Generalizing Conversational Dense Retrieval via LLM-Cognition Data Augmentation

Anonymous ACL submission

Abstract

Conversational search utilizes muli-turn natural language contexts to retrieve relevant passages. Existing conversational dense retrieval models mostly view a conversation as a fixed sequence of questions and responses, overlooking the severe data sparsity problem – that is, users can perform a conversation in various ways, and these alternate conversations are unrecorded. Consequently, they often struggle to generalize to diverse conversations in real-world scenarios. In this work, we propose a framework for generalizing **Conv**ersational dense retrieval via LLM-cognition data **Aug**mentation (ConvAug). We first generate multi-level augmented conversations to capture the diverse nature of conversational contexts. Inspired by human cognition, we devise a cognition-aware prompting process to mitigate the generation of false positives, false negatives, and hallucinations. Moreover, we develop a difficulty-adaptive sample filter that selects challenging samples for complex conversations, thereby giving the model a larger learning space. A contrastive learning objective is then employed to train a better conversational context encoder. Extensive experiments conducted on four public datasets, under both normal and zero-shot settings, demonstrate the effectiveness, generalizability, and applicability of ConvAug.

1 Introduction

Conversational search is anticipated to become the leading form of ad-hoc search engines in the future (Gao et al., 2023a). This approach, utilizing multi-turn natural language interactions, offers a user-friendly experience, particularly for complex information-seeking tasks.

There are two typical approaches for conversational search. One way is conversational query rewriting (CQR) (Mo et al., 2023a; Wu et al., 2022). CQR models convert a conversational query into a de-contextualized search query suitable for ad-hoc retrieval. However, CQR models either perform poorly because they cannot be optimized by downstream retrieval task (Mao et al., 2023c), or have unacceptable search latency when using large language models (LLMs) during inference (Mao et al., 2023b). Another approach is to perform conversational dense retrieval (CDR) in an end-to-end manner. It typically uses the entire conversational context to train the context encoder within CDR models for passage retrieval. This approach has been demonstrated to be more effective than CQR models on the downstream retrieval task of conversational search (Jin et al., 2023; Mao et al., 2023c).

Existing CDR approaches typically utilize conversations as fixed multi-turn natural language texts to train the context encoder. However, in real-world scenarios, users can express conversations in various ways. The conversational search data often lack the diversity to support training for such variability due to the severe data sparsity issue. In other words, numerous alternative conversations with the same intent (or with similar expressions but different intents) as a specific data sample are unrecorded. As a result, CDR models trained on these limited and fixed data often struggle to adapt to diverse real-world conversations. Some works have tried to compensate for the deficiency of multi-turn texts. However, these efforts often rely on basic rule-based strategies (Zhu et al., 2021) or human annotations to augment conversations (Mao et al., 2022a). Furthermore, comprehending turn dependencies in multi-turn conversations poses a significant challenge for simple language models.

To tackle these problems, we propose an LLM-based data augmentation framework to mimic how users perform diverse conversations. Specifically, we design multi-level augmentation strategies to generate positive (similar intents but different expressions, denoted as $+$) and hard negative conversations (similar expressions but different intents, denoted as $-$): (1) Token level. To mitigate the model's overreliance on specific tokens, we ran-

domly mask some tokens of conversations (+). Besides, we identify and replace the entities (−) to help the model focus on key information. (2) Turn level. To prevent the model from depending on specific turns or the order of turns within conversations, we mask (+) and reorder (+) turns to generate diverse conversations. We also generate a noisy turn (+) to enhance the model's denoising ability. To avoid generating false positives, we identify the turn dependency structure to guide the turn-level augmentations. (3) Conversation level. We paraphrase the conversation (+) to introduce linguistic variations. We also shift the intent of conversations to help the model detect subtle intent changes (−).

However, LLMs may generate false positives or negatives and be prone to generate texts with hallucinations (Li et al., 2023). To produce high-quality conversations, we propose a three-step prompting process inspired by human cognition. Initially, we prompt an LLM to get a comprehensive understanding of the conversation (*e.g.*, its intent and theme) in the first step (Van Dijk et al., 1983). Subsequently, the LLM associates existing elements, such as expressions, intents, and entities, with new yet related ones (Collins and Loftus, 1975). Finally, the LLM can conclude final outputs based on former outputs. These outputs are less prone to be false positives, false negatives, or hallucinations, as the LLM has a deeper understanding of the original conversation (Step 1) and the generated elements are associated based on existing ones (Step 2).

Subsequently, we employ contrastive learning to bring together augmented positive conversations and push them away from negative ones. Through this, we aim to train a more robust and generalized conversational context encoder, capable of accurately interpreting users' search intents of diverse conversations. To enhance the contrastive learning process, we go beyond basic random sampling methods (Zhu et al., 2021), and introduce a difficulty-adaptive sample filter to select more challenging augmented samples for more difficult conversations. We believe that complex conversations offer a larger learning space for the model. More challenging data can thus provide the model with richer information, enabling it to understand these complex conversations better.

Extensive experiments on four public datasets demonstrate that ConvAug can consistently improve the performance of various conversational dense retrievers across various complexity levels of conversational turns.

The contributions of our work are as follows:

(1) We propose an LLM-based multi-level data augmentation framework ConvAug for conversational search. It manages to comprehensively improve the effectiveness and generalizability of conversational retrievers.

(2) To obtain high-quality data, a cognition-aware prompting process is designed to prevent false positives/negatives and mitigate the hallucination problem of LLMs in conversation generation.

(3) We develop a difficulty-adaptive sample filter to select challenging samples for complex conversations to improve the model's understanding of those with large learning spaces.

2 Related Work

Conversational search. CQR models usually utilize the context to rewrite the conversation into a standalone query (Lin et al., 2020; Qian and Dou, 2022; Mo et al., 2023a). Some researchers attempt to connect the downstream retrieval task to the rewriting task (Wu et al., 2022; Chen et al., 2022; Mao et al., 2023a). On the other hand, CDR models try to utilize the whole conversation to train a conversational context encoder. Some works use a few-shot manner to train the CDR model (Yu et al., 2021; Mao et al., 2022b; Mo et al., 2024). Some design delicate denoising approaches for better CDR models (Mao et al., 2022a; Mo et al., 2023b; Mao et al., 2023c). However, none of these models focus on developing a context encoder that can comprehend diverse conversations smoothly.

Data augmentation for Information Retrieval. Because of the limited nature of relevance judgments, researchers of Information Retrieval (IR) (Zhu et al., 2023; Mao et al., 2020) have resorted to data augmentation. Some use LLMs to generate queries from a document (Bonifacio et al., 2022), or documents from a query (Gao et al., 2023b; Mackie et al., 2023; Wang et al., 2023) in ad-hoc retrieval. For multi-turn ranking, some use basic rule-based approaches to generate variance of sequences for session search (Zhu et al., 2021), personalized search (Zhou et al., 2021), and product search (Dai et al., 2023). COTED (Mao et al., 2022a) generates conversations based on human-annotated necessary historical turns.

