Generative Query Augmentation with Dual-view Contrastive Learning for Dense Retrieval in Conversational Search

Wenyu Yan, Aoran Gan, Xukai Liu, Yanjiang Chen, Kai Zhang(✉), and Qi Liu

State Key Laboratory of Cognitive Intelligence, University of Science and Technology of China, Hefei, China
{yanwy,gar,chthollylxk,yjchen}@mail.ustc.edu.cn,
{kkzhang,qiliuql}@ustc.edu.cn

Abstract. Conversational search enables complex information retrieval by accurately understanding the user's actual search intent through multiple interactions between the user and the system. Existing conversational query rewriting methods are difficult to optimize directly for downstream tasks. Traditional conversational dense retrieval methods directly utilize the entire session as input, which introduces redundant noise from irrelevant conversation turns. To address these limitations, we propose the **G**enerative **Q**uery **A**ugmentation with **D**ual-View **C**ontrastive Learning for Conversational Dense **R**etrieval (**GQADCR**). Initially, we leverage the fact that the response for each conversation turn is available during training and use large language models to generate multiple standalone search queries. Subsequently, using the Best-of-N sampling strategy, we select the optimal rewritten query based on the actual retrieval performance. Finally, we propose a novel dual-view contrastive learning strategy that aligns the representation of the conversational session with those of self-contained rewritten queries and target retrieval passages, thereby facilitating the training of dense retrievers. Extensive experiments on two public datasets demonstrate the effectiveness, efficiency, and generalizability of our approach (Our source code is available at https://github.com/ywy5516/GQADCR).

Keywords: Conversational search · Large language model · Contrastive learning

1 Introduction

Conversational search enables users to engage in multi-turn interactions to satisfy their information needs by retrieving relevant passages from a collection of passages, based on the current query and its conversation context that includes previous queries and responses [6,12]. Unlike traditional single-turn ad-hoc retrieval, which relies primarily on keyword and phrase matching, conversational search requires modeling the whole conversation context to accurately

H. Jung et al. (Eds.): DASFAA 2026, LNCS 16540, pp. 170–185, 2026.
https://doi.org/10.1007/978-981-92-0378-9_11

capture the underlying search intent, as this intent may be distributed across the entire conversation context [23,28]. Therefore, conversational search is much more challenging than ad-hoc retrieval. Existing methods can be roughly categorized into two groups: Conversational Query Rewriting (CQR) and Conversational Dense Retrieval (CDR), as illustrated in Fig. 1.

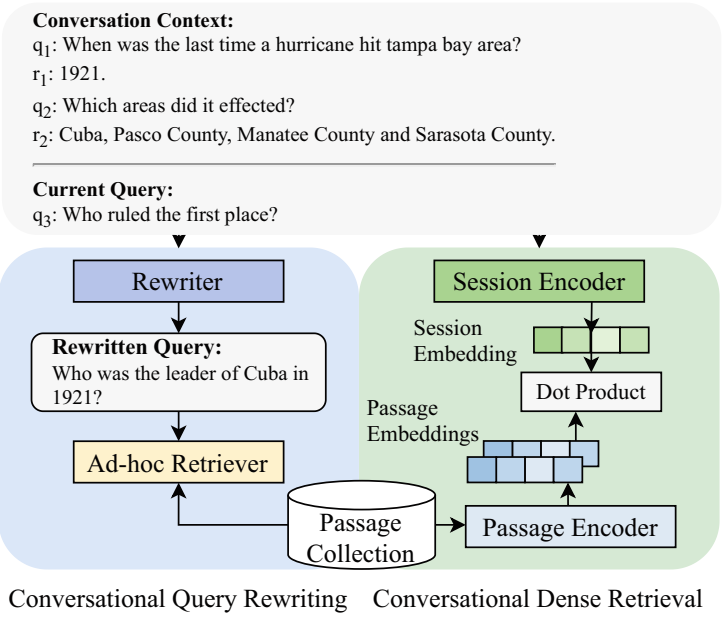

Fig. 1. A conceptual illustration for the CQR and CDR.

To capture real information needs in multi-turn conversation, CQR aims to reformulate conversational queries into stand-alone queries that can be submitted to any off-the-shelf retrievers [5,24]. Previous studies often fine-tune a pre-trained language model, such as T5. However, these methods rely on manually rewritten queries as supervision signals to train the rewrite model, yet obtaining large-scale manually annotated data for training remains challenging in practice. Furthermore, this rewrite-then-retrieve pipeline prevents CQR models from being directly optimized for downstream retrieval tasks [21,25], as the two-stage process hinders end-to-end training.

In contrast, CDR leverages a pre-trained ad-hoc retriever to encode the entire conversation context and candidate passages into a unified embedding space, followed by end-to-end fine-tuning on conversational data [14,19,29]. End-to-end CDR models can be directly optimized for better retrieval performance [4]. Previous studies have demonstrated the effectiveness of CDR. However, previous studies often treat the entire conversation context as input, while prior queries and responses in the conversation may be ambiguous or irrelevant to the current

query. These approaches inevitably introduce noise into the training process of CDR models [27]. Moreover, fine-tuning the retriever typically requires a large volume of labeled context-passage pairs. However, in practice, obtaining accurate annotations for such pairs is significantly more challenging than collecting conversational data itself.

