AdaCQR: Enhancing Query Reformulation for Conversational Search via Sparse and Dense Retrieval Alignment

Anonymous ACL submission

Abstract

Conversational Query Reformulation (CQR) has significantly advanced in addressing the challenges of conversational search, particularly those stemming from the latent user intent and the need for historical context. Recent works aimed to boost the performance of CRQ through alignment. However, they are designed for one specific retrieval system, which potentially results in poor generalization. To overcome this limitation, we present a novel framework AdaCQR. By aligning reformulation models with both *term*-based and *semantic*-based retrieval systems, AdaCQR enhances the generalizability of information-seeking queries across diverse retrieval environments through a two-stage training strategy. We also developed two effective approaches for acquiring superior labels and diverse input candidates, boosting the efficiency and robustness of the framework. Experimental evaluations on the TopiOCQA and QReCC datasets demonstrate that AdaCQR significantly outperforms existing methods, offering both quantitative and qualitative improvements in conversational query reformulation.[1]

1 Introduction

Conversational search extends traditional information retrieval paradigms by addressing complex information-seeking requirements through multi-turn interactions (Radlinski and Craswell, 2017; Qu et al., 2020; Gao et al., 2023). A fundamental challenge in conversational search is to discover the latent user intent within the current query and historical context, which complicates the application of off-the-shelf retrievers due to issues such as omissions, ambiguity, and coreference (Anantha et al., 2021; Adlakha et al., 2022).

Existing methods to address this challenge can be broadly categorized into two types: dense retriever-based and query reformulation-based. For dense retrievers-based approaches (Qu et al., 2020; Lin et al., 2021b; Kim and Kim, 2022; Mo et al., 2024; Chen et al., 2024), long dialogue contexts can be effectively grasped while incurring retraining costs and lacking the adaptability to sparse retrieval systems like BM25 (Robertson et al., 2009). Query reformulation-based approaches leverage a language model to decontextualize the query of user into a stand-alone query, a process known as conversational query reformulation (CQR), as shown in Figure 1. Previous studies have demonstrated the effectiveness of CQR (Wu et al., 2022; Mo et al., 2023a; Ye et al., 2023).

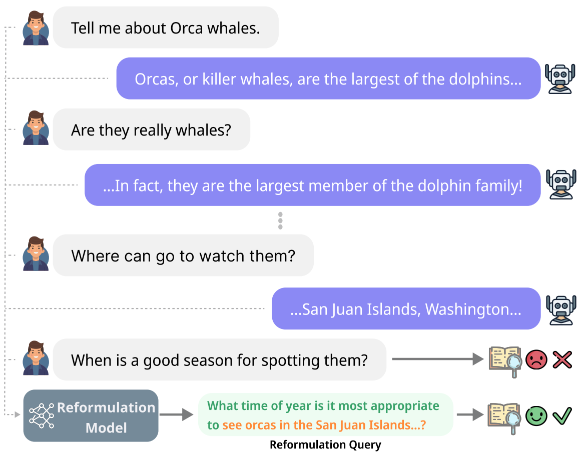

Figure 1: An example of CQR which takes the context and current query as input and generates a decontextualized query as output.

Due to the limitation that the training objectives do not align with task targets, *i.e.*, minimizing cross-entropy loss for teacher forcing generation during training while expecting to maximize retrieval metric during inference, subsequent works have aimed to enhance the performance of CQR through alignment. In detail, Jang et al. (2023) utilize Minimum Bayes Risk (MBR) (Smith and Eisner, 2006) based on semantic similarity between

[1] The code and datasets of this paper will be publicly available upon the acceptance of the paper.

the query and gold passage to achieve alignment. Yoon et al. (2024) create binarized comparisons based on retriever feedback and optimize the reformulation model via Direct Preference Optimization (DPO) (Rafailov et al., 2023). They also tackle the reliance on sub-optimal and costly human-annotated reformulation labels by using Large Language Models (LLMs) to generate labels via iterative prompting or multi-perspective prompting.

However, previous alignment methods are designed for one specific retrieval system, which may fail to be generalized to multiple retrieval systems simultaneously. For an information-seeking query to generalize well across both sparse and dense retrieval systems, it must have: (1) precise *term* overlap (*e.g.*, the presence of key entities in the query) and (2) high *semantic* similarity between the document and the query (Luan et al., 2021). Focusing on only one of these aspects would lead to performance degradation. In addition, previous works for achieving alignment exhibit stability issues when using reinforcement learning (Jang et al., 2023) and require an explicit reference model (Jang et al., 2023; Yoon et al., 2024).

Therefore, in this paper, we introduce AdaCQR, a novel framework that aligns the training objective with the task target. In specific, AdaCQR aligns the reformulation model and the retrievers from both *term*-based and *semantic*-based perspectives to achieve strong generalization abilities in sparse and dense retrieval. Furthermore, to address the issues of high complexity and instability inherent in MBR (Jang et al., 2023), we employ a two-stage training strategy to achieve alignment (Liu et al., 2022), where the reformulation model serves both as a `generation` model using cross-entropy loss for teacher forcing generation and a `reference-free evaluation` model using contrastive loss.

The framework works as follows: 1) An advanced generation model is initialized with superior reformulation labels. Specifically, a few-shot LLM prompting method is employed inspired by the principles of contrastive learning (Paranjape et al., 2021; He et al., 2022) to get superior labels; 2) Unlike previous methods that rely on human or LLMs annotations, Diverse Beam Search (Vijayakumar et al., 2016) is used to generate multiple candidates simultaneously. Among the generated candidates, one oracle candidate exhibits exceptional performance, while the remaining candidates are relatively ranked based on a fusion metric; 3) We employ a two-stage training, where the reformulation model can be aligned using the contrastive loss from both *term*-based and *semantic*-based perspectives.

AdaCQR achieves excellent performance on two widely used conversation search datasets, TopiOCQA (Adlakha et al., 2022) and QReCC (Anantha et al., 2021). Notably, AdaCQR achieves the performance comparable to those approaches fine-tuned on the LLaMA-7B backbone, despite of utilizing only the T5-base. Experimental results demonstrate the quantitative and qualitative improvements of our proposed framework.

The contributions of this work are as follows:

- We propose AdaCQR to align reformulation models with *term*-based and *semantic*-based retrievers, simultaneously.
- Two effective approaches are developed: one to acquire superior labels for generation and another to gather diverse input candidates for reference-free evaluation.
- Extensive experiments on two benchmark datasets conclusively demonstrate our proposed AdaCQR significantly outperforms existing methods, establishing its superiority in performance.

2 Related Work

2.1 Conversational Search

Conversational search improves traditional information retrieval by using iterative, multi-turn interactions to address the complex information needs of user (Gao et al., 2023). A key challenge is understanding the the implicit intent of user, requiring attention to both the current query and its historical context. Two main approaches to this problem are conversational dense retrieval (CDR) and conversational query reformulation (CQR).

CDR (Qu et al., 2020; Yu et al., 2021; Lin et al., 2021b) aims to improve the representation of the current query along with its historical context by training dense retrievers. Recent advancements in CDR have focused on mitigating the influence of irrelevant historical contexts (Kim and Kim, 2022; Mo et al., 2023b, 2024; Chen et al., 2024) and enhancing interpretability (Mao et al., 2023b; Cheng et al., 2024). However, this approach incurs additional training costs and lacks the adaptability to sparse retrieval systems like BM25 (Robertson et al., 2009).

