ChatRetriever: Adapting Large Language Models for Generalized and Robust Conversational Dense Retrieval

Anonymous ACL submission

Abstract

Conversational search requires accurate interpretation of user intent from complex multi-turn contexts. This paper presents ChatRetriever, which inherits the strong generalization capability of large language models to robustly represent complex conversational sessions for dense retrieval. To achieve this, we propose a simple and effective dual-learning approach that adapts LLM for retrieval via contrastive learning while enhancing the complex session understanding through masked instruction tuning on high-quality conversational instruction tuning data. Extensive experiments on five conversational search benchmarks demonstrate that ChatRetriever substantially outperforms existing conversational dense retrievers, achieving state-of-the-art performance on par with LLM-based rewriting approaches. Furthermore, ChatRetriever exhibits superior robustness in handling diverse conversational contexts. Our work highlights the potential of adapting LLMs for retrieval with complex inputs like conversational search sessions and proposes an effective approach to advance this research direction. The code and checkpoints are anonymously released at `https://anonymous.4open.science/r/ChatRetriever-8156`.

1 Introduction

Conversational search is rapidly gaining prominence and reshaping how users interact with search engines to foster a more natural information-seeking experience. At the heart of a conversational search system lie two key components: retrieval and generation (Gao et al., 2022; Zhu et al., 2023). The retrieval process is tasked with sourcing relevant passages, which the generation component then uses to craft the final response. Conversational retrieval plays a crucial role in ensuring the accuracy and reliability of the system responses by providing relevant passages (Liu et al., 2023).

Compared to traditional ad-hoc web search, conversational retrieval requires an accurate understanding of the user's real search intent within longer, noisier, and more complex conversational contexts. A "shortcut" approach is to transform the conversational session into a standalone query rewrite, enabling the usage of ad-hoc retrievers for conversational retrieval. However, the additionally introduced rewriting process is hard to directly optimize towards better retrieval, and it also introduces extra search latency from the rewriting step (Yu et al., 2021). In contrast, the end-to-end conversational dense retrieval appears to be more promising, as it directly encodes the original conversational search session and passages into dense representations without additional input processing and can enjoy the efficiency benefit from advanced approximate nearest neighbor search algorithms (e.g. Faiss (Johnson et al., 2021)).

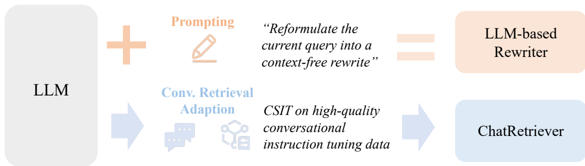

Figure 1: Illustration of adapting LLM for query rewriting and conversational dense retrieval.

Nonetheless, the effectiveness of existing conversational dense retrievers largely trails behind state-of-the-art conversational query rewriting approaches, which leverage large language models (LLMs). Owing to their strong text understanding and generation capabilities, LLM-based rewriters (Mao et al., 2023b; Ye et al., 2023) have demonstrated exceptional effectiveness, even outperforming human rewrites. Given that LLMs are inherently generative models, they can naturally serve as a high-quality conversational rewriter just through prompting (Figure 1). The question that remains is: *whether the potent capabilities of LLMs can be harnessed to substantially enhance the performance of conversational dense retrievers.*

Several studies have explored tuning LLMs for dense retrieval but with a primary focus on ad-hoc search (Asai et al., 2023; Su et al., 2023; Ma et al., 2023; Wang et al., 2024; Muennighoff et al., 2024). While in conversational search, the multi-turn sessions exhibit greater diversity, complex expressions, and longer-tail intents compared to single-turn ad-hoc queries, posing severe challenges to the session representation learning. Additionally, these approaches often rely on manually designed and fixed instruction templates, which can considerably limit their ability to generalize and handle intricate conversational scenarios.

In this work, we propose adapting LLM itself to serve as a powerful conversational dense retriever. To achieve this, we select high-quality conversational instruction tuning data (Ding et al., 2023) as our training data and propose a simple dual-learning approach called *Contrastive Session-Masked Instruction Tuning (CSIT)* for the model training. Specifically, we adopt the classical contrastive ranking loss function (Izacard et al., 2022) to fine-tune LLM from a generative model to a retrieval (or representational) model on the multi-turn instruction (i.e., session)-response pairs, using the special tokens at the end of the input text to represent the entire text. Meanwhile, we mix the basic contrastive learning with a session-masked instruction tuning objective, where we mask all tokens except the special tokens of the session when computing the language modeling loss of the response tokens. The incorporation of this generative instruction tuning loss forces a strong enhancement in the learning of the complex session representation since the response tokens have to be generated solely based on the special tokens representing the session. Furthermore, it also helps retain the strong generalization capability of LLM for retrieval.

Our resulting model, which we call **ChatRetriever**, can inherit the strong generalization capability of LLM to robustly represent complex conversational sessions for dense retrieval. We conducted extensive experiments across five conversational search benchmarks, where ChatRetriever substantially outperforms existing conversational dense retrievers. Notably, it achieves absolute NDCG@3 improvements of 6.8% and 12.2% on CAsT-20 and CAsT-21, respectively, matching the performance of the leading LLM-based conversational query rewriting methods. Beyond standard evaluations using fixed conversational trajectories, we also developed two robustness evaluation methods to assess the resilience of conversational retrieval approaches by altering the historical context. ChatRetriever demonstrates markedly more stable performance in our robustness test, showcasing its superior robustness in comparison to baselines when faced with varied contexts.

Our contributions can be summarized as:

(1) We introduce ChatRetriever, the first LLM-adapted conversational dense retriever, which substantially outperforms existing conversational dense retrievers and achieves performance comparable to LLM-based rewriting approaches.

(2) We propose Contrastive Session-Masked Instruction Tuning for such a retrieval-oriented adaption for LLM, which can help achieve better complex session representation and generalization.

(3) We design two robustness evaluation methods for conversational retrieval by systematically varying the conversation contexts. Results highlight ChatRetriever's superior generalization capability in handling diverse conversational search scenarios.