To tackle these challenges, in this paper, we propose **G**enerative **Q**uery **A**ugmentation with **D**ual-View **C**ontrastive Learning for Conversational Dense **R**etrieval (**GQADCR**), a novel framework that integrates the strengths of both CQR and CDR, while eliminating reliance on manual query rewriting. Firstly, GQADCR fully leverages prior knowledge from the training data and the few-shot reasoning capabilities of large language models (LLMs) to resolve ambiguities in conversational context, thereby improving the informativeness of context and the quality of generated search queries. Based on the current turn's dialogue question-answer pairs and the conversation context, it uses LLMs to generate multiple candidate rewrites. Subsequently, by incorporating retrieval ranking signals, we distill high-quality rewritten queries using the Best-of-N sampling strategy. Finally, GQADCR further aligns the representations of rewritten queries and relevant passages with the retriever through a carefully designed dual-view contrastive learning module.

Our contributions are summarized as follows:

- We propose GQADCR, an end-to-end architecture of conversational search. To solve key challenges, a novel contrasting module is designed to fuse semantic data from diverse perspectives into conversational context.
- We highlight the key contribution of integrating self-contained search queries into the CDR framework. Our approach bridges the gap between CQR and CDR by leveraging their complementary strengths, enabling more coherent and context-aware retrieval.
- Quantitative and qualitative experimental results on two CS-based datasets and three different base retrievers demonstrate that GQADCR exhibits superior performance, along with strong efficiency and generalizability.

2 Related Work

2.1 Conversational Dense Retrieval

Conversational Dense Retrieval (CDR) [6,8,22,23,29] leverages conversational search sessions to fine-tune an end-to-end, ad-hoc retriever that enables the encoding of conversation sessions into a shared embedding space for dense retrieval. Considering that the context of the entire conversation may be lengthy and contain a significant amount of noise, some studies [14,19] design sophisticated context-denoising approaches for better CDR models.

Our approach belongs to the family of CDR methods. To address this key challenge, this paper incorporates a query rewriting module into the architecture of our GQADCR framework. After obtaining high-quality training data, a novel contrastive learning module is employed to align the session input with the representations of semantically complete rewritten queries and target passages.

2.2 Conversational Query Rewriting

Conversational Query Rewriting (CQR) aims to enhance conversational search performance by transforming context-dependent queries into standalone ad-hoc queries [3,16,24,25,28]. To optimize query rewriting, some studies have leveraged reinforcement learning or incorporated ranking signals during model training [21, 23]. However, these approaches rely on manually annotated rewritten queries for training CQR models, which are difficult to obtain in practice. Recently, LLMs have been demonstrated to be capable of rewriting conversational queries, and some studies [7,18,27] employ LLMs to perform query rewriting via a training-free, purely prompt-based method.

In this work, we design a prompt template that leverages prior knowledge from training data to effectively resolve ambiguities in conversation context by prompting LLMs, thereby reducing the need for manually rewritten queries. We then utilize the generated rewritten queries as training data to assist in training CDR models. By leveraging the strengths of CDR models in implicit context modeling, we aim to enhance the semantic correlation between the retrieval system and downstream tasks.

2.3 Contrastive Learning

Contrastive Learning (CL) [30–35] has long been recognized as an effective approach for learning meaningful representations by comparing different samples. Typically, CL operates on pairs of examples that are semantically related, treating them as positive (similar) instances, while other samples serve as negatives (dissimilar). The objective of training is thus to pull the representations of positive pairs closer together in the embedding space while pushing those of negative pairs apart.

In our method, we first leverage LLMs to generate standalone search queries. We then align the corresponding semantic units derived from heterogeneous data signals through a dual-view contrastive learning framework.

3 Methodology

This section presents a novel CDR framework GQADCR, as depicted in Fig. 2.

3.1 Task Formulation

Given a new query q_t and the conversation context $C_{t-1} = \{(q_i, r_i)\}_{i=1}^{t-1}$, where q_i and r_i denote the query and the system response to each previous turn, respectively. The i-th historical turn is denoted as (q_i, r_i, p_i^*), where p_i^* is the ground-truth passage corresponding to q_i. For given the current query q_t and the conversation context C_{t-1}, our task is to retrieve the gold passage p_t^* from a passage collection $\mathcal{D}$. The challenge of conversational search lies in the context-dependent nature of queries, which demands natural language understanding techniques to accurately infer the user's real search intent and thus find the gold passage p_t^*.