Conversely, CQR (Elgohary et al., 2019) concentrates on decontextualizing the query of user into a stand-alone query suitable for use with off-the-shelf retrievers. Numerous prior studies have demonstrated the effectiveness of CQR by utilizing human annotations in supervised methods (Lin et al., 2020a; Yu et al., 2020; Vakulenko et al., 2021) and integrating query expansion models (Mo et al., 2023a). However, human-annotated labels are costly and reported to be sub-optimal (Lin et al., 2021b; Wu et al., 2022). In the era of LLMs, several studies have utilized LLMs to generate query reformulations directly (Ye et al., 2023; Mao et al., 2023a) and obtain reformulation labels for distillation (Jang et al., 2023; Yoon et al., 2024).

This paper focuses on conversational query reformulation, proposing a novel framework ADACQR to align with *term*-based and *semantic*-based retrieval systems. To overcome the limitations of human annotation, we also developed two effective methods for obtaining superior labels and diverse input candidates.

2.2 Aligning LMs using Feedback

Aligning language models with feedback involves adjusting their behavior and outputs based on evaluation feedback (Wang et al., 2023), employing various reward learning methodologies to provide accurate supervised signals (Schulman et al., 2017; Rafailov et al., 2023).

Recent studies have enhanced conversational query reformulation by aligning language models with retriever feedback (Jang et al., 2023; Yoon et al., 2024). Jang et al. (2023) achieve the alignment through minimizing Bayes Risk based on semantic similarity between the query and the gold passage. Yoon et al. (2024) leverages LLMs to generate numerous reformulations via multi-perspective prompting, creating binarized comparisons based on retriever feedback and optimizing the reformulation model using DPO (Rafailov et al., 2023). However, previous methods struggle with the high cost of generating reformulations with LLMs (Yoon et al., 2024), or the instability of MBR (Jang et al., 2023; Finkelstein and Freitag, 2023).

In contrast, our framework utilizes a contrastive loss (Liu et al., 2022) to achieve alignment with retrievers. To the best of our knowledge, we are the first to employ the language model as a reference-free evaluation model to align retrievers, thereby enhancing stability and reducing complexity.

3 Method

3.1 Task Formulation

The conversational search task discussed in this paper involves finding the passage most relevant to the intent of user from a large collection of passages $\mathcal{C}$, given the current query of the user and historical context. To achieve this goal, the CQR task is proposed to utilize a language model G_θ to condense the current query q_k and historical context $\mathrm{H}_{k-1} = \{q_i, r_i\}_{i=1}^{k-1}$ into a stand-alone query $\hat{Q}_k$, where q_i and r_i denote the query and system answer of the i-th turn conversation, with k indicating the current turn. This decontextualized query $\hat{Q}_k$ is subsequently input into an *off-the-shelf* retrieval system $\mathcal{R}$, which returns a ranked list of the top-k relevant passages.

For the sake of convenience, we define the CQR task as the problem $\mathcal{P} = \{q, \mathrm{H}\}$, where q represents the current query of user and H denotes the historical context. The task is to generate a reformulated query $\hat{Q}$, as discussed in the following sections.

3.2 Overall framework

Our framework begins to leverage LLM to generate superior reformulation labels $Q^\star$ via few-shot learning (§3.4), where we select representative examples and implicitly guide LLM to generate labels that meet the needs of retrievers. Subsequently, we employ a two-stage training strategy using these labels to align the reformulation model with the retrievers. In the first stage, we train the reformulation model with a cross-entropy loss $\mathcal{L}_g$ to acquire the basic ability to generate reformulation queries using the superior $Q^\star$. (§3.5.1). Afterwards, we use this model to create a diverse set $\mathcal{S}$, including candidate queries $C_{(1)}, \cdots, C_{(n)}$ (§3.5.2). These candidates are then evaluated through sparse and dense retrieval, assessing their performance from both term-based and semantic-based perspectives. We utilize a proposed fusion metric (§3.3) to synthesize these evaluations and obtain the relative order of the candidates. In the second stage, leveraging the relative order of the candidates, we apply a contrastive loss $\mathcal{L}_c$ (§3.5.3) to achieve alignment between the reformulation model and the retrievers, where the reformulation model is treated as an evaluation model. The overall framework is depicted in Figure 2.

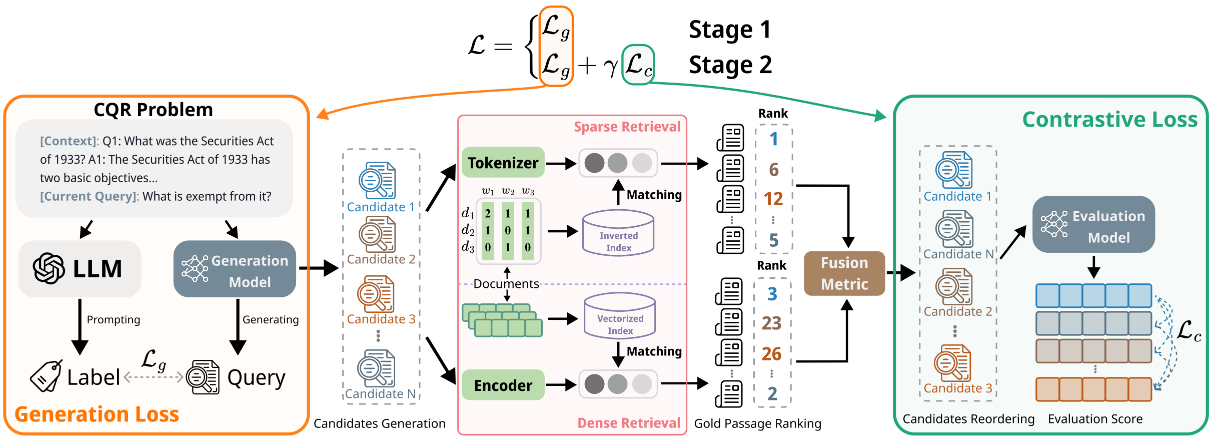

Figure 2: The framework of the proposed AdaCQR. A two-stage training is employed, where Stage 1 involves minimizing generation loss $\mathcal{L}_g$, followed by Stage 2 employing contrastive loss $\mathcal{L}_c$. The evaluation score is a distribution vector defined in Eq. (7).

3.3 Fusion Metric for Sparse and Dense Retrieval

A good information-seeking query must have precise term overlap and high semantic similarity between the document and the query to generalize well across sparse and dense retrieval (Luan et al., 2021).

To measure the generalization ability of the reformulation queries, we input them into sparse and dense retrieval systems and assess their performance based on the ranking of the corresponding gold passages, as illustrated in the central part of Figure 2.

In sparse retrieval, the inverted index is constructed using the sparse vectors of the transformed documents. The query is then tokenized into terms and matches passage based on term overlap. In contrast, dense retrieval involves creating a vectorized index using the dense vectors of the transformed documents. Subsequently, the query is converted into an embedding by the encoder, and the passage is searched based on semantic similarity.

Leveraging the performance of the reformulation query in both retrieval systems, we propose a fusion metric to evaluate the performance of reformulation query $\hat{Q}$ more comprehensively, similar to reciprocal rank fusion (Cormack et al., 2009):

$$\mathrm{M}(\hat{Q}, d) = \frac{r_s(\hat{Q}, d) + r_d(\hat{Q}, d)}{r_s(\hat{Q}, d) \times r_d(\hat{Q}, d)} \quad (1)$$

where $\hat{Q}$ is a reformulation query, d is the gold passage. $r_s(q, d)$ and $r_d(q, d)$ represent the rank of the gold passage d within the sparse and dense retrieval results for query q, respectively. The ranking r_s and r_d starts from 1, indicating the highest-ranked passage. Based on Eq (1), a larger $\mathrm{M}(\hat{Q}, d)$ indicates better generalization performace for reformulation query $\hat{Q}$ on sparse and dense retrieval systems.