2 Related Work

Conversational search has seen the development of two primary approaches: conversational query rewriting (CQR) and conversational dense retrieval (CDR). The former approach transforms the conversational search problem into a traditional ad-hoc search problem by reformulating the conversational context into a standalone query. Techniques in this area range from selecting useful tokens from the context (Voskarides et al., 2020; Lin et al., 2021b) to training generative rewriters based on session-rewrite pairs (Yu et al., 2020; Wu et al., 2022; Mao et al., 2023a; Mo et al., 2023a). Inspired by the strong language generation capability of LLMs, some studies (Mao et al., 2023b; Ye et al., 2023; Yoon et al., 2024) propose to leverage LLMs as query rewriters and achieve amazing performance. Conversational dense retrieval (CDR), on the other hand, directly encodes the entire conversational session for end-to-end dense retrieval (Yu et al., 2021). Efforts in this direction have focused on improving session representation through various perspectives such as context denoising (Mao et al., 2022a; Mo et al., 2023b; Mao et al., 2023c), data augmentation using other corpus and LLMs (Lin et al., 2021a; Mao et al., 2022b; Dai et al., 2022; Jin et al., 2023; Chen et al., 2024; Mo et al., 2024b), and hard negative mining (Kim and Kim, 2022; Mo et al., 2024a).

LLM-based and instruction-aware retrieval. Existing research has demonstrated that similar to the scaling laws (Kaplan et al., 2020) observed in LLMs, increasing the scale of models, data, and computing resources can also enhance the performance of retrieval models (Ni et al., 2022). To incorporate the ability to follow instructions into retrievers, some studies (Su et al., 2023; Asai et al., 2023) propose the creation of fixed instruction templates for various retrieval tasks, and use these instruction-enhanced datasets to train the retrievers. Moreover, there have been efforts to adapt LLMs for retrieval purposes by training on improved search data (Ma et al., 2023; Wang et al., 2024) or developing new search-oriented training objectives (Li et al., 2023). However, these approaches often rely on manually designed and fixed instruction templates, which can limit the generalization capabilities of the retrievers across diverse instructions. Additionally, they are typically designed for single-turn ad-hoc search, lacking the capability to comprehend long and complex search sessions. In contrast to LLMs, which can smoothly understand a wide range of complex user inputs, existing LLM-based retrievers still exhibit a large gap in their generalization capabilities, particularly in the context of conversational search.

3 Methodology

We describe our simple and effective dual-learning approach, *Contrastive Session-Masked Instruction Tuning (CSIT)*, which is designed to adapt LLM to a generalized and robust conversational dense retriever. An overview is shown in Figure 2.

Contrastive instruction tuning. Recent works have demonstrated the effectiveness of simply using the contrastive ranking loss to adapt LLM to a retriever (Asai et al., 2023; Su et al., 2023; Ma et al., 2023; Wang et al., 2024; Muennighoff et al., 2024). However, their generalization capability can be limited as they overfit the narrow distribution of ad-hoc queries and fixed instruction templates they were trained on. We fine-tune LLM on diverse conversational instruction tuning data for more general conversational retrieval adaption. Specifically, given a training sample $\{(x, y^+)\}$ from conversational instruction tuning dataset, where x comprises all historical turns and the current instruction (we call x a *session*) and y is the response, we fine-tune LLM with the contrastive ranking loss:

$$\mathcal{L}_{\text{C}} = -\log\frac{\phi(x, y^+)}{\phi(x, y^+) + \sum_{y^- \in D^-} \phi(x, y^-)}, \quad (1)$$

where $\phi(x, y) = \exp((E(x) \cdot E(y))/\tau)$, $E(\cdot)$ is the shared text encoder of the retriever. D^- is a negative response collection for x. τ is a hyperparameter temperature.

To encode text with LLM, we append t special tokens ($[\text{EMB}_1]$, ..., $[\text{EMB}_t]$) to the end of the input text and utilize the representation of the last token ($[\text{EMB}_t]$) as the comprehensive representation of the entire text. This approach is analogous to the text-level chain-of-thought (CoT) (Wei et al., 2020) for LLMs. We hypothesize that these t consecutive special tokens act as a representational chain-of-thought, expanding and guiding the learning space to achieve a more effective representation.

Session-masked instruction tuning. To enhance the generalized encoding of complex search sessions, we integrate a session-masked instruction tuning objective with the fundamental contrastive learning. Given a training sample (x, y^+), we concatenate the instruction and the response to form one input sequence s:

$$\begin{aligned} s = [&x_1, ..., x_N, [\text{EMB}_1], ..., [\text{EMB}_t], y_1^+, \\ &..., y_M^+, [\text{EMB}_1], ..., [\text{EMB}_t]], \end{aligned} \quad (2)$$

where x_i and y_i^+ represent the i-th token of the session and the response, respectively. N and M denote the total number of tokens in the session and the response, respectively. We then input this sequence into the LLM to obtain the token representations. Specifically, the representations for the $(N + t)$ **session** tokens are obtained through a standard auto-regressive process. However, for the subsequent $(M+t)$ **response** token representations, we mask the N session token representations and allow only the attention of t special session tokens and their preceding response tokens. We achieve it by applying a customized attention mask matrix illustrated on the right side of Figure 1. Correspondingly, the loss function of the session-masked instruction tuning is defined as:

$$\mathcal{L}_{\text{S}} = -\frac{1}{M}\sum_{i=1}^{M} \log p(y_i^+ | y_1^+, ..., y_{i-1}^+, \mathbf{x}_{1:t}), \quad (3)$$

where $\mathbf{x}_{1:t}$ are the representations of the t session special tokens, which have been contextualized by the N session tokens.

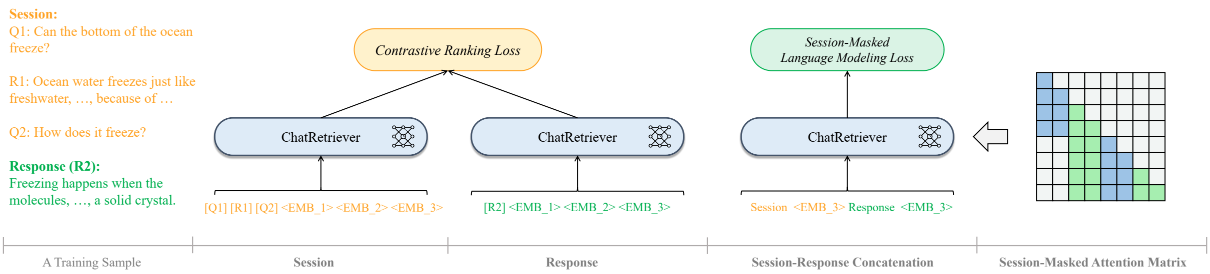

Figure 2: Overview of CSIT. We fine-tune LLM to be ChatRetriever using dual learning objectives. We use the last special token (i.e., <EMB_3>) to represent the input text, which can be session or response. In the session-masked attention matrix, the blue squares denote the session or the response tokens while the green squares denote their special tokens.

