AdaRewriter: Unleashing the Power of Prompting-based Conversational Query Reformulation via Test-Time Adaptation

Anonymous ACL submission

Abstract

Prompting-based conversational query reformulation has emerged as a powerful approach for conversational search, refining ambiguous user queries into standalone search queries. Best-of-N reformulation over the generated candidates via prompting shows impressive potential scaling capability. However, both the previous tuning methods (training time) and adaptation approaches (test time) can not fully unleash their benefits. In this paper, we propose AdaRewriter, a novel framework for query reformulation using an outcome-supervised reward model via test-time adaptation. By training a lightweight reward model with contrastive ranking loss, AdaRewriter selects the most promising reformulation during inference. Notably, it can operate effectively in black-box systems, including commercial LLM APIs. Experiments on five conversational search datasets show that AdaRewriter significantly outperforms the existing methods across most settings, demonstrating the potential of test-time adaptation for conversational query reformulation.[1]

1 Introduction

The rapid advancement of Large Language Models (LLMs) has driven significant innovations in information retrieval (Zhao et al., 2023). Notably, conversational AI search engines (*e.g.*, `Perplexity` and `SearchGPT`) have attracted considerable attention due to their potential to shape the next generation of information retrieval (Mo et al., 2024b).

A fundamental challenge of conversational search is understanding user intent by considering the historical context and the current query, as user inputs are often vague, ambiguous, or incomplete (Gao et al., 2023; Mo et al., 2024b). Two types of approaches have been proposed to tackle this challenge: (1) Conversation dense retrieval involves training a dense encoder to generate conversational session embeddings (Lin et al., 2021b; Mo et al., 2023b, 2024c; Mao et al., 2024). However, it can not be compatible with sparse retrieval systems like BM25 and may suffer from limited interpretability (Cheng et al., 2024). (2) Conversational query reformulation is explored to derive the user's search intent by turning the conversational context and current query into a standalone query. With the advancement of LLMs, prompting-based query reformulation has emerged as a powerful way (Mao et al., 2023b; Ye et al., 2023; Mo et al., 2024a). Previous studies have demonstrated the strong capability of the reformulation candidates generated through prompting, which have impressive potential scaling capability (Mo et al., 2024a; Lai et al., 2025).

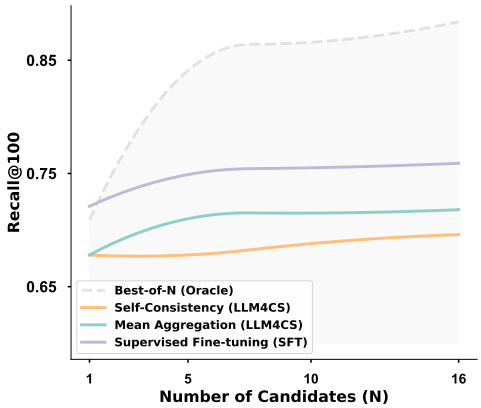

Figure 1: Comparison of training time and test-time adaptation strategies on the TopiOCQA using `LLaMA3.1-8B`. Best-of-N (Oracle) refers to prompting the model N times and selecting the best-performing reformulation result.

As illustrated in Figure 1, Best-of-N prompting-based reformulation demonstrates strong scalability. However, simply supervised fine-tuning on the best reformulation at the training time has not yielded consistent performance gains, as described in Sec 4.4. Another approach is to

[1]The code are available in `https://anonymous.4open.science/r/AdaRewriter-anonymous-3177/`

scale up during test time, leveraging increased computational resources to enhance model performance (Snell et al., 2024). Mao et al. (2023b) investigate mean aggregation and self-consistency strategy (Wang et al., 2023) during test time; they still exhibit a significant gap from the upper bound, as shown in Figure 1. This suggests the potential of test-time scaling has yet to be fully realized. Based on these empirical observations, a natural question arises: *How to design the appropriate* ***test-time scaling paradigm*** *to unleash the power of prompting-based query reformulation?*

In this work, we introduce **AdaRewriter**, leveraging an outcome-supervised reward model to unleash the power of prompting-based conversational query reformulation. Inspired by the effectiveness of the reward model at test time (Uesato et al., 2022; Shi et al., 2024), a lightweight, BERT-sized reward model is proposed and trained using a contrastive ranking loss as the reward of reformulation in CQR is implicit. During the inference stage, it serves as a scoring function to select the most promising reformulation. It should be pointed out that AdaRewriter can be seamlessly applied in black-box conversational search systems, particularly those utilizing commercial LLMs via API services.

AdaRewriter achieves excellent performance on five widely used conversation search datasets, including TopiOCQA (Adlakha et al., 2022), QReCC (Anantha et al., 2021), and TREC CAsT 2019, 2020 & 2021 (Dalton et al., 2020, 2021, 2022). Extensive experiments and analytical evaluations validate the effectiveness and robustness of AdaRewriter.

The contributions of this paper are threefold:

- To the best of our knowledge, we are the first to uncover and analyze the prompting-based query reformulation at test time under the Best-of-N paradigm.
- We propose AdaRewriter, a framework to unleash the power of prompting-based conversational query reformulation through an outcome-supervised reward model.
- Extensive experiments on several benchmark datasets demonstrate our proposed AdaRewriter outperforms existing methods across most settings, establishing its superiority in performance.

2 Preliminaries

2.1 Task Formulation

Conversational search systems aim to satisfy users' information-seeking needs in a multi-turn conversational form (Gao et al., 2023; Mo et al., 2024b). Formally, given the current query q^k and historical context $\mathrm{H}^{k-1} = \{q^i, r^i\}_{i=1}^{k-1}$, the objective of these systems is to generate responses using the passages set P^k retrieved by an off-the-shelf retrieval system, where k is the k-th turn of a conversation[2].

The conversational query reformulation task clarifies user intent by transforming the current query q and historical context H into a standalone query $\mathcal{S}$. Recent advancements in LLMs have made prompting-based CQR a promising approach, offering simplicity and superior performance. In this method, the reformulated query $\hat{q}$ and the pseudo-response $\hat{r}$ are generated by LLM based on the task instructions $\mathcal{I}$ and few-shots examples $\mathcal{D}$, where each example consists of the whole conversation history and human-written turn-level query reformulation:

$$\{\hat{q}, \hat{r}\} = \mathrm{LLM}(\mathcal{I}, \mathcal{D}, \{q, \mathrm{H}\}) \quad (1)$$

2.2 Potential of Best-of-N in CQR

Oracle We concatenate the reformulated query $\hat{q}$ with the pseudo-response $\hat{r}$ to form the reformulation query $\mathcal{S} = \hat{q} \oplus \hat{r}$, representing the user's search intent (Mo et al., 2023a). To fully explore the potential of multiple candidates, we generate a set of reformulation queries $\{\mathcal{S}_1, \ldots, \mathcal{S}_N\}$ and evaluate them using the Best-of-N paradigm, aiming to investigate the upper bound performance based on gold passage labels. Figure 1 presents our preliminary results, indicating that the number of candidates improves performance.

Training Time Fine-tuning Supervised fine-tuning(SFT) with the best-performing oracle reformulation via rejection sampling is a straightforward approach to further enhance the performance of prompting-based query reformulation. However, it does not consistently lead to performance gains based on our practices, as shown in Sec 4.4.

Test Time Adaptation Previous work (Mao et al., 2023b) proposes a simple yet effective method that generates multiple candidates query-response pairs $\{\hat{q}_1, \hat{r}_1\}, \{\hat{q}_2, \hat{r}_2\}, \ldots, \{\hat{q}_N, \hat{r}_N\}$

[2]For sake of convenience, we omit the superscript k in the following sections.