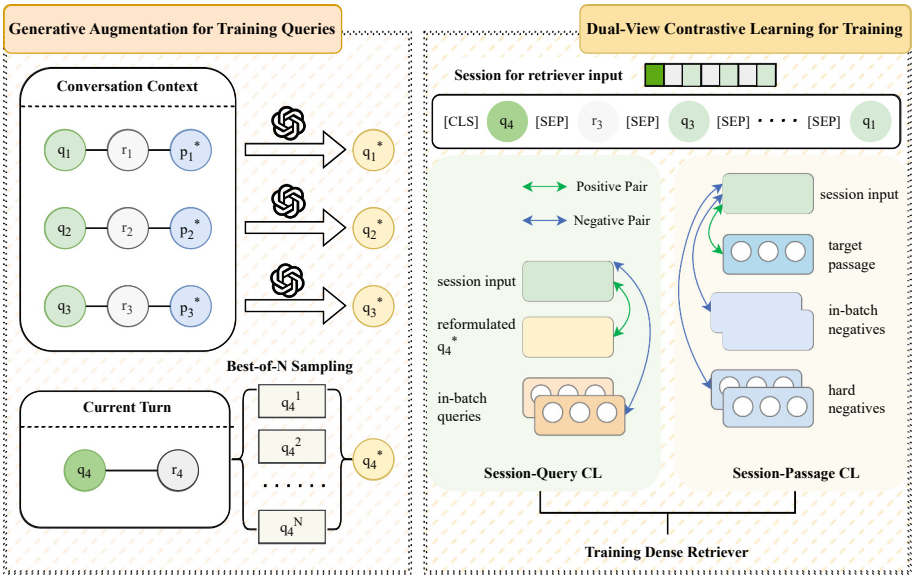

Fig. 2. Illustration of the proposed GQADCR. From left to right, the modules are LLM-based Generative Augmentation of Training Data and Dual-view Contrastive Learning.

3.2 Generative Augmentation for Training Queries

Recent studies [7,18] have demonstrated that open-source LLMs with strong language understanding and text generation capabilities can be effectively applied to real-world scenarios for query rewriting, without requiring fine-tuning. In this work, we leverage the known dialog responses in the training data to generate multiple candidate queries by invoking LLMs. The optimal rewritten query is then selected from these candidates based on retrieval performance.

Multi-candidate Query Generation from Responses. In this section, we propose the Rewrite-with-Response (RWR) instruction to address the challenges of omission, ambiguity and co-reference in multi-turn dialogues. During the training phase, the current t-th turn query q_t, the conversation context C_{t-1} from the first $t-1$ turns and the response for the current turn r_t are known. As shown in Eq. 1, for any turn t, we concatenate q_t, r_t, and C_{t-1} into a prompt, which is then fed into an LLM to generate the collection of de-contextualized search queries Q_t:

$$q_t^i = \mathcal{LLM}(\mathcal{I}^{RWR}; C_{t-1}, q_t, r_t) \tag{1}$$

$$Q_t = \{q_t^1, q_t^2, \cdots, q_t^N\} = \{q_t^i\}_{i=1}^N \tag{2}$$

where $\mathcal{LLM}(\cdot)$ denotes invoking LLMs, $\mathcal{I}^{RWR}$ is the prompt template we propose to generate self-contained search queries as shown in Fig. 3, and N is the parameter[1] that controls how many completions to generate for each prompt.

Prompt Template

<SYSTEM>
- Role: Expert in conversation understanding and information retrieval.
- Background: When users retrieve information, they often omit or use pronouns in the conversation, which may lead to incomplete or vague query sentences. In order to improve the accuracy and efficiency of retrieval, it is necessary to complete the user's query statement, make it independent of the session context, and correctly handle the topic conversion.
- Profile: You are an expert proficient in natural language processing and information retrieval. You can deeply analyze the dialog context, understand the user's real intention, and convert fuzzy query sentences into accurate and context independent queries.
- Skills: You have the key abilities of conversation analysis, semantic understanding, anaphora resolution, topic tracking and information retrieval. You can accurately complete the omitted parts, handle the pronouns, and maintain the consistency and accuracy of the query during topic conversion.
- Goals: According to the historical dialog context, the current user question and the system reply, complete the omitted and referred parts of the user query sentence, correctly handle the topic conversion phenomenon, and finally output an independent query with complete information and unrelated to the session context for information retrieval.
- Constrains: The complete query sentence must be clear and accurate to avoid ambiguity; When dealing with topic conversion, it is necessary to ensure that the query statement is consistent with the user's real intention; The output query statement shall be concise and clear, which is convenient for the information retrieval system to understand and execute.
- OutputFormat: Note that you should always try to rewrite it. Output a complete and standalone query statement, in the format of `Rewrite: $Rewrite`.

<USER>
Context Begin
{{ context }}
Context End

Current Question: {{ cur_query }}
Current Response: {{ cur_response }}

Now, you should give me the rewriter of the **Current Question** under the **Context** and the **Current Response**. Do not add any explanation, the output format should always be `Rewrite: $Rewrite`.

Fig. 3. The prompt template through LLMs.

Filtering Through Retrieval Ranking Signals. After obtaining the reformulated query set Q_t, we apply it to an off-the-shelf retrieval system to obtain retrieval ranking feedback. BM25 is a widely used sparse retrieval algorithm that scores and ranks documents based on term matches between queries and documents, where "terms" typically refer to keywords extracted through tokenization. Given its effectiveness in assessing query self-containment via keyword matching, BM25 is selected as the off-the-shelf retriever in our work.

Based on the retrieval ranking signals, we select the optimal rewritten query q_t^* from Q_t for the current t-th turn of dialogue:

$$q_t^* = \arg\max_{Q_t} \{\delta(\mathrm{R}(q_t^i), p_t^*)\} \tag{3}$$

where $\mathrm{R}(\cdot)$ is the retriever that returns passages given a query, and $\delta(\cdot,\cdot)$ is an evaluation function that calculates the relevance between the retrieved passages and the ground-truth passage.