LLM for Information Retrieval. LLMs have been widely used in various modules of the IR pipeline (Zhu et al., 2023), such as retriever (Asai

et al., 2023a), reranker (Ma et al., 2023), and reader (Asai et al., 2023b). In conversational search, some employ LLMs to aid the training (Ye et al., 2023) and the inference (Mao et al., 2023b) stage of CQR. Instructor (Jin et al., 2023) uses LLMs to generate pseudo passage labels to facilitate unsupervised CDR models. However, these models fail to utilize LLMs to alternate the contexts for a generalized context encoder.

3 Methodology: ConvAug

In this section, we present our two-stage framework ConvAug, as illustrated in Figure 1. In the first stage, we leverage an LLM to perform our data augmentation strategies tailored for conversational search. A three-step cognition-aware prompting process is developed to guide the LLM to generate multi-level augmented conversations. The second stage is to utilize the augmented data to optimize the conversational context encoder. We propose to select more challenging samples for more complex conversations to facilitate model learning.

3.1 Problem Formulation

In this work, we focus on the conversational passage retrieval task. The context of a conversation is denoted as $C_n = \{q_1, r_1, ..., q_{n-1}, r_{n-1}, q_n\}$, where q_i and r_i are the query and response of the i-th turn (t_i) in C_n, and q_n is the current query. Given C_n, our goal is to retrieve the relevant passage d^+ from the passage collection $\mathcal{D}$. For convenience, we will omit the subscript n in the rest of this paper.

3.2 LLM-enhanced Data Augmentation

Conversational search suffers from a severe data sparsity issue, *i.e.*, varying expressions of recorded conversations are inadequate, leading to insufficient training of context encoders. As shown in Figure 2, we propose to mimic the diverse ways users might express conversations by developing data augmentation strategies. We propose both positive (+) and hard negative (−) generation strategies to produce conversations with similar (C^+) and different intents (C^-), respectively. Furthermore, the LLM-based generation is prompted by a three-step cognition-aware process to mitigate hallucinations and enhance the data quality.

3.2.1 Multi-level Conversation Alteration

• Token-level alteration

Firstly, we propose to perform fine-grained token-level alterations on C to help the model learn nuanced information.

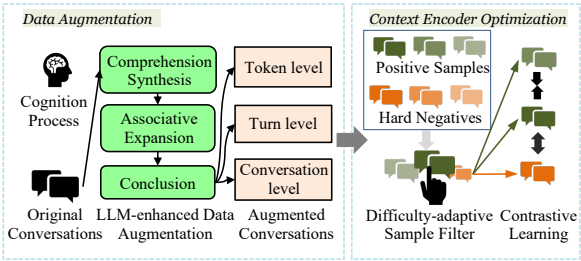

Figure 1: The training workflow of our framework.

Token Masking (+). To prevent the model from relying too much on specific tokens, we employ a rule-based strategy. A context is treated as a sequence of tokens: $C = \{w_1, w_2, \ldots, w_M\}$, where M is the total number of tokens. We randomly mask a proportion r_{w} of the tokens in C with a special token "[token_mask]". By this, we aim to produce a similar context C^+_{tom} as it only has little differences from C in some tokens.

Entity Replacing (−). In real-world scenarios, the same conversational flow can occur with different entities. We use the LLM to identify and replace entities in C to generate C^-_{ent}, which is contextually similar to C but differs in critical details. By contrasting it to other C^+, the model can pay closer attention to the key information in the context rather than the superficial aspects.

• Dependecy-aware turn-level alteration

Secondly, we propose more coarse-grained alterations at the turn level. As shown in Figure 2, the understanding of $t_2 = (q_2, r_2)$ and $t_3 = (q_3)$ both depend on t_1 since they all need the information "train". Therefore, the dependencies within conversations are important if we want to alternate them without changing their search intents, *i.e.*, avoiding producing false positives. Utilizing the ability of LLMs, we can identify the necessary historical turns of t_i automatically. After performing this sequentially on all turns of C, we can construct a turn dependency Directed Acyclic Graph (DAG) $\mathcal{G}$, as shown in the right part of Figure 2.

Turn Masking (+). For all historical turns $T_{\text{h}} = \{t_1, t_2, \ldots, t_{n-1}\}$ of C, we mask a proportion r_{t} of the turns with a special token "[turn_mask]" to generate C^+_{tum}. With this, ConvAug is forced to not rely on specific turns and get a more robust understanding of the whole conversation. To maintain the dependency structure of C, we can only mask the turns that are not the ancestors of t.

Turn Reordering (+). We select a pair of historical turns (t_i, t_j) in T_{h} and swap their positions to

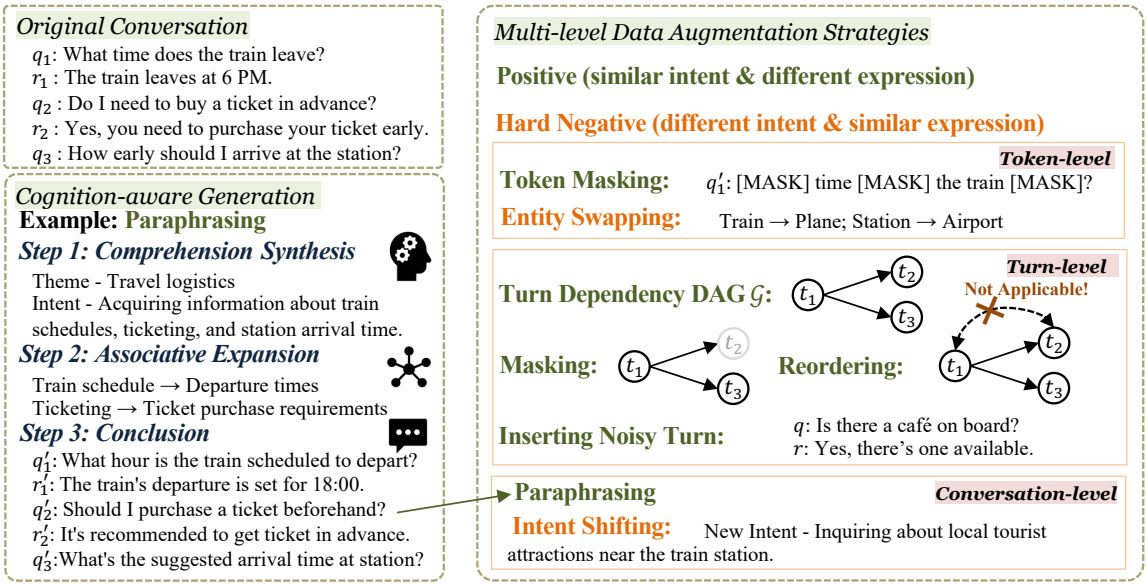

Figure 2: An example to illustrate our cognition-aware prompting process and multi-level augmented data.

produce C^{+}_{reo}. We can only choose turns that the topological ordering of $\mathcal{G}$ remains the same after the swapping. Through this restriction, C^{+}_{reo} will have a different order of expression while maintaining the logical chain as C. This process challenges the model to focus more on the content of each turn rather than just the order.

