Think Then Rewrite: Reasoning Enhanced Query Rewriting for Domain Specific Retrieval

Ang Li[1*], Yufei Shi[2*], Yuxuan Si[1], Yiquan Wu[3†], Ming Cai[1], Xu Tan[4]
Yi Wang[5], Changlong Sun[3], Xiaozhong Liu[6], Kun Kuang[1]
[1]College of Computer Science and Technology, Zhejiang Unirersity,
[2]The Hong Kong Polytechnic University,
[3]Guanghua Law School, Zhejiang University,
[4]Zhejiang University of Science and Technology,
[5] Chongqing Ant Consumer Finance Co,. Ltd , Ant Group,
[6]Worcester Polytechnic Institute, Worcester, USA
{leeyon, syx_sue, wuyiquan, cm, 11921173, kunkuang}@zju.edu.cn,
yufei1999.shi@connect.polyu.hk, tanxu@zust.edu.cn, haonan.wy@myxiaojin.cn, xliu14@wpi.edu

Abstract

Query rewriting is a crucial task for improving retrieval, especially in professional domains such as law and medicine, where user queries are often underspecified and ambiguous. While large language models (LLMs) offer strong understanding and generation capabilities, existing LLM-based approaches reduce the task to text transformation or expansion, neglecting reasoning to disambiguate queries, which fails to bridge the cognitive gap between user queries and specialized documents. In this paper, we propose Think-Then-Rewrite (TTR), a reinforcement learning based framework that unleashes LLMs' reasoning ability for domain-specific query rewriting. TTR introduces a contrastive mutual information reward to encourage the LLM to generate reasoning processes that effectively distinguish confusing distractors. To boost early-stage training, TTR also constructs golden query rewrites as off-policy data, providing strong guidance for RL learning. A mixed-policy optimization then combines on-policy and off-policy signals, ensuring both effectiveness and stability. Extensive experiments on legal and medical retrieval benchmarks demonstrate that TTR achieves state-of-the-art performance.

1 Introduction

Information Retrieval (IR) technologies are a cornerstone among data processing techniques when it comes to acquiring information (Manning, Raghavan, and Schütze 2008). Given an input query, retrieval aims to obtain relevant documents from external data collections (Kobayashi and Takeda 2000; Singhal 2001). Retrieval is critical in professional domains, including medicine (Nadkarni 2000), legal (Giri et al. 2017; Li et al. 2025b), and finance (Sumithra and Sridhar 2020) areas, where accurate information access directly impacts decision-making and outcomes.

According to Anomalous State of Knowledge (ASK) theory (Belkin, Oddy, and Brooks 1982), users often search under ambiguous knowledge, so retrieval systems should support intent exploration rather than merely matching explicit queries. This challenge is amplified in professional domains, where terminology is dense, conceptual structures are complex, and ambiguity is inherent, significantly degrading retrieval performance (Zheng et al. 2025a). To address this, query rewriting aims to better describe user information needs and ensure the accuracy of retrieval. Given the strong understanding and generation capabilities of large language models (LLMs), recent approaches rely on LLMs (Wei et al. 2022; Gao et al. 2023; Ma et al. 2023; Wang, Yang, and Wei 2023; Zheng et al. 2025a) to rewrite queries. However, they typically treat query rewriting as a direct text transformation or expansion, without modeling the user's cognitive gap or the underlying exploration trajectory explicitly. As shown in Fig. 1, when a query fails in direct retrieval, rewriting it via LLM prompting or supervised fine-tuning may enrich it with additional relevant details, but still fails to resolve key am-

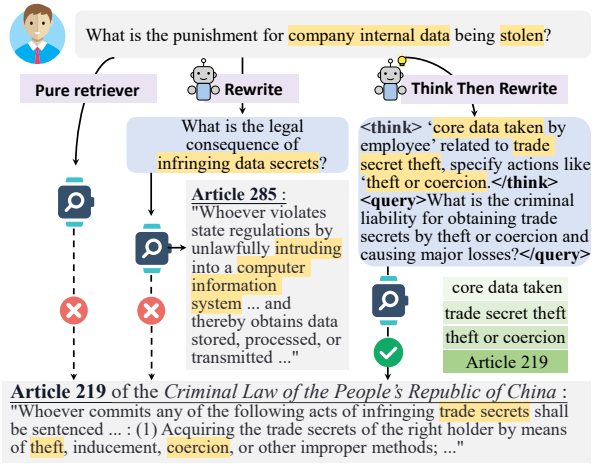

Figure 1: A real-world retrieval example of a legal query and its applicable law article, key information is highlighted in yellow and the reasoning chain is in green.

* These authors contributed equally.
† Corresponding authors.

biguities in user intent. In contrast, reasoning before rewriting gives the LLM room to explore and clarify the user's intent, helping bridge the gap between surface queries and true information needs through a progressively constructed reasoning chain. Therefore, we aim to unleash LLM's reasoning ability on the query rewriting for domain retrieval.

Reinforcement learning (RL) technology has shown strong potential in enhancing model reasoning (Havrilla et al. 2024), yet its application to the reasoning process preceding domain-specific query rewriting remains challenging. Traditional retrieval-based rewards such as Recall@k (Nogueira and Cho 2017; Wu et al. 2022) offer only coarse, outcome-oriented feedback, leaving the reasoning process unsupervised and overlooking distinctions between easily confused domain documents. This limitation is exacerbated when RL methods rely solely on-policy data, which weakens domain-specific reasoning in early training and increases the risk of spurious local optima.