3.3 Dual-view Contrastive Learning

Compared to ad-hoc search, the queries in conversational search are context-dependent, thus requiring the retriever to reformulate the current search query based on its conversational history. Following prior work [29], we construct a

[1] https://platform.openai.com/docs/api-reference/completions/create.

session-specific query by concatenating the current query with its preceding context, which serves as input for training the conversational dense retriever:

$$s_t = [\text{CLS}] \circ q_t \circ [\text{SEP}] \circ r_{t-1} \circ [\text{SEP}] \circ q_{t-1} \circ [\text{SEP}] \circ \cdots \circ q_1 \circ [\text{SEP}] \quad (4)$$

Additionally, we adopt the architecture of the Dense Passage Retriever [11,12], which consists of two encoders: E_Q and E_P. The passage encoder E_P maps each passage to a d-dimensional real-valued vector and builds an index over all passages in the collection $\mathcal{D}$. These passage embeddings can be computed offline and indexed for efficiency. At inference time, the query encoder E_Q encodes the conversational input into a d-dimensional vector and retrieves the k most similar passages, those whose vectors are closest to the input vector. The similarity between a query and a passage is measured by the dot product of their embeddings:

$$\text{sim}(q, p) = E_Q(q)^T \cdot E_P(p) \quad (5)$$

Session-Passage Contrastive Learning. Contrastive learning has become a prevalent approach for training dense passage retrievers in recent studies [12, 29]. From the perspective of the target passage, we train the query encoder E_Q session-passage contrastive learning to minimize the distance between the representation of the session input and that of the relevant passage in the shared embedding space. The following positive and negative samples are used during training:

- p_t^*: The ground-truth passage corresponding to the current query q_t, used as the positive sample.
- $\mathcal{P}_b^-$: In-batch negative passages sampled from other instances within the same training batch.
- $\mathcal{P}_r^-$: These are retrieved passages that serve as hard negatives. They can be obtained by using the top-ranked passages retrieved for s_t by an off-the-shelf retriever after excluding p_t^* if it is present [11,19,20]. In this work, we employ a sparse retriever (BM25) to obtain these hard negatives.

Formally, the complete set of negative samples is defined as $\mathcal{P}^- = \mathcal{P}_b^- \cup \mathcal{P}_r^-$. The session-passage contrastive loss for the dense passage retriever is defined in Eq. 6, where $\phi(\cdot, \cdot) = \exp(\text{sim}(\cdot, \cdot)/\tau)$ and τ denotes the temperature parameter used to scale the whole distribution.

$$\mathcal{L}_{SP} = -\log \frac{\phi(s_t, p_t^*)}{\phi(s_t, p_t^*) + \sum_{p^- \in \mathcal{P}^-} \phi(s_t, p^-)} \quad (6)$$

Session-Query Contrastive Learning. CDR typically treats the entire conversation context as input to encode contextual dependencies across dialogue turns, aiming to retrieve relevant passages for the current user query. However, this approach often incorporates redundant or irrelevant information from earlier turns, which can introduce noise and distract the model during training and inference. Such historical noise may lead to suboptimal representations,

particularly when prior utterances address topics only loosely related or entirely unrelated to the current information need.

To mitigate this issue, we propose aligning the semantic representation of the full conversation session with that of a semantically complete, context-independent reformulated query. Such a query explicitly resolves co-references, incorporates implicit user intents, and encapsulates the standalone meaning of the current turn, thereby serving as a denoised and focused target for representation learning. Instead of directly fusing these representations, we design a session-query contrastive learning objective, encouraging the semantic spaces of sessions and queries to gradually converge before fusion. This objective is formalized as follows:

$$\mathcal{L}_{SQ} = -\log \frac{\phi(s_t, q_t^*)}{\phi(s_t, q_t^*) + \sum_{q^- \in \mathcal{Q}_b^-} \phi(s_t, q^-)} \tag{7}$$

We adopt in-batch negatives to construct the negative set $\mathcal{Q}_b^-$.

3.4 Overall Objective

In line with previous work, we combine our loss term $\mathcal{L}_{SQ}$ with the original contrastive ranking loss $\mathcal{L}_{SP}$ to form a joint training objective. Both losses are computed using the cross-entropy loss function. Thus, the overall training objective for the query encoder E_Q is defined as:

$$\mathcal{L} = \mathcal{L}_{SP} + \lambda \mathcal{L}_{SQ} \tag{8}$$

where λ is a hyperparameter that controls the trade-off of the two terms.

4 Experiments

4.1 Experimental Setup

Datasets. Following prior work [7,29], we evaluate our approach on two widely-used conversational search datasets. QReCC [2] focuses on the query rewriting task, with most queries within a session centered on the same topic. In contrast, TopiOCQA [1] exhibits complex topic-switching phenomena within sessions and features a large number of conversation turns compared to QReCC, making it more challenging. The statistics of each dataset are provided in Table 1.

Evaluation Metrics. In this study, to enable a thorough comparison with prior work, we employ three metrics to evaluate the performance of our GQADCR framework: Mean Reciprocal Rank (MRR), Normalized Discounted Cumulative Gain at rank 3 (NDCG@3) and Recall at rank 10 (Recall@10). These metrics provide a comprehensive assessment of ranking and retrieval effectiveness.