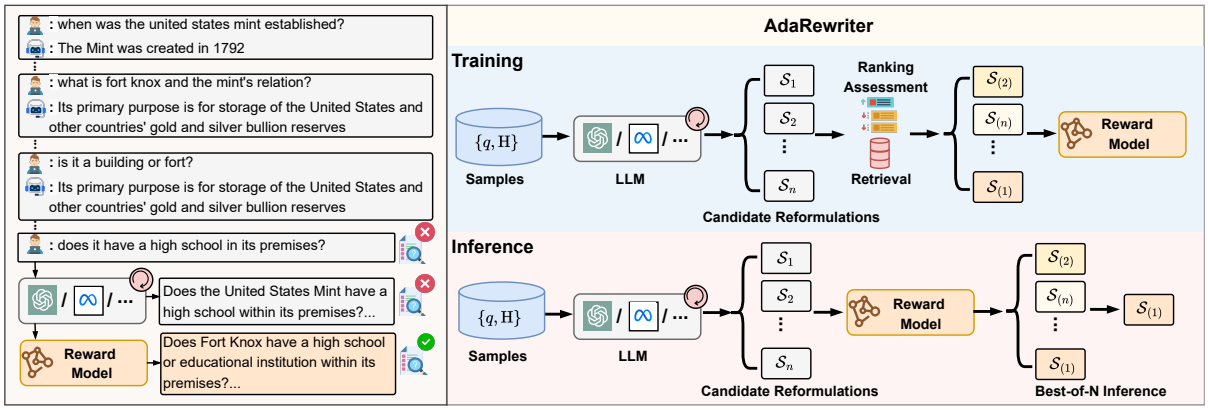

Figure 2: Overview of AdaRewriter.

and obtain the aggregated representation s in embedding space. Subsequently, the aggregated representation s, treated as the standalone query $\mathcal{S}$, is utilized in dense retrieval systems to retrieve relevant passages. However, this method and self-consistency do not consistently lead to performance gains as the number of candidates increases, as shown in Figure 1.

This motivates us to investigate prompting-based query reformulation further from the Best-of-N perspective. Building on these insights and recent advancements in test-time scaling, we propose AdaRewriter, which leverages an outcome-supervised reward model to unleash the full potential of prompting-based query reformulation.

3 Methodology

To uncover the potential of prompting-based query reformulation under the Best-of-N paradigm, we propose AdaRewriter as presented in Figure 2. Specifically, we leveraged a vanilla LLM to generate reformulation candidates and construct implicit reward signals to train the reward model based on end-to-end performance assessment, as detailed in §3.1. §3.2 introduces the improved prompting-based query reformulation approach under the Best-of-N paradigm during inference.

3.1 Reward Model Training

Constrative Ranking Loss Unlike traditional outcome-based methods that rely on binary classification labels, training a reward model for conversational query reformulation is non-trivial due to the absence of binary evaluation metrics in conversational search reformulation[3]. Without explicit reward, we leverage contrastive ranking loss, which is well-suited for tasks where relative ordering signals are much easier to obtain (Liu and Liu, 2021; Chuang et al., 2023). Specifically, the loss function targets to assign higher scores to top-ranked reformulations and lower scores to bottom-ranked ones:

$$\mathcal{L} = \sum_{i=1}^{n} \sum_{j>i} \max(0, r_j - r_i + (j-i) \times \lambda) \quad (2)$$

where r_i is the score of candidate reformulation $\mathcal{S}_i$ with rank i assigned by the trained reward model, λ is a hyperparameter controls the margin between the candidates. Despite the lack of explicit labels, this loss function can effectively optimize the model to distinguish the most promising reformulation $\mathcal{S}$ based on the assigned score among candidate reformulations.

Candidates Generation To construct candidate reformulations $\{\mathcal{S}_1, \mathcal{S}_2, \cdots, \mathcal{S}_n\}$ described in Eq. (2), an vanilla LLM is employed, which generate multiple candidtates $\{S_{(1)}, S_{(2)}, \cdots, \mathcal{S}_{(n)}\}$ conditioned on a conversational session $\{q, \mathrm{H}\}$. The generation process is guided by instructions $\mathcal{I}$ and few-shot examples $\mathcal{D}$:

$$\{S_{(1)}, S_{(2)}, \cdots, \mathcal{S}_{(n)}\} = \mathrm{LLM}(\mathcal{I}, \mathcal{D}, \{q, \mathrm{H}\}) \quad (3)$$

Ranking Assessment To rank the candidates, we utilize an end-to-end scoring function that combines multiple factors into a fusion score (Cormack et al., 2009; Lai et al., 2025):

$$\mathrm{M}(\mathcal{S}_{(i)}) = \frac{1}{r_s(\mathcal{S}_{(i)}, p)} + \frac{1}{r_d(\mathcal{S}_{(i)}, p)} \quad (4)$$

where $r_s(\mathcal{S}_{(i)}, p)$ denotes the corresponding rank with the gold passage p giving query $\mathcal{S}_{(i)}$ in a dense

[3]We considered from end-to-end retrieval performance, as human-written labels are labor-intensive to collect and not always lead to the best performance.

retrieval system, and $r_s(\mathcal{S}_{(i)}, p)$ represents the rank in a sparse retrieval system. The candidate reformulation $\mathcal{S}_{(i)}$ is subsequently assigned a rank j based on its performance according to the metric in Eq. (4), with higher ranks corresponding to better performance.

Therefore, the trained outcome-supervised reward model g_θ based on a BERT-sized model can be trained by the contrastive ranking Loss. It can assess the quality of query $\mathcal{S}$ generated by LLM conditioned on a conversational session $\{q, \mathrm{H}\}$ and return a score r:

$$r = g_\theta(\mathcal{S}, \{q, \mathrm{H}\}) \tag{5}$$

3.2 Best-of-N Inference

Leveraging the outcome-supervised reward model g_θ, our framework functions as a plug-and-play module to enhance prompting-based CQR during inference, adhering to the Best-of-N paradigm. Owing to test-time scalability, this module can be seamlessly integrated into a wide range of conversational search systems, regardless of whether the underlying large language model is deployed locally or accessed through commercial API services.

Specifically, given a conversational session $\{q, \mathrm{H}\}$, the LLM generates multiple reformulation candidates $\{S_{(1)}, S_{(2)}, \cdots, \mathcal{S}_{(N)}\}$, as described in Eq. (3), where N is the budget parameter that is adjustable during inference. The reward model g_θ then assigns scores to each candidate, and the highest-scoring candidate is selected as the most promising reformulation $\mathcal{S}$:

$$\mathcal{S} \leftarrow \mathcal{S}_{(k)}, k = \underset{j=1,\cdots,N}{\arg\max}\, g_\theta(\mathcal{S}_{(j)}, \{q, \mathrm{H}\}) \tag{6}$$

The selected reformulation $\mathcal{S}$ is subsequently treated as the refined representation of the user's intent, leveraging the enhanced reasoning capabilities unlocked by our framework. The reformulation is then used to retrieve relevant passages, thereby improving the performance of conversational search systems.

4 Experiments

Datasets & Evaluation Metrics The training data for the outcome-supervised reward model is derived from two widely used conversational search datasets: TopiOCQA (Adlakha et al., 2022) and QReCC (Anantha et al., 2021). For evaluation, we use the test sets of TopiOCQA and QReCC. Additionally, to assess the zero-shot reformulation performance of our method, we conduct experiments on the TREC CAsT 2019, 2020, and 2021 datasets (Dalton et al., 2020, 2021, 2022). To evaluate the reformulation results, we adopt four standard metrics from information retrieval: MRR, NDCG@3, and Recall@10, which align with previous studies (Dalton et al., 2021; Yu et al., 2021; Mo et al., 2023a). Metric computation uses the `pytrec_eval` tool (Van Gysel and de Rijke, 2018). Further details about the datasets can be found in the Appendix B.1.