Inserting Noisy Turn (+). Conversations are often interrupted by unrelated interjections. Corrupting the current context can help the model handle conversational dynamics. We extend the existing context for one additional noisy turn t_{noi} and randomly insert it into T_{h}. Since we prompt the LLM to generate a turn that is relevant to the main background of C but introduces a slightly divergent element, the generated turn can be inserted into any position in T_{h} to produce C^{+}_{noi} without disrupting the dependency structure.

• **Conversation-level alteration**

At last, we apply more high-level changes to the whole conversation.

Paraphrasing (+). To mimic users' various expressions of similar intents, we aim to use the LLM to expand the linguistic diversity by paraphrasing the whole C to produce C^{+}_{para}. This can help reduce the model's tendency to overfit specific phrasings or patterns of C, which enhances the model's ability to generalize to unseen conversations.

Intent Shifting (−). The intent behind a dialogue can shift subtly without significant changes in the expression of the conversation. Therefore, we utilize the LLM to produce the intent-shifted conversations C^{-}_{int}. By contrasting them to C^{+}, our model is trained to detect and adapt to subtle intent shifts in real conversations.

3.2.2 Cognition-aware Prompting Process

To enhance the data quality, we propose a three-step prompting method inspired by human cognition theory, including *Comprehension Synthesis (Step 1)*, *Associative Expansion (Step 2)*, and *Conclusion (Step 3)*. As shown in Figure 2, we take the paraphrasing strategy as an example for illustration:

Step 1: Comprehension Synthesis. When we have a conversation, our brains initially construct a comprehensive representation of the text (Van Dijk et al., 1983). This step allows the LLM to have a comprehensive understanding of the whole conversation. Specifically, we prompt the LLM using "*Step 1: Comprehension Synthesis: [Identify key themes and intents of the conversation]*". The understanding of these core aspects will prevent the LLM from generating C^{+}_{para} that deviates from the theme and search intents (false positive).

Step 2: Associative Expansion. The human mind often uses spreading activation in semantic networks, where one concept triggers related concepts (Collins and Loftus, 1975). Inspired by this theory, the prompt we give the LLM is "*Step 2: Associative Expansion: [Generate alternative expressions based on existing ones]*". This step serves as an intermediate process that leverages LLM's creativity to think of novel elements while preventing it from hallucinating unrelated elements.

Step 3: Conclusion. In the final step, we prompt the LLM as: "*Step 3: Conclusion: [Paraphrase the*

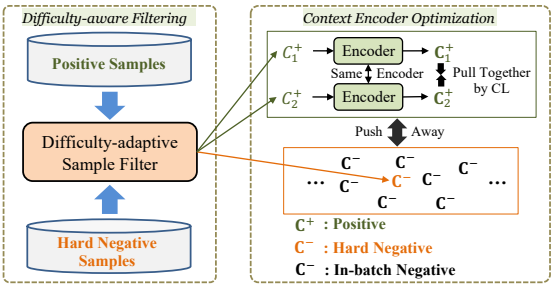

Figure 3: The optimization of context encoder.

conversation based on outputs of last two steps]". In our example, the output is a paraphrased conversation that maintains C's search intent (Step 1) while introducing new but related (Step 2) expressions, avoiding false positives and hallucinations.

We manually write several demonstrations for each step to prompt an LLM to do in-context generation. The complete prompts are in Appendix C.

3.3 Training Conversational Context Encoder

Through our proposed data augmentation strategies, we can generate a set of positive samples $\mathcal{C}^+ = \{C^+_{\text{tom}}, C^+_{\text{tum}}, C^+_{\text{reo}}, C^+_{\text{noi}}, C^+_{\text{para}}\}$ and hard negative samples $\mathcal{C}^- = \{C^-_{\text{ent}}, C^-_{\text{int}}\}$ for an original conversation C in the dataset. Then, to enhance model learning, we develop a difficulty-adaptive sample filter to keep samples of matching difficulty for original conversations. Finally, we train the conversational context encoder on these augmented samples with multi-task contrastive learning.

3.3.1 Difficulty-adaptive Sample Filter

Considering that simple augmentations for complex C may result in underfitting, and complex augmentations for simple C can cause overfitting, we develop a difficulty-adaptive sample filter. It selects difficult samples for difficult conversations to enhance the training process.

Specifically, the difficulty of the original conversations is defined as: $\text{Diff}(C) = |T_{\text{h}}| + \Big(|\text{Topic}(C)| * \overline{\text{PPL}(C)}\Big)$, where $|T_{\text{h}}|$ denotes the number of the historical turns, $|\text{Topic}(C)|$ is the number of topics , and $\overline{\text{PPL}(C)}$ denotes the average perplexity of C. The detailed calculation of these components can be found in Appendix D. To give the diversity of topics and the linguistic challenges more emphasis, we compute $|\text{Topic}(C)| * \overline{\text{PPL}(C)}$ and use $|T_{\text{h}}|$ as an indicator of rich information within long conversations.

For the difficulty of the augmented conversations, we first obtain paired positive samples: $\mathcal{P}_{\mathcal{C}^+} = \{(C^+_i, C^+_j) \mid C^+_i, C^+_j \in \mathcal{C}^+, i \neq j\}$. We then use a sentence-transformers model to compute the similarity of each pair, the difficulty is denoted as $\text{Diff}^+(C^+_i, C^+_j) = 1 - \text{BERTSim}(C^+_i, C^+_j)$, where $\text{BERTSim}(\cdot)$ is the cosine similarity of encoded conversations.

For the diversity of used augmented samples, we divide all training conversations into $|\mathcal{P}_{\mathcal{C}^+}|$ buckets based on $\text{Diff}(C)$. We then filter and select one positive pair with matching $\text{Diff}^+(C^+_i, C^+_j)$ for each conversation. As for hard negatives, we pair each negative with selected positive samples: $\text{Diff}^-(C^-_h) = (\text{BERTSim}(C^+_i, C^-_h) + \text{BERTSim}(C^+_j, C^-_h))/2$. We select k hard negatives with higher $\text{Diff}^-(C^-_h)$ for difficult C.

3.3.2 Multi-task Contrastive Learning

For the ranking task, we apply a standard ranking loss based on contrastive learning of passages:

$$\mathcal{L}_{\text{rank}} = -\log \frac{e^{(\mathbf{C}\cdot\mathbf{d}^+)}}{e^{(\mathbf{C}\cdot\mathbf{d}^+)} + \sum_{d^- \in \mathcal{D}} e^{(\mathbf{C}\cdot\mathbf{d}^-)}}, \quad (1)$$

where $\mathbf{C} = \text{CCE}(s)$ denotes C encoded by the conversational context encoder and $s = [\text{CLS}] \circ q_n \circ r_{n-1} \circ \ldots \circ r_1 \circ q_1 \circ [\text{SEP}]$ is the concatenated sequence of C. $\mathbf{d}^+$ and $\mathbf{d}^-$ are encoded by the frozen passage encoder $\mathbf{d} = \text{PE}(d)$.