Table 1. Statistics of conversational search datasets.

Dataset	Split	#Conv.	#Turns(Qry.)	#Collection
TopiOCQA	Train	3,509	45,450	25M
	Test	205	2,514	
QReCC	Train	10,823	63,501	54M
	Test	2,775	16,451	

Implementation Details. Our implementation is based on PyTorch, Huggingface's Transformers and Faiss [10]. We adopt ANCE [26] as the backbone model for the Dual-View Contrastive Learning. The AdamW [17] optimizer is used with default parameters, and the learning rate is initialized to 5×10^{-5} with a cosine decay scheduler. Training uses a batch size of 32 and gradient accumulation over 2 steps, resulting in an effective batch size of 64. The input sequence lengths for the query, concatenated session, and passage are truncated to 64, 512, and 384 tokens, respectively. The temperature parameter τ in contrastive learning is set to the default value of 0.07. The hyperparameter λ is fixed at 1.0.

For the Generative Augmentation of Training Data, we deploy the instruct-tuned Llama3.2-3B[2] model using vLLM [13]. In all experiments, text generation employs sampling with a temperature of 0.7 for query rewriting. The LLMs are solely employed for offline training data augmentation, while inference relies entirely on the dense retriever. In Best-of-N sampling, $N = 10$ is used throughout, and the best rewritten query is selected based on its BM25 retrieval score. Following established practice, BM25 parameters are set to $k1 = 0.9, b = 0.4$ for TopiOCQA and $k1 = 0.82, b = 0.68$ for QReCC, respectively. All experiments are conducted on two NVIDIA A100 40G GPUs.

Baselines. To validate the effectiveness of our approach, we compare it with two categories of state-of-the-art approaches in conversational search. The first category leverages generative language models to perform conversational query rewriting (CQR), followed by retrieval using pre-trained dense retrievers. This group includes Human-Rewritten [2], T5QR [15], CONQRR [25], ConvGQR [21], LLM4CS [18] and IterCQR [7]. The second category focuses on end-to-end conversational dense retrieval (CDR), where models are initialized from standard ad-hoc search retrievers and subsequently fine-tuned on conversational search data. Representative methods in this category include ConvDR [29], SDRConv [12], Conv-ANCE [20], InstructoR [9] and ConvSDG [22]. To ensure a fair comparison, all selected baseline methods are implemented to evaluate on ANCE [26].

4.2 Main Results

The evaluation results on the TopiOCQA and QReCC datasets are presented in Table 2. We have the following observations:

[2] https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct.

Table 2. Experiment results (%) on two benchmark datasets. Only the QReCC dataset has manually rewritten queries. All compared models are initialized from ANCE. **Bold** and underline indicate the best and the second-best results, respectively.

Category	Method	TopiOCQA			QReCC		
		MRR	NDCG@3	Recall@10	MRR	NDCG@3	Recall@10
CQR	Human-Rewrite [2]	-	-	-	38.4	35.6	58.6
	T5QR [15]	23.0	22.2	37.6	34.5	31.8	53.1
	CONQRR [25]	-	-	-	41.8	-	65.1
	ConvGQR [21]	25.6	24.3	41.8	42.0	39.1	63.5
	LLM4CS [18]	27.7	26.7	43.3	44.8	42.1	66.4
	IterCQR [7]	26.3	25.1	42.6	42.9	40.2	65.5
CDR	ConvDR [29]	27.2	26.4	43.5	38.5	35.7	58.2
	SDRConv [12]	26.1	25.4	44.4	47.3	43.6	69.8
	Conv-ANCE [20]	22.9	20.5	43	47.1	45.6	**71.5**
	InstructoR [9]	25.3	23.7	45.1	43.5	40.5	66.7
	ConvSDG [22]	21.4	19.9	37.8	-	-	-
CDR	**GQADCR**	**29.5**	**28.2**	**46.5**	**48.6**	**46.5**	65.6

We observe that on both datasets, GQADCR outperforms the vast majority of compared baseline methods across all three evaluation metrics. On the TopiOCQA dataset, GQADCR improves MRR by 6.5% and NDCG@3 by 5.6% over the second-best method. On the QReCC dataset, although GQADCR achieves lower Recall@10 than some baselines, it surpasses all baselines in terms of MRR and NDCG@3.

The performance advantages of GQADCR can be attributed to two key factors. First, GQADCR integrates the query rewriting capability of CQR with the passage-level context modeling capability of CDR, enabling effective capture of intent shifts and incorporation of multi-turn conversational context in dynamic scenarios. Second, GQADCR enhances its training through an optimized negative sampling strategy that incorporates additional semantic information, thereby improving the model's ability to distinguish contextually relevant passages. The generalizability and efficiency of GQADCR are further validated through experiments in Sect. 4.4 and Section 4.5, respectively.

4.3 Ablation Study

Table 3 presents the results of ablation studies conducted on the TopiOCQA and QReCC datasets, aiming to analyze the contribution of each key component in the GQADCR framework. The results show that removing any component leads to performance degradation, confirming the effectiveness of each module.

Specifically, eliminating the session-query contrastive loss $\mathcal{L}_{SQ}$ results in noticeable performance drops on both datasets. This indicates that $\mathcal{L}_{SQ}$ plays

Table 3. Ablation study of different components.

