen3-4B	71.08	23.09	35.94	31.45	35.36	31.22	28.08
	+ RL-QR (Ours)	92.43	28.34	47.27	25.29	33.42	33.58	74.86
	Qwen3-8B	84.21	27.70	50.72	30.47	36.71	36.40	27.86
	+ RL-QR (Ours)	91.92	**32.15**	**51.69**	**33.76**	**41.54**	**39.79**	73.97
	Qwen3-14B	70.73	29.60	49.28	31.77	26.82	34.37	69.00
	+ RL-QR (Ours)	**92.67**	28.97	49.22	33.69	37.69	37.39	**80.61**
Lexical	Raw query	80.22	6.76	10.47	3.90	8.66	7.45	72.55
	Qwen3-4B	61.46	9.13	31.72	10.50	22.78	18.53	21.26
	+ RL-QR (Ours)	84.84	12.57	29.31	**11.50**	**33.79**	21.79	76.45
	Qwen3-8B	71.15	11.25	36.29	9.53	20.56	19.41	23.67
	+ RL-QR (Ours)	82.85	**15.17**	38.93	10.75	26.94	**22.95**	78.76
	Qwen-14B	58.43	8.72	**39.70**	9.67	20.18	19.57	60.03
	+ RL-QR (Ours)	**85.10**	12.44	38.66	8.14	24.72	20.99	**80.84**
Multimodal	Raw query	-	45.23	51.38	57.48	51.04	51.28	73.84
	Qwen3-4B	-	24.17	32.08	37.30	23.49	29.26	68.19
	+ RL-QR (Ours)	-	48.48	41.46	41.95	37.52	42.35	78.23
	Qwen3-8B	-	48.30	53.69	55.10	51.14	52.06	70.80
	+ RL-QR (Ours)	-	55.44	57.82	**63.46**	59.29	59.00	79.83
	Qwen3-14B	-	10.80	8.28	19.84	12.00	12.73	77.54
	+ RL-QR (Ours)	-	**58.83**	**62.43**	62.47	**60.02**	**60.94**	**81.08**

Table 2: Information retrieval benchmark results with NDCG@3 metrics. We report performance across varying model sizes (4B, 8B, 14B) to demonstrate the robustness of RL-QR. Note that RL-QR consistently improves performance over Raw Query even with smaller model backbones.

Retriever	Rewriter	Length
Semantic	Raw query	36 ± 14
	Qwen3-14B	95 ± 104
	+RL-QR	38 ± 15
Lexical	Raw query	36 ± 14
	Qwen3-14B	95 ± 104
	+RL-QR	36 ± 14

Table 3: Query length statistics on MS MARCO benchmark.

Retriever	Rewriter	Length
Semantic	Raw query	84 ± 23
	Qwen3-14B	157 ± 199
	+RL-QR	164 ± 113
Lexical	Raw query	84 ± 23
	Qwen3-14B	190 ± 258
	+RL-QR	143 ± 168
Multimodal	Raw query	84 ± 23
	Qwen3-14B	43 ± 171
	+RL-QR	116 ± 48

Table 4: Query length statistics on MTEB Vidore V2 benchmark.

the proposed RL-QR approach, not merely the inherent strength of the baseline models.

5.2 Domain-Specific Performance

Text-Modal Benchmark: On the MS MARCO v2.1 experiment, RL-QR advances the retrieval performance by more than 5% on both the *Semantic* and *Lexical* retrievers. This is in stark contrast to the baseline rewriter, which fails significantly, degrading performance by up to -27% in the worst case and -4% in the best case. For the semantic retriever (Qwen3-14B), RL-QR achieves an NDCG@3 of 92.67 compared to 87.62 for the raw query.