3.4 Superior Reformulation Annotation

To mitigate the dependency on costly and suboptimal human-annotated reformulation labels, we utilize LLMs to generate superior reformulation labels, offering a more robust foundation for our framework.

Our intuitive approach to obtaining superior reformulation labels is to convey the LLMs of the characteristics of an effective query reformulation for retrieval. However, defining a good reformulation query or providing explicit instructions for generating one that meets the needs of retrievers poses significant challenges. Leveraging excellent in-context learning capabilities of LLMs (Brown et al., 2020; Dong et al., 2022), we propose a prompting strategy that implicitly selects representative demonstrations and guides LLMs to generate reformulation labels aligned with the requirements of retrieval systems.

Our method begins with a `vanilla generative model` G_π with basic query reformulation ability. We employ G_π to generate a reformulation candidates set $\mathcal{S}_\pi = \{C_{(1)}, C_{(2)}, \cdots, C_{(n)}\}$ for each the reformulation problem $\mathcal{P} = \{q, \mathrm{H}\}$, on the validation set. To select representative demonstrations, we use a score $\hat{R}$ to describe the difficulty of the query reformulation problem $\mathcal{P}$:

$$\hat{R} = \frac{1}{n}\sum_{i=1}^{n}[\mathrm{M}(C_{(i)}, d) - \frac{1}{n}\sum_{j=1}^{n}\mathrm{M}(C_{(j)}, d)]^2 \quad (2)$$

where $C_{(i)}$ is a reformulation candidate generated by G_π, and d is the corresponding gold passage for reformulation problem $\mathcal{P}$. The metric $\mathrm{M}(\cdot)$ is defined according to Eq. (1).

Subsequently, we selected the top-m reformulation problems exhibiting the highest $\hat{R}$ scores from the validation set. For each selected reformulation problem $\mathcal{P} = \{q, \mathrm{H}\}$, we identified the best and worst reformulation candidates from set $\mathcal{S}_\pi$ based on the metric in Eq. (1). These candidates denoted as C_{best} and C_{worst}, serve to implicitly guide the LLMs in generating labels aligned with the needs of the retrieval system, inspired by contrastive learning (Paranjape et al., 2021; He et al., 2022). We then concatenate the demonstration $P = (q, \mathrm{H}, C_{\text{best}}, C_{\text{worst}})$ and task instruction $\mathcal{I}$ to form the final prompt $\mathcal{D} = \mathcal{I}\,||P_1|| \cdots ||P_m$, where $||$ donates concatenation. Finally, we employed the LLM to obtain the superior reformulation labels $Q^\star$ through in-context learning. The details for the annotation are presented in Appendix D.

3.5 Align LMs with Retrievers

After getting superior reformulation labels using a defined fusion metric, we can align LMs with retrievers through two-stage training. The reformulation model serves as a standard generation model at the training stage 1. (§3.5.1) Then we develop a method to generate multiple candidate queries using this trained model. (§3.5.2) By learning the relative order of these candidates, we implicitly guide the language model to generate queries that meet the requirements of the retrievers. Lastly, in training stage 2, the reformulation model serves both as a generation model using cross-entropy loss and a reference-free evaluation model using contrastive loss to achieve alignment. (§3.5.3)

3.5.1 Training Stage 1 for Initialization

In the first training stage, we train a language model using the superior reformulation labels to endow it basic capability of query reformulation. To encourage more diverse generation results, a label smooth cross-entropy loss is used:

$$\mathcal{L}_1 = \mathcal{L}_g = \sum_{j=1}^{l} \sum_{x} p_s(x \mid \mathcal{P}, Q^\star_{<j}) \log p_{G_\theta}(x \mid \mathcal{P}, Q^\star_{<j}; \theta) \tag{3}$$

where $\mathcal{P}$ is the reformulation problem including current query q and historical context H, $Q^\star_{<j}$ is the first j tokens of the reformulation label $Q^\star$. p_s is a label smooth distribution, defined as follows:

$$p_s(x \mid \mathcal{P}, Q^\star_{<j}) = \begin{cases} 1-\beta & x = Q^\star_j \\ \frac{\beta}{N-1} & x \neq Q^\star_j \end{cases} \tag{4}$$

where β is the probability mass parameter, and N is the size of the dictionary. Now we have a trained language model G_θ using cross-entropy loss, which can be used for candidate generation and serves as a reference-free evaluation model during the training at stage 2.

3.5.2 Candidates Generation for Alignment

To efficiently generate a variety of candidates, we utilized Diverse Beam Search (Vijayakumar et al., 2016), an extension of the beam search strategy designed to generate a more diverse set of beam sequences for selection. Formally, given trained language model G_θ and reformulation problem $\mathcal{P}$, we generate candidates set $\mathcal{S} = \{C_{(1)}, \cdots, C_{(n)}\}$ with diverse beam search, where $C_{(i)}$ is the candidate of reformulation query, n is the number of candidates.

To align the retrievers from both term-based and semantic-based perspectives with the language model, we define the relative rank order as implicitly supervised signals, utilizing the metric proposed in Eq.(1), which simultaneously considers both types of retrievers, as follows:

$$C_{(i)} \succ C_{(j)} \iff \mathrm{M}(C_{(i)}, d) > \mathrm{M}(C_{(j)}, d) \tag{5}$$

where d is the gold passage of reformulation problem $\mathcal{P}$.

For reformulation problem $\mathcal{P}$, we now have candidates set $S = \{C_1, \cdots, C_n\}$ and their relative rank order $C_1 \succ C_2 \succ \cdots \succ C_n$, where C_i represents the i-th candidate in the sorted order.

3.5.3 Training Stage 2 for Alignment

Now we have sorted candidates S and trained model G_θ to perform training at stage 2. Leveraging the candidates set $\mathcal{S}$ and their relative rank order $C_1 \succ C_2 \succ \cdots \succ C_n$, a contrastive loss (Liu et al., 2022) for alignment:

$$\mathcal{L}_c = \sum_{i=1}^{n} \sum_{j>i} \max(0, f(C_j) - f(C_i) + (j-i) \times \lambda) \tag{6}$$

where j and i are the rank order in the candidates, and λ is the margin parameter. $f(C)$ represents the length-normalized estimated log-probability, where the language model serves as a reference-free evaluation model:

$$f(C) = \frac{1}{|C|^{\alpha}} \sum_{t=1}^{l} \log p_{G_\theta}(c_t \mid \mathcal{P}, C_{<t}; \theta) \quad (7)$$

where $|C|$ and l is the length of candidate, c_t is the generated t-th token given reformulation problem and previous $t-1$ tokens, and α is the length penalty parameter.

To ensure the stability of the training process, we employed a multi-task learning loss function, where the language model served as both a generation model and an evaluation model:

$$\mathcal{L}_2 = \mathcal{L}_g + \gamma \mathcal{L}_c \quad (8)$$

where γ is the weight of the contrastive loss.