By masking the session text and forcing correct generation for the response tokens, we build a closer connection between the session representation and the response token representations. The model has to perform a more nuanced understanding of the complex session and accurately encode them into the t session special tokens.

We combine the contrastive instruction tuning and the session-masked instruction tuning to form the final training objective of ChatRetriever:

$$\mathcal{L} = \mathcal{L}_{\text{C}} + \alpha \mathcal{L}_{\text{S}}, \tag{4}$$

where α is a hyperparameter to balance the two losses.

Discussion. Our dual-learning approach CSIT takes inspiration from several notable works in LLM-based retrieval and input compression such as RepLLaMA (Ma et al., 2023), E5$_{\text{mistral-7b}}$ (Wang et al., 2024), GRIT (Muennighoff et al., 2024), Gisting (Mu et al., 2023), and AutoCompressor (Chevalier et al., 2023). However, CSIT distinguishes from them in the following key aspects: (1) RepLLaMA and E5$_{\text{mistral-7b}}$ primarily focus on contrastive learning using (synthetic) ad-hoc search data with pre-defined instruction templates, which is hard to generalize to complex conversational search scenarios. (2) GRIT aims to build a unified model for both retrieval and generation, incorporating vanilla instruction tuning and using different training data for its contrastive learning and instruction tuning. (3) The mechanism of our session-masked instruction tuning shares similarities with Gisting and AutoCompressor, but they are for a completely different target: improving long-context language modeling, not retrieval. In contrast, CSIT stands out from these works by specifically addressing the challenges of adapting LLM generalized to complex conversational retrieval.

4 Experiments

4.1 Setup

Training data. We fine-tune LLM to be ChatRetriever on high-quality conversational instruction tuning datasets. We select training samples that are informative, diverse, and exhibit information-seeking intents. Our final training data comprises two sources: (1) The *Question About the World* subset of UltraChat (Ding et al., 2023) and (2) MSMARCO (Nguyen et al., 2016) passage ranking dataset. Ultrachat is a multi-turn instruction tuning dataset while MSMARCO can be deemed as a single-turn search-oriented instruction tuning dataset by treating the query as the instruction and the positive passage as the response. We find that incorporating MSMARCO is important to improve the basic (ad-hoc) retrieval performance.

Evaluation data and metrics. We conduct evaluations on five public conversational search benchmarks, including QReCC (Anantha et al., 2021), TopiOCQA (Adlakha et al., 2022), CAsT-19 (Dalton et al., 2020), CAsT-20 (Dalton et al., 2021), and CAsT-21 (Dalton et al., 2022). The retrieval corpus sizes of these five datasets are in the tens of millions. Among them, the large-scale QReCC and TopiOCQA have training sets, while the other three CAsT datasets are small datasets that only have test sets. We mainly report NDCG@3 to evaluate the retrieval performance, as conversational search is more concerned with the top results (Dalton et al., 2021).

Baselines. We compare ChatRetriever against the following three types of retrieval baselines. The first is CQR baselines, including T5QR (Lin et al., 2020), ConvGQR (Mo et al., 2023a), and LLM4CS (Mao et al., 2023b). The original

Model	Base Model	#Model Parameter	QReCC	TopiOCQA	CAsT-19	CAsT-20	CAsT-21
			Conversational Query Rewriting				
T5QR	T5-base (Raffel et al., 2020)	250M	31.8	22.2	41.7	29.9	33.0
ConvGQR	T5-base (Raffel et al., 2020)	250M	41.0	24.3	43.4	33.1	27.3
LLM4CS (REW)	ChatGPT-3.5 (OpenAI)	Unknown	-	-	43.1	35.7	40.4
LLM4CS (RAR)	ChatGPT-3.5 (OpenAI)	Unknown	-	-	45.3	39.5	44.9
LLM4CS	ChatGPT-3.5 (OpenAI)	Unknown	-	-	51.5	**45.5**	49.2
			LLM-based Retrieval				
LLM Embedder	BGE (Xiao et al., 2023)	110M	50.5	22.4	36.6	15.3	31.2
INSTRCUTOR	GTR-XL (Ni et al., 2022)	1.5B	42.3	12.3	26.8	17.3	32.4
RepLLaMA	LLaMA-2 (Touvron et al., 2023)	7B	31.8	15.0	31.6	18.3	32.7
$\text{E5}_{\text{mistral-7b}}$	Mistral (Jiang et al., 2023)	7B	32.9	16.9	31.3	15.4	32.4
GRIT	Mistral (Jiang et al., 2023)	7B	33.5	17.3	30.9	19.3	33.6
			Conversational Dense Retrieval				
Conv-ANCE	ANCE (Xiong et al., 2021)	110M	45.6	20.5	34.1	27.5	34.2
ConvDR	ANCE (Xiong et al., 2021)	110M	35.7	26.4	43.9	32.4	37.4
DialogInpainter	T5-Large (Raffel et al., 2020)	770M	-	-	47.0	33.2	-
LeCoRE	SPLADE (Formal et al., 2022)	110M	48.5	31.4	42.2	29.0	32.3
ChatRetriever	Qwen (Bai et al., 2023)	7B	**52.5**†	**40.1**†	**52.1**†	40.0†	**49.6**†

Table 1: Results of the normal evaluation on five conversational search benchmarks. The base models of CQR methods are their rewriters and the model parameters are also counted as the rewriter's parameters. † denotes significant differences to baselines ($p < 0.05$). The best results are bold and the second-best results are underlined.

LLM4CS has three prompting methods: REW, RAR, and RTR, and it requires multiple rounds of generation, which is time-consuming. For efficiency consideration, we additionally compare with its two single-generation variants based on RAR and REW; The second is CDR baselines, including ConvDR (Yu et al., 2021), Conv-ANCE (Mao et al., 2023c), DialogInpainter (Dai et al., 2022), and LeCoRE (Mao et al., 2023c); The third is the LLM-based retriever baselines, including INSTRUCTOR (Su et al., 2023), LLM Embedder (Zhang et al., 2023), RepLLaMA (Ma et al., 2023), $\text{E5}_{\text{mistral-7b}}$ (Wang et al., 2024), and GRIT (Muennighoff et al., 2024). More baseline details on in Appendix A.