	TopiOCQA			QReCC		
	MRR	NDCG@3	Recall@10	MRR	NDCG@3	Recall@10
GQADCR	**29.5**	**28.2**	**46.5**	**48.6**	**46.5**	**65.6**
w/o $\mathcal{L}_{SQ}$	27.9	26.9	43.8	41.8	39.5	59.5
w/o $\mathcal{L}_{SP}$	20.2	18.6	37.6	39.1	36.5	57.0

a crucial role in guiding the model to extract contextually consistent query representations from the full conversational context. Without this loss, the model struggles to filter out irrelevant information from the dialogue context.

More significantly, removing the session-passage contrastive loss $\mathcal{L}_{SP}$ causes a substantial performance decline, particularly on TopiOCQA. This demonstrates that $\mathcal{L}_{SP}$ is essential for aligning the semantic space of the conversation with that of the target passages, enabling the retriever to more accurately identify relevant passages. The larger performance drop further underscores the importance of establishing a direct semantic link between the conversational context and external knowledge, which is critical in multi-turn conversation scenario.

Furthermore, while the relative contribution of each component varies across datasets, the overall trend remains consistent. This suggests that $\mathcal{L}_{SQ}$ and $\mathcal{L}_{SP}$ are functionally complementary, and their synergistic interaction enhances the model's robustness and retrieval capability in complex conversational settings.

4.4 Application to Other Base Retrivers

Table 4. Retrieval performance of various base retrievers trained using our framework.

Method	Backbone	TopiOCQA			QReCC		
		MRR	N@3	R@10	MRR	N@3	R@10
GQADCR	ANCE	29.5	28.2	46.5	**48.6**	**46.5**	65.6
	BGE-base-en	29.2	27.9	47.3	45.0	42.6	64.8
	Contriever-msmarco	**31.1**	**29.7**	**48.3**	47.2	44.9	**69.5**

While we adopt ANCE as the base model for our GQADCR given its popularity and widespread use as a foundation in existing CDR systems, our training framework is general and readily applicable to other dense retrieval models. To validate the generalizability of our approach, we further instantiate it with two additional base models: BGE[3] and Contriever-msmarco[4]. As shown in Table 4,

[3] https://huggingface.co/BAAI/bge-base-en-v1.5.
[4] https://huggingface.co/facebook/contriever-msmarco.

our method consistently improves retrieval performance across all base models and surpasses previous baseline approaches. These results confirm the strong transferability and architectural generality of our approach.

4.5 Low-Resource Evaluation

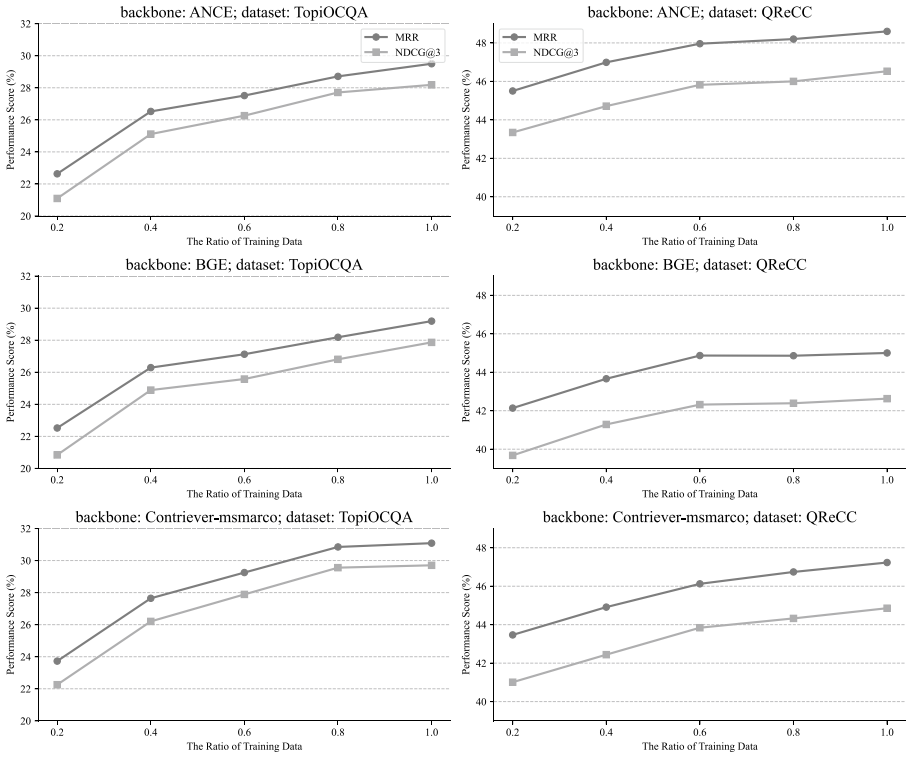

Fig. 4. Evaluation of GQADCR framework with different backbone retrievers under low-resource settings.