In this paper, we propose Think-Then-Rewrite (TTR), a novel RL framework designed to enhance query rewriting for domain-specific retrieval. Specifically, we introduce two core techniques to tackle the challenges above. First, we reformulate the retrieval task as a mutual information maximization problem and propose the contrastive mutual information (CMI) reward, computed using a Noise Contrastive Estimation (Gutmann and Hyvärinen 2010) approach. In this formulation, the reward for each rewritten query is computed by contrasting its score for the relevant document with those for irrelevant ones, which are adaptively sampled to balance training stability and discriminative power. Second, to overcome early-stage local optima in RL training, we introduce off-policy guidance using golden rewrites with professional reasoning processes. These are generated by the reasoning LLM (e.g., DeepSeek-R1 (Guo et al. 2025)), conditioned on the relevant document, and filtered by our reward function to retain only high-scoring ones. Finally, following a mixed-policy RL paradigm (Yan et al. 2025), the model is optimized using both on-policy rewrites and off-policy golden rewrites. These are jointly supervised by the CMI reward, recall reward, and format reward, with sampling weights dynamically adjusted throughout training.

Extensive experiments on challenging domain-specific retrieval tasks, including legal and medical domains, demonstrate that our method TTR achieves state-of-the-art performance and consistently improves performance across both sparse and dense retrievers. We summarize the major contributions of the paper as follows:

- We reconceptualize query rewriting as a reasoning process to overcome the cognitive gap in domain retrieval.
- We introduce a reasoning-enhanced query rewriting method with contrastive mutual information reward and off-policy guidance, which provides fine-grained reward signals and enhances training stability.
- Our approach consistently outperforms baselines across two challenging domain-specific retrieval tasks and generalizes well across diverse retrievers and LLMs. All data and code are publicly available [1].

[1]https://github.com/LIANG-star177/TTR/tree/master

2 Related Work

2.1 LLM-based Query Rewriting

Query rewriting aims to refine ambiguous user queries in professional domains. With the rise of LLMs, early work focused on using them as external knowledge sources for query expansion. HyDE (Gao et al. 2023) instructs an LLM to generate a hypothetical document, which is then encoded into a vector to find similar real documents via semantic search. Similarly, Query2doc (Wang, Yang, and Wei 2023) generates a pseudo-document and expands the original query by concatenating it with the generated text. To directly optimize retrieval performance, later methods adopt reinforcement learning (RL). Rewrite-Retrieve-Read (Ma et al. 2023) uses RL to train a small rewriter model, which adapts to a black-box LLM reader by using its performance as a reward signal. DeepRetrieval (Jiang et al. 2025) applies RL to train LLMs for query generation by using retrieval metrics as direct rewards, eliminating the need for supervised data. CoEvo (Li et al. 2025a) further jointly optimizes LLM-rewritten queries and retrievers through an alternating coevolution framework. These works primarily focus on expanding queries or directly optimizing the rewritten text for retrieval, while our work focuses on the critical reasoning process that precedes the rewrite, leveraging an LLM's rich knowledge to fully extrapolate the user's intent.

2.2 Reinforcement Learning

Reinforcement learning (RL) has shown great success in enhancing the complex reasoning behaviors of LLMs. Proximal Policy Optimization (PPO) (Schulman et al. 2017) is a foundational on-policy algorithm that stabilizes training by using a clipped surrogate objective to constrain policy updates. Group Relative Policy Optimization (GRPO) (Shao et al. 2024) is a PPO that estimates relative advantages, significantly reducing training resources for reasoning tasks. DAPO (Yu et al. 2025) further refines this paradigm for large-scale RL, introducing techniques to address issues like entropy collapse and improve training stability. LUFFY (Yan et al. 2025) introduces a framework that augments on-policy RL with high-quality off-policy traces, helping the model overcome the limitations of its own exploration by learning from external demonstrations. Recent work also explores stepwise supervision with MCTS to enhance domain-specific reasoning capabilities (Liu et al. 2025). These works highlight a growing shift toward specialized optimization strategies for advancing LLM reasoning. Building on ASK theory and mutual information, our work proposes a novel RL framework, specifically tailored for query rewriting in domain-specific retrieval.

3 Preliminaries

Mutual Information. Mutual information quantifies the amount of information shared between two random variables. Formally, given variables X and Y, the mutual information $I(X;Y)$ is defined as:

$$I(X;Y) = \mathbb{E}_{p(x,y)}\left[\log(p(x,y)/p(x)p(y))\right]. \quad (1)$$

This definition has an intuitive interpretation: mutual information measures how much knowledge of one variable reduces the uncertainty about the other.

Mutual Information Estimation via InfoNCE. Direct computation of MI is often intractable in high-dimensional settings. A practical solution is InfoNCE (Oord, Li, and Vinyals 2018). Given a positive sample pair (x, y) and N negative samples $\{y_i^-\}$, InfoNCE defines the loss

$$\mathcal{L}_{\text{InfoNCE}} = -\mathbb{E}\Big[\log \frac{\exp(f(x,y))}{\sum_{y' \in \{y\} \cup y^-} \exp(f(x,y'))}\Big], \quad (2)$$

where $f(x, y)$ measures similarity (e.g., dot-product of embeddings). It provides a lower bound on MI:

$$I(X;Y) \geq \log(N) - \mathcal{L}_{\text{InfoNCE}}. \quad (3)$$

Intuitively, InfoNCE pulls positive pairs together while pushing apart negative pairs, effectively capturing the dependency between X and Y in a tractable way.

4 Methodology

4.1 Problem Setting

The goal of our task is as follows: given an original user-issued query q_0, which is often underspecified or colloquial, the objective is to generate a rewritten query q that is better aligned with the target corpus, thereby improving retrieval performance. To achieve this, the generation process is explicitly divided into two stages: a reasoning sequence $\mathbf{t}$, and a final rewritten query q, both generated by a model parameterized by θ:

$$P(\mathbf{t}, q \mid q_0; \theta) = P(\mathbf{t} \mid q_0; \theta) \cdot P(q \mid q_0, \mathbf{t}; \theta). \quad (4)$$

Once q is generated, it is passed to a retrieval module $\mathcal{R}$ to retrieve the most relevant document from a target collection $\mathcal{D}$, i.e., $\hat{d} = \mathcal{R}(q, \mathcal{D})$. The final objective is to maximize the likelihood that the top-ranked document $\hat{d}$ matches the gold document $d^* \in \mathcal{D}$. This setting is particularly important for domain-specific retrieval tasks, where effective retrieval requires careful reasoning beyond surface-level query reformulation due to the cognitive gap between colloquial user queries and formal domain documents.