Implementations. We initialize ChatRetriever with Qwen-7B-Chat (Bai et al., 2023) and train it on eight 40G A100 GPUs using LoRA (Hu et al., 2022) with a maximum input sequence length of 1024. The training process involves 2500 steps with a learning rate of 1e-4, a gradient accumulation of 4 steps, a batch size of 64, and 4 hard negatives per sample. For consistency, we adopt the *chatml* input format of Qwen-Chat to form the input of ChatRetriever. We add three special tokens (i.e., <|*extra_1*|>, <|*extra_2*|>, and <|*extra_3*|>) at the end of the instructions and responses. For baseline comparisons, we adhere to the implementation settings specified in their original papers.

4.2 Normal Evaluation

The retrieval performance comparisons on the five datasets are reported in Table 1. Our proposed ChatRetriever outperforms all the baseline methods across these datasets. Existing conversational dense retrievers are constrained by limited model capacity and data quality, resulting in suboptimal performance for conversational retrieval tasks. Prior to ChatRetriever, there was a considerable performance gap between existing conversational dense retrieval methods and the state-of-the-art LLM-based conversational query rewriter (i.e., LLM4CS). Specifically, the absolute gaps between the best existing CDR model and LLM4CS were 1.6%, 12.2%, and 11.8% on the three CAsT datasets, respectively. However, ChatRetriever can achieve comparable or even superior performance to LLM4CS, highlighting the high potential of end-to-end conversational dense retrieval compared to the two-stage approach of conversational query rewriting methods. If we force LLM4CS to generate a single output (RAR) or only consider query rewriting (REW) for efficiency, the advantages of ChatRetriever become even more pronounced, with over 4% absolute gains. We also observe that existing LLM-based retrievers do not perform well on conversational retrieval tasks. This can be at-

Model	Partial Response Modification						Full Context Modification					
	CAsT-19		CAsT-20		CAsT-21		CAsT-19		CAsT-20		CAsT-21	
	NDCG@3↑	Diff.↓	NDCG@3↑	Diff.↓	NDCG@3↑	Diff.↓	Mean↑	SD↓	Mean↑	SD↓	Mean↑	SD↓
LLM4CS	50.4	1.1	43.8	1.7	49.4	0.2	49.7	1.5	44.0	1.1	48.4	1.4
ConvDR	44.3	0.4	31.0	1.4	34.8	2.6	39.3	3.4	30.2	2.6	35.8	2.9
LeCoRE	44.5	2.3	25.4	3.6	29.9	2.4	42.0	1.9	28.3	2.2	31.0	2.3
ChatRetriever	52.2	0.1	39.5	0.5	48.9	0.7	51.5	1.6	45.8	1.7	48.8	1.8

Table 2: Results of the robust evaluation. *Diff.* represents the absolute difference compared to the results in Table 1 and *SD* represents the standard deviation, where a smaller value means more stable.

tributed to the fact that they are fine-tuned solely on templated instructions, which fails to fully leverage the generalization capabilities of LLMs to handle complex and diverse conversational scenarios.

4.3 Robustness Evaluation

Existing evaluations for conversational retrieval are mainly conducted on fixed conversation trajectories. In this section, we evaluate the robustness of conversational retrievers in different contexts. Our principle is modifying the context but fixing the current query (i.e., search intents) for each turn so that the original relevance labels can be re-used. Specifically, we propose the following two types of context modification:

(1) *Partial response modification:* We do not use the provided responses in the evaluation dataset. Instead, for each turn, we input the current query, the context, and the top-3 passages retrieved by the conversational retriever, and prompt LLM to generate the response. The simulated online nature of generating responses turn-by-turn better matches how conversational retrieval systems are used in practice. However, a problem with this online evaluation manner is that the query of the next turn in the original dataset may become unreasonable after modifying its last response (Li et al., 2022). We propose a simple heuristic method to tackle this problem with LLM. Specifically, we prompt LLM to judge whether the current query is reasonable given the context. If not, we replace the current query with its human rewrite to make it stand on its own without needing external context. Otherwise, we can use the original query. The prompts can be found in Appendix B.

(2) *Full context modification:* For each turn, we supply the original query and its human-modified version to the LLM, prompting it to generate new contexts (See Appendix C). We finally got five different contexts for each turn.

We evaluate conversational retrievers based on different contexts generated by these two modification methods using ChatGPT 3.5. For the partial response modification setting, we report the retrieval performances and their absolute differences (*Diff.*) compared to the original counterpart results reported in Table 1. For the full context modification setting, we report the *Mean* performance of different runs and their *standard deviation (SD)*. The robust evaluation results are shown in Table 2.

For the partial response modification setting, it shows that the performance changes of ChatRetriever are the smallest. By referring to Table 1, we also observe a general degradation in retrieval performance compared to the original context. This degradation may stem from the retrieved passages being inaccurate, consequently leading to inaccurate responses, and then affecting the retrieval performance of the subsequent turns.

For the full context modification setting, the robustness of ChatRetriever is further highlighted by its small average standard deviation of 1.7, which is lower compared to the 3.0 and 2.1 standard deviations observed for ConvDR and LeCoRE, respectively. These results demonstrate the strong robustness of ChatRetriever to different conversational search contexts. In contrast, the LLM4CS, which utilizes ChatGPT for query rewriting, shows an even lower standard deviation of 1.3, demonstrating the superior robustness of ChatGPT for conversational query rewriting.

4.4 Ablation Studies

We build four ablations to study the effects of our proposed training approach: (1) *w/o R-CoT*: removing the representational CoT; (2) *w/o SIT*: removing the session-masked instruction tuning; (3) *with Vanilla IT*: replacing the session-masked instruction tuning with vanilla instruction tuning.

Table 4 shows the ablation results. We find that either removing the representational CoT or removing or replacing session-masked instruction tun-

Base LLM	Model Parameter	Base/Chat	Training	CAsT-19	CAsT-20	CAsT-21
Qwen	1.8B	Chat	Full	38.8	33.7	45.2
Qwen	1.8B	Chat	LoRA	35.1	31.9	42.4
Qwen	7B	Base	LoRA	46.9	37.7	46.5
Qwen	7B	Chat	LoRA	50.5	40.0	49.6
LLaMA-2	7B	Chat	LoRA	47.3	38.4	49.1
Mistrial	7B	Chat	LoRA	49.5	39.2	49.6

Table 3: Performance comparisons of ChatRetrievers under different settings with different backbone LLMs.