4 Experiments

Datasets We train and evaluate our model using two widely utilized conversational search datasets: QReCC (Anantha et al., 2021) and TopiOCQA (Adlakha et al., 2022). The details of these datasets are shown in Appendix B.1.

Retrieval Systems Following prior works in the CQR task (Wu et al., 2022; Mo et al., 2023a; Jang et al., 2023; Yoon et al., 2024), we evaluate AdaCQR using sparse and dense retrieval systems[2]. The sparse retrieval system used is BM25 (Robertson et al., 2009). For dense retrieval, we use ANCE (Xiong et al., 2020), trained on the MS MARCO (Nguyen et al., 2016) retrieval task.

Baselines In our study, we compare AdaCQR with the following representative baselines in the CQR task: (1) T5QR (Lin et al., 2020b) is a vanilla baseline that utilizes the T5-base (Raffel et al., 2020) model to perform CQR tasks. (2) CONQRR (Wu et al., 2022) aligns the reformulation model with retrievers through direct optimization using reinforcement learning. (3) ConvGQR (Mo et al., 2023a) enhances retrieval performance by employing two generative models, one for query reformulation and the other for query expansion. (4) InfoCQR (Ye et al., 2023) employs ChatGPT (OpenAI, 2022) to conduct query reformulation via a "rewrite-then-edit" process. (5) IterCQR (Jang et al., 2023) achieves alignment by minimizing Bayes Risk based on the semantic similarity between the query and the gold passage. (6) RetPO (Yoon et al., 2024) utilizes large language models to generate multiple reformulations through multi-perspective prompting, creates binarized comparisons based on retriever feedback, and optimizes LLaMA2-7B (Touvron et al., 2023) using direct preference optimization (DPO) (Rafailov et al., 2023).

[2]We do not fine-tune the retrievers within our framework, thus excluding consideration of such CDR work in baselines.

The details regarding **implementation** and **evaluation metrics** are provided in the Appendix C and Appendix B.2, respectively.

4.1 Main Results

To evaluate the efficacy of our framework, we conducted comprehensive experiments on the QReCC and TopiOCQA datasets, employing the AdaCQR model trained individually on each dataset, presented in Table 1. We consider three kinds of backbones as baselines: the T5-based, the LLaMA2-7B-based, and the ChatGPT-based. The results demonstrate that AdaCQR significantly outperforms previous models utilizing T5-base as the backbone. Furthermore, AdaCQR exceeds the performance of RetPO, which uses LLaMA2-7B as the backbone, in both the QReCC and the dense retrieval section of TopiOCQA, underscoring the effectiveness of our approach. Additionally, the enhanced performance of RetPO is attributed to the inherently strong common-sense reasoning capabilities of the backbone model. To a fair comparison with RetPO, we employed the vanilla LLaMA2-7B, generating potential answers or keywords as query expansions for reformulation queries. AdaCQR with query expansion achieves results comparable to RetPO in sparse retrieval on the TopiOCQA, while significantly outperforming RetPO in other settings. Specifically, AdaCQR with expansion shows superior performance in dense retrieval on the TopiOCQA, attaining the best MRR (38.5), NDCG (37.6), R@10 (58.4), and R@100 (75.0).

These results underscore the efficacy and generalizability of AdaCQR in enhancing retrieval performance across different retrieval systems.

4.2 Ablation Study

In this study, we employ a contrastive loss to align the retrievers. We also introduce a fusion metric to evaluate query performance across semantic and term perspectives. Additionally, we leverage LLM to obtain superior labels $Q^\star$ through prompting, reducing reliance on sub-optimal and costly human

Type	Query Reform.	TopiOCQA MRR	TopiOCQA NDCG	TopiOCQA R@10	TopiOCQA R@100	QReCC MRR	QReCC NDCG	QReCC R@10	QReCC R@100
Sparse (BM25)	`Human Rewrite`	-	-	-	-	39.8	36.3	62.7	98.5
	T5QR (T5-base)	11.3	9.8	22.1	44.7	33.4	30.2	53.8	86.1
	CONQRR (T5-base)	-	-	-	-	38.3	-	60.1	88.9
	IterCQR (T5-base)	16.5	14.9	29.3	54.1	46.7	44.1	64.4	85.5
	InfoCQR (ChatGPT)	-	-	-	-	49.4	-	67.1	88.2
	ConvGQR (T5-base)†	12.4	10.7	23.8	45.6	44.1	41.0	64.4	88.0
	RETPO (LLaMA2-7B)†	**28.3**	**26.5**	48.3	**73.1**	50.0	47.3	69.5	89.5
	ADACQR (*Ours*, T5-base)	17.8	15.8	34.1	62.1	52.4	49.9	70.9	91.0
	`+Expansion`†	**28.3**	**26.5**	**48.9**	71.2	**55.1**	**52.5**	**76.5**	**93.7**
Dense (ANCE)	`Human Rewrite`	-	-	-	-	38.4	35.6	58.6	78.1
	T5QR (T5-base)	23.0	22.2	37.6	54.4	34.5	31.8	53.1	72.8
	CONQRR (T5-base)‡	-	-	-	-	41.8	-	65.1	**84.7**
	IterCQR (T5-base)	26.3	25.1	42.6	62.0	42.9	40.2	65.5	84.1
	ConvGQR (T5-base)†	25.6	24.3	41.8	58.8	42.0	39.1	63.5	81.8
	RETPO (LLaMA2-7B)†	30.0	28.9	49.6	68.7	44.0	41.1	66.7	84.6
	ADACQR (*Ours*, T5-base)	32.8	31.5	54.6	73.0	45.1	42.4	66.3	83.4
	`+Expansion`†	**38.5**	**37.6**	**58.4**	**75.0**	**45.8**	**42.9**	**67.3**	83.8

Table 1: Evaluation results of various retrieval system types on the test sets of QReCC and TopiOCQA. The best results among all methods are **bolded,** and the second-best results are underlined. † denotes the baseline involved using query expansion. ‡ denotes the baselines utilizing another dual encoder dense retrieval. `+Expansion` denotes the addition of query expansion, details in Appendix F.

Type	Query Reform.	QReCC MRR	QReCC R@10
Sparse	Superior labels $Q^\star$	45.4	65.5
	ADACQR (*Ours*)	**52.4**	**70.9**
	w/o. Contrastive Loss	43.3	62.8
	w/o. Fusion Metric	50.5	67.7
	w/o. Sparse Rank	50.9	69.7
	w/o. Dense Rank	51.6	70.5
	w/o. Labels $Q^\star$	44.9	63.7
Dense	Superior labels $Q^\star$	40.1	60.2
	ADACQR (*Ours*)	**45.1**	**66.3**
	w/o. Contrastive Loss	38.5	58.9
	w/o. Fusion Metric	42.4	63.7
	w/o. Sparse Rank	43.5	64.2
	w/o. Dense Rank	42.9	63.0
	w/o. Labels $Q^\star$	41.0	60.5

Table 2: Ablation study for each component of ADACQR. We also report the performance of the superior labels $Q^\star$ which are obtained by prompting LLMs through in-context learning, as detailed in Section 3.4.

labels. To investigate the impact of each component on the performance of ADACQR, we conducted ablation experiments focusing on these three specific modules in Table 2. To assess the influence of contrastive loss, we executed a single-stage training process without alignment. To determine the effect of the fusion metric, we substituted it with the evaluation method used in previous work (Jang et al., 2023), which only relies on the cosine similarity between the query and the gold passage. To further investigate the effectiveness of considering both perspectives in the fusion metric, we separately remove sparse ranking r_s and dense ranking r_d within it for analysis. To examine the impact of superior labels $Q^\star$, we trained ADACQR using human rewrite labels instead. The results of these ablation experiments reveal that the exclusion of any of these modules greatly affects the performance of ADACQR, showing the importance of these components for ADACQR. In particular, the most notable decline in performance occurs upon the removal of Contrastive Loss. Its decline in performance is followed by the impact of the labels $Q^\star$ and the Fusion Metric. Removing any rank degrades performance for both retrievers, more significantly for the corresponding retriever. This confirms the rationale behind considering both perspectives simultaneously. It is worth noting that superior labels $Q^\star$ can achieve comparable performance both in sparse and dense retrievals, which validates the effectiveness of the proposed fusion metric and the annotation method. The results indicate that queries reformulated by ADACQR significantly outperform superior labels $Q^\star$, demonstrating the advantages of an aligned model in CQR.