MS MARCO v2.1 @ Semantic Retriever	
Raw query	(0.47) cost of finishing basement which is half done'
Qwen3-14B	(0.30) Average cost to finish a partially completed basement
RLQR-14B	(0.70) cost to finish a basement that is halfway completed
Raw query	(0.47) how is chicken distributed to consumers
Qwen3-14B	(0.00) Explain the process of chicken distribution from farms to consumers, including supply chain, transportation, packaging, and retail channels.
RLQR-14B	(0.77) chicken distribution to consumers
MS MARCO v2.1 @ Lexical Retriever	
Raw query	(0.77) computer networking definition and examples
Qwen3-14B	(0.77) Explain the concept of computer networking and provide real-world examples of its applications
RLQR-14B	(1.0) computer network definition and examples
Raw query	(0.0) granby missouri high and low temp by month calendar
Qwen3-14B	(0.0) Monthly average high and low temperatures in Granby, Missouri, with climate data organized by month for a yearly temperature calendar
RLQR-14B	(0.47) granby missouri monthly high and low temperature calendar
MTEB VIDORE V2 @ Semantic Retriever	
Raw query	(0.0) What are the specific outcomes of using autologous chondrocyte implantation in canine studies?
Qwen3-14B	(0.0) What are the clinical outcomes, efficacy, and long-term results of autologous chondrocyte implantation in canine cartilage repair studies, including veterinary research and experimental trials?
RLQR-14B	(0.47) What are the clinical outcomes, efficacy, and complications associated with autologous chondrocyte implantation (ACI) in canine cartilage repair studies?
Raw query	(0.0) What are the differences in results obtained from two-photon microscopy and confocal microscopy?
Qwen3-14B	(0.0) What is the difference between two-photon microscopy and confocal microscopy in terms of their results?
RLQR-14B	(1.0) What are the key differences in resolution, penetration depth, phototoxicity, and image quality between two-photon microscopy and confocal microscopy results?
MTEB VIDORE V2 @ Lexical Retriever	
Raw query	(0.0) What role do biomaterials play in the development of joint replacement prostheses?
Qwen3-14B	(0.0) What are the key contributions of biomaterials to the design, functionality, and longevity of joint replacement prostheses, including their biocompatibility and interaction with human tissues?
RLQR-14B	(0.31) What is the role and function of biomaterials in the design and development of joint replacement prostheses, including their biocompatibility, durability, and contribution to tissue integration?
Raw query	(0.0) How do the roles of integrins in cell-matrix interactions compare to their roles in signal transmission within the cell?
Qwen3-14B	(0.63) What are the differences in the roles of integrins in extracellular matrix (ECM) interactions versus their involvement in intracellular signaling pathways and mechanisms?
RLQR-14B	(1.0) Compare the functions of integrins in cell-extracellular matrix interactions with their roles in intracellular signal transduction pathways.
MTEB VIDORE V2 @ Multimodal Retriever	
Raw query	(0.23) What role do scaffolds play in tissue engineering?
Qwen3-14B	(0.30) What is the role of a scaffold in tissue engineering and why is it important?
RLQR-14B	(0.47) What is the function and importance of scaffolds in tissue engineering, including their structural support, cell interaction, and role in tissue regeneration
Raw query	(0.0) What are the key factors influencing bioadhesion in biomaterials?
Qwen3-14B	(0.5) What is bioadhesion in the context of biomaterials, and what factors determine its effectiveness?
RLQR-14B	(0.63) What are the primary factors that affect bioadhesion in biomaterials, including molecular interactions, surface properties, and biological compatibility?

Figure 1: Qualitative results with the raw query, the baseline rewriting model, and the proposed method rewriter, RL-QR. Each sample is in form of '(NDCG@3 score) query'.

Unstructured Multimodal Benchmarks: RL-QR notably boosts the recall for the unstructured multimodal settings. We observe gains of $3.9\times$ and $3.5\times$ on *Lexical*-VidoreV2 and *Semantic*-VidoreV2, respectively, with improvements ranging from 5% to 20% for the other combinations. The baselines partially enhance performance on specific tasks but suffer dramatic degradation on others, such as a -71% drop on *Lexical*-Internal.

5.3 Qualitative Analysis of Rewriting Strategies

To understand the mechanism behind the performance gains, we analyzed the rewritten queries shown in Figure 1. We observe that RL-QR does not apply a uniform rewriting policy; rather, it

adapts its strategy based on the domain and the nature of the retriever.