Ablation	CAsT-19	CAsT-20	CAsT-21
w/o SIT	49.5	36.8	45.8
w/o R-CoT	49.9	38.5	47.5
with Vanilla IT	51.1	39.3	48.4
CSIT	52.1	40.0	49.6

Table 4: Results of ablation studies.

ing can lead to performance degradation. By contrast, the session-masked instruction tuning, which achieves 6.6% relative performance gains across the three CAsT datasets on average, is shown to be more effective than representational CoT, which achieves 3.4% relative performance gains on average. The results suggest that our two techniques have positive effects in helping adapt LLMs for conversational retrieval. We also studied the influence of the number of special CoT tokens, which can be found in Appendix D.

4.5 Influence of LLMs

Table 3 shows the comparisons between different settings about the backbone LLM of ChatRetriever.

(1) **Base vs. Chat.** Our results indicate that the Chat model outperforms the Base model, which aligns with our expectations. We hypothesize that the ability to follow instructions well is indicative of strong generalization capabilities, which are crucial for complex conversational search tasks. Therefore, the Chat model, having been fine-tuned for conversational instructions, provides a more appropriate foundation for this task.

(2) **Different LLMs.** We find that different LLMs have similar performance under our training recipe. The relatively worst variation based on LLaMA-2 still largely outperforms existing conversational dense retrieval baselines on the more complex CAsT-20 and CAsT-21 datasets, and also outperforms smaller ChatRetrievers.

(3) **LoRA vs. full parameter tuning.** Due to constraints in computing resources, our investigation into training modes (i.e., LoRA vs. full parameter tuning) was limited to the 1.8B scale model. Our findings indicate that employing LoRA training yields inferior performance compared to full parameter tuning. However, this may be attributed to the LoRA parameter capacity being insufficient for the 1.8B model.

4.6 Influence of Training Data

Fine-tuning on different data sources. Table 6 presents the performance of ChatRetriever when trained solely on UltraChat, solely on MSMARCO, and on a combination of QReCC+MSMARCO (i.e., replacing UltraChat with the QReCC's training set). The model performance is evaluated using both session inputs and human rewrite inputs (i.e., converted to ad-hoc search). We find that training exclusively on UltraChat leads to a decline in performance for both input types, with a more pronounced degradation observed for the rewrite input. Conversely, training solely on MSMARCO yields comparable results for the rewrite input but considerably worse performance for the session input. These results suggest that MSMARCO effectively enhances the ad-hoc retrieval capabilities of LLMs, possibly due to its well-curated hard negatives. However, ad-hoc search data from MSMARCO alone is insufficient for transferring the generalization capability of LLMs to the more complex context of conversational search. The traditional conversational QA data (i.e., QReCC) is also not highly effective for LLMs in learning a diverse range of complex conversational patterns. To optimize LLM to be a universal conversational retriever, we recommend combining general conversational instruction tuning data (e.g., UltraChat) with ad-hoc search-oriented instruction tuning data (e.g., MSMARCO).

Continually fine-tuning baselines on the same training data of ChatRetriever. In Table 1, we follow the original training settings of the

Methods	QReCC		TopiOCQA		CAsT-19		CAsT-20		CAsT-21	
	Original	New	Original	New	Original	New	Original	New	Original	New
GRIT	33.5	48.3	17.3	36.0	30.9	47.1	19.3	35.7	33.6	45.3
Conv-ANCE	45.6	44.8	20.5	21.6	34.1	35.0	27.5	30.5	34.2	36.0
ConvDR	35.7	36.0	26.4	24.9	43.9	43.2	32.4	30.9	37.4	35.5
LeCoRE	48.5	46.1	31.4	31.0	42.2	42.9	29.0	30.1	32.3	33.4
ChatRetriever	52.5		40.1		52.1		40.0		49.6	

Table 5: Results of continually fine-tuning baselines on the training data of ChatRetriever. “Original” and “New” denote the performance before and after fine-tuning, respectively.

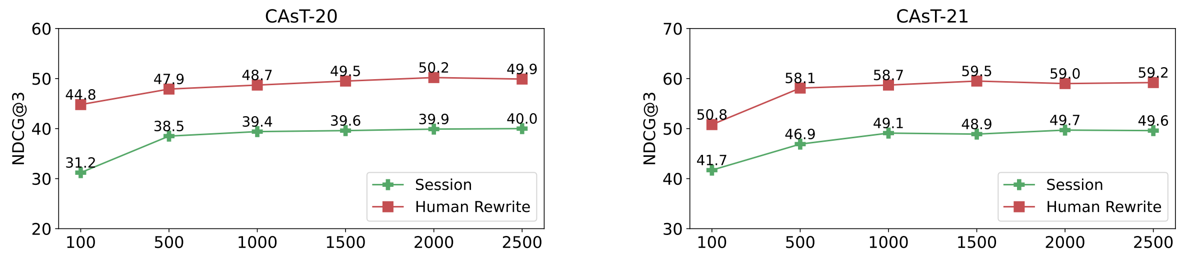

Figure 3: Performance of ChatRetriever at different training steps.

Data Source	CAsT-20		CAsT-21	
	Session	Rewrite	Session	Rewrite
Only U	39.5	43.7	46.5	50.0
Only M	18.3	49.8	34.1	58.9
Q+M	31.5	46.9	42.4	47.9
U+M	40.0	49.9	49.6	59.2

Table 6: Comparisons of using different data sources combinations for training. U, M, and Q represent UltraChat, MSMARCO, and QReCC, respectively.

baselines. Here, we further fine-tune baselines on the training data of ChatRetriever. Results are shown in Table 5 and we find: (1) GRIT, a unified retrieval and generation model based on LLM, showed substantial performance improvement after fine-tuning on conversational instruction tuning data. Its performance approached that of ChatRetriever without session-masked instruction tuning, although it still lagged behind the final ChatRetriever. (2) The performance of Conv-ANCE, ConvDR, and LeCoRE did not show noticeable improvements and even experienced declines in QReCC and TopiOCQA. This may be because that the newly introduced training data disrupted their original in-domain training-test settings, as they were initially trained on the in-domain training sets of QReCC and TopiOCQA. This also highlights the robust generalization of ChatRetriever, which, when trained only on general conversational instruction tuning data, can effectively adapt to various conversational search test sets.

Data volume. Figure 3 shows the performance of ChatRetriever across various training steps. It is observed that the performance attains a relatively high level at 500 steps and subsequently experiences marginal improvements as the number of training steps increases. The performance stabilizes upon reaching 2500 steps. Furthermore, the trends for inputs with sessions and human rewrites are similar. These findings suggest that, under our framework, adapting LLMs to function effectively as conversational retrievers may require only a small amount of high-quality data.