4.3 Robustness to Topic Shifts in Conversation

In the conversational search task, the frequent topic changes during the dialogue pose challenges for

Model	Topic-Concentrated MRR	R@10	Topic-Shifted MRR	R@10
T5QR	35.2	54.4	25.2	45.1
CONQRR	41.9	63.1	25.2	45.9
IterCQR	54.4	72.4	24.9	49.7
Human Rewrite	44.0	66.7	31.8	56.7
ADACQR	**66.0**	**82.4**	**34.1**	**58.3**

Table 3: Performance of ADACQR on topic-concentrated and topic-shifted samples on QReCC, MRR and R@10 are reported. The result is reported on BM25 Retrieval System.

Coefficient(γ)	QReCC MRR	NDCG	R@10	R@100
0	43.3	41.0	62.8	88.5
0.1	45.3	42.7	65.2	90.2
1	48.8	46.1	68.7	**91.2**
10	50.2	47.7	68.8	89.0
100	**52.4**	**49.9**	**70.9**	91.0
1000	49.4	46.7	68.6	90.7
$+\infty$	44.5	41.8	65.5	90.9

Table 4: ADACQR performance with different γ coefficients weighting of the contrastive loss in Eq. (8). $+\infty$ indicates only using the contrastive loss. 0 indicates only using the cross-entropy loss. BM25 is used as the retriever for experiments.

CQR. To evaluate the robustness of ADACQR in handling topic shifts, we divided the QReCC dataset into two parts: Topic-Concentrated and Topic-Shifted. Following previous work (Jang et al., 2023), we determine whether a topic shift has occurred in the current conversation by checking if the gold passage ID associated with the current query appears in the gold passage IDs corresponding to the previous context. The results presented in Table 3 indicate that ADACQR substantially outperforms previous models in both parts of conversations. Additionally, ADACQR exceeds human rewrites in topic-shifted dialogues, showing the robustness of our approach in query reformulation when addressing topic shiftings.

5 Analysis

5.1 Effect of the Multi-Task Loss

The multi-task loss defined in Eq. (8) is designed to align with retrievers by incorporating both cross-entropy loss and contrastive loss. We conducted experiments with various γ coefficients, as shown in Table 4. The results indicate that increasing γ improves the performance of ADACQR within a certain range, highlighting the crucial role of contrastive loss for alignment. However, the importance of cross-entropy loss is also evident: when γ is excessively high or cross-entropy loss is omitted, the performance declines. Therefore, it concludes that including cross-entropy loss is essential to prevent excessive model variation, illustrating its necessity in the design of this multi-task loss.

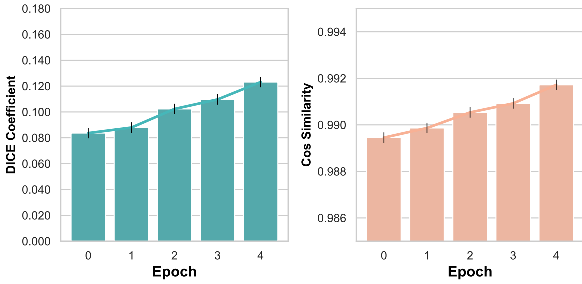

Figure 3: Analysis of the aligned reformulation query across different epochs in Stage 2 training, focusing on the term overlap with the gold passage (DICE coefficient), and semantic similarity to the gold passage (cosine similarity).

5.2 Analysis of the Aligned Query

To evaluate the effectiveness of the aligned reformulation queries, we analyzed the reformulation queries across the first 5 epochs during Stage 2 training in Figure 3. We conducted analyses focusing on the average term overlap and semantic similarity between the queries and the gold passages. The DICE Coefficient (Dice, 1945) is utilized to assess term overlap, while cosine similarity is employed to measure semantic similarity. This analysis indicates that both term overlap and semantic similarity between the reformulated queries and the gold passages exhibit an increasing trend with each epoch in Stage 2, demonstrating the effectiveness of our method in considering both perspectives.

6 Conclusion

In this paper, to achieve alignment between the reformulation model and both term-based and semantic-based retrieval systems, ADACQR is proposed to enhance the generalizability of information-seeking queries across diverse retrieval environments. We developed effective techniques to acquire superior reformulation labels and generate diverse input candidates, boosting the efficiency and robustness of the ADACQR framework. Extensive experiments on two datasets demonstrate the superiority of ADACQR, achieving performance comparable with the LLaMA2-7B model while using only the T5-base.

Limitations

Although ADACQR demonstrates remarkable performance in experimental evaluations, it also has several limitations.

During the ADACQR training process, we leverage ChatGPT for superior reformulation label annotation, and our annotation prompt requires training a basic model, which incurs additional costs and training expenses. Furthermore, due to budget constraints, we did not use more powerful LLMs, such as GPT-4 to obtain reformulation labels, although it is obvious that employing a more powerful LLM would yield better reformulation labels.

Although no further costs are introduced during reformulation model inference, aligning AdaCQR with retrievers introduces additional training time. Furthermore, generating the sorted candidate set for alignment demands extra retrieval time and increased storage capacity.

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.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Haonan Chen, Zhicheng Dou, Kelong Mao, Jiongnan Liu, and Ziliang Zhao. 2024. Generalizing conversational dense retrieval via llm-cognition data augmentation. *arXiv preprint arXiv:2402.07092*.

Yiruo Cheng, Kelong Mao, and Zhicheng Dou. 2024. Interpreting conversational dense retrieval by rewriting-enhanced inversion of session embedding. *arXiv preprint arXiv:2402.12774*.

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.

Lee R Dice. 1945. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*.

Ahmed Elgohary, Denis Peskov, and Jordan Boyd-Graber. 2019. Can you unpack that? learning to rewrite questions-in-context. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5918–5924, Hong Kong, China. Association for Computational Linguistics.

Mara Finkelstein and Markus Freitag. 2023. Mbr and qe finetuning: Training-time distillation of the best and most expensive decoding methods. In *The Twelfth International Conference on Learning Representations*.

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

Xuanli He, Islam Nassar, Jamie Kiros, Gholamreza Haffari, and Mohammad Norouzi. 2022. Generate, annotate, and learn: NLP with synthetic text. *Transactions of the Association for Computational Linguistics*, 10:826–842.

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

Yunah Jang, Kang-il Lee, Hyunkyung Bae, Seungpil Won, Hwanhee Lee, and Kyomin Jung. 2023. Itercqr: Iterative conversational query reformulation without human supervision. *arXiv preprint arXiv:2311.09820*.