Based on experimental results using two benchmark datasets for conversational search, this study systematically evaluates the performance of the proposed GQADCR method under various base retrievers and low-resource training conditions. As shown in Fig. 4, the performance trends of the GQADCR method are consistent across both datasets, and it achieves performance comparable to baseline methods even when only 60% of the training data is used. More importantly, GQADCR consistently improves performance across base models with diverse architectures and training strategies, demonstrating that the introduced mechanism is not tied to the structural characteristics of specific

retrievers, but instead exhibits strong modular compatibility and generalizability. These results suggest that the GQADCR method effectively reduces reliance on large-scale annotated data, particularly in scenarios with limited training data, thereby highlighting its practical applicability to real-world low-resource conversational search tasks.

4.6 The Impact of LLMs

Table 5. Dense retrieval results for ANCE trained on data augmented by various LLMs.

$\mathcal{LLM}$	N	TopiOCQA			QReCC		
		MRR	NDCG@3	Recall@10	MRR	NDCG@3	Recall@10
LLama-3.2-3B	5	28.7	27.5	46.0	47.6	45.3	64.0
	10	**29.5**	**28.2**	**46.5**	**48.6**	**46.5**	**65.6**
Qwen-2.5-3B	5	28.2	26.9	45.7	47.2	45.0	63.7
	10	28.3	27.0	46.4	47.8	45.6	64.3

As shown in Table 5, this study systematically evaluates the impact of LLMs with different architectures and the sampling size N for query rewriting on the proposed GQADCR framework across benchmark datasets. The experiment employs two instruction-fine-tuned models, Llama3.2-3B and Qwen2.5-3B[5], with comparable parameter counts but distinct architectural designs to investigate the sensitivity of framework performance to model architecture. The results demonstrate that the choice of LLM significantly affects the overall performance of GQADCR. Notably, configurations using Llama3.2-3B consistently outperform those using Qwen2.5-3B, suggesting that Llama3 models may possess stronger generalization capabilities or superior instruction-following behavior in tasks involving conversational context understanding and query disambiguation.

Furthermore, the impact of increasing the number of rewrites is examined. When N increases from 5 to 10, the Llama3.2-3B-based model achieves sustained and significant performance gains on both datasets. In contrast, for Qwen2.5-3B, the performance improvement from $N = 10$ relative to $N = 5$ is marginal, with only a slight increase in Recall@10 and no notable gains in MRR or NDCG@3. This indicates that different LLMs exhibit varying sensitivity to the number of samples. Consequently, in practical applications, the sampling size should be carefully selected based on the specific characteristics of the LLM to balance computational cost and performance.

[5] https://huggingface.co/Qwen/Qwen2.5-3B-Instruct.

5 Conclusion

In this paper, we present GQADCR, an innovative framework for CDR. Firstly, GQADCR leverages both prior knowledge from training data and the few-shot reasoning capabilities of LLMs to enhance the informativeness of conversational context and improve the quality of generated search queries. Subsequently, by incorporating retrieval ranking signals, we employ a Best-of-N sampling strategy to distill high-quality rewritten queries. Lastly, GQADCR further aligns the rewritten queries and target retrieval passages with the retriever through a carefully designed contrastive learning module. Extensive experiments on two public datasets demonstrate the effectiveness of our approach, which outperforms all baseline models in retrieval performance (Table 2). Furthermore, GQADCR exhibits robustness and generalizability across various backbone retrievers and LLMs (Tables 4 and 5) and maintains strong performance in low-resource settings (Fig. 4).

Acknowledgements. This research was partially supported by the National Natural Science Foundation of China (Grants No.62406303) and Anhui Province Science and Technology Innovation Project (202423k09020010).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Adlakha, V., Dhuliawala, S., Suleman, K., de Vries, H., Reddy, S.: TopiOCQA: open-domain conversational question answering with topic switching. Trans. Assoc. Comput. Linguist. **10**, 468–483 (2022)
2. Anantha, R., Vakulenko, S., Tu, Z., Longpre, S., Pulman, S., Chappidi, S.: 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, pp. 520–534. Association for Computational Linguistics (2021)
3. Chen, Z., Zhao, J., Fang, A., Fetahu, B., Rokhlenko, O., Malmasi, S.: Reinforced question rewriting for conversational question answering. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track, pp. 357–370. Association for Computational Linguistics (2022)
4. Cheng, Y., Mao, K., Dou, Z.: Interpreting conversational dense retrieval by rewriting-enhanced inversion of session embedding. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 2879–2893. Association for Computational Linguistics (2024)
5. Fang, H.C., Hung, K.H., Huang, C.W., Chen, Y.N.: Open-domain conversational question answering with historical answers. In: Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022, pp. 319–326. Association for Computational Linguistics (2022)
6. Huang, C.W., Hsu, C.Y., Hsu, T.Y., Li, C.A., Chen, Y.N.: CONVERSER: few-shot conversational dense retrieval with synthetic data generation. In: Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pp. 381–387. Association for Computational Linguistics (2023)