Suppose a minibatch contains N conversations, we use our difficulty-adaptive sample filter to select two positive samples for each C to form $\{\mathcal{X}\}$ comprising $2N$ sequences. The two sequences derived from the same C are considered a similar pair, whereas the remaining $2(N-1)$ serve as in-batch negative samples $\{\mathcal{X}\}^-$. Besides, we select k hard negative samples for each C to form $\{\mathcal{H}\}$ comprising kN sequences. The contrastive learning loss for a positive pair (C^+_i, C^+_j) and negatives $C^- \in \{\{\mathcal{X}\}^- \cup \mathcal{H}\}$ of C is formulated as follows:

$$\mathcal{L}_{\text{CL}}(i,j) = -\log \frac{\phi(\mathbf{C}^+_i, \mathbf{C}^+_j)}{\phi(\mathbf{C}^+_i, \mathbf{C}^+_j) + \sum \phi(\mathbf{C}^+_i, \mathbf{C}^-)} \quad (2)$$

where $\phi(\cdot) = \exp(\cos(\cdot)/\tau)$, $\cos(\cdot)$ is cosine similarity and τ is a hyperparameter temperature.

We optimize these two tasks together as: $\mathcal{L} = \mathcal{L}_{\text{rank}} + \alpha\mathcal{L}_{\text{CL}}$, where α is used to balance losses.

4 Experiments

4.1 Datasets and Metrics

We evaluate our model with both normal and zero-shot evaluation. Following previous CDR

Category	Model	QReCC			TopiOCQA		
		MRR	NDCG@3	Recall@10	MRR	NDCG@3	Recall@10
CQR Models	T5QR	34.5	31.8	53.1	23.0	22.2	37.6
	ConQRR	41.8	-	65.1	-	-	-
	ConvGQR	42.0	41.0	64.4	25.6	24.3	41.8
	ED	49.4	-	67.0	-	-	-
CDR Models	ConvDR	38.5	35.7	58.2	27.2	26.4	43.5
	InstructoR-ANCE	43.5	40.5	66.7	25.3	23.7	45.1
	Conv-ANCE	49.0	46.6	71.4	30.4	28.5	52.6
	Conv-SPLADE	50.0	46.6	69.9	30.7	29.5	52.1
	LeCoRE	51.1	48.5	73.9	32.0	31.4	54.3
	ConvAug (Ours)	**52.7**†	**50.4**†	**75.6**†	**35.0**†	**33.3**†	**57.9**†

Table 1: The results of the normal evaluation. "†" denotes our model outperforms all baselines significantly except CONQRR and ED. The best performance is in bold and the second-best performance is underlined.

works (Mao et al., 2023c; Jin et al., 2023), we train ConvAug on **QReCC** (Anantha et al., 2021) and **TopiOCQA** (Adlakha et al., 2022). Additionally, we test ConvAug that has been trained on QReCC in a zero-shot setting on **CAsT-20** (Dalton et al., 2020) and **CAsT-21** (Dalton et al., 2021). We omit the CAsT-19 dataset since it is less challenging and realistic compared to CAsT-20 and CAsT-21 (Mao et al., 2023b). More details are in Appendix A.

Following previous works (Ye et al., 2023), we use some popular metrics for normal evaluation: MRR, NDCG@3, Recall@10. For zero-shot setting, we use metrics suggested by CAsT (Dalton et al., 2021): MRR, NDCG@3. All significant tests are done using paired t-tests at $p < 0.05$ level with Bonferroni correction.

4.2 Implementation Details

We adopt ANCE (Xiong et al., 2021) as the base model of ConvAug. For the large language model, we use Llama 2-Chat (7B) (Touvron et al., 2023) to perform our data augmentation tasks. We use $k = 1$ augmented negative conversations as hard negatives. More details about training and hyperparameters are in the anonymous version of our code (`https://anonymous.4open.science/r/ConvAug-7CB5/`) and Appendix B.

4.3 Baselines

We compare ConvAug with two kinds of models:

Conversational query rewriter. • T5QR (Lin et al., 2020) trains the rewriter with the human rewrites. • ConQRR (Wu et al., 2022) employs reinforcement learning to train CQR models. • ConvGQR (Mo et al., 2023a) reformulates better conversational queries by relating to the retrieval task. • ED (Ye et al., 2023) distills the rewriting capabilities of LLMs into smaller models. Note we do not compare those using black-boxed LLMs (*e.g.*, ChatGPT) during inference (Mao et al., 2023b) since these models require significant resources and time to invoke API numerous times during inference.

Conversational dense retriever. • ConvDR (Yu et al., 2021) distills knowledge for few-shot learning. • Conv-ANCE (Lin et al., 2020) & Conv-SPLADE (Formal et al., 2021) are ANCE and SPLADE fine-tuned on the training conversations with only the training loss. • ConvDR (Yu et al., 2021) distills knowledge for few-shot learning. • LeCoRE (Mao et al., 2023c) extends SPLADE by generating denoised and interpretable lexical session representation. • InstructoR (Jin et al., 2023) employs LLMs to estimate the session-passage relevance score to guide the retriever's training. We use the "ANCE+InstructoR$_{\text{QRPG}}$" version for fair comparisons with ConvAug.

4.4 Overall Results

4.4.1 Normal Evaluation

We conduct the normal evaluation on QReCC and TopiOCQA, and the results are presented in Table 1. We can make these observations: (1) ConvAug outperforms all baseline models significantly on both datasets. This demonstrates the effectiveness of our LLM-enhanced data augmentation and context encoder optimization. Furthermore, based on the model ANCE, whose performance is comparable to SPLADE, ConvAug still manages to gain superior performance than the SPLADE-based model LeCoRE. This further indicates that our ap-

Model	CAsT-20		CAsT-21	
	MRR	NDCG@3	MRR	NDCG@3
InstructoR-ANCE	43.7	29.6	53.0	34.9
Conv-ANCE	42.2	27.7	52.3	34.2
Conv-SPLADE	36.9	28.1	47.9	29.9
LeCoRE	37.7	29.0	50.8	32.3
CONVAUG (Ours)	**45.0**†	**30.7**†	**54.8**†	**36.8**†

Table 2: The performances of CDR models at zero-shot setting. "†" denotes our model outperforms all baselines significantly. The best performance is in bold and the second-best performance is underlined.