4.2 Vanilla Implementation

Although instruction-tuned LLMs show promising zero-shot rewriting ability, their reasoning is often suboptimal for domain-specific tasks due to the lack of fine-tuning on reasoning-intensive retrieval instructions, while collecting large-scale human-annotated data for such supervision is costly and infeasible (Zhuang et al. 2025). To address this, we employ the Group Relative Policy Optimization (GRPO) RL algorithm (Shao et al. 2024), which unlocks the model's reasoning ability by leveraging reward signals derived directly from retrieval results.

We follow a straightforward reward combining Recall@k and format compliance (Nogueira and Cho 2017; Wu et al. 2022). A reward of 1 is assigned if the rewritten query retrieves the gold document d^* within the top-k results and the output adheres to the expected reasoning format (i.e., the tokens correctly fills in the <think></think> and <query></query> spans). Otherwise, a reward of 0 is given.

Formally, let $\pi_{\theta_{\text{old}}}$ and π_θ denote the policy before and after the RL update, both representing token-level distributions over rewritten queries. For each original query q_0, we sample N rewritten queries $\{q_i\}_{i=1}^N$ from $\pi_{\theta_{\text{old}}}$. The advantage of each sample is computed by normalizing rewards within the group:

$$A_i = \frac{R(q_i) - \text{mean}(\{R(q_j)\}_{j=1}^N)}{\text{std}(\{R(q_j)\}_{j=1}^N)}. \quad (5)$$

The policy is then updated using a PPO-style (Schulman et al. 2017) clipped objective:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \frac{1}{\sum_{i=1}^N |q_i|} \sum_{i=1}^N \sum_{t=1}^{|q_i|} \text{CLIP}(r_{i,t}(\theta), A_i, \epsilon) - \beta \cdot \mathbb{D}_{\text{KL}}[\pi_\theta \| \pi_{\text{ref}}], \quad (6)$$

where $r_{i,t}(\theta) = \pi_\theta(q_{i,t} \mid q_0, q_{i,<t}) / \pi_{\theta_{\text{old}}}(q_{i,t} \mid q_0, q_{i,<t})$ is the importance weight for each token, and $\text{CLIP}(\cdot)$ is the clipped surrogate loss. The KL penalty keeps the updated policy close to the original, stabilizing training.

Despite its simplicity, this vanilla setup has two limitations in query rewriting for domain-specific retrieval: (1) Recall@k offers coarse feedback that misses supervision over the reasoning process. (2) GRPO's relative reward design can assign advantages to uniformly poor rewrites early on, causing spurious local optima.

4.3 TTR Framework

To address the limitations above, we propose the **Think Then Rewrite** (TTR) framework. An overview is shown in Fig. 2. In the following, we elaborate on its two core designs: the CMI reward and off-policy guided learning.

Contrastive Mutual Information (CMI) Reward. Let's first revisit the objective of query rewriting tasks. The ultimate goal is to generate a rewritten query q that enhances retrieval performance, typically by increasing the likelihood of retrieving the relevant document d^*. Formally, this corresponds to maximizing the retrieval probability $P(d^* \mid q)$ under a fixed retriever. Let q_0 denote the original query, and let $q \sim p_\theta(q \mid q_0)$ be a rewritten query sampled from the query rewriting model. The retrieval objective can be expressed as:

$$\max_\theta \mathbb{E}_{(q_0, d^*) \sim D} \left[\mathbb{E}_{q \sim p_\theta(\cdot \mid q_0)} \left[\log P(d^* \mid q)\right]\right]. \quad (7)$$

Now, consider the mutual information between the rewritten query and the relevant document:

$$I(q; d^*) = \mathbb{E}_{q, d^*} \left[\log P(d^* \mid q)\right] - \mathbb{E}_{d^*} \left[\log P(d^*)\right]. \quad (8)$$

Since the marginal $P(d^*)$ is independent of the rewriting model parameters θ, maximizing $I(q; d^*)$ is equivalent to maximizing $\log P(d^* \mid q)$. Therefore, we can interpret the query rewriting objective as implicitly maximizing the mutual information between them:

$$\max_\theta I(q; d^*). \quad (9)$$

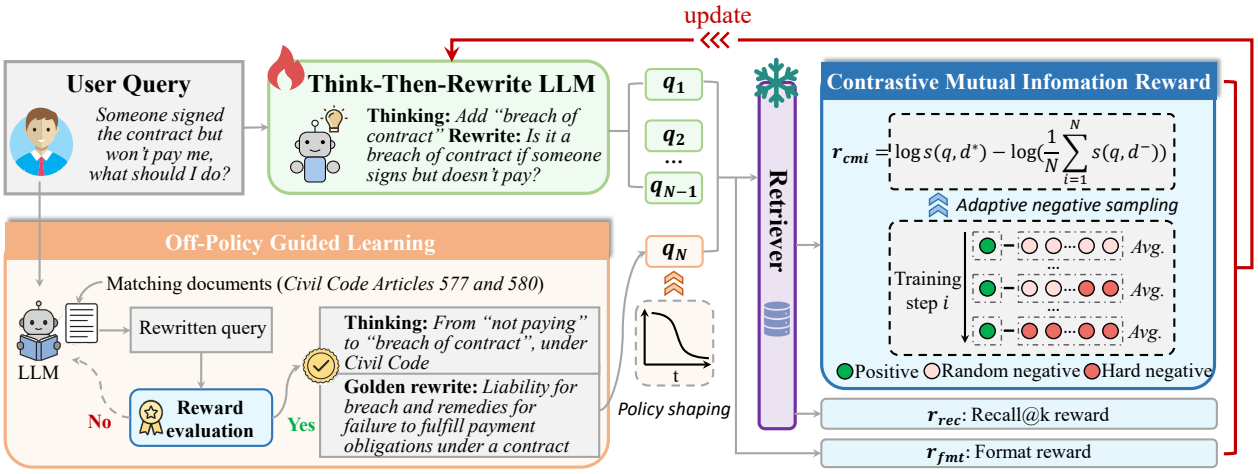

Figure 2: The architecture of TTR. It introduces two core designs beyond the vanilla implementation: (1) a fine-grained CMI reward that supervises not only the rewritten query but also the intermediate reasoning process; (2) an off-policy guided learning strategy that provides babysitting-style guidance in early training to help the model acquire domain-specific reasoning skills.