7. Jang, Y., Lee, K.i., Bae, H., Lee, H., Jung, K.: IterCQR: iterative conversational query reformulation with retrieval guidance. In: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pp. 8121–8138. Association for Computational Linguistics (2024)
8. Jeong, S., Baek, J., Hwang, S.J., Park, J.: Phrase retrieval for open domain conversational question answering with conversational dependency modeling via contrastive learning. In: Findings of the Association for Computational Linguistics: ACL 2023, pp. 6019–6031. Association for Computational Linguistics (2023)
9. Jin, Z., Cao, P., Chen, Y., Liu, K., Zhao, J.: InstructoR: instructing unsupervised conversational dense retrieval with large language models. In: Findings of the Association for Computational Linguistics: EMNLP 2023, pp. 6649–6675. Association for Computational Linguistics (2023)
10. Johnson, J., Douze, M., Jégou, H.: Billion-scale similarity search with GPUs. IEEE Trans. Big Data **7**(3), 535–547 (2019)
11. Karpukhin, V., et al.: Dense passage retrieval for open-domain question answering. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 6769–6781. Association for Computational Linguistics (2020)
12. Kim, S., Kim, G.: Saving dense retriever from shortcut dependency in conversational search. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pp. 10278–10287. Association for Computational Linguistics (2022)
13. Kwon, W., et al.: Efficient memory management for large language model serving with pagedattention. In: Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles (2023)
14. Lin, S.C., Yang, J.H., Lin, J.: Contextualized query embeddings for conversational search. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 1004–1015. Association for Computational Linguistics (2021)
15. Lin, S.C., Yang, J.H., Nogueira, R., Tsai, M.F., Wang, C.J., Lin, J.J.: Conversational question reformulation via sequence-to-sequence architectures and pretrained language models. arXiv preprint arXiv:2004.01909 (2020)
16. Liu, L., Hill, B., Du, B., Wang, F., Tong, H.: Conversational question answering with language models generated reformulations over knowledge graph. In: Findings of the Association for Computational Linguistics: ACL 2024, pp. 839–850. Association for Computational Linguistics, Bangkok, Thailand (2024)
17. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2017)
18. Mao, K., Dou, Z., Mo, F., Hou, J., Chen, H., Qian, H.: Large language models know your contextual search intent: a prompting framework for conversational search. In: Findings of the Association for Computational Linguistics: EMNLP 2023, pp. 1211–1225. Association for Computational Linguistics (2023)
19. Mao, K., Dou, Z., Qian, H.: 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, pp. 176–186. Association for Computing Machinery (2022)
20. Mao, K., et al.: Learning denoised and interpretable session representation for conversational search. In: Proceedings of the ACM Web Conference 2023, pp. 3193–3202. Association for Computing Machinery, New York, NY, USA (2023)

21. Mo, F., Mao, K., Zhu, Y., Wu, Y., Huang, K., Nie, J.Y.: ConvGQR: generative query reformulation for conversational search. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 4998–5012. Association for Computational Linguistics (2023)
22. Mo, F., Yi, B., Mao, K., Qu, C., Huang, K., Nie, J.Y.: Convsdg: session data generation for conversational search. In: Companion Proceedings of the ACM Web Conference 2024, pp. 1634–1642. Association for Computing Machinery (2024)
23. Qian, H., Dou, Z.: Explicit query rewriting for conversational dense retrieval. In: Goldberg, Y., Kozareva, Z., Zhang, Y. (eds.) Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pp. 4725–4737. Association for Computational Linguistics (2022)
24. Vakulenko, S., Longpre, S., Tu, Z., Anantha, R.: Question rewriting for conversational question answering. In: Proceedings of the 14th ACM International Conference on Web Search and Data Mining, pp. 355–363. Association for Computing Machinery (2021)
25. Wu, Z., et al.: CONQRR: Conversational query rewriting for retrieval with reinforcement learning. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pp. 10000–10014. Association for Computational Linguistics (2022)
26. Xiong, L., et al.: Approximate nearest neighbor negative contrastive learning for dense text retrieval. In: International Conference on Learning Representations (ICLR) (2021)
27. Ye, F., Fang, M., Li, S., Yilmaz, E.: Enhancing conversational search: large language model-aided informative query rewriting. In: Findings of the Association for Computational Linguistics: EMNLP 2023, pp. 5985–6006. Association for Computational Linguistics (2023)
28. Yu, S., et al.: Few-shot generative conversational query rewriting. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1933–1936. Association for Computing Machinery (2020)
29. Yu, S., Liu, Z., Xiong, C., Feng, T., Liu, Z.: Few-shot conversational dense retrieval. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, p. 829–838. Association for Computing Machinery (2021)
30. Zhang, K., et al.: Graph adaptive semantic transfer for cross-domain sentiment classification. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1566–1576 (2022)
31. Zhang, K., et al.: EATN: an efficient adaptive transfer network for aspect-level sentiment analysis. IEEE Trans. Knowl. Data Eng. **35**(1), 377–389 (2021)
32. Zhang, K., et al.: Multi-interactive attention network for fine-grained feature learning in ctr prediction. In: Proceedings of the 14th ACM International Conference on Web Search and Data Mining, pp. 984–992 (2021)
33. Zhang, K., Zhang, H., Liu, Q., Zhao, H., Zhu, H., Chen, E.: Interactive attention transfer network for cross-domain sentiment classification. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 5773–5780 (2019)
34. Zhang, K., et al.: Incorporating dynamic semantics into pre-trained language model for aspect-based sentiment analysis. arXiv preprint arXiv:2203.16369 (2022)
35. Zhou, Y., Zhou, K., Zhao, W.X., Wang, C., Jiang, P., Hu, H.: C^2-crs: Coarse-to-fine contrastive learning for conversational recommender system. In: Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining, pp. 1488–1496. Association for Computing Machinery (2022)