Implementation Details In our prompting-based conversational query reformulation approach, we adopt the prompt used in Mao et al. (2023b), specifically the "rewrite-and-response" setting with chain-of-thought, which represents the most advanced configuration. For the backbone selection in Sec 3.1, we utilize `Llama2-7B` and `Llama3.1-8B` with a candidate size of $N = 16$ and a temperature setting of 0.7, in line with previous studies (Mao et al., 2023b; Mo et al., 2024a). The outcome-supervised reward model is based on a lightweight BERT variant, `deberta-v3-base`. For retrieval, we employ BM25 (Robertson et al., 2009) for sparse retrieval and ANCE (Xiong et al., 2020) for dense retrieval, consistent with prior work (Mo et al., 2023a; Mao et al., 2023b). The margin parameter λ in Eq. (2) is set to 0.1, determined through grid search. Further details about the implementation can be found in the Appendix B.2.

4.1 Baselines

We conducted the primary experiments utilizing open-source large language models (LLMs) `Llama2-7B` and `Llama3.1-8B` to demonstrate the effectiveness of AdaRewriter.

Our approach is compared with various conversational query reformulation frameworks, which can be categorized into **fine-tuning** and **prompting-based** methods. The fine-tuning-based methods include T5QR (Lin et al., 2020), CONQRR (Wu et al., 2022), EDIRCS (Mao et al., 2023a), ConvGQR (Mo et al., 2023a), IterCQR (Jang et al., 2024), RetPO (Yoon et al., 2024), and AdaCQR (Lai et al., 2025), while the prompting-based methods comprise LLM-Aided (Ye et al., 2023), CHIQ (Mo et al., 2024a), and LLM4CS (Mao et al., 2023b). Following Mo et al. (2024a), we also compare with the framework that fine-tuned LLM-based retrievers, including RepLLama (Ma et al., 2024), E5-Mistral (Wang

Type	Framework	Backbone	TopiOCQA MRR	TopiOCQA NDCG@3	TopiOCQA R@10	QReCC MRR	QReCC NDCG@3	QReCC R@10
Sparse (BM25)	T5QR	T5-base	11.3	9.8	22.1	33.4	30.2	53.8
	CONQRR	T5-base	-	-	-	38.3	-	60.1
	EDIRCS	T5-base	-	-	-	41.2	-	62.7
	ConvGQR	T5-base	12.4	10.7	23.8	44.1	41.0	64.4
	IterCQR	T5-base	16.5	14.9	29.3	46.7	44.1	64.4
	AdaCQR	T5-base	17.8	15.8	34.1	52.4	49.9	70.9
	RETPO	Llama2-7B	28.3	26.5	48.3	50.0	47.3	69.5
	AdaCQR+Expansion	Llama2-7B*	28.3	26.5	48.9	55.1	52.5	76.5
	LLM-Aided	GPT3.5-Turbo	-	-	-	49.4	46.5	67.1
	CHIQ-AD	Llama2-7B	22.5	20.5	40.4	53.1	50.7	77.2
	CHIQ-Fusion	Llama2-7B*	25.6	23.5	44.7	54.3	51.9	78.5
	LLM4CS	Llama3.1-8B	24.5	22.6	42.1	49.7	46.9	73.8
	AdaRewriter (N=5)	Llama3.1-8B	28.2	26.2	48.3	54.0	51.3	77.4
	AdaRewriter (N=16)	Llama2-7B	27.8	25.9	47.6	55.2	52.8	78.0
	AdaRewriter (N=16)	Llama3.1-8B	**30.7**†	**28.8**†	**51.3**†	**56.2**†	**53.8**†	**78.8**†
Dense (ANCE)	T5QR	T5-base	23.0	22.2	37.6	34.5	31.8	53.1
	CONQRR	T5-base	-	-	-	41.8	-	65.1
	EDIRCS	T5-base	-	-	-	42.1	-	65.6
	IterCQR	T5-base	26.3	25.1	42.6	42.9	40.2	65.5
	ConvGQR	T5-base	25.6	24.3	41.8	42.0	39.1	63.5
	AdaCQR	T5-base	32.8	31.5	54.6	45.1	42.4	66.3
	RETPO	Llama2-7B	30.0	28.9	49.6	44.0	41.1	66.7
	AdaCQR+Expansion	Llama2-7B*	38.5	37.6	58.4	45.8	42.9	67.3
	LLM-Aided	GPT3.5-Turbo	-	-	-	43.5	41.3	65.6
	CHIQ-AD	Llama2-7B	33.2	32.2	53.0	47.0	44.6	70.8
	CHIQ-Fusion	Llama2-7B*	38.0	37.0	61.6	47.2	44.2	**70.7**
	LLM4CS(N=5)	Llama3.1-8B	34.6	33.5	54.3	42.6	40.0	64.0
	LLM4CS(N=16)	Llama2-7B	33.5	33.1	53.0	43.0	40.5	64.8
	LLM4CS(N=16)	Llama3.1-8B	35.4	34.5	55.1	43.2	40.7	64.6
	AdaRewriter (N=5)	Llama3.1-8B	38.9	37.9	59.6	46.1	43.4	69.2
	AdaRewriter (N=16)	Llama2-7B	38.2	37.1	58.0	47.2	44.4	69.0
	AdaRewriter (N=16)	Llama3.1-8B	**40.3**†	**39.7**†	**61.9**†	**47.5**	**44.7**†	69.8

Table 1: Evaluation results of various retrieval system types on the QReCC and TopiOCQA. The best results among all methods are **bolded**, and the second-best results are underlined. * denotes including fused results from a trained T5-based model. † denotes significant improvements with t-test at $p < 0.05$ over all compared baselines.

et al., 2024), and LLM-Embedder (Zhang et al., 2023). Additionally, we reproduce LLM4CS with the same LLM backbones of our method, using varying budget parameters N, to facilitate a fair and comprehensive comparison.

The Appendix C presents comprehensive details of all the baseline methods. We also include the comparison with the Conversational Dense Retrieval(CDR) methods in Appendix A.3.

4.2 Main Results

We evaluate our method on two benchmarks, TopiOCQA and QReCC, under both sparse and dense retrieval settings. As shown in Table 1, AdaRewriter consistently outperforms baseline models across almost all scenarios.

On TopiOCQA with sparse retrieval, AdaRewriter (N=16) achieves MRR of 30.7, significantly outperforming LLM4CS's 24.5. In the dense setting (ANCE), it also surpasses LLM4CS with an MRR of 40.3 vs. 35.4. Performance further improves with larger candidate sets. For example, on QReCC (sparse), MRR increases from 54.0 (N=5) to 56.2 (N=16). This suggests that AdaRewriter effectively utilizes candidate reformulations, thereby enhancing the model's ability to select the most promising one. Similar trends are observed on the `Llama2-7B`.

Overall, AdaRewriter demonstrates strong adaptability to different retrieval conditions and benefits from scaling the number of candidate reformulations, offering an advantage in tasks requiring broader data exploration.

4.3 Zero-shot Results

In the zero-shot experiments conducted on the TREC CAsT 2019, 2020, and 2021 datasets, our proposed AdaRewriter consistently outperforms existing baselines across various budget parameters N, as shown in Table 2.