Building on the variational bound discussed in Sec. 3, we adopt a contrastive variational lower bound on mutual information, inspired by the InfoNCE formulation (Oord, Li, and Vinyals 2018). Let $s(q, d)$ denote the relevance score from a frozen retriever approximating the conditional likelihood of document d given rewritten query q. Then the mutual information $I(q; d^*)$ can be lower bounded as:

$$I(q; d^*) \geq \mathbb{E}\left[\log s(q, d^*) - \log\left(\frac{1}{N}\sum_{i=1}^{N} s(q, d_i^-)\right)\right], \tag{10}$$

where d^* is the target document and d_i^- are negative documents. The contrastive objective naturally defines our RL reward:

$$r_{\text{cmi}}(q) = \log s(q, d^*) - \log\left(\frac{1}{N}\sum_{i=1}^{N} s(q, d_i^-)\right). \tag{11}$$

To balance training stability and discrimination difficulty, we employ adaptive negative sampling. Let $\alpha_0 \in [0, 1]$ denote the initial proportion of BM25-based negatives, with $1 - \alpha_0$ for random negatives. At training step i out of total T, we compute the BM25 ratio as:

$$\alpha_i = \alpha_0 + (1 - 2\alpha_0) \cdot \frac{i}{T}. \tag{12}$$

At each step, we sample $N \cdot \alpha_i$ BM25-based negatives and $N \cdot (1 - \alpha_i)$ random negatives. Early training emphasizes diversity, while later training increases difficulty by focusing more on semantically relevant negatives.

According to Eq. 4, the reasoning sequence $\mathbf{t}$ directly influences the generation of the final query q. We therefore apply the CMI reward to the entire generation $\mathbf{x} = [\mathbf{t}; q]$, ensuring that retrieval signals are captured throughout the reasoning and rewriting process. We separately employ the recall reward and format reward in Sec. 4.2, both defined as binary indicators:

$$r_{\text{rec}}(q) = \mathbb{1}[d^* \in \mathcal{R}(q, \mathcal{D}, k)], \tag{13}$$

$$r_{\text{fmt}}(\mathbf{x}) = \mathbb{1}[f(\mathbf{x}) = 1], \tag{14}$$

where $f(\mathbf{x})$ checks whether the reasoning and rewritten query are correctly wrapped in the expected spans. The final reward combines all terms:

$$r_{\text{total}}(\mathbf{x}) = \lambda_{\text{cmi}} \cdot \sigma(r_{\text{cmi}}(\mathbf{x})) + \lambda_{\text{rec}} \cdot r_{\text{rec}}(q) + \lambda_{\text{fmt}} \cdot r_{\text{fmt}}(\mathbf{x}), \tag{15}$$

where λ_{MI}, λ_{rec}, λ_{fmt} are tunable weights. CMI reward keeps reasoning on-topic via a contrastive fine-grained signal, while recall and format rewards ensure correctness and structured reasoning, collectively reducing reward hacking.

Off-Policy Guided Learning. To address the local optima in GRPO, caused by early-stage models generating uniformly poor outputs, we construct off-policy golden rewrites to bootstrap policy training, which consists of three steps:

Step 1: Rewrite Generation. We leverage the strong reasoning LLM DeepSeek-R1 (Guo et al. 2025) to generate candidate rewrites. For each training pair (q_0, d^*), we prompt the model with both the original query and its corresponding relevant document to generate reasoning-driven rewrites:

$$q_c \sim P_{\text{LLM}}(\cdot \mid q_0, d^*). \tag{16}$$

Step 2: Reward-based Filtering. To ensure high-quality candidates for off-policy learning, we generate up to $M = 5$ rewrites per instance and select those meeting strict reward criteria. If no candidate satisfies all conditions, the relevant document d^* serves as the golden rewrite. Formally, the filtered candidate set is defined as:

$$\mathcal{Q}_{\text{ref}} = \{q_c \mid r_{\text{fmt}}(q_c) = 1, r_{\text{rec}}(q_c) = 1\}. \tag{17}$$

The curated rewrite pool $\mathcal{Q}_{\text{ref}}$ serves as off-policy golden rewrites in RL training. The pseudocode of the above construction process is shown in Alg. 1.

Step 3: Mixed-Policy RL. Following the mixed-policy GRPO (Yan et al. 2025), we integrate the golden rewrites into our training process through importance sampling from

Algorithm 1: Golden Rewrites Construction.