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

Sungdong Kim and Gangwoo Kim. 2022. Saving dense retriever from shortcut dependency in conversational search. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10278–10287, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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*.

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. 2020a. Conversational question reformulation via sequence-to-sequence architectures and pretrained language models. *arXiv preprint arXiv:2004.01909*.

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

Yixin Liu, Pengfei Liu, Dragomir Radev, and Graham Neubig. 2022. BRIO: Bringing order to abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2890–2903, Dublin, Ireland. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

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

Kelong Mao, Zhicheng Dou, Fengran Mo, Jiewen Hou, Haonan Chen, and Hongjin Qian. 2023a. 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. 2023b. 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, 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. 2024. History-aware conversational dense retrieval. *arXiv preprint arXiv:2401.16659*.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human-generated machine reading comprehension dataset.

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

Bhargavi Paranjape, Julian Michael, Marjan Ghazvininejad, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2021. Prompting contrastive explanations for commonsense reasoning tasks. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4179–4192, Online. Association for Computational Linguistics.

Chen Qu, Liu Yang, Cen Chen, Minghui Qiu, W Bruce Croft, and Mohit Iyyer. 2020. Open-retrieval conversational question answering. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 539–548.

Filip Radlinski and Nick Craswell. 2017. A theoretical framework for conversational search. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, CHIIR '17, page 117–126, New York, NY, USA. Association for Computing Machinery.

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.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

David A. Smith and Jason Eisner. 2006. Minimum risk annealing for training log-linear models. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 787–794, Sydney, Australia. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Svitlana Vakulenko, Shayne Longpre, Zhucheng Tu, and Raviteja Anantha. 2021. Question rewriting for conversational question answering. In *Proceedings of the 14th ACM international conference on web search and data mining*, pages 355–363.

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

Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2016. Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv:1610.02424*.

Yufei Wang, Wanjun Zhong, Liangyou Li, Fei Mi, Xingshan Zeng, Wenyong Huang, Lifeng Shang, Xin Jiang, and Qun Liu. 2023. Aligning large language models with human: A survey. *arXiv preprint arXiv:2307.12966*.

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, Jiahua Liu, Jingqin Yang, Chenyan Xiong, Paul Bennett, Jianfeng Gao, and Zhiyuan Liu. 2020. Few-shot generative conversational query rewriting. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 1933–1936.

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.

A Discussion

A.1 Effectiveness of Prompt Setting

Type	Prompt Setting	QReCC MRR	R@10
Sparse	0-shot	36.3	54.9
	3-shot (**Random**)	39.1	58.0
	3-shot (**Representative**)	**45.4**	**65.5**
Dense	0-shot	34.5	52.6
	3-shot (**Random**)	37.2	56.0
	3-shot (**Representative**)	**40.1**	**60.2**

Table 5: The annotation results generated by ChatGPT under different prompt settings on the QReCC test set. **Random** denotes examples randomly chosen from the validation set, while **Representative** refers to select examples as described in Section 3.4.

To evaluate the effectiveness of the prompt design method proposed in Section 3.4, we applied our prompt design method for reformulation label annotation on the QReCC test set.

We compared the results with the 0-shot approach (*i.e.*, using only the Instruction and Annotated Sample parts from Table 8) and the 3-shot-random approach (*i.e.*, randomly selecting 3 examples from the validation set). The results are shown in Table 5.

Based on these results, our prompt setting significantly improves performance in both sparse and dense retrieval compared to the 0-shot and 3-shot-random methods, showing the effectiveness of our prompt setting.

A.2 Generalization on Out-Of-Domain (OOD) Dataset

Model	Sparse MRR	Sparse NDCG	Dense MRR	Dense NDCG
ConvGQR(**ID**)	12.4	10.7	25.6	24.3
IterCQR(**OOD**)	13.7	12.2	17.8	16.4
AdaCQR(**OOD**)	14.0	12.6	20.1	18.6

Table 6: Performance of AdaCQR on out-of-distribution dataset. We use the AdaCQR model trained on QReCC and test on TopiOCQA dataset.

To measure the generalization performance of AdaCQR, we trained the model on the QReCC dataset and evaluated it on the TopiOCQA dataset, with the results presented in Table 6. As indicated by the results in Table 6, AdaCQR demonstrates superior generalization performance, outperforming IterCQR in both sparse and dense Retrieval, and surpassing the in-domain model ConvCQR in sparse retrieval.

	QReCC Train	QReCC Valid	QReCC Test	TopiOCQA Train	TopiOCQA Valid	TopiOCQA Test
# Digalogue	10822	769	2775	3509	720	205
# Turns	62701	800	16451	44650	800	2514
# Turns with Gold	28796	800	8209	44650	800	2514

Table 7: The statistics of QReCC and TopiOCQA datasets.

B Experimental Details

B.1 Datasets Details

The QReCC dataset comprises 14K conversations with 80K question-answer pairs, and we aim to retrieve the gold passage from a collection containing 54M passages. Conversely, the TopiOCQA dataset includes 3.9K topic-switching conversations with 51K question-answer pairs, where the passage collection is sourced from Wikipedia and contains about 20M passages. Notably, a few examples from the QReCC and TopiOCQA training sets were randomly partitioned to create respective validation sets. The datasets details are described in Table 7.

B.2 Evaluation Metrics

We evaluate AdaCQR's retrieval performance using several widely used metrics, such as Mean Reciprocal Rank (MRR), Normalized Discounted Cumulative Gain (NDCG), Recall@10, and Recall@100. MRR is a ranking quality metric that considers the position of the first relevant passage among the ranked passages. NDCG@3 evaluates the retrieval results by considering the relevance and the rank of the top three results. Recall@K measures whether the gold passage is present within the top-K results.

C Implementation Details

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

C.1 AdaCQR Details

For the implementatio of AdaCQR, we use Huggingface *transformers* library [3] and *Pytorch Lightning* [4] framework.

We use T5-base[5] (Raffel et al., 2020) as the backbone of AdaCQR. After conducting a comprehen-

[3]https://github.com/huggingface/transformers
[4]https://github.com/Lightning-AI/pytorch-lightning
[5]https://huggingface.co/google-t5/t5-base

sive grid search, we configured the number of candidates $n = 32$, the margin parameter $\lambda = 0.1$, the weight of the contrastive loss $\gamma = 100$, the length penalty parameter $\alpha = 0.6$, and the probability mass parameter in label smooth distribution $\beta = 0.1$. The model parameters are optimized by the `AdamW` optimizer (Loshchilov and Hutter, 2018).

AdaCQR is trained for 10 epochs in Stage 1 with a learning rate set to `2e-5` and 8 epochs in Stage 2 with a learning rate adjusted to `5e-6`. Both stages incorporate linear learning rate schedulers with a warm-up ratio of 0.1.

The vanilla reformulation model G_π in Section 3.4 is trained on reformulation labels of the QReCC dataset acquired by zero-shot prompting with ChatGPT, and the prompt is shown in Appendix D. This model is trained in 10 epochs, and the learning rate is set to `2e-5` with a linear learning rate scheduler with a warm-up ratio of 0.1.

For candidate generation in Section 3.5.2, we used diverse beam search with a diverse penalty of 2.0. The minimum token length for generated candidates is set to 8, and the maximum token length is set to 64. For the generation of reformulation queries, we employed beam search with a beam size of 5, and the maximum token length is set to 64 for generated queries.