5 Conclusion

In this paper, we introduce ChatRetriever, a large conversational retrieval model adapted from LLM. We propose a novel contrastive session-masked instruction tuning approach for this adaptation and fine-tune LLM on high-quality conversational instruction tuning data. Experimental results on five conversational retrieval datasets demonstrate the superior performance and robustness of ChatRetriever. Looking ahead, we aim to further explore and expand the generalization capabilities of ChatRetriever in a broader range of complex IR scenarios beyond conversational search, such as legal case retrieval, product search, and other instruction-followed search tasks. We envision ChatRetriever to be as versatile as LLMs, capable of accepting and understanding any conversational inputs and retrieving useful information for those inputs.

Limitations

Efficiency. As indicated in Table 1, ChatRetriever is a 7B model which is much larger than existing CDR models. Our preliminary findings (Section 4.5) suggest that the large model size is a crucial factor for ChatRetriever's exceptional performance. However, this also raises efficiency concerns. With an embedding dimension of 4096, ChatRetriever incurs higher time and storage costs for indexing and retrieval than existing CDR models. Nevertheless, on the one hand, ChatRetriever's enhanced retrieval accuracy potentially reduces the need for extensive passage re-ranking, which could, in real-world applications, offset the initial higher costs by ultimately reducing the total time spent on ranking. On the other hand, we view ChatRetriever as a promising research direction in leveraging the potent capabilities of LLMs for more complex and potentially universal retrieval tasks. We are exploring the possibility of distilling ChatRetriever into a more efficient, smaller model.

Hard Negatives. Unlike typical search datasets that provide a large retrieval corpus, the conversational instruction tuning dataset we used (i.e., UltraChat) consists of only multi-turn instructions (i.e., sessions) and responses. In this work, we simply chose the CAsT-21 corpus for the hard negative mining of UltraChat (see Appendix A.3). However, as existing studies have shown, hard negatives are crucial for improving retrieval performance (Zhan et al., 2021; Zhou et al., 2022). Therefore, a better strategy for mining hard negatives tailored to instruction tuning data is desirable. We plan to explore using LLMs to generate hard negatives for instructions similar to (Wang et al., 2024).

Generalizability. ChatRetriever substantially outperforms existing CDR models in understanding and retrieving information for complex multi-turn inputs and achieves comparable performance to state-of-the-art LLM-based rewriting, showcasing its strong generalization capability. However, it has not yet achieved the same level of generalization as LLMs, particularly in following complex retrieval instructions, addressing very detailed information needs, or performing in-context learning across various specific domains. It is worth noting that existing instruction-aware retrievers (Su et al., 2023; Zhang et al., 2023; Muennighoff et al., 2024) also have limitations in perceiving complex (multi-turn) instructions that largely fall short of the generality of LLMs, as highlighted in this work (Table 1) and also in recent studies (Oh et al., 2024; Weller et al., 2024). As stated in our conclusion, we are committed to further advancing ChatRetriever's generalization capabilities to match those of LLMs.

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 *NAACL-HLT*, pages 520–534. Association for Computational Linguistics.

Akari Asai, Timo Schick, Patrick S. H. Lewis, Xilun Chen, Gautier Izacard, Sebastian Riedel, Hannaneh Hajishirzi, and Wen-tau Yih. 2023. Task-aware retrieval with instructions. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 3650–3675. Association for Computational Linguistics.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

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*.

Alexis Chevalier, Alexander Wettig, Anirudh Ajith, and Danqi Chen. 2023. Adapting language models to compress contexts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 3829–3846. Association for Computational Linguistics.

Zhuyun Dai, Arun Tejasvi Chaganty, Vincent Y Zhao, Aida Amini, Qazi Mamunur Rashid, Mike Green, and Kelvin Guu. 2022. Dialog inpainting: Turning documents into dialogs. In *International Conference on Machine Learning*, pages 4558–4586. PMLR.

Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. 2020. Trec 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*.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Enhancing chat language models by scaling high-quality instructional conversations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 3029–3051. Association for Computational Linguistics.

Thibault Formal, Carlos Lassance, Benjamin Piwowarski, and Stéphane Clinchant. 2022. From distillation to hard negative sampling: Making sparse neural IR models more effective. In *SIGIR*, pages 2353–2359. ACM.

Jianfeng Gao, Chenyan Xiong, Paul Bennett, and Nick Craswell. 2022. Neural approaches to conversational information retrieval. *arXiv preprint arXiv:2201.05176*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Hamish Ivison, Yizhong Wang, Valentina Pyatkin, Nathan Lambert, Matthew E. Peters, Pradeep Dasigi, Joel Jang, David Wadden, Noah A. Smith, Iz Beltagy, and Hannaneh Hajishirzi. 2023. Camels in a changing climate: Enhancing LM adaptation with tulu 2. *CoRR*, abs/2311.10702.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Trans. Mach. Learn. Res.*, 2022.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *CoRR*, abs/2310.06825.

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, Singapore, December 6-10, 2023*, pages 6649–6675. Association for Computational Linguistics.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2021. Billion-scale similarity search with gpus. *IEEE Trans. Big Data*, 7(3):535–547.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *CoRR*, abs/2001.08361.

Sungdong Kim and Gangwoo Kim. 2022. Saving dense retriever from shortcut dependency in conversational search.

Chaofan Li, Zheng Liu, Shitao Xiao, and Yingxia Shao. 2023. Making large language models A better foundation for dense retrieval. *CoRR*, abs/2312.15503.

Huihan Li, Tianyu Gao, Manan Goenka, and Danqi Chen. 2022. Ditch the gold standard: Re-evaluating conversational question answering. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 8074–8085. Association for Computational Linguistics.

Sheng-Chieh Lin, Jheng-Hong Yang, and Jimmy Lin. 2021a. Contextualized query embeddings for conversational search. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

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*.

Sheng-Chieh Lin, Jheng-Hong Yang, Rodrigo Nogueira, Ming-Feng Tsai, Chuan-Ju Wang, and Jimmy Lin. 2021b. Multi-stage conversational passage retrieval: An approach to fusing term importance estimation and neural query rewriting. *ACM Transactions on Information Systems (TOIS)*, 39(4):1–29.

Nelson F. Liu, Tianyi Zhang, and Percy Liang. 2023. Evaluating verifiability in generative search engines. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 7001–7025. Association for Computational Linguistics.