Specifically, AdaRewriter achieves significant improvements on most metrics across all three

Framework	Backbone	CAsT-19 NDCG@3	CAsT-19 R@10	CAsT-20 MRR	CAsT-20 NDCG@3	CAsT-20 R@10	CAsT-21 NDCG@3	CAsT-21 R@10
T5QR	T5-base	41.7	-	42.3	29.9	-	33.0	-
ConvGQR	T5-base	43.4	-	46.5	33.1	-	27.3	-
RepLLama	Llama2-7B	31.6	10.6	26.8	18.3	10.4	32.7	19.6
E5-Mistral	Mistral2-7B	31.3	9.5	22.0	15.4	8.4	32.5	20.5
LLM-Embedder	Llama2-7B	36.6	11.4	25.2	15.4	8.7	31.2	17.3
AdaCQR+Expansion	Llama2-7B*	48.5	**13.0**	56.6	38.5	19.2	45.6	25.0
CHIQ-Fusion	Llama2-7B*	**50.5**	12.9	54.0	38.0	19.3	46.5	25.2
LLM4CS (N=5)	Llama3.1-8B	44.4	11.5	61.7	44.8	23.0	50.5	25.7
LLM4CS (N=10)	Llama3.1-8B	45.5	11.9	61.9	46.0	23.2	**51.5**	25.8
AdaRewriter (N=5)	Llama3.1-8B	46.6	12.6	62.0	45.6	22.6	49.5	26.5
AdaRewriter (N=10)	Llama2-7B	48.0	12.7	59.3	44.5	20.2	47.7	25.9
AdaRewriter (N=10)	Llama3.1-8B	48.3	**13.0**	**63.0**†	**46.5**†	21.6	49.7	**27.2**†

Table 2: Zero-shot experiment results on TREC CAsT 2019, 2020 & 2021 datasets. The best results among all methods with similar settings are **bolded**, and the second-best results are underlined. * denotes including fused results from a trained T5-based model. † denotes significant improvements with t-test at $p < 0.05$ over all compared baselines.

datasets. For CAsT 2021, AdaRewriter yields strong gains in R@10, although its NDCG@3 performance is slightly lower. Despite this, our framework continues to exhibit considerable strength and robustness, confirming its capability to excel in retrieval performance and highlighting its robustness and adaptability across various datasets.

4.4 Comparison with Training-time Tuning

To fully investigate the benefit of test-time adaptation, we compare our proposed AdaRewriter with two strong training-time baselines: supervised fine-tuning (SFT) and direct preference optimization(DPO) (Rafailov et al., 2023). All methods generate $N = 16$ candidate reformulations on the TopiOCQA dataset for a fair comparison. SFT employs rejection sampling by selecting the best-performing candidates for fine-tuning. DPO treats the best and worst candidates as chosen and rejected samples, respectively.

As shown in Table 3, AdaRewriter consistently outperforms the strong baselines in the datasets. Notably, on CAsT 2020, it achieves an MRR of 63.0, compared to 59.1 for SFT and 60.7 for DPO, demonstrating its robustness, especially on out-of-domain data. These results highlight the effectiveness of test-time adaptation and confirm AdaRewriter's advantage in generating more relevant query reformulations. We provide some details for the setup of SFT and DPO in the Appendix B.3.

	TopiOCQA MRR	TopiOCQA R@10	CAsT 19 R@10	CAsT 20 R@10
SFT	39.2	59.4	70.0	59.1
DPO	39.1	59.8	66.4	60.7
AdaRewriter	40.3	61.9	71.4	63.0

Table 3: Comparison with Training-time Tuning

5 Analysis

In this section, we present a series of comprehensive experiments that aim to provide an in-depth analysis of the proposed AdaRewriter. Specifically, we investigate its effectiveness in addressing the following Research Questions (**RQs**):

- **RQ1:** Can AdaRewriter be applied to black-box commercial LLMs?
- **RQ2:** Does the conversational context H influence the score assigned to a reformulation query $\mathcal{S}$?
- **RQ3:** How do the components (*e.g.*, ranking loss, ranking assessment) impact the learning objectives of AdaRewriter?
- **RQ4:** Does AdaRewriter enhance the robustness of CQR in long conversations?

We also provide further discussions in Appendix A.

5.1 Adaptation in Black-Box Models

Building on the concept of test-time adaptation, our proposed AdaRewriter framework seamlessly integrates with conversational search systems that leverage commercial black-box LLMs, particularly those utilizing API services.

To answer **RQ1**, Figure 3 presents evaluation results on the TopiOCQA, QReCC, and zero-shot datasets to validate AdaRewriter's effectiveness. Experimental results show that AdaRewriter consistently enhances the performance of commercial LLMs, such as GPT4o-mini, across most evaluation metrics, even when trained on data generated

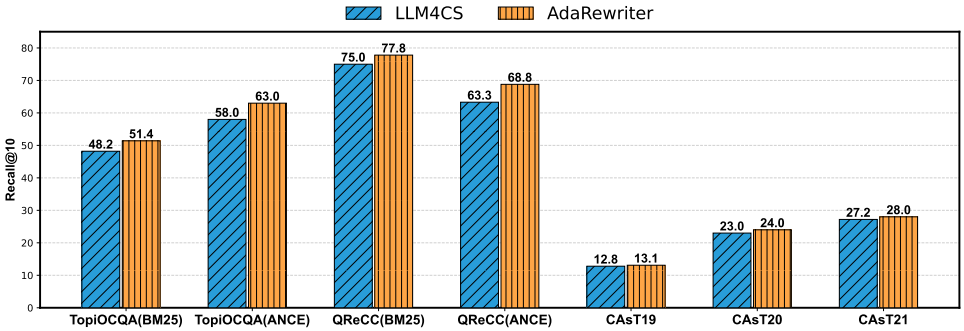

Figure 3: Performance comparsion on black-box model `GPT4o-mini`. We use $N = 5$ for inference.

by open-source LLMs. For instance, compared to the baseline, AdaRewriter boosts the R@10 from 48.2 to 51.4 in sparse retrieval and from 58.0 to 63.0 in dense retrieval on the TopiCOQA dataset. Additionally, our framework demonstrates robust improvements on zero-shot datasets using commercial LLMs, as shown in Figure 3.

These results prove that AdaRewriter effectively boosts the commercial LLMs like GPT4o-mini, even with training data from open-source models, highlighting the robustness and promise of test-time adaptation for conversational query reformulation.

5.2 Contextual Dependency in Scoring

To investigate **RQ2**, we begin by examining the relationship between conversational history and reformulation query scoring. In conversational search systems, the meaning and relevance of a query can vary significantly depending on the context in which it is presented. Specifically, the conversational context H provides essential information about the ongoing conversation, such as user intent and topics, which may influence how a reformulated query is assessed.

To assess the impact of context H in our proposed framework, we conduct an ablation study in Table 4 (w/o. Context H), in which the conversational context H is removed from the outcome-supervised reward model during both training and inference. The results reveal a significant drop in model performance when the context is excluded, showing the pivotal role of conversational context in guiding the outcome-supervised reward model's scoring of reformulated queries.

5.3 Influence of the Learning Objective

To investigate the individual contributions of our reward model's learning objectives as addressed in **RQ3**, we conduct an ablation study.

Specifically, we evaluate two variants: (1) w/o Ranking Loss , where the ranking loss is replaced by a cross-entropy loss assigning the true label the top rank and the false label to the bottom; and (2) w/o Ranking Assessment , where candidate reformulations are randomly ordered instead of ranked.