Query Refinement vs. Expansion: In the MS MARCO benchmark (Web Search), RL-QR primarily functions as a query *refiner*. As seen in the "basement cost" example, the model corrects grammatical structures and sharpens the intent (0.70 NDCG) without significantly increasing length. Conversely, in the MTEB VIDORE V2 benchmarks (Technical/Medical), the model shifts to *query expansion*. For instance, when querying about "autologous chondrocyte implantation," RL-QR explicitly injects related technical terms such as "clinical outcomes," "efficacy," and "complications". This expansion, which achieved a 0.47 NDCG score compared to 0.00 for the baseline, bridges the vocabulary gap in specialized corpora.

5.4 Behavioral Alignment for Retrieval

A critical failure mode of the baseline Qwen3 model is its tendency to interpret search queries as generative instructions. As shown in Figure 1, for the query "how is chicken distributed," the baseline generates a verbose command: "Explain the process of chicken distribution..." resulting in a score of 0.00. This suggests the pre-trained model is misaligned for retrieval tasks, prioritizing conversational fluency over keyword matching.

In contrast, RL-QR effectively "unlearns" this chat-oriented behavior. It strips away conversational artifacts and focuses on keyword density and search intent. For the same chicken distribution query, RL-QR outputs "chicken distribution to consumers" (0.77 NDCG), demonstrating that the RL optimization successfully realigned the model's output distribution from *instruction following* to *index-oriented retrieval*.

5.5 Effect of Query Length and Index-Awareness

RL-QR achieved retrieval enhancements irrespective of whether the query length was preserved or enlarged. Cross-referencing the length statistics in Table 4 with the samples in Figure 1 reveals a dynamic, index-aware adaptability.

In precision-oriented environments like MS MARCO, the model retains brevity, with query lengths remaining close to the raw input (e.g., 38 ± 15 tokens). For example, for the "granby missouri" query, it merely reorders keywords to match a likely document title format (0.47 NDCG), avoiding the excessive verbosity seen in the baseline (0.00 NDCG).

Conversely, in recall-oriented multimodal environments, RL-QR significantly expands the query length (116 ± 48 tokens). For the multimodal query on "scaffolds in tissue engineering," RL-QR hallucinates visual and structural context ("structural support," "cell interaction"), boosting the score from 0.23 to 0.47. This behavior results from the training objective (Eq. 3.2), which does not explicitly penalize length, allowing the model to learn an implicit representation of the corpus distribution and adjust its verbosity accordingly.

6 Conclusion

In this work, we presented RL-QR, an annotation-free reinforcement learning framework for query rewriting that eliminates the need for expensive human-annotated training data in Retrieval-Augmented Generation systems. By synthesizing index-aligned queries and directly optimizing the rewriter with verifiable search rewards derived from NDCG, RL-QR achieves robust and substantial retrieval improvements across diverse retrievers and modalities—up to 3.9× on lexical and 3.5× on semantic retrievers for unstructured visual documents on the MTEB VIDORE V2 benchmark, alongside consistent gains of 5%–10% on MS MARCO v2.1 and internal industrial datasets.

The proposed approach is retriever- and index-agnostic, modular, and readily deployable in production environments, significantly reducing the maintenance overhead associated with domain-specific retriever tuning or re-indexing. Qualitative analysis further reveals that RL-QR adaptively learns index-aware rewriting strategies—ranging from concise refinement in text-heavy domains to targeted expansion in technical and multimodal corpora—effectively aligning user queries with the representation space of the underlying index.

RL-QR thus offers a scalable, cost-effective solution to one of the central bottlenecks in modern RAG systems. Future directions include extending the framework to multi-turn conversational retrieval, incorporating richer reward signals from downstream generation quality, and exploring its integration with emerging multimodal foundation models.

7 Limitations

Restriction to Single-Document Grounding. The current implementation of Index-Aligned Query Synthesis (Section 3.1) establishes a one-to-one mapping between a synthesized query and a single positive document. While effective for standard retrieval tasks, this formulation does not explicitly model complex scenarios where a single query necessitates aggregating information from multiple documents (e.g., multi-hop reasoning or multi-aspect retrieval). Future iterations of this work could address this by clustering semantically related documents prior to generation, thereby training the model to synthesize queries that target document groups rather than individual passages.