Input: Training pairs $\{(q_0, d^*)\}$, Max attempts M
Output: Off-policy rewrite set $\mathcal{Q}_{\text{ref}}$

Initialize $\mathcal{Q}_{\text{ref}} \leftarrow \emptyset$
for each training pair (q_0, d^*) **do**
  $rewrite \leftarrow$ null
  for $i = 1$ to M **do**
    $q_c^i \sim P_{\text{LLM}}(\cdot \mid q_0, d^*)$
    if $r_{\text{fmt}}(q_c^i) = 1, r_{\text{rec}}(q_c^i) = 1$ **then**
      $rewrite \leftarrow q_c^i$
      break
    end if
  end for
  if $rewrite =$ null **then**
    $rewrite \leftarrow d^*$
  end if
  $\mathcal{Q}_{\text{ref}} \leftarrow \mathcal{Q}_{\text{ref}} \cup \{rewrite\}$
end for

$\mathcal{Q}_{\text{ref}}$. Our objective function combines both on-policy and off-policy components:

$$\mathcal{J}_{\text{Mixed}}(\theta) = \frac{1}{Z}\left[\sum_{i=1}^{N_{\text{on}}} \text{CLIP}(r_{i,t}(\theta), A_{i,t}) + \sum_{j=1}^{N_{\text{off}}} \frac{\pi_\theta(q_j)}{\pi_{\text{off}}(q_j)} \text{CLIP}(r_{j,t}(\theta), A_{j,t})\right], \quad (18)$$

where N_{on} and N_{off} represent the number of on-policy and off-policy samples respectively, π_{off} denotes the off-policy reasoning model, and Z is a normalization factor. The importance ratio $\frac{\pi_\theta(q_j)}{\pi_{\text{off}}(q_j)}$ ensures proper weighting between on-policy and off-policy learning.

5 Experiments

5.1 Experimental Setup

Datasets. We conduct experiments on two retrieval benchmarks from the legal and medical domains. (1) The **EQUALS** (Chen et al. 2023) dataset consists of user queries and corresponding applicable law articles from a legal consultation system. We select the first substantive law article for each query to focus on core legal provisions. Following Zheng et al. (2025a), we filter data to include only query-document pairs with lexical overlap ratios below 10%, calculated as the proportion of overlapping tokens relative to the total query length after tokenization. (2) The **KUAKE-IR** (Long et al. 2022a) dataset is derived from the medical paragraph retrieval task in the CBLUE benchmark (Zhang et al. 2022), using data from real medical consultation scenarios of Alibaba Kuake Search. The input is the user query, and the output is the relevant medical content. We apply the same lexical overlap filtering criterion as EQUALS, retaining only pairs with overlap ratios below 10%. The statistics of these two datasets are shown in Tab. 1.

Type	EQUALS	KUAKE-IR
# of Training Samples	2683	13121
# of Test Samples	281	222
# of Document Base	17043	13343
Avg. # of Tokens in Query	17.14	12.19
Avg. # of Tokens in Document	90.44	59.43
Lexical Overlap Ratio (%)	3.92	3.93

Table 1: Statistics of the dataset.

Metrics. Following the evaluation approach employed in prior work (Long et al. 2022b), we evaluate the retrieval performance by Recall precision at top 1, 5, 10, 100 (Recall@1, Recall@5, Recall@10, Recall@100) and Mean Reciprocal Rank at 10 passages (MRR@10).

Baselines. We implement a comprehensive set of baselines covering two main categories: (1) **Prompting methods**, including Chain-of-Thought (Wei et al. 2022), Structured Reasoning (Zheng et al. 2025b), Hyde (Gao et al. 2023), and Query2doc (Wang, Yang, and Wei 2023). (2) **Training methods**, including supervised fine-tuning (SFT) and reinforcement learning (RL) approaches such as Rewrite-Retrieval-Read (Ma et al. 2023), DeepRetrieval (Jiang et al. 2025), DAPO (Yu et al. 2025), and LUFFY (Yan et al. 2025), which optimize query rewriting with different reward designs and policy learning strategies. The dataset and baseline details are presented in Appendix A.

5.2 Implementation Details

Our experiments are conducted on six A100 GPUs. For all training methods, we train for 3 epochs with a batch size of 8, learning rate of 1e-6. For GRPO, we use N=8 on-policy samples. Following Yan et al. (2025), we use N_{off}=1 off-policy and N_{on}=7 on-policy samples to ensure fairness. All main experimental results utilize BGE-M3 (Chen et al. 2024) as the retriever. Since rewards are computed based on retriever feedback, following Langchain Chachat (Liu et al. 2024), we employ FAISS to pre-vectorize the knowledge base, enabling efficient computation of similarity scores and recall@k metrics. Reward weights are set as $\lambda_{\text{fmt}} = 0.4$, while λ_{cmi} and λ_{rec} sum to 0.8, and their impact is further explored in Fig. 3. The LLM prompts are provided in Appendix B. All reported results represent the best hyperparameter configuration obtained through validation, where $\lambda_{\text{cmi}} = 0.4$ and $\alpha_0 = 0.3$. We repeat all experiments five times with different random seeds and report the average. We further apply the Fisher randomization test to verify statistical significance.

5.3 Experiments Results

Main Results. Tab. 2 presents the performance comparison of our method against various baselines. We obtain that: (1) Our method TTR significantly outperforms all baselines on both datasets. Remarkably, our TTR-trained 1.5B LLM achieves better retrieval through query rewriting than the far larger API LLM (Qwen2.5-turbo). (2) Retrieving professional documents with colloquial queries is highly challenging, i.e., the vanilla retriever achieves only 7.47 R@1

Model	EQUALS					KUAKE-IR				
	R@1	R@5	R@10	R@100	MRR@10	R@1	R@5	R@10	R@100	MRR@10
Pure Retriever	7.47	19.93	27.05	61.57	12.84	18.02	35.14	43.24	74.32	25.02
Prompting Methods (with Qwen2.5-turbo)										
Chain-of-Thought	10.32	23.49	33.10	67.26	16.27	18.92	29.73	38.74	72.07	23.99
Structured Reasoning	16.73	30.60	39.86	72.24	22.76	19.37	30.63	39.19	75.68	24.97
Hyde	11.74	30.25	37.37	71.17	19.02	14.86	30.18	37.39	77.03	21.33
Query2doc	17.08	34.88	45.20	71.53	24.87	19.37	35.14	42.34	78.38	26.07
Training Methods (with Qwen2.5-1.5B)										
Rewrite-Retrieval-Read	9.25	21.35	29.18	63.35	14.29	18.47	33.78	42.34	74.32	25.40
DeepRetrieval	9.61	23.49	32.38	62.99	15.76	15.32	31.08	39.19	73.42	22.05
DAPO	8.54	21.35	28.54	61.57	13.98	16.67	31.98	40.99	75.23	23.45
LUFFY	14.59	32.38	40.91	71.89	21.58	19.32	34.83	42.39	76.22	25.76
SFT	12.10	27.40	35.59	65.48	18.34	12.62	28.38	36.04	73.42	19.21
Vanilla Implementation	11.54	26.93	34.71	61.98	18.50	15.77	32.43	40.54	72.97	22.57
TTR	**21.35**	**38.43**	**48.62**	**75.09**	**28.60**	**21.41**	**36.13**	**46.49**	**78.92**	**30.15**

Table 2: Performance comparison of various retrieval methods. The best is **bolded**, the second best is underlined.