Model	MRR	NDCG@3
CONVAUG (Full)	**52.7**†	**50.4**†
w/o. Token Masking (C^{+}_{tom})	51.2	48.9
w/o. Turn Masking (C^{+}_{tum})	51.9	49.6
w/o. Turn Reordering (C^{+}_{reo})	52.0	49.5
w/o. Noisy Turn (C^{+}_{noi})	52.3	49.9
w/o. Dependency-aware	52.0	49.6
w/o. Paraphrasing (C^{+}_{para})	52.1	49.8
w/o. Entity Replacing (C^{-}_{ent})	50.8	48.5
w/o. Intent Shifting (C^{-}_{int})	52.4	50.0
w/o. Cognition-aware	51.1	49.0
w/o. Filter (rand)	51.7	49.5
w/o. Filter (easy)	51.6	49.3

Table 3: Performances of ablated models on QReCC. "†" denotes CONVAUG outperforms ablated models significantly.

proach can train a more robust and generalized context encoder. (2) CDR models generally outperform CQR models. We can observe that even the simply fine-tuned model Conv-ANCE still outperforms the LLM-aided CQR model ED. This indicates the importance of the ranking signal and the effectiveness of our multi-task learning approach.

4.4.2 Zero-shot Evaluation

We also evaluate our model's generalization ability by conducting a zero-shot test of CDR models trained on QReCC on two challenging datasets CAsT-20 and CAsT-21. From the results in Table 2, we can make the following observations: (1) CONVAUG consistently outperforms all CDR baseline models in terms of both metrics on all datasets. Specifically, CONVAUG maintains its superiority over ANCE-based CDR models (Conv-ANCE and InstructoR-ANCE), which further demonstrates the generalization ability of CONVAUG. (2) The unsupervised model InstructoR-ANCE gains the second-best performance in the zero-shot setting. For example, it gains a performance of 43.7 in terms of MRR on CAsT-20. However, its performance is poor in the normal setting. This indicates that this unsupervised approach might not align well with labeled tasks but it can be effectively applied to unseen datasets.

4.5 Ablation Study

To evaluate the effectiveness of each component, we conduct ablation studies on CONVAUG:

Data augmentation strategies. We first conduct ablation experiments on our multi-level data augmentation strategies. As shown in Table 3, the performance of CONVAUG drops significantly after discarding each kind of alteration. Specifically, the performance of CONVAUG drops most when we discard the strategy Entity Replacing (C^{-}_{ent}). This demonstrates that teaching our model to pay more attention to key information in conversations is effective for understanding search intents. Additionally, we find that CONVAUG's performance decreases if we do not mask or reorder turns based on the dependency graph constructed by the LLM. All these results demonstrate the effectiveness of our designed data augmentation strategies.

Cognition-aware prompting process. We also replace the three-step prompting process with a naive prompt template (Appendix C.2) and train "CONVAUG w/o. Cognition-aware" on data generated by this prompt. The performance of CONVAUG decreases by about 3% in terms of MRR when we replace the cognition prompting process. This demonstrates that our cognition-aware prompting process can produce data with higher quality.

Difficulty-adaptive sample filter. We replace our filter with a random selector (CONVAUG w/o. Filter (rand)) and one that selects easy samples for difficult conversations (CONVAUG w/o. Filter (easy)). The decrease in CONVAUG's performance demonstrates that selecting challenging augmented samples for difficult conversations can help the model understand them better. Specifically, the performance of CONVAUG decreases if we assign easy samples to difficult conversations (even worse than randomly selecting). This further demonstrates that we will underfit CONVAUG if we do not give harder conversations enough learning space.

4.6 Performance on Different Turns

To investigate the performance of CONVAUG at a more fine-grained level, we compare it with LeCoRE and Conv-ANCE at the turn level using TopiOCQA (normal) and CAsT-21 (zero-shot)

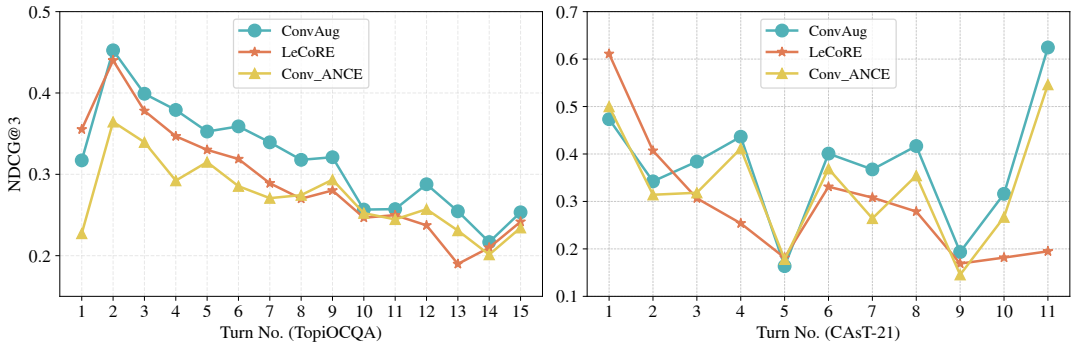

Figure 4: Turn-level performance comparisons on TopiOCQA (normal) and CAsT-21 (zero-shot).

Ratio	QReCC		CAsT-21	
	MRR	NDCG@3	MRR	NDCG@3
$k = 0$	50.8	48.4	53.3	35.3
$k = 1$	**52.7**†	**50.4**†	**54.8**†	**36.8**†
$k = 2$	51.5	49.0	50.8	34.3

Table 4: Performances of CONVAUG with different ratios k of hard negative samples. "†" indicates the result is significantly better than others.

Model	MRR	NDCG@3
Conv-SPLADE	50.0	46.6
Conv-SPLADE + CONVAUG	**52.4**†	**49.8**†
LeCoRE	51.1	48.5
LeCoRE + CONVAUG	**53.1**†	**50.7**†

Table 5: The performances of the base models and the models with our training framework (CONVAUG) on the QReCC dataset. "†" indicates the result in bold is significantly better than the base model.

datasets. The results, as shown in Figure 4, indicate that CONVAUG surpasses both baselines in the majority of turns, underscoring its effectiveness and generalizability again. Specifically, CONVAUG shows more significant improvements in later conversation turns (*e.g.*, from the 2nd to the 15th turns on TopiOCQA and the 3rd to the 11th turns on CAsT-21). This is because longer conversations often contain more diverse information and our augmented data can help CONVAUG to generalize to these complex conversations. Besides, our difficulty-adaptive sample filter can challenge CONVAUG to learn more about complex conversations.

4.7 Influence of Augmented Hard Negatives

We use k generated hard negative contexts to facilitate the training of CONVAUG's context encoder. The performances of CONVAUG with different ks are in Table 4. We can observe that CONVAUG performs best on QReCC with $k = 1$ hard negative. We believe there is a trade-off. The lack of hard negatives limits the model's ability to benefit from challenging comparisons, leading to a less robust feature representation. On the other hand, incorporating multiple hard negatives may introduce noise or ambiguity, potentially corrupting the learning process. Besides, we can observe that CONVAUG ($k = 0$) performs better on zero-shot than on normal evaluation. This further demonstrates that too many hard negative samples will introduce noise and harm the model's generalizability.