Type	Abaltion Variants	MRR	R@10
Sparse	AdaRewriter (*Ours*)	**30.7**	**51.3**
	w/o. Context H	27.3	44.9
	w/o. Ranking Loss	24.6	43.0
	w/o. Rank Assessment	23.8	41.8
Dense	AdaRewriter (*Ours*)	**40.3**	**61.9**
	w/o. Context H	36.2	56.4
	w/o. Ranking Loss	34.4	53.2
	w/o. Ranking Assessment	32.8	51.5

Table 4: Ablation study for the learning objective and contextual dependency of AdaRewriter on TopiOCQA dataset. We use `LLama3.1-8B` and $N = 16$ for inference.

Table 4 shows the results of these variants. Notably, the MRR in the dense retrieval drops from 40.3 to 34.4 when the ranking loss is removed, and also decreases to 32.8 when the ranking assessment is omitted. These findings demonstrate that both the contrastive loss and the ranking assessment are crucial for achieving strong performance, highlighting the importance of our proposed learning objectives for the reward model.

5.4 Robustness in Long Conversation

One of the primary challenges in conversational search systems is sustaining performance in extended conversation, as highlighted by **RQ4**. To answer this question, we assess the robustness of our proposed method across three datasets, which include TopiOCQA, QReCC, and TREC CAsT 2020. The results, presented in Figure 4, reveal that as the length of the conversation increases, performance across all methods experiences a notable decline.

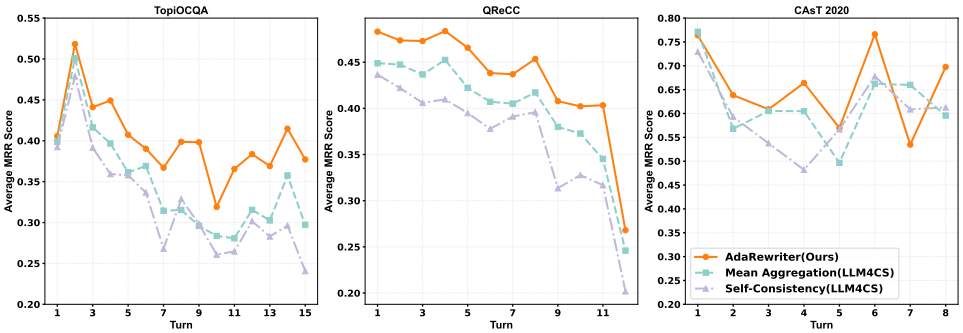

Figure 4: Turn-round performance comparison on TopiOCQA, QReCC, and TREC CAsT 2020.

This suggests that long conversations still present a challenge for current CQR methods.

Despite this general decline in performance, AdaRewriter consistently outperforms the other baselines across all conversation turns. Notably, even as the dialogue length increases, AdaRewriter maintains a higher performance compared to Mean Aggregation and Self-Consistency proposed by Mao et al. (2023b), which demonstrates a more substantial drop in effectiveness. This behavior suggests that AdaRewriter is more robust to the degradation typically observed in long conversations.

6 Related Works

Conversational Query Reformulation Query reformulation plays a crucial role in conversational search systems, addressing the inherent complexity of user intent, which often involves semantic challenges such as anaphora and ellipsis (Gao et al., 2023; Mo et al., 2024b). Current conversational query reformulation adopts hybrid approaches that combine query rewriting and query expansion, as exemplified by Mo et al. (2023a). In the era of LLMs, prompting-based query reformulation has garnered significant attention due to its simplicity and superior performance. Ye et al. (2023) treats LLMs as both query rewriters and rewrite editors, following a "*rewrite-then-edit*" paradigm to refine reformulations. Mao et al. (2023b) further explores advanced prompting strategies, such as few-shot learning, chain-of-thought reasoning, and self-consistency, demonstrating the remarkable efficacy of prompting-based approaches. Building on these developments, Mo et al. (2024a) proposed a two-step method that leverages the basic capabilities of open-source LLMs to enhance the conversational history for conducting query reformulation.

Test-time Supervision and Scaling Enhancing LLMs through test-time supervision and scaling test-time computation represents a promising direction for building robust and self-improving agent systems (Snell et al., 2024). A series of works have focused on improving the reasoning capabilities of LLMs by incorporating reward model supervision during test-time inference (Uesato et al., 2022; Lightman et al., 2023). In addition to these methods, test-time supervision has been proposed to improve the performance of LLMs in specific target domains using lightweight adapters (Sun et al., 2024b; Zhuang et al., 2024; Shi et al., 2024). For example, Shi et al. (2024) employs a lightweight model to rank outputs generated by LLMs in the medical domain, enhancing the domain-specific performance.

However, based on our empirical observations, the ability of LLMs in the context of conversational search remains insufficiently explored. To address this limitation, we propose leveraging a contrastive ranking loss to effectively train a lightweight reward model, unlocking LLM's reasoning capability in conversational search. To the best of our knowledge, we are the first to uncover and analyze the prompting-based conversational query reformulation at test time under the Best-of-N paradigm.

7 Conclusion

In this paper, we aim to unleash the power of prompting-based query reformulation at test time within the Best-of-N paradigm. Therefore, we propose AdaRewriter, a framework that effectively uses a lightweight outcome-supervised reward model as a scoring function to select the most promising reformulation. Extensive experimental evaluations across several benchmark datasets demonstrate that AdaRewriter consistently outperforms existing methods in most settings. These contributions advance the understanding of user intent in conversational search systems and improve the effectiveness of prompting-based query reformulation.

Limitation

We identify the below limitations in AdaRewriter:

Although the reward model is lightweight and the latency of AdaRewriter is comparable to that of previous work (Mao et al., 2023b), the primary latency bottleneck stems from the process of generating multiple reformulation candidates using LLMs. Despite this, we believe that improving prompting-based query reformulation through test-time adaptation shows considerable potential, as it combines both simplicity and effectiveness. This approach may reduce the need for extensive passage re-ranking. Additionally, test-time adaptation and scaling offer promising results, particularly with the Best-of-N paradigm, which has demonstrated superior performance across various tasks (Snell et al., 2024).

To further reduce latency, our method could benefit from applying existing inference acceleration techniques (Sun et al., 2024a; Wang et al., 2025). A key trade-off also exists between computational cost and latency, specifically when increasing the number of candidates N. A more efficient strategy may involve dynamically allocating computational resources based on reformulation task difficulty, *i.e.*, generating more candidates for complex scenarios and fewer for simpler ones.

Lastly, due to budget constraints, while we have demonstrated the effectiveness of AdaRewriter on black-box commercial LLMs, we have been unable to evaluate its performance with a larger candidate set N.

References

Vaibhav Adlakha, Shehzaad Dhuliawala, Kaheer Suleman, Harm de Vries, and Siva Reddy. 2022. TopiOCQA: Open-domain conversational question answering with topic switching. *Transactions of the Association for Computational Linguistics*, 10:468–483.

Raviteja Anantha, Svitlana Vakulenko, Zhucheng Tu, Shayne Longpre, Stephen Pulman, and Srinivas Chappidi. 2021. Open-domain question answering goes conversational via question rewriting. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 520–534, Online. Association for Computational Linguistics.

Haonan Chen, Zhicheng Dou, Kelong Mao, Jiongnan Liu, and Ziliang Zhao. 2024. Generalizing conversational dense retrieval via LLM-cognition data augmentation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2700–2718, Bangkok, Thailand. Association for Computational Linguistics.

Yiruo Cheng, Kelong Mao, and Zhicheng Dou. 2024. Interpreting conversational dense retrieval by rewriting-enhanced inversion of session embedding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2879–2893, Bangkok, Thailand. Association for Computational Linguistics.

Yung-Sung Chuang, Wei Fang, Shang-Wen Li, Wen-tau Yih, and James Glass. 2023. Expand, rerank, and retrieve: Query reranking for open-domain question answering. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12131–12147, Toronto, Canada. Association for Computational Linguistics.