Xueguang Ma, Liang Wang, Nan Yang, Furu Wei, and Jimmy Lin. 2023. Fine-tuning llama for multi-stage text retrieval. *CoRR*, abs/2310.08319.

Kelong Mao, Zhicheng Dou, Bang Liu, Hongjin Qian, Fengran Mo, Xiangli Wu, Xiaohua Cheng, and Zhao Cao. 2023a. Search-oriented conversational query editing. In *ACL (Findings)*, volume ACL 2023 of *Findings of ACL*. 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, Singapore, December 6-10, 2023*, pages 1211–1225. Association for Computational Linguistics.

Kelong Mao, Zhicheng Dou, and Hongjin Qian. 2022a. Curriculum contrastive context denoising for few-shot conversational dense retrieval. In *Proceedings of the 45th International ACM SIGIR conference on research and development in Information Retrieval (SIGIR)*.

Kelong Mao, Zhicheng Dou, Hongjin Qian, Fengran Mo, Xiaohua Cheng, and Zhao Cao. 2022b. Convtrans: Transforming web search sessions for conversational dense retrieval. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

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*, pages 3193–3202.

Fengran Mo, Kelong Mao, Yutao Zhu, Yihong Wu, Kaiyu Huang, and Jian-Yun Nie. 2023a. ConvGQR: generative query reformulation for conversational search. In *ACL*, volume ACL 2023. 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 *29th ACM SIGKDD Conference On Knowledge Discover and Data Mining (SIGKDD)*.

Fengran Mo, Chen Qu, Kelong Mao, Tianyu Zhu, Zhan Su, Kaiyu Huang, and Jian-Yun Nie. 2024a. History-aware conversational dense retrieval. *arXiv preprint arXiv:2401.16659*.

Fengran Mo, Bole Yi, Kelong Mao, Chen Qu, Kaiyu Huang, and Jian-Yun Nie. 2024b. Convsdg: Session data generation for conversational search. *arXiv preprint arXiv:2403.11335*.

Jesse Mu, Xiang Li, and Noah D. Goodman. 2023. Learning to compress prompts with gist tokens. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Niklas Muennighoff, Hongjin Su, Liang Wang, Nan Yang, Furu Wei, Tao Yu, Amanpreet Singh, and Douwe Kiela. 2024. Generative representational instruction tuning. *CoRR*, abs/2402.09906.

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. In *CoCo@ NIPS*.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernández Ábrego, Ji Ma, Vincent Y. Zhao, Yi Luan, Keith B. Hall, Ming-Wei Chang, and Yinfei Yang. 2022. Large dual encoders are generalizable retrievers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 9844–9855. Association for Computational Linguistics.

Hanseok Oh, Hyunji Lee, Seonghyeon Ye, Haebin Shin, Hansol Jang, Changwook Jun, and Minjoon Seo. 2024. Instructir: A benchmark for instruction following of information retrieval models. *arXiv preprint arXiv:2402.14334*.

OpenAI. https://platform.openai.com/docs/models/gpt-3-5-turbo.

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. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Hongjin Su, Weijia Shi, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen-tau Yih, Noah A. Smith, Luke Zettlemoyer, and Tao Yu. 2023. One embedder, any task: Instruction-finetuned text embeddings. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 1102–1121. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas

Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Nikos Voskarides, Dan Li, Pengjie Ren, Evangelos Kanoulas, and Maarten de Rijke. 2020. Query resolution for conversational search with limited supervision. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval (SIGIR)*, pages 921–930.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Improving text embeddings with large language models. *CoRR*, abs/2401.00368.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, Quoc Le, and Denny Zhou. 2020. Chain of thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*.

Orion Weller, Benjamin Chang, Sean MacAvaney, Kyle Lo, Arman Cohan, Benjamin Van Durme, Dawn Lawrie, and Luca Soldaini. 2024. Followir: Evaluating and teaching information retrieval models to follow instructions. *arXiv preprint arXiv:2403.15246*.

Zeqiu Wu, Yi Luan, Hannah Rashkin, David Reitter, and Gaurav Singh Tomar. 2022. Conqrr: Conversational query rewriting for retrieval with reinforcement learning.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighof. 2023. C-pack: Packaged resources to advance general chinese embedding. *CoRR*, abs/2309.07597.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. 2021. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*.

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, Singapore, December 6-10, 2023*, pages 5985–6006. 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. *CoRR*, abs/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 (SIGIR)*, 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)*.

Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Jiafeng Guo, Min Zhang, and Shaoping Ma. 2021. Optimizing dense retrieval model training with hard negatives. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, pages 1503–1512. ACM.

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

Kun Zhou, Yeyun Gong, Xiao Liu, Wayne Xin Zhao, Yelong Shen, Anlei Dong, Jingwen Lu, Rangan Majumder, Ji-Rong Wen, and Nan Duan. 2022. Simans: Simple ambiguous negatives sampling for dense text retrieval. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: EMNLP 2022 - Industry Track, Abu Dhabi, UAE, December 7 - 11, 2022*, pages 548–559. Association for Computational Linguistics.

Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Zhicheng Dou, and Ji-Rong Wen. 2023. Large language models for information retrieval: A survey. *arXiv preprint arXiv:2308.07107*.

Appendix

A More Details of Experimental Setup

A.1 Evaluation Datasets

The basic statistics of these five evaluation datasets are shown in Table 7. All the datasets except TopiOCQA provide the human rewrite for each turn. The relevance annotations in the CAsT datasets are made by experts, making them more detailed.

Statistics	QReCC	TopiOCQA	CAsT-19	CAsT-20	CAsT-21
#Conversation	2,775	205	50	25	26
#Turns	16,451	2,514	479	208	239
#Passages	54M	25M	38M		40M

Table 7: Basic statistics of the five evaluation datasets.

A.2 Baselines

We provide a more detailed introduction to the baselines:

T5QR (Lin et al., 2020): a T5-based query rewriting method trained with human rewrites as the supervised signals.

ConvGQR (Mo et al., 2023a): A unified framework for query reformulation that integrates rule-based query rewriting with a generative model to expand queries.

LLM4CS (Mao et al., 2023b): A state-of-the-art LLM-based prompting method for conversational query rewriting. LLM4CS has two three prompting methods: REW, RAR, and RTR. REW only generates a rewrite and RAR additionally generates a hypothetical response. While RAR generates a rewrite and response in a two-step manner. For LLM4CS (REW) and LLM4CS (RAR), we only generate once for efficiency consideration and thus do not need aggregation.