4.8 Application to Other Base Retrivers

We use ANCE as the base model of CONVAUG since it is a popular dense retriever that has been the base model of many CDR models. However, our training framework can be easily applied to other CDR models. We choose Conv-SPLADE and LeCoRE as the base models and apply our approach to them. From the results shown in Table 5, we can observe that our method can bring significant improvements across different base CDR models (even sparse retrievers). This demonstrates the broad applicability of our approach.

5 Conclusion

In this work, we present CONVAUG to augment conversational search data with LLMs. We design a three-step cognition-aware prompting process to generate multi-level augmented conversations. We also develop a difficulty-adaptive sample filter to assign challenging samples to difficult conversations for larger learning space. A contrastive learning objective is employed to train a generalized conversational context encoder. Extensive experiments on four public datasets at both normal and zero-shot settings validate the effectiveness, generalization ability, and applicability of CONVAUG.

Limitations

For future studies, our work has the following limitations that we plan to address:

1. The equation we developed to assess the complexity of conversations is relatively basic. We plan to design a more sophisticated equation of our three components in the future.

2. We use an LLM to augment the training conversations in the pre-processing stage. Although the inference time remains the same as base retrievers, the augmentation process takes quite a long time because of the data amount we need to generate (millions of conversations) and the limited computational resources (4 NVIDIA A100 GPUs).

3. We only conduct experiments using one LLM Llama 2 (7B) due to the cost of augmenting such a large number of data. Performances of other LLMs will be experimented with in the future.

4. There is also a potential risk involved. Since we are using LLMs to generate conversations, the original data should not contain sensitive or private information that may cause LLMs to produce risky texts.

References

Vaibhav Adlakha, Shehzaad Dhuliawala, Kaheer Suleman, Harm de Vries, and Siva Reddy. 2022. Topiocqa: Open-domain conversational question answering with topic switching. *Trans. Assoc. Comput. Linguistics*, 10:468–483.

Raviteja Anantha, Svitlana Vakulenko, Zhucheng Tu, Shayne Longpre, Stephen Pulman, and Srinivas Chappidi. 2021. Open-domain question answering goes conversational via question rewriting. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, 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. 2023a. 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.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023b. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *CoRR*, abs/2310.11511.

Luiz Henrique Bonifacio, Hugo Queiroz Abonizio, Marzieh Fadaee, and Rodrigo Frassetto Nogueira. 2022. Inpars: Data augmentation for information retrieval using large language models. *CoRR*, abs/2202.05144.

Zhiyu Chen, Jie Zhao, Anjie Fang, Besnik Fetahu, Oleg Rokhlenko, and Shervin Malmasi. 2022. Reinforced question rewriting for conversational question answering. 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 357–370. Association for Computational Linguistics.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. Quac: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2174–2184. Association for Computational Linguistics.

Allan M Collins and Elizabeth F Loftus. 1975. A spreading-activation theory of semantic processing. *Psychological review*, 82(6):407.

Shitong Dai, Jiongnan Liu, Zhicheng Dou, Haonan Wang, Lin Liu, Bo Long, and Ji-Rong Wen. 2023. Contrastive learning for user sequence representation in personalized product search. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2023, Long Beach, CA, USA, August 6-10, 2023*, pages 380–389. ACM.

Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. 2020. Cast 2020: The conversational assistance track overview. In *Proceedings of the Twenty-Ninth Text REtrieval Conference, TREC 2020, Virtual Event [Gaithersburg, Maryland, USA], November 16-20, 2020*, volume 1266 of *NIST Special Publication*. National Institute of Standards and Technology (NIST).

Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. 2021. TREC cast 2021: The conversational assistance track overview. In *Proceedings of the Thirtieth Text REtrieval Conference, TREC 2021, online, November 15-19, 2021*, volume 500-335 of *NIST Special Publication*. National Institute of Standards and Technology (NIST).

Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021. SPLADE: sparse lexical and expansion model for first stage ranking. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, pages 2288–2292. ACM.

Jianfeng Gao, Chenyan Xiong, Paul Bennett, and Nick Craswell. 2023a. *Neural Approaches to Conversational Information Retrieval*, volume 44 of *The Information Retrieval Series*. Springer.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023b. Precise zero-shot dense retrieval without relevance labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 1762–1777. Association for Computational Linguistics.

Zhuoran Jin, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. 2023. Instructor: Instructing unsupervised conversational dense retrieval with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 6649–6675. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur P. Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. *Trans. Assoc. Comput. Linguistics*, 7:452–466.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. Halueval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 6449–6464. Association for Computational Linguistics.

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

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

Iain Mackie, Shubham Chatterjee, and Jeffrey Dalton. 2023. Generative relevance feedback with large language models. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023*, pages 2026–2031. ACM.

Kelong Mao, Zhicheng Dou, Bang Liu, Hongjin Qian, Fengran Mo, Xiangli Wu, Xiaohua Cheng, and Zhao Cao. 2023a. Search-oriented conversational query editing. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 4160–4172. 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 *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 176–186. ACM.

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 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 2935–2946. Association for Computational Linguistics.

Kelong Mao, Hongjin Qian, Fengran Mo, Zhicheng Dou, Bang Liu, Xiaohua Cheng, and Zhao Cao. 2023c. Learning denoised and interpretable session representation for conversational search. In *Proceedings of the ACM Web Conference 2023, WWW 2023, Austin, TX, USA, 30 April 2023 - 4 May 2023*, pages 3193–3202. ACM.

Kelong Mao, Xi Xiao, Jieming Zhu, Biao Lu, Ruiming Tang, and Xiuqiang He. 2020. Item tagging for information retrieval: A tripartite graph neural network based approach. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2327–2336.

Fengran Mo, Kelong Mao, Yutao Zhu, Yihong Wu, Kaiyu Huang, and Jian-Yun Nie. 2023a. Convgqr: Generative query reformulation for conversational search. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 4998–5012. Association for Computational Linguistics.

Fengran Mo, Jian-Yun Nie, Kaiyu Huang, Kelong Mao, Yutao Zhu, Peng Li, and Yang Liu. 2023b. Learning to relate to previous turns in conversational search. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2023, Long Beach, CA, USA, August 6-10, 2023*, pages 1722–1732. ACM.

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

Hongjin Qian and Zhicheng Dou. 2022. Explicit query rewriting for conversational dense retrieval. 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 4725–4737. 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.

Teun Adrianus Van Dijk, Walter Kintsch, et al. 1983. Strategies of discourse comprehension.

Liang Wang, Nan Yang, and Furu Wei. 2023. Query2doc: Query expansion with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 9414–9423. Association for Computational Linguistics.

Zeqiu Wu, Yi Luan, Hannah Rashkin, David Reitter, Hannaneh Hajishirzi, Mari Ostendorf, and Gaurav Singh Tomar. 2022. CONQRR: conversational query rewriting for retrieval with reinforcement learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 10000–10014. Association for Computational Linguistics.

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*. OpenReview.net.

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.