Gordon V Cormack, Charles LA Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 758–759.

Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. 2020. Cast 2019: The conversational assistance track overview. In *In Proceedings of TREC*.

Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. 2021. Cast 2020: The conversational assistance track overview. In *In Proceedings of TREC*.

Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. 2022. Trec cast 2021: The conversational assistance track overview. In *In Proceedings of TREC*.

Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021. Splade: Sparse lexical and expansion model for first stage ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 2288–2292, New York, NY, USA. Association for Computing Machinery.

Jianfeng Gao, Chenyan Xiong, Paul Bennett, and Nick Craswell. 2023. *Neural approaches to conversational information retrieval*, volume 44. Springer Nature.

Yunah Jang, Kang-il Lee, Hyunkyung Bae, Hwanhee Lee, and Kyomin Jung. 2024. IterCQR: Iterative conversational query reformulation with retrieval guidance. 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 8121–8138, Mexico City, Mexico. Association for Computational Linguistics.

Zhuoran Jin, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. 2023. InstructoR: Instructing unsupervised

conversational dense retrieval with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6649–6675, Singapore. Association for Computational Linguistics.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Yilong Lai, Jialong Wu, Congzhi Zhang, Haowen Sun, and Deyu Zhou. 2025. AdaCQR: Enhancing query reformulation for conversational search via sparse and dense retrieval alignment. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7698–7720, Abu Dhabi, UAE. Association for Computational Linguistics.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. *arXiv preprint arXiv:2305.20050*.

Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021a. Pyserini: A Python toolkit for reproducible information retrieval research with sparse and dense representations. In *Proceedings of the 44th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2021)*, pages 2356–2362.

Sheng-Chieh Lin, Jheng-Hong Yang, and Jimmy Lin. 2021b. Contextualized query embeddings for conversational search. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1004–1015, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sheng-Chieh Lin, Jheng-Hong Yang, Rodrigo Nogueira, Ming-Feng Tsai, Chuan-Ju Wang, and Jimmy Lin. 2020. Conversational question reformulation via sequence-to-sequence architectures and pretrained language models. *arXiv preprint arXiv:2004.01909*.

Yixin Liu and Pengfei Liu. 2021. SimCLS: A simple framework for contrastive learning of abstractive summarization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 1065–1072, Online. Association for Computational Linguistics.

Xueguang Ma, Liang Wang, Nan Yang, Furu Wei, and Jimmy Lin. 2024. Fine-tuning llama for multi-stage text retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 2421–2425, New York, NY, USA. Association for Computing Machinery.

Kelong Mao, Chenlong Deng, Haonan Chen, Fengran Mo, Zheng Liu, Tetsuya Sakai, and Zhicheng Dou. 2024. ChatRetriever: Adapting large language models for generalized and robust conversational dense retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1227–1240, Miami, Florida, USA. Association for Computational Linguistics.

Kelong Mao, Zhicheng Dou, Bang Liu, Hongjin Qian, Fengran Mo, Xiangli Wu, Xiaohua Cheng, and Zhao Cao. 2023a. Search-oriented conversational query editing. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4160–4172, Toronto, Canada. Association for Computational Linguistics.

Kelong Mao, Zhicheng Dou, Fengran Mo, Jiewen Hou, Haonan Chen, and Hongjin Qian. 2023b. Large language models know your contextual search intent: A prompting framework for conversational search. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1211–1225, Singapore. Association for Computational Linguistics.

Kelong Mao, Hongjin Qian, Fengran Mo, Zhicheng Dou, Bang Liu, Xiaohua Cheng, and Zhao Cao. 2023c. Learning denoised and interpretable session representation for conversational search. In *Proceedings of the ACM Web Conference 2023*, WWW '23, page 3193–3202, New York, NY, USA. Association for Computing Machinery.

Fengran Mo, Abbas Ghaddar, Kelong Mao, Mehdi Rezagholizadeh, Boxing Chen, Qun Liu, and Jian-Yun Nie. 2024a. CHIQ: Contextual history enhancement for improving query rewriting in conversational search. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2253–2268, Miami, Florida, USA. Association for Computational Linguistics.

Fengran Mo, Kelong Mao, Ziliang Zhao, Hongjin Qian, Haonan Chen, Yiruo Cheng, Xiaoxi Li, Yutao Zhu, Zhicheng Dou, and Jian-Yun Nie. 2024b. A survey of conversational search. *arXiv preprint arXiv:2410.15576*.

Fengran Mo, Kelong Mao, Yutao Zhu, Yihong Wu, Kaiyu Huang, and Jian-Yun Nie. 2023a. ConvGQR: Generative query reformulation for conversational search. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4998–5012, Toronto, Canada. Association for Computational Linguistics.

Fengran Mo, Jian-Yun Nie, Kaiyu Huang, Kelong Mao, Yutao Zhu, Peng Li, and Yang Liu. 2023b. Learning to relate to previous turns in conversational search. In

Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD '23, page 1722–1732, New York, NY, USA. Association for Computing Machinery.

Fengran Mo, Chen Qu, Kelong Mao, Tianyu Zhu, Zhan Su, Kaiyu Huang, and Jian-Yun Nie. 2024c. History-aware conversational dense retrieval. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 13366–13378, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

OpenAI. 2022. Introducing chatgpt. https://openai.com/blog/chatgpt. Accessed: 2024-02-06.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pages 53728–53741. Curran Associates, Inc.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

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

Wenqi Shi, Ran Xu, Yuchen Zhuang, Yue Yu, Haotian Sun, Hang Wu, Carl Yang, and May Dongmei Wang. 2024. MedAdapter: Efficient test-time adaptation of large language models towards medical reasoning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22294–22314, Miami, Florida, USA. Association for Computational Linguistics.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*.

Hanshi Sun, Momin Haider, Ruiqi Zhang, Huitao Yang, Jiahao Qiu, Ming Yin, Mengdi Wang, Peter Bartlett, and Andrea Zanette. 2024a. Fast best-of-n decoding via speculative rejection. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Haotian Sun, Yuchen Zhuang, Wei Wei, Chao Zhang, and Bo Dai. 2024b. BBox-adapter: Lightweight adapting for black-box large language models. In *Forty-first International Conference on Machine Learning*.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. 2022. Solving math word problems with process-and outcome-based feedback. *arXiv preprint arXiv:2211.14275*.

Christophe Van Gysel and Maarten de Rijke. 2018. Pytrec_eval: An extremely fast python interface to trec_eval. In *SIGIR*. ACM.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Improving text embeddings with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11897–11916, Bangkok, Thailand. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Zhenglin Wang, Jialong Wu, Yilong Lai, Congzhi Zhang, and Deyu Zhou. 2025. SEED: Accelerating reasoning tree construction via scheduled speculative decoding. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4920–4937, Abu Dhabi, UAE. Association for Computational Linguistics.

Jialong Wu, Wenbiao Yin, Yong Jiang, Zhenglin Wang, Zekun Xi, Runnan Fang, Deyu Zhou, Pengjun Xie, and Fei Huang. 2025. Webwalker: Benchmarking llms in web traversal. *arXiv preprint arXiv:2501.07572*.

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.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *International Conference on Learning Representations*.

Fanghua Ye, Meng Fang, Shenghui Li, and Emine Yilmaz. 2023. Enhancing conversational search: Large language model-aided informative query rewriting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5985–6006, Singapore. Association for Computational Linguistics.

Chanwoong Yoon, Gangwoo Kim, Byeongguk Jeon, Sungdong Kim, Yohan Jo, and Jaewoo Kang. 2024. Ask optimal questions: Aligning large language models with retriever's preference in conversational search. *arXiv preprint arXiv:2402.11827*.