Models	R@1	R@5	R@10	R@100	MRR@10
Reward Design					
w/o r_{cmi}	13.88	34.88	43.26	70.11	23.32
w/o r_{fmt}	15.39	35.56	46.41	70.77	24.04
w/o r_{rec}	14.68	30.64	40.98	69.79	21.06
w/o ANS	18.89	36.42	47.18	73.92	28.13
w/ r_{sim}	16.90	36.78	46.54	72.57	26.09
Off-policy Learning					
w/o off	14.10	32.40	39.52	68.63	22.00
w/ doc-off	20.90	37.40	44.52	72.63	26.74
TTR	**21.35**	**38.43**	**48.62**	**75.09**	**28.60**

Table 3: Ablation study on different variants.

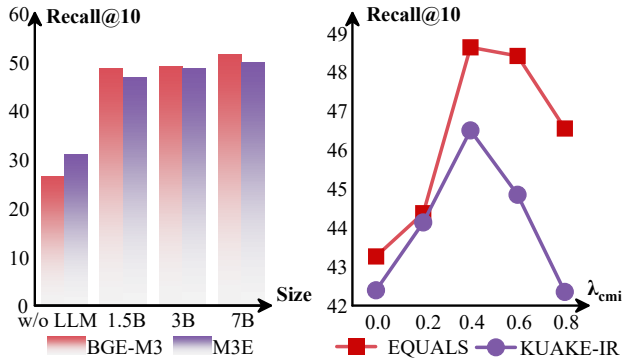

Figure 3: Performance of TTR across different LLM backbone sizes and retrievers.

on EQUALS. (3) Query2doc achieves second-best performance with moderate gains over vanilla retrieval, benefiting from powerful LLM APIs, but remains limited by the lack of targeted optimization from retrieval feedback. (4) Replacing RL algorithms (e.g., PPO, GRPO, DAPO) to train query rewriting models does not yield significant retrieval improvements. (5) LUFFY achieves strong performance, demonstrating the benefit of incorporating off-policy data. However, its coarse recall reward design still limits the overall potential. (6) SFT uses all of the off-policy data to learning reasoning processes but yields minimal improvement, as next-token prediction causes rigid overfitting. In contrast, RL-style signals better enable flexible reasoning. (7) Query rewriting yields weaker retrieval gains on medical data than legal data because medical paragraphs have more flexible structures than rigid law articles.

Ablation Studies. We compare five variants on reward design: **w/o** r_{cmi} removes CMI reward. **w/o** r_{fmt} removes format reward. **w/o** r_{rec} removes recall reward. **w/o ANS** removes adaptive negative sampling, only uses random negative. **w/** r_{sim} replaces CMI with similarity-based reward. We also compare two variants on off-policy learning: **w/o off** removes off-policy learning. **w/ doc-off** uses target documents as off-policy data directly instead of constructed golden rewrites. Results in Tab. 3 indicate: (1) Removing any reward leads to performance decline, with **w/o** r_{cmi} and **w/o** r_{rec} showing significant drops. This is because CMI reward provides fine-grained contrastive feedback to distinguish rewrites, while recall reward directly aligns with evaluation metrics. (2) **w/o ANS** causes performance degradation, demonstrating that dynamic negative selection is effective in CMI reward. (3) **w/** r_{sim} performs worse than TTR's reward design. Although similarity rewards are fine-grained, they strip away the contrastive structure, making the signal less informative. (4) **w/o off** performs worse than the full model, showing off-policy learning can promote knowledge-intensive reasoning in early stages. (5) **w/ doc-off** shows inferior performance. We infer that golden rewrites have smaller format differences with rewritten queries than raw

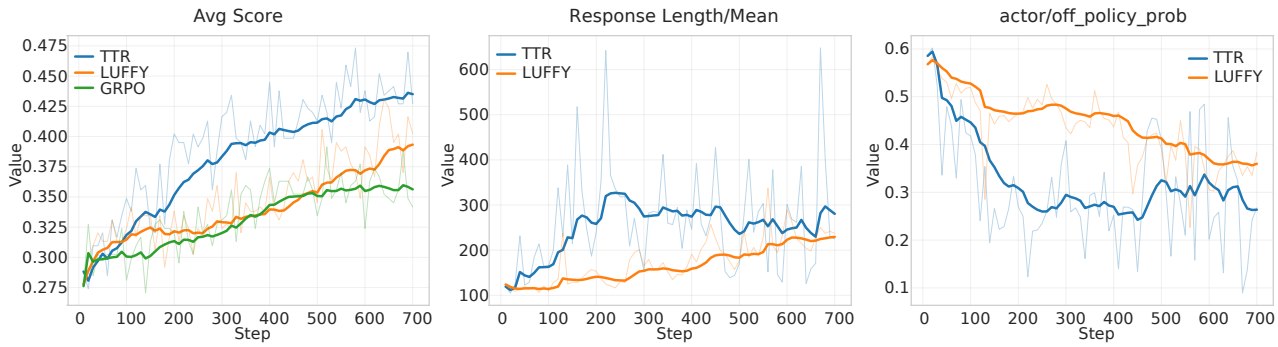

Figure 4: Training dynamics of TTR.