Dependency on White-Box Indices. A core prerequisite of RL-QR is the calculation of a verifiable search reward, which requires direct access to the target index to determine if the intended document was successfully retrieved. Consequently, our method is currently inapplicable to "black-box" retrieval scenarios, such as commercial Internet search engines (e.g., Google or Bing), where the underlying index is private and the ranking mechanism is opaque. RL-QR is instead optimized for "white-box" settings, such as enterprise search, private RAG (Retrieval-Augmented Generation) systems, or domain-specific vertical search where the index is accessible to the developer.

Computational Overhead of RL Training. Compared to standard supervised fine-tuning, the proposed RL framework incurs higher computational costs during the training phase. Since the reward calculation involves executing a retrieval operation for every generated query, the training throughput is constrained by the latency of the retrieval engine. While inference latency remains unaffected, optimizing the training efficiency for large-scale indices remains an area for future optimization.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

Sungguk Cha, Jusung Lee, Younghyun Lee, and Cheoljong Yang. 2024. Visually dehallucinative instruction generation. *Preprint*, arXiv:2402.08348.

Chi-Min Chan, Chunpu Xu, Ruibin Yuan, Hongyin Luo, Wei Xue, Yike Guo, and Jie Fu. 2024. Rq-rag: Learning to refine queries for retrieval augmented generation. *arXiv preprint arXiv:2404.00610*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. 2024. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. 2024. Colpali: Efficient document retrieval with vision language models. In *The Thirteenth International Conference on Learning Representations*.

Hao Feng, Shu Wei, Xiang Fei, Wei Shi, Yingdong Han, Lei Liao, Jinghui Lu, Binghong Wu, Qi Liu, Chunhui Lin, and 1 others. 2025. Dolphin: Document image parsing via heterogeneous anchor prompting. *arXiv preprint arXiv:2505.14059*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2023. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 24(251):1–43.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. 2025. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick SH Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *EMNLP (1)*, pages 6769–6781.

Jonathan Larson and Steven Truitt. 2024. Graphrag: Unlocking llm discovery on narrative private data.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Zhicong Li, Jiahao Wang, Zhishu Jiang, Hangyu Mao, Zhongxia Chen, Jiazhen Du, Yuanxing Zhang, Fuzheng Zhang, Di Zhang, and Yong Liu. 2024. Dmqr-rag: Diverse multi-query rewriting for rag. *arXiv preprint arXiv:2411.13154*.

Hang Liu, Meng Chen, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2021. Conversational query rewriting with self-supervised learning. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7628–7632. IEEE.

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.

Quentin Macé, António Loison, and Manuel Faysse. 2025. Vidore benchmark v2: Raising the bar for visual retrieval. *Preprint*, arXiv:2505.17166.

Duy A Nguyen, Rishi Kesav Mohan, Van Yang, Pritom Saha Akash, and Kevin Chen-Chuan Chang. 2025. Rl-based query rewriting with distilled llm for online e-commerce systems. *arXiv preprint arXiv:2501.18056*.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. *CoRR*, abs/1611.09268.

Stephen Robertson, Hugo Zaragoza, and 1 others. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Yujing Wang, Hainan Zhang, Liang Pang, Binghui Guo, Hongwei Zheng, and Zhiming Zheng. 2025. Maferw: Query rewriting with multi-aspect feedbacks for retrieval-augmented large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25434–25442.

Haoran Wei, Chenglong Liu, Jinyue Chen, Jia Wang, Lingyu Kong, Yanming Xu, Zheng Ge, Liang Zhao, Jianjian Sun, Yuang Peng, and 1 others. 2024. General ocr theory: Towards ocr-2.0 via a unified end-to-end model.

Zhangchen Xu, Fengqing Jiang, Luyao Niu, Yuntian Deng, Radha Poovendran, Yejin Choi, and Bill Yuchen Lin. 2024. Magpie: Alignment data synthesis from scratch by prompting aligned llms with nothing. *Preprint*, arXiv:2406.08464.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

Nana Zhu, Xueting Li, Lei Xiong, and Han Xue. 2016. Query rewriting for archived information retrieval. In *Proceedings of the International Conference on Internet Multimedia Computing and Service*, pages 323–326.