Shi Yu, Zhenghao Liu, Chenyan Xiong, Tao Feng, and Zhiyuan Liu. 2021. Few-shot conversational dense retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and*

Development in Information Retrieval, SIGIR '21, page 829–838, New York, NY, USA. Association for Computing Machinery.

Peitian Zhang, Shitao Xiao, Zheng Liu, Zhicheng Dou, and Jian-Yun Nie. 2023. Retrieve anything to augment large language models. *arXiv preprint arXiv:2310.07554*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand. Association for Computational Linguistics.

Yuchen Zhuang, Haotian Sun, Yue Yu, Rushi Qiang, Qifan Wang, Chao Zhang, and Bo Dai. 2024. HYDRA: Model factorization framework for black-box LLM personalization. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

A Discussion

A.1 AdaRewriter as Reformulation Evaluator

The evaluation of query reformulation primarily relies on two main approaches (Mo et al., 2024b): (1) **lexical overlapping**, which assesses the accuracy of the reformulated query relative to the reference query by computing token-level precision, recall, and F1 score, and (2) **end-to-end evaluation**, which measures the effectiveness of the reformulated query based on its final retrieval performance. However, these evaluation methods have their limitations: while lexical overlapping is efficient, it provides only an indirect measure and does not reflect the real effectiveness of the reformulated queries in downstream tasks; end-to-end assessment, though comprehensive, is computationally intensive and may be influenced by model biases.

In contrast, our proposed neural-based reward model could be a trade-off evaluation suite between efficiency and accuracy, demonstrating its robustness across both sparse and dense retrieval systems. Based on our practices, the reward model could serve as an effective proxy for assessing the quality of query reformulation, with potential applications in retrieval-augmented generation (RAG) systems (Wu et al., 2025), conversational search systems, and hard negative mining for retriever training.

A.2 Comparsion with AdaCQR

AdaCQR (Lai et al., 2025) aims to improve the performance of conversational query reformulation through a two-stage training paradigm. In the first stage, the model is trained using a large set of pseudo-labels generated by a large language model. The second stage further refines the model via iterative self-training with a contrastive ranking loss.

Despite demonstrating effectiveness, AdaCQR faces two notable limitations:

- AdaCQR exhibits a performance gap compared to LLM-based methods. To enable a fair comparison with such methods, an additional query expansion step using an LLM is required (i.e., the AdaCQR+Expansion setting proposed in the original paper).

- AdaCQR functions primarily as a training-time alignment approach, which restricts its applicability in real-world scenarios, particularly in environments where LLMs are accessed as black-box systems.

To address these limitations, AdaRewriter is proposed as a lightweight framework that employs a reward model to select the most promising candidate reformulations by combining query rewriting and expansion. It retains simplicity while benefiting from the concept of test-time scaling.

Moreover, AdaRewriter demonstrates the potential of leveraging test-time scaling and test-time adaptation in the context of conversational query reformulation. We believe this could offer some insights for future research in the field of conversational search.

A.3 Comparsion with CDR Methods

Conversational Dense Retrieval(CDR) represents an orthogonal approach to conversational query reformulation in the context of conversational search. This methodology focuses on training dense retrievers to improve the representation of both the current query and its associated historical context. Although a direct comparison may not be appropriate, we present a performance comparison between our proposed AdaRewriter and several CDR methods evaluated across the QReCC, TopiOCQA, and TREC CAsT datasets, as shown in Table 5.

We compare AdaRewriter with the following representative CDR methods: Conv-ANCE (Xiong et al., 2020), ConvDR (Yu et al., 2021), Conv-SPLADE (Formal et al., 2021), InstructorR-ANCE (Jin et al., 2023), LeCoRE (Mao et al., 2023c), ConvAug (Chen et al., 2024), and ChatRetriever (Mao et al., 2024). Among these, ChatRetriever stands out as one of the most representative works in the era of LLMs, which fine-tunes an LLM using contrastive learning and leverages the conversational session's embeddings to retrieve relevant passages. The results in Table 5 demonstrate that our proposed method achieves consistently strong performance across all five datasets, highlighting the robustness and effectiveness of AdaRewriter.

Moreover, conversational query reformulation-based approaches, such as AdaRewriter, offer superior explainability compared to CDR methods. This is valuable for enhancing user intent understanding and shows promise for improving conversational search systems.

B Experimental Details

B.1 Datasets Details

This paper uses five datasets: TopiOCQA (Adlakha et al., 2022), QReCC (Anantha et al., 2021), and

Framework	TopiOCQA	QReCC	CAsT-19	CAsT-20	CAsT-21	Avg.
Conv-ANCE (Xiong et al., 2020)	20.5	45.6	34.1	27.5	34.2	32.4
ConvDR (Yu et al., 2021)	26.4	35.7	43.9	32.4	37.4	35.2
Conv-SPLADE (Formal et al., 2021)	29.5	46.6	-	28.1	29.9	-
InstructoR-ANCE (Jin et al., 2023)	23.7	40.5	-	29.6	34.9	-
LeCoRE (Mao et al., 2023c)	32.0	51.1	42.2	37.7	50.8	42.8
ConvAug (Chen et al., 2024)	33.3	50.4	-	30.7	36.8	-
ChatRetriever (Mao et al., 2024)	40.1	52.5	52.1	40.0	49.6	46.9
AdaRewriter (LLama3.1-8B, N=5)	37.9	51.3	46.6	45.6	49.5	46.2
AdaRewriter (LLama3.1-8B, N=16)	39.7	53.8	48.3	46.5	49.7	47.6
AdaRewriter (GPT4o-mini, N=5)	40.4	51.5	49.0	47.3	52.5	**48.1**

Table 5: NDCG@3 performance comparison of our proposed AdaRewriter and Conversational Dense Retrieval(CDR) methods. The best average results among all methods are **bolded**, and the second-best results are underlined.

	QReCC		TopiOCQA	
	Train	Test	Train	Test
# Dialogues	10823	2775	3509	205
# Turns	29596	8209	45450	2514
# Collections	54M		25M	

Table 6: The statistics of QReCC and TopiOCQA datasets.

	CAsT-19	CAsT-20	CAsT-21
# Dialogues	50	25	26
# Turns	479	208	239
# Collections	38M	38M	42M

Table 7: The statistics of TREC CAsT 2019, 2020, and 2021 datasets.

TREC CAsT 2019 (Dalton et al., 2020), 2020 (Dalton et al., 2021), and 2021 (Dalton et al., 2022). TopiOCQA and QReCC contain both training and testing data, while TREC CAsT datasets provide only testing data for zero-shot experiments.

The QReCC dataset consists of 14K conversations with 80K question-answer pairs, and we aim to retrieve relevant passages from a collection of 54M passages. The TopiOCQA dataset contains 3.9K topic-switching conversations with 51K question-answer pairs, with a passage collection of 25M passages. Detailed statistics for both datasets are shown in Table 6.

TREC CAsT 2019, 2020, and 2021 are known for their complexity in conversational search under a zero-shot setting. Table 7 provides more details.

B.2 Implementation Details

All experiments are conducted on a server with four Nvidia GeForce 3090 GPUs.

Our framework is implemented using the Huggingface Transformers[4] and PyTorch Lightning[5]. The AdamW optimizer is used with a learning rate of 5e-6, following a cosine learning rate schedule with a warmup ratio of 0.1. Training is carried out for 10 epochs, and model checkpoints are saved at the end of each epoch. We employed the vLLM (Kwon et al., 2023) framework for candidate construction and inference, ensuring reproducibility by saving the results for inference. The retrieval systems were implemented using Faiss (Johnson et al., 2019) and Pyserini (Lin et al., 2021a). For the BM25 algorithm, we set the parameters as follows: $k_1 = 0.82, b = 0.68$ in QReCC, and $k_1 = 0.9, b = 0.4$ in TopiOCQA. Here, k_1 controls non-linear term frequency normalization, while b adjusts the scaling of the inverse document frequency. The query length was set to 32, and the concatenated reformulation query length was set to 256, following prior works (Mao et al., 2023b).