Query: Married for 9 years but got a fake marriage certificate. What legal liabilities?
Target document: Article 280 of the Criminal Law: Whoever forges ... official documents or certificates of a state organ shall be sentenced ...

Method	Rewritten query	Hit
Query2doc	Article 1049 of the Civil Code ...	✗
TTR	**<think>** 'Fake marriage certificate' is a criminal issue. Use terms like 'forge' and 'state organ documents' ... **</think>** **<query>** What is the criminal liability for forging and selling state organ documents? **</query>**	✓

Query: Persistent low-grade fever for a month, what could be the reason?
Target document: Reasons of long-term fever are tuberculosis, typhoid fever ...

Method	Rewritten query	Hit
Query2doc	A month-long low-grade fever may indicate tumors... Causes include: 1. Infectious diseases: such as brucellosis ...	✗
TTR	**<think>** 'persistent for a month' suggests a complex underlying cause. Focus specific diseases like 'tuberculosis' ... **</think>** **<query>** What causes long-term low-grade fever? Consider investigating tuberculosis and typhoid fever.**</query>**	✓

Figure 5: Case study.

documents, enabling more stable off-policy learning.

5.4 Additional Results

Robustness Analysis. We further test TTR across different LLM backbone sizes and retrievers in Fig. 3a. Results indicate that: (1) TTR consistently performs well across various LLM sizes, with larger backbones achieving better results, demonstrating that larger models possess stronger reasoning capacity and are better suited for domain query rewriting. (2) For both BGE-M3 and M3E retrievers, TTR substantially improves performance, confirming its flexibility in adapting to different retrievers by effectively leveraging CMI reward and off-policy learning.

Hyperparameter Exploration. We explore the effect of varying the weight of the reward combination in Fig. 3b. Results show that: (1) increasing the CMI weight initially improves performance but eventually leads to a decline, with the best result achieved at a weight of 0.4. (2) A weak CMI signal produces insufficiently fine-grained feedback, while an excessively strong signal suppresses the utility of the recall reward, highlighting the need for a balanced trade-off.

Training Dynamics. Fig. 4 shows TTR's training on the EQULAS dataset, tracking recall@10, reasoning length, and off-policy weights, and compares the results with baselines. The observations are as follows: (1) TTR consistently achieves higher recall rewards than LUFFY and GRPO, indicating its superior ability to rewrite queries for improved retrieval performance. (2) TTR generates longer reasoning chains, which provide sufficient space to leverage the LLM's internal knowledge and reasoning capacity; notably, TTR surpasses LUFFY significantly in the early phase, showing that off-policy learning effectively guides the reasoning process. (3) The off-policy data weights in TTR decrease more rapidly, suggesting that its on-policy data quality improves quickly and reduces dependence on off-policy data; this benefit stems from the CMI reward, which offers fine-grained feedback and clearer optimization directions.

Case Study. We present case studies on two datasets in Fig. 5: (1) In the legal dataset, Query2doc focused on civil aspects and missed the criminal nature of forging a marriage certificate. TTR identified the act as a criminal offense, used precise legal terms, and correctly retrieved the relevant Criminal Law article, demonstrating its ability to capture criminal liability. (2) In the medical dataset, Query2doc listed possible causes but ignored diagnostic uncertainty. TTR highlighted persistent fever and treatment ineffectiveness, emphasizing the need for thorough investigation rather than speculative diagnosis, and retrieved the correct guidance to seek comprehensive evaluation. These cases show that TTR enables nuanced and cautious query rewriting. Its strong reasoning supports dialectical analysis and sound judgment in high-stakes domains.

6 Conclusion

This work investigates the query rewriting task in domain retrieval, with experiments conducted on two professional datasets (law and medicine). Following ASK theory, we find that the reasoning process plays a central role in bridging the cognitive gap between queries and documents. To stimulate reasoning, we design a novel reward function from the perspective of mutual information and introduce off-policy learning to escape local optima. The proposed framework, TTR, demonstrates the ability to generate high-quality reasoning chains and achieve superior retrieval performance.

Acknowledgements

This work was supported in part by the National Key Research and Development Program of China (2024YFE0203700), the "Pioneer" and "Leading Goose" R&D Program of Zhejiang (2025C02037), and the National Natural Science Foundation of China (62376243, 62406287). It was also supported by the Key R&D Program of Hangzhou (2025SZDA0254), Ant Group, and Chongqing Ant Consumer Finance Co. All opinions in this paper are those of the authors and do not necessarily reflect the views of the funding agencies.

References

Belkin, N. J.; Oddy, R. N.; and Brooks, H. M. 1982. ASK for information retrieval: Part I. Background and theory. *Journal of documentation*, 38(2): 61–71.

Chen, A.; Yao, F.; Zhao, X.; Zhang, Y.; Sun, C.; Liu, Y.; and Shen, W. 2023. EQUALS: A Real-world Dataset for Legal Question Answering via Reading Chinese Laws. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, ICAIL '23, 71–80. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701979.

Chen, J.; Xiao, S.; Zhang, P.; Luo, K.; Lian, D.; and Liu, Z. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.

Gao, L.; Ma, X.; Lin, J.; and Callan, J. 2023. Precise zero-shot dense retrieval without relevance labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1762–1777.

Giri, R.; Porwal, Y.; Shukla, V.; Chadha, P.; and Kaushal, R. 2017. Approaches for information retrieval in legal documents. In *2017 Tenth International Conference on Contemporary Computing (IC3)*, 1–6. IEEE.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Gutmann, M.; and Hyvärinen, A. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 297–304. JMLR Workshop and Conference Proceedings.

Havrilla, A.; Du, Y.; Raparthy, S. C.; Nalmpantis, C.; Dwivedi-Yu, J.; Zhuravinskyi, M.; Hambro, E.; Sukhbaatar, S.; and Raileanu, R. 2024. Teaching large language models to reason with reinforcement learning. *arXiv preprint arXiv:2403.04642*.