Shi Yu, Zhenghao Liu, Chenyan Xiong, Tao Feng, and Zhiyuan Liu. 2021. Few-shot conversational dense retrieval. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, pages 829–838. ACM.

Yujia Zhou, Zhicheng Dou, Yutao Zhu, and Ji-Rong Wen. 2021. PSSL: self-supervised learning for personalized search with contrastive sampling. In *CIKM '21: The 30th ACM International Conference on Information and Knowledge Management, Virtual Event, Queensland, Australia, November 1 - 5, 2021*, pages 2749–2758. ACM.

Yutao Zhu, Jian-Yun Nie, Zhicheng Dou, Zhengyi Ma, Xinyu Zhang, Pan Du, Xiaochen Zuo, and Hao Jiang. 2021. Contrastive learning of user behavior sequence for context-aware document ranking. In *CIKM '21: The 30th ACM International Conference on Information and Knowledge Management, Virtual Event, Queensland, Australia, November 1 - 5, 2021*, pages 2780–2791. ACM.

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

Appendix

A Dataset Details

In this part, we will introduce more details of the four datasets we use.

QReCC represents the large-scale, open-domain conversational question-answering (QA) dataset featuring human-annotated question rewrites. It integrates conversations from QuAC (Choi et al., 2018), TREC CAsT, and Natural Questions (Kwiatkowski et al., 2019). The text corpus used for retrieval contains 54 million passages.

TopiOCQA comprises conversations coming from a real search query found in Natural Questions, with subsequent interactions simulated using a wizard-of-oz approach.

QReCC	Training	Testing
# Conversations	10,823	2,775
# Turns	63,501	16,451
# Passages	54M	
TopiOCQA	**Training**	**Testing**
# Conversations	3,509	205
# Turns	45,450	2,514
# Passages	25M	

Table 6: Statistics of QReCC and TopiOCQA.

CAsT-20 and **CAsT-21** were released by the TREC Conversational Assistance Track (CAsT). Their limited number of conversations often makes them evaluation datasets. Each query turn in both CAsT-20 and CAsT-21 has a corresponding human rewrite a canonical response passage.

Dataset	CAsT-20	CAsT-21
# Conversations	25	18
# Turns	208	157
# Passages	38M	40M

Table 7: Statistics of the CAsT datasets.

B Implementation Details

We use ANCE provided by Huggingface as the base model[1]. We use $k = 1$ augmented negative conversations as hard negative. We set the temperatures as 0.0012 and 0.001 for training conversational context encoders on QReCC and TopiOCQA, respectively. The token mask ratio r_{w} and turn mask ratio r_{t} are tuned and established as 0.5 and 0.5, respectively for the QReCC dataset and 0.9 and 0.5, respectively for the TopiOCQA dataset. The learning rates are set as 1e-5 and 1.5e-5 for training on QReCC and TopiOCQA, respectively. The weight α is set as 1.0 and 0.1 for QReCC and TopiOCQA, respectively. The model is trained with a batch size of 12. More details can be found in the anonymous version of our code only for review[2].

C Prompt Templates

C.1 Multi-level Data Augmentaion

Prompt: Entity Replacing

Task Overview:
Your task is to replace entities in the current conversation context while keeping the expressions as similar as possible to the original. This involves identifying key entities, replacing them with suitable alternatives, and ensuring the conversation remains coherent. Use the following structured approach:
Example to Illustrate the Process (Demonstration):
Original Conversation:
Query1: "How long is the Golden Gate Bridge?"
Response1: "The Golden Gate Bridge is about 1.7 miles long."
Query2: "When was it opened to the public?"
Response2: "It was opened in May 1937."
Step 1: Comprehension Synthesis (Identify key entities in the conversation)
Output: Key Entities - Golden Gate Bridge, 1.7 miles, May 1937.
Step 2: Associative Expansion (Find suitable replacements for the identified entities)
Output: Brooklyn Bridge, 1.1 miles, December 1883.
Step 3: Conclusion (Reconstruct the conversation with new entities)
Entity-replaced Conversation:
Query1: "How long is the Brooklyn Bridge?"
Response1: "The Brooklyn Bridge is about 1.1 miles long."
Query2: "When was it opened to the public?"
Response2: "It was opened in December 1883."
Now, it's your turn. Please replace entities in the following conversation using the same process:
Original Conversation:
{Input Conversation}
Step 1: Comprehension Synthesis:
[Identify entities of the conversation]
Step 2: Associative Expansion:
[Find suitable replacements for the identified entities]
Step 3: Conclusion:
[Reconstruct the conversation with new entities based on the outputs of the last two steps]
{Output}

[1] https://huggingface.co/castorini/ance-msmarco-passage

[2] https://anonymous.4open.science/r/ConvAug-7CB5/

Prompt: Dependency Identifying

Task Overview:
Your task is to analyze a given conversation and identify the turns that are necessary for understanding the current search intent. Each turn includes one query and one response. Follow this structured approach:
Example to Illustrate the Process (Demonstration):
Original Conversation Context:
Turn1:
Query1: "What are the main attractions in Paris?"
Response1: "The Eiffel Tower and the Louvre are among the top attractions."
Turn2:
Query2: "Is the Louvre open on Sundays?"
Response2: "Yes, it's open from 9 AM to 6 PM."
Current Search Intent (New Query):
Query3: "How can I get tickets to the Louvre?"
Step 1: Comprehension Synthesis (Identify key themes and intents)
Output: Theme - Paris attractions; Intent - Acquiring information about attractions and related logistics.
Step 2: Associative Expansion (Evaluate the importance of each turn in relation to the current intent)
Output: Turn1: General information about attractions; not directly relevant to ticket acquisition. Turn2: Specific information about the Louvre; more relevant to planning a visit, potentially linked to ticketing information.
Step 3: Conclusion (Select turns crucial for the current intent)
Necessary Turns:
Turn2.
Now, it's your turn. Please identify the necessary turns in the following conversation using the same process:
Original Conversation:
{Input Conversation}
Step 1: Comprehension Synthesis:
[Identify key themes and intents of the conversation]
Step 2: Associative Expansion:
[Evaluate the importance of each turn in relation to the current intent]
Step 3: Conclusion:
[Select turns crucial for the current intent based on the outputs of the last two steps]
{Output}