B.3 Training-time Tuning Details

We use `Llama-Factory` (Zheng et al., 2024) to conduct experiments on supervised fine-tuning (SFT) and direct preference optimization (DPO). To accommodate our hardware constraints, we adopt the LoRA technique with the rank $r = 16$. The training is performed for 3 epochs with a learning rate of 1e-4.

[4] https://github.com/huggingface/transformers
[5] https://github.com/Lightning-AI/pytorch-lightning

C Baseline Details

We compare AdaRewriter with the following representative baselines in the CQR task:

- **T5QR** (Lin et al., 2020) is a vanilla baseline that train the T5-base (Raffel et al., 2020) model to perform CQR tasks.
- **CONQRR** (Wu et al., 2022) aligns the T5-base reformulation model with retrievers through direct optimization using reinforcement learning.
- **ConvGQR** (Mo et al., 2023a) improves retrieval performance by utilizing two fine-tuned T5-base models, with one dedicated to query reformulation and the other to query expansion.
- **EDIRCS** (Mao et al., 2023a) effectively generates reformulation queries by combining non-autoregressive text-selection techniques with autoregressive token generation, utilizing a fine-tuned T5-base model.
- **LLM-Aided** (Ye et al., 2023) employs ChatGPT (OpenAI, 2022) to conduct query reformulation via a "rewrite-then-edit" prompting strategy.
- **IterCQR** (Jang et al., 2024) aligns the T5-base reformulation model with the dense retriever by minimizing Bayesian risk, which is driven by the semantic similarity between the query and the gold passage.
- **RETPO** (Yoon et al., 2024) leverages large language models to produce diverse reformulations through multi-perspective prompting, generates binarized comparisons informed by retriever feedback, and fine-tunes LLama2-7B via direct preference optimization (Rafailov et al., 2023).
- **ADACQR** (Lai et al., 2025) aligns the reformulation model with sparse and dense retrieval systems through a fusion metric, demonstrating the effectiveness of guiding reformulation using hybrid retrieval signals. The reformulation model leverages the T5-base and uses a vanilla LLama2-7B for query expansion.
- **CHIQ** (Mo et al., 2024a) utilizes the fundamental capabilities of LLMs to improve the quality of contextual history, thereby enhancing query rewriting performance. For comparison, we employ the most advanced CHIQ-Fusion, which combines reformulated queries generated by a fine-tuned T5-based model and the LLama2-7B model, utilizing result-level fusion techniques to derive the final retrieval outcomes.
- **LLM4CS** (Mao et al., 2023b) is our primary comparison method, exploring various strategies for aggregating multiple reformulation candidates to improve conversational search. The candidates are generated by a large language model through direct prompting. We employ the most advanced settings, including Rewrite-and-Response (RAR), Mean Aggregation, and Chain-of-Thought (CoT). To ensure a fair comparison, we present results based on our reproduction of the officially released code.

D Case Study

To show the effectiveness of AdaRewriter, we provide a detailed case in Table 9.

Prompt for Conversational Query Reformulation

Instruction

For an information−seeking dialog, please help reformulate the question into rewrite that can fully express the user's information needs without the need of context, but also generate an informative response to answer the question. I will give you several example multi−turn dialogs, where each turn contains a question as well as a rewrite and a response that you need to generate. The rewrite part begins with a sentence explaining the reason for the generated rewrite.

Demonstrations

Example #1:
Question: What should I consider when buying a phone?
Rewrite: This is the first turn. So the question should be rewritten as: What should I consider when buying a phone?
Response: The design of the phone and the overall look and feel of the phone are very important. You should be comfortable with the way the phone looks and feels when you hold it in your hand. In addition, don't be afraid to get a phone from a different manufacturer than you're used to. Consider an older, used, or refurbished phone to save money. Sometimes a year−old or even a two−year−old device might give you everything you need.

Question: Cool. Which one would you recommend?
Rewrite: Based on Turn 1, you are inquiring about what should be considered when buying a phone. So the question should be rewritten as: Cool. Which smartphone would you recommend for me?
Response: Just because a phone has everything and the kitchen sink doesn't mean you have to buy it − it's just an option for those who might have a larger budget. The same logic applies to pretty much anything you can buy. If you want something more affordable, cheaper, but has the latest in Android and that comes with 5000mAh of battery, then the Motorola Moto G7 Power is an excellent phone to consider. If you want an Android smartphone that's cheaper and more affordable, and you still find the Moto G7 Power a bit off your budget, then I highly recommend the Samsung Galaxy A10e.

<— Omit Following Demonstrations —>

Annotated Sample

Context:{{current_context}}

Current Question: {{current_query}}

Now, you should give me the rewrite and response of the **Current Question** under the ** Context**. The output format should always be: "Rewrite: $Reason. So the question should be rewritten as: $Rewrite\nResponse: $Response." Note that you should always try to rewrite it and generate an informative response. Never ask for clarification or say you don't understand it in the generated rewrite and response. Go ahead!

Table 8: The prompt used to obtain rewritten query $\hat{q}$ and pseudo-response $\hat{r}$.

Conversation:
Q1: what is roosevelt dime?
A1: Is the current dime, or ten-cent piece, of the United States
Q2: who designed it?
A2: John R. Sinnock
Q3: who is he?
A3: Eighth Chief Engraver of the United States Mint
Q4: mention a few controversies of his?
A4: The letters "JS" actually stood not for John Sinnock, but for Joseph Stalin. The urban folk story coincided with the Second Red Scare. Another controversy was an allegation that Sinnock copied or borrowed the design of the President's profile from a bronze bas relief created by sculptress Selma H. Burke
Q5: did the aforementioned mint struck the roosevelt dime?
A5: UNANSWERABLE
Q6: who made the dime?
A6: John R. Sinnock
Q7: when was the united states mint established?
A7: The Mint was created in 1792
Q8: where is its location?
A8: In Philadelphia
Q9: what is fort knox and the mint's relation?
A9: Its primary purpose is for storage of the United States and other countries' gold and silver bullion reserves.
Q10: is it a building or fort?
A10: Is a fortified vault building
Q11: any recent incident happened over there related to shooting?
A11: On 18 October 1993, Arthur Hill went on a shooting rampage, killing three and wounding two before attempting suicide, shooting and severely wounding himself.
Q12: how does air corps utilize it?
A12: As a training base during World War II.
Q13: is it used for protecting valuable objects?
A13: For protection after the Japanese attack on Pearl Harbor in 1941, the Declaration of Independence, the Constitution of the United States and the Gettysburg Address were all moved for safekeeping
Original Query: does it have a high school in its premises? (**rank: Not Found**)

Max-prob Rewritten Query: Does the United States Mint have a high school within its premises? The United States Mint does not have...(**rank: Not Found**)

AdaRewriter(*Ours*): Does Fort Knox have a high school or educational institution within its premises? Fort Knox does not have a high school ...(**rank: 1**)

Gold Passage: Fort Knox is one of only four Army posts (the others being Fort Campbell, Kentucky, Fort Meade, Maryland, and Fort Sam Houston, Texas) that still has a high school located on-post. Fort Knox High School was built in 1958 and has undergone only a handful of renovations...

Table 9: Successful case study on TopiOCQA (id: 126_14). The underline part shows the decontextualized information in the reformulation query.