Jiang, P.; Lin, J.; Cao, L.; Tian, R.; Kang, S.; Wang, Z.; Sun, J.; and Han, J. 2025. Deepretrieval: Hacking real search engines and retrievers with large language models via reinforcement learning. *arXiv preprint arXiv:2503.00223*.

Kobayashi, M.; and Takeda, K. 2000. Information retrieval on the web. *ACM Comput. Surv.*, 32(2): 144–173.

Li, A.; Wu, Y.; Hu, Y.; Qing, L.; Wang, S.; Liu, C.; Wu, T.; Jatowt, A.; Cai, M.; Wu, F.; and Kuang, K. 2025a. CoEvo: Coevolution of LLM and Retrieval Model for Domain-Specific Information Retrieval. In Christodoulopoulos, C.; Chakraborty, T.; Rose, C.; and Peng, V., eds., *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 14991–15010. Suzhou, China: Association for Computational Linguistics. ISBN 979-8-89176-332-6.

Li, A.; Wu, Y.; Liu, Y.; Cai, M.; Qing, L.; Wang, S.; Kang, Y.; Liu, C.; Wu, F.; and Kuang, K. 2025b. UniLR: Unleashing the Power of LLMs on Multiple Legal Tasks with a Unified Legal Retriever. In Che, W.; Nabende, J.; Shutova, E.; and Pilehvar, M. T., eds., *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 11953–11967. Vienna, Austria: Association for Computational Linguistics. ISBN 979-8-89176-251-0.

Liu, C.; Wang, S.; Qing, L.; Song, K.; Cao, J.; Lin, J.; Zhang, J.; Li, A.; Kuang, K.; and Wu, F. 2025. Towards Stepwise Domain Knowledge-Driven Reasoning Optimization and Reflection Improvement. arXiv:2504.09058.

Liu, Q.; Song, J.; Huang, Z.; Zhang, Y.; glide the; and liunux4odoo. 2024. langchain-chatchat. https://github.com/chatchat-space/Langchain-Chatchat.

Long, D.; Gao, Q.; Zou, K.; Xu, G.; Xie, P.; Guo, R.; Xu, J.; Jiang, G.; Xing, L.; and Yang, P. 2022a. Multi-CPR: A Multi Domain Chinese Dataset for Passage Retrieval.

Long, D.; Gao, Q.; Zou, K.; Xu, G.; Xie, P.; Guo, R.; Xu, J.; Jiang, G.; Xing, L.; and Yang, P. 2022b. Multi-cpr: A multi domain chinese dataset for passage retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 3046–3056.

Ma, X.; Gong, Y.; He, P.; Zhao, H.; and Duan, N. 2023. Query rewriting in retrieval-augmented large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 5303–5315.

Manning, C. D.; Raghavan, P.; and Schütze, H. 2008. *Introduction to Information Retrieval*. USA: Cambridge University Press. ISBN 0521865719.

Nadkarni, P. 2000. Information retrieval in medicine: Overview and applications. *Journal of Postgraduate Medicine*, 46(2): 116–122.

Nogueira, R.; and Cho, K. 2017. Task-Oriented Query Reformulation with Reinforcement Learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 574–583.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y.; Wu, Y.; et al. 2024. Deepseekmath:

Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Singhal, A. 2001. Modern Information Retrieval: A Brief Overview. *IEEE Data Eng. Bull.*, 24: 35–43.

Sumithra, M. K.; and Sridhar, R. 2020. Information retrieval in financial documents. In *Evolving Technologies for Computing, Communication and Smart World: Proceedings of ETCCS 2020*, 265–274. Springer.

Wang, L.; Yang, N.; and Wei, F. 2023. Query2doc: Query Expansion with Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 9414–9423.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.

Wu, Z.; Luan, Y.; Rashkin, H.; Reitter, D.; Hajishirzi, H.; Ostendorf, M.; and Tomar, G. S. 2022. CONQRR: Conversational Query Rewriting for Retrieval with Reinforcement Learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 10000–10014.

Yan, J.; Li, Y.; Hu, Z.; Wang, Z.; Cui, G.; Qu, X.; Cheng, Y.; and Zhang, Y. 2025. Learning to reason under off-policy guidance. *arXiv preprint arXiv:2504.14945*.

Yu, Q.; Zhang, Z.; Zhu, R.; Yuan, Y.; Zuo, X.; Yue, Y.; Dai, W.; Fan, T.; Liu, G.; Liu, L.; et al. 2025. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*.

Zhang, N.; Chen, M.; Bi, Z.; Liang, X.; Li, L.; Shang, X.; Yin, K.; Tan, C.; Xu, J.; Huang, F.; Si, L.; Ni, Y.; Xie, G.; Sui, Z.; Chang, B.; Zong, H.; Yuan, Z.; Li, L.; Yan, J.; Zan, H.; Zhang, K.; Tang, B.; and Chen, Q. 2022. CBLUE: A Chinese Biomedical Language Understanding Evaluation Benchmark. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 7888–7915. Dublin, Ireland: Association for Computational Linguistics.

Zheng, L.; Guha, N.; Arifov, J.; Zhang, S.; Skreta, M.; Manning, C. D.; Henderson, P.; and Ho, D. E. 2025a. A Reasoning-Focused Legal Retrieval Benchmark. In *Proceedings of the Symposium on Computer Science and Law on ZZZ*, CSLAW '25, 169–193. ACM.

Zheng, L.; Guha, N.; Arifov, J.; Zhang, S.; Skreta, M.; Manning, C. D.; Henderson, P.; and Ho, D. E. 2025b. A reasoning-focused legal retrieval benchmark. In *Proceedings of the 2025 Symposium on Computer Science and Law*, 169–193.

Zhuang, S.; Ma, X.; Koopman, B.; Lin, J.; and Zuccon, G. 2025. Rank-r1: Enhancing reasoning in llm-based document rerankers via reinforcement learning. *arXiv preprint arXiv:2503.06034*.