C.2 Retrieval Systems Details

We implement the retrieval systems using Faiss (Johnson et al., 2019) and Pyserini (Lin et al., 2021a). For BM25, as in previous work (Mo et al., 2023a; Jang et al., 2023; Yoon et al., 2024), we set $k_1 = 0.82$, $b = 0.68$ in QReCC, and $k_1 = 0.9$, $b = 0.4$ in TopiOCQA. The k_1 controls the non-linear term frequency normalization and b is the scale of the inverse document frequency. For ANCE[6], the maximum token length is set to 128 tokens for reformulation query and 384 tokens for passage.

For both sparse and dense retrieval systems, we retrieved the top 100 relevant passages for each query and obtained the result of evaluation metrics with *pytrec_eval* (Van Gysel and de Rijke, 2018).

D ChatGPT Annotation Details

We use `gpt-3.5-turbo-0125` (OpenAI, 2022)[7] to obtain the initial and superior reformulation labels via zero-shot and few-shots prompting.

For initial reformulation labels of G_π, we use the "Instruction" and "Annotated Sample" parts shown in Table 8, *i.e.*, zero-shot.

For superior reformulation labels for AdaCQR, we utilize the top-3 most challenging demonstrations (*i.e.*, $m = 3$) for the QReCC dataset and the top-5 most challenging demonstrations (*i.e.*, $m = 5$) for the TopiOCQA dataset, *i.e.*, few-shots. The prompts to annotate the QReCC dataset and the TopiOCQA dataset are shown in Table 8 and Table 9, respectively.

To encourage a more deterministic output, we set the temperature to 0.1 and seed is set to 42 for reproductivity. The total consumption to annotate QReCC and TopiOCQA datasets for initial and superior reformulation labels is about 151M tokens, which cost about 120$.

E Case Study

In this section, we present several examples of how AdaCQR succeeded or failed on the QReCC and TopiOCQA datasets.

Table 12 demonstrates a case where AdaCQR successfully retrieved the gold passage through query rewriting, whereas human rewrites failed, showing the superiority of AdaCQR over human rewrites. After being written by AdaCQR, the query is decontextualized, resulting in overlaps while concurrently offering more specific information. This enhanced specificity aids the retriever toward the most relevant passages effectively. Additionally, in Tables 13 and 14, we also show examples of how the AdaCQR and AdaCQR with Expansion models successfully retrieved the gold passage.

F Query Expansion Details

For query expansion, we leverage `LLaMA2-7B-Chat`[8] as the backbone for a fair comparison with prior work (Yoon et al., 2024). The query expansion process involves directly answering the given query (Mo et al., 2023a) and generating relevant keywords (Jagerman

[6] https://huggingface.co/sentence-transformers/msmarco-roberta-base-ance-firstp

[7] https://platform.openai.com/docs/models/gpt-3-5-turbo

[8] https://huggingface.co/meta-llama/Llama-2-7b-chat-hf

et al., 2023). Then the reformulation queries are concated with the generated answers and keywords for retrieval. The prompts employed for query expansion are presented in Table 10 and Table 11.

vLLM framework (Kwon et al., 2023) is used for inference, with the temperature parameter set to 0.5 and the maximum token limit set to 50 during the generation process.

Prompt for QReCC Annotation

Instruction

Given a question and its context, decontextualize the question by addressing coreference and omission issues. The resulting question should retain its original meaning and be as informative as possible, and should not duplicate any previously asked questions in the context.

Demonstrations

Context: [Q: What was Ridley Scott's directing approach to directing? A: Russell Crowe commented about Ridley Scott's directing, I like being on Ridley's set because actors can perform and the focus is on the performers. Q: Were there others who commented about Scott's approach as a director and producer? A: Charlize Theron praised the Ridley Scott's willingness to listen to suggestions from the cast for improvements in the way their characters are portrayed on screen. Q: What was Ridley Scott's style? A: In Ridley Scott's visual style, he incorporates a detailed approach to production design and innovative, atmospheric lighting Q: How did that translate into his films? A: In his movies, Ridley Scott commonly uses slow pacing until the action sequences. Q: What popular movies did he take this approach and use this style? A: Examples of Ridley Scott's directing style include Alien and Blade Runner.]
Question: Is there anything else interesting about his style?
Good Rewrite: Is there anything else interesting about Ridley Scott's style besides his slow pacing until the action sequences?
Bad Rewrite: is there anything else interesting about Ridley Scott's directing style?

Context: [Q: What was the health issues did Bad Brains frontman H.R. have? A: On March 15, 2016, Bad Brains frontman H.R. was reportedly diagnosed with a rare type of headache called Short–lasting unilateral neuralgiform headache with conjunctival injection and tearing (SUNCT syndrome) Q: Was there anything to cure it? A: As diagnostic criteria have been indecisive and its pathophysiology remains unclear, no permanent cure is available for short–lasting unilateral neuralgiform headache with conjunctival injection and tearing (SUNCT syndrome) Q: Are there any other interesting aspects about this article? A: On November 3, 2015, Bad Brains announced on their Facebook page that Dr. Know (Gary Miller) was hospitalized and on life support, after many other musicians reported so.]
Question: What did they do in 2015?
Good Rewrite: What did Bad Brains do in 2015 after Dr. Know (Gary Miller) was hospitalized and on life support?
Bad Rewrite: What do the Bad Brains do in 2015?

<— Omit One demonstration —>

Annotated Sample

Context: [{{current_context}}]
Question: {{current_query}}
Good Rewrite:

Table 8: The prompt used to obtain QReCC annotated labels.

Prompt for TopiOCQA Annotation

Instruction

Given a question and its context, decontextualize the question by addressing coreference and omission issues. The resulting question should retain its original meaning and be as informative as possible, and should not duplicate any previously asked questions in the context.

Demonstrations

Context: [Q: what is the fallacy of the argumentum ad hominem A: That it is not always fallacious, and that in some instances, questions of personal conduct, character, motives, etc., are legitimate and relevant to the issue, as when it directly involves hypocrisy, or actions contradicting the subject's words. Q: what does that last phrase mentioned above mean? A: It is an argumentum(a quarrel; altercation) ad hominem, refers to several types of arguments, not all are fallacious. Q: where does this phrase come from? A: The ancient Greek. Q: are there any philosophers who have written about this? A: Yes, Greeks. Aristotle, Sextus Empiricus, John Locke, Charles Leonard Hamblin, Douglas N. Walton. Q: who is the first mentioned person? A: He was a Greek philosopher and polymath during the Classical period in Ancient Greece. Q: has he written any book? A: He has written on subjects including physics, biology, zoology, metaphysics, logic, ethics, aesthetics, poetry, theatre, music, rhetoric, psychology, linguistics, economics, politics, and government. Q: what did he theorize about dreaming? A: He explained that dreams do not involve actually sensing a stimulus. In dreams, sensation is still involved, but in an altered manner.He also explains that when a person stares at a moving stimulus such as the waves in a body of water, and then look away, the next thing they look at appears to have a wavelike motion. Q: who is the second philosopher mentioned earlier? A: Sextus Empiricus was a Pyrrhonist philosopher and a physician mostly involved in ancient Greek and Roman Pyrrhonism.]
Question: do his teachings/work have any similarities with buddhism?
Good Rewrite: do sextus empiricus' teachings/work have any similarities with buddhism?
Bad Rewrite: there are any similarities between the philosophers mentioned above and buddhism.