Conv-ANCE (Mao et al., 2023c), which uses the classical ranking loss to train the session embeddings based on ANCE (Xiong et al., 2021).

ConvDR (Yu et al., 2021), which uses knowledge distillation to learn the session embeddings from rewrites.

DialogInpainter (Dai et al., 2022), which is fine-tuned from the T5-large model using information seeking dialogues generated from large web corpora.

LeCoRE (Mao et al., 2023c), which extends SPLADE (Formal et al., 2022) to be a conversational lexical retriever using multi-level denoising methods.

Generate a response to the current query given the context and retrieved passages. If the passages are relevant and useful, referring to their information when forming your response. Otherwise, you may disregard them.

Context:
{Context}

Current Query:
{query}

Retrieved Passages:
{context}

Figure 4: The prompt to generate the response in the experiment of partial response modification.

INSTRUCTOR (Su et al., 2023), a general retriever tailored to various tasks and domains by trained with various task-specific instructions.

LLM Embedder (Zhang et al., 2023): a unified retrieval model that can support diverse retrieval augmentation needs of LLMs. It is fine-tuned on various tasks and datasets such as MS-MARCO, NQ, ToolLLM, QReCC, FLAN, Books3, and Multi-Session Chat.

RepLLaMA (Ma et al., 2023), a large ad-hoc retriever fine-tuned from LLaMA-7B on the MS-MARCO dataset.

E5$_{\text{mistral-7b}}$ (Wang et al., 2024), a large ad-hoc retriever fine-tuned from Mistral-7B on the synthetic dataset generated by ChatGPT and MSMARCO.

GRIT (Muennighoff et al., 2024), a unified model for retrieval and generation. It is fine-tuned based on Mistral-7B. The retrieval part is fine-tuned on the E5 (Wang et al., 2024) dataset with task-specific instructions while the generation part is fine-tuned on the Tulu 2 (Ivison et al., 2023) dataset.

A.3 Hard Negatives

For UltraChat, we first use in-context learning with Qwen-7B-Chat, similar to the approach in (Mao et al., 2023b), to generate a query rewrite for each turn. We then obtain hard negatives by randomly sampling from the top-15 to top-30 retrieval results using the LLM Embedder on the CAsT-21 corpus with rewrites. The hard negatives for MSMARCO are consistent with those used in (Ma et al., 2023).

B Prompts in Partial Response Modification

The prompts to generate the response and judge whether the current query is reasonable are shown

Given the context of a conversation, evaluate whether the subsequent query is reasonable. A query is considered unreasonable if we cannot figure out its real search intent based on the context. For example:

\# Context:
Query: Who achieved 40,000 points in the NBA?
Response: Michael Jordan.

\# Next Query:
Which team drafted James?

This query is unreasonable because it is unclear who "James" is, as he was not mentioned in the context. The confusion arises because the response to the previous query is incorrect; the correct answer should be "LeBron James."

Now, it's your turn to assess the reasonableness of the query in the following context:
\# Context:
{context}

\# Next Query
{query}

Figure 5: The prompt to judge whether the current query is reasonable in the experiment of partial response modification.

in Figure 4 and Figure 5, respectively.

Given a conversational query, its context-independent rewrite, and its response, generate *two* turns of conversational context for it.

This turn:
 # Query: *How much does it cost for someone to fix it?*
 # Rewrite: *How much does it cost for someone to repair a garage door opener?*
 # Response: *Garage door opener repair can cost between $100 and $300 depending on the extent of the problem. Return to Top. The type of garage door you select -- and any extra pieces or labor required -- will influence how much you pay to have it professionally...*

\# Synthetic Conversation Context:
Query1: How much does a new garage door opener cost?
Response1: The cost of a new garage door opener can range from $150 to $500, depending on the brand, features, and installation requirements.

Query2: What are some common problems with garage door openers?
Response2: Some common problems with garage door openers include issues with the remote control, the motor, the sensors, or the door itself.

Figure 7: An example prompt to generate synthetic conversation text in the experiment of full context modification. *Italicized contents* are filled into the placeholders of the prompt. The green content is the model output.

C Prompts in Full Context Modification

The prompt to generate synthetic conversation text in the experiment of full context modification is shown in Figure 7. The green content is the output of ChatGPT3.5 for the above prompt.

D Influence of the Number of Special CoT Tokens

In Figure 6, we present the performance of ChatRetriever when varying the number of special tokens used for text representation. Our findings suggest that the inclusion of additional special tokens generally enhances retrieval performance. This improvement may be attributed to the fact that a sequence of consecutive special tokens can serve as a form of representational-level CoT, effectively expanding the learning space. However, we observe that performance plateaus when the number of special tokens exceeds three. Consequently, we finally append three special tokens in our implementation.

E Settings of Continually Fine-tuning Baselines

Since the training data of ChatRetriever only contains session-response pairs but does not contain human rewrites, we use in-context learning with Qwen-7B-Chat, similar to the approach in (Mao et al., 2023b), to generate query rewrite for each turn and use them for the training of ConvDR and LeCoRE. GRIT and Conv-ANCE are fine-tuned with their original contrastive ranking loss.

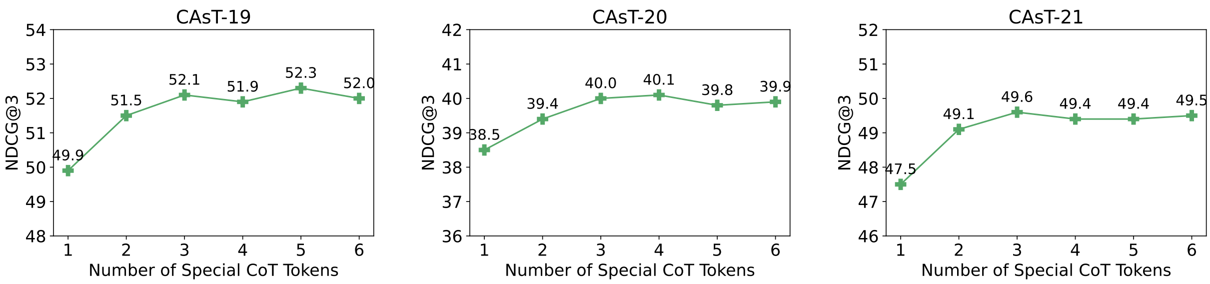

Figure 6: Performance comparisons when using different numbers of special CoT tokens.