Context: [Q: who was the french leader the diplomats were trying to meet with A: French foreign minister Talleyrand Q: what was this affair about? A: Confrontation between the United States and Republican France that led to the Quasi–War. Q: what was this confrontation about? A: To negotiate a solution to problems that were threatening to break out into war. Q: can you name any one who attended the previous meetings? A: Charles Cotesworth Pinckney Q: who was he? A: He was an early American statesman of South Carolina, Revolutionary War veteran, and delegate to the Constitutional Convention. Q: where was he born? A: Charleston, South Carolina]
Question: what was his views regarding slaves?
Good Rewrite: what was charles cotesworth pinckney's views regarding slaves?
Bad Rewrite: whatever were carlos castellanos' views regarding slaves?

<— Omit Three Demonstrations —>

Annotated Sample

Context: [{{current_context}}]
Question: {{current_query}}
Good Rewrite:

Table 9: The prompt used to obtain TopiOCQA annotated labels.

Prompt for Query Expansion (Answer)

Instruction

Given a question, please answer the question in a sentence. The answer should be as informative as possible.

Demonstrations

Question: and by whom was the game the last of us established?
Answer: Andy Gavin and Jason Rubin. Naughty Dog, LLC (formerly JAM Software, Inc.) is an American first–party video game developer based in Santa Monica, California. Founded by Andy Gavin and Jason Rubin in 1984 as an independent developer.

Question: is chelsea a club?
Answer: Yes, chelsea is an English professional football club.

Question: is call me by your name a movie?
Answer: Yes, based on a book of the same name. Call Me by Your Name is a 2017 coming–of–age romantic drama film directed by Luca Guadagnino. Its screenplay, by James Ivory, who also co–produced, is based on the 2007 novel of the same name by Andr Aciman.

Question: where was alan menken born?
Answer: lan Irwin Menken was born on July 22, 1949, at French Hospital in Manhattan, to Judith and Norman Menken.

Question: where was ulysses s. grant from?
Answer: Hiram Ulysses Grant was born in Point Pleasant, Ohio, on April 27, 1822, to Jesse Root Grant, a tanner and merchant, and Hannah Simpson Grant.

Annotated Sample

Question: {{reformulation_query}}
Answer:

Table 10: The prompt for query expansion by directly answering the question.

Prompt for Query Expansion (Keywords)

Instruction

Write a few keywords for the given query.

Annotated Sample

Query: {{reformulation_query}}
Keywords:

Table 11: The prompt for query expansion by giving keywords.

Conversation:
Q1: What was the Securities Act of 1933?
A1: The Securities Act of 1933 has two basic objectives: To require that investors receive financial and other significant information concerning securities being offered for public sale; and. To prohibit deceit, misrepresentations, and other fraud in the sale of securities.
Q2: What is exempt from it?
A2: However, there are exempt securities, under Section 4 of the Securities Act of 1933. These securities are financial instruments that carry government backing and typically have a government or tax-exempt status
Q3: Why was it needed?
A3: The act took power away from the states and put it into the hands of the federal government. The act also created a uniform set of rules to protect investors against fraud.
Q4: What was the reason for creating the 1934 act?
A4: The SEA of 1934 was enacted by Franklin D. Roosevelt's administration as a response to the widely held belief that irresponsible financial practices were one of the chief causes of the 1929 stock market crash.
Q5: What is the largest securities exchange in the world?
A5: The New York Stock Exchange founded on May 17, 1792, is the world's biggest stock exchange in trader value and has a capitalization of $19.223 Trillion USD.
Original Query: How and when when was it created? (**rank: Not Found**)

Human Rewrite: How and when was the largest securities exchange in the world created? (**rank: Not Found**)

AdaCQR(*Ours*): How and when was <u>the Securities Act of 1933</u> created? The New York Stock Exchange founded on May 17, 1792? (**rank: 2**)

Table 12: Case study on QReCC (id: 8_6) when using BM25. The <u>underline</u> part shows the decontextualized information in the reformulation query. We do not show the gold passage of this case because it is too long.

Conversation:
Q1: when did the first episode of sesame street air?
A1: 10 November 1969.
Q2: is it a series?
A2: Yes, an educational children's television series.
Q3: is the series still running?
A3: Yes, it is on its second run.
Q4: when was the pilot episode done?
A4: July 1969.
Q5: is it a puppet show?
A5: Yes.
Q6: name a few characters from the series?
A6: Human characters included Susan and muppet characters included Elmo.
Q7: can you name more muppet characters?
A7: Big Bird and Oscar the Grouch.
Q8: how do the latter look like?
A8: He has a green body with no visible nose.
Q9: does the muppet perform any oscar related play?
A9: UNANSWERABLE
Q10: who performed the aforementioned bird muppet?
A10: It was performed by Caroll Spinney till his retirement.
Q11: who is he by profession?
A11: He was an American puppeteer, cartoonist, author and speaker.
Original Query: did he do comics too? (**rank: Not Found**)

ADACQR(*Ours*): did Caroll Spinney do Caroll comics? (**rank: 1**)

Gold Passage: Caroll Spinney Comics and cartoons While in the Air Force, Spinney wrote and illustrated "Harvey", a comic strip about military life. He also animated a series of black-and-white cartoons called "Crazy Crayon".

Table 13: Successful case study on TopiOCQA (id: 16_12) when using BM25. The underline part shows the decontextualized information in the reformulation query.

Conversation:
Q1: does callie baby die in season 7 episode 18?
A1: No.
Q2: who plays the character mentioned above?
A2: Sara Ramirez.
Q3: apart from acting, does she have a career in any other profession?
A3: She is a singer and songwriter.
Q4: name some of her songs ?
A4: Silent Night.
Q5: what is the significance of the above song?
A5: It is a popular Christmas carol.
Q6: who has written it?
A6: Joseph Mohr
Q7: the above mentioned episode is from which series?
A7: "Grey's Anatomy"
Q8: name some characters of it.
A8: Meredith Grey, Alex Karev, Miranda Bailey and Richard Webber
Q9: what is the real name of the third character mentioned in the above list?
A9: Chandra Wilson
Q10: which movie did she debute in?
A10: "Philadelphia"
Original Query: what was it about? (**rank: Not Found**)

ADACQR: what was the movie "Philadelphia" about? (**rank: Not Found**)

AdaCQR + Expansion: what was the movie "Philadelphia" about? Philadelphia is a 1993 American drama film directed by Jonathan Demme and starring Tom Hanks and Denzel Washington. The movie tells the story of Andrew Beckett, a gay lawyer who is fired from his job because of his sexual orientation, and his subsequent fight for justice and equality in the legal system. Philadelphia, movie, Tom Hanks, Denzel Washington, AIDS, discrimination, lawsuit. (**rank: 1**)

Gold Passage: Philadelphia (film) Introduction Philadelphia is a 1993 American legal drama film written by Ron Nyswaner, directed by Jonathan Demme and starring Tom Hanks and Denzel Washington. It was one of the first mainstream Hollywood films to acknowledge HIV/AIDS, homosexuality, and homophobia. For his role as Andrew Beckett, Hanks won the Academy Award for Best Actor at the 66th Academy Awards, while the song "Streets of Philadelphia" by Bruce Springsteen won the Academy Award for Best Original Song. Nyswaner was also nominated for the Academy Award for Best Original Screenplay, but lost to Jane Campion for "The Piano".

Table 14: Successful case study with query expansion on TopiOCQA (id: 55_11) when using BM25. The part and the part represent the answers and keywords generated by LLM, respectively. These components furnish additional information that assists the retriever in enhancing its performance.