Prompt: Noisy Turn

Task Overview:
Your task is to introduce a noisy turn (one query and one response) into an existing conversation. This turn should be relevant to the main background of the original conversation but introduce a new, slightly divergent element. Use the following structured approach:
Example to Illustrate the Process (Demonstration):
Original Conversation:
Query1: "Can you tell me about the history of the Sydney Opera House?"
Response1: "Certainly, it was designed by Jørn Utzon and opened in 1973."
Query2: "Is it true that Utzon faced challenges during its construction?"
Response2: "Yes, there were significant design and financial challenges that led to his resignation."
Step 1: Comprehension Synthesis (Identify key themes and intents)
Output: Theme - Sydney Opera House's history; Intent - Learning about design, construction challenges, and historical events.
Step 2: Associative Expansion (Generate a related but distinct element)
Output: Exploring Utzon's architectural style or other famous works.
Step 3: Conclusion (Introduce the new turn)
Noisy Turn:
Query: "Apart from the Sydney Opera House, did Utzon design other notable buildings?"
Response: "Yes, he also designed the Bagsværd Church in Denmark, known for its unique roof structure."
Now, it's your turn. Please introduce a noisy turn into the following conversation using the same process:
Original Conversation:
{Input Conversation}
Step 1: Comprehension Synthesis:
[Identify key themes and intents of the conversation]
Step 2: Associative Expansion:
[Generate a related but distinct element of existing ones]
Step 3: Conclusion:
[Generate a noisy turn based on the outputs of the last two steps]
{Output}

Prompt: Paraphrasing

Task Overview:
Your task is to paraphrase the provided conversation while preserving the original intent and meaning. Each turn in the conversation, including queries and responses, should be paraphrased thoughtfully.
Example to Illustrate the Process (Demonstration):
Original Conversation:
Query1: "What time does the train leave?"
Response1: "The train leaves at 6 PM."
Query2: "Do I need to buy a ticket in advance?"
Response2: "Yes, you need to purchase your ticket early."
Query3: "How early should I arrive at the station?"
Step 1: Comprehension Synthesis (Identify key themes and intents)
Output: Theme - Travel logistics; Intent - Acquiring information about train schedules, ticketing, and station arrival time.
Step 2: Associative Expansion (Generate alternative expressions based on existing ones)
Output: Train schedule -> Queries about departure times Ticketing -> Questions about ticket purchase requirements
Step 3: Conclusion (Paraphrase the conversation based on outputs of last two steps)
Paraphrased Conversation:
Query1: "What hour is the train scheduled to depart?"
Response1: "The train's departure is set for 18:00."
Query2: "Should I purchase a ticket beforehand?"
Response2: "It's recommended to get ticket in advance."
Query3: "What's the suggested arrival time at station?"
Now, it's your turn. Please paraphrase the following conversation using the same process:
Original Conversation:
{Input Conversation}
Step 1: Comprehension Synthesis:
[Identify key themes and intents of the conversation]
Step 2: Associative Expansion:
[Generate alternative expressions based on existing ones]
Step 3: Conclusion:
[Paraphrase the conversation based on outputs of the last two steps]
{Output}

Prompt: Intent Shifting

Task Overview:
Your task is to modify the current conversation by shifting its search intent. The new conversation should retain similar expressions to the original but embody a distinctly different intent. Follow this structured approach:
Example to Illustrate the Process (Demonstration):
Query1: "Can you recommend some good Italian restaurants in New York City?"
Response1: "Sure, one popular option is L'Artusi in the West Village."
Query2: "Do they offer vegetarian dishes?"
Response2: "Yes, they have a variety of vegetarian options."
Step 1: Comprehension Synthesis (Identify key themes and intents)
Output: Theme - Italian restaurants; Intent - Seeking recommendations in New York City.
Step 2: Associative Expansion (Choose a distinctly different intent)
Output: New Intent - Inquiring about Italian cooking classes in New York City.
Step 3: Conclusion (Reconstruct the conversation with the new intent)
Intent-Shifted Conversation:
Query1: "Can you suggest some places to learn Italian cooking in New York City?"
Response1: "Certainly, one well-known place is the Culinary Institute in Lower Manhattan."
Query2: "Do they offer classes for beginners?"
Response2: "Yes, they have a variety of courses for beginners."
Now, it's your turn. Please shift the intent of the following conversation using the same process:
Original Conversation:
{Input Conversation}
Step 1: Comprehension Synthesis:
[Identify key themes and intents of the conversation]
Step 2: Associative Expansion:
[Shift the intent based on existing ones]
Step 3: Conclusion:
[Shift the conversation's intent based on the outputs of the last two steps]
{Output}

C.2 Other Prompts

Prompt: Simple QA to Caluculate Perplexity

Task Overview:
I'm going to provide you with a conversation context and a current query. Your task is to answer the current query based on the information of the context:
Example to Illustrate the Process (Demonstration):
Conversation Context:
Query1: Can you recommend an Italian restaurant for me in New York City?
Response1: Giovanni's Veggie Delight is a popular Italian restaurant in NYC.
Query2: What's the weather like in New York today?
Response2: It's currently sunny and warm in New York.
Current Query:
Great, does Giovanni's have outdoor seating?
Response:
Yes, they have a beautiful patio area.
Now, it's your turn. Please answer the following conversation:
Conversation Context:
{Input Conversation Context}
Current Query:
{Input Current Query}
Response:
{Output}

Prompt: Naive Paraphrasing

Task Overview:
I'm going to provide you with a conversation. Your task is to paraphrase the conversation while keeping the original intent and meaning intact. Each turn in the conversation, including a query and a response, should have a corresponding paraphrased version. Here's an example to illustrate:
Example to Illustrate the Process (Demonstration):
Original Conversation:
Query1: "What time does the train leave?"
Response1: "The train leaves at 6 PM."
Query2: "Do I need to buy a ticket in advance?"
Response2: "Yes, you need to purchase your ticket early."
Query3: "How early should I arrive at the station?"
Paraphrased Conversation:
Query1: "What hour is the train scheduled to depart?"
Response1: "The train's departure is set for 18:00."
Query2: "Should I purchase a ticket beforehand?"
Response2: "It's recommended to get ticket in advance."
Query3: "What's the suggested arrival time at station?"
Now, it's your turn. Please paraphrase the following conversation:
Original Conversation:
{Input Conversation}
Paraphrased Conversation:
{Output}

D Details of Calculating Difficulty

To estimate a conversation's complexity, we use $\text{Diff}(C) = |T_\text{h}| + \Big(|\text{Topic}(C)| * \overline{\text{PPL}(C)}\Big)$. This equation is comprised of three components: (1) The number of the historical turns T_h. Longer conversations often contain richer information (Mao et al., 2022a). (2) The number of topics. Each new topic introduces potential contextual shifts. We apply a topic model to count C's topics (more details are in Appendix D). The topic model we used was pre-trained on Wikipedia (https://huggingface.co/MaartenGr/BERTopic_Wikipedia). We illustrate the process of counting topics for a conversation C in Alg. 1. Intuitively, we assume the first turn of C has one topic and each turn can only add at most one topic to its previous turn. To ensure we only count new topics, we only add a topic if our topic model is more confident of identifying this new topic than its last identified topic. (3) The average perplexity of C. Perplexity is a measure to quantify how well an LM predicts a sample. We prompt an LLM (Appendix C.2) to predict the response based on the context and compute the average perplexity of all turns. A higher $\overline{\text{PPL}(C)}$ indicates that the conversation contains a more challenging language.

The sentence-transformer model we use to calculate the similarity between augmented samples is all-MiniLM-L6-v2 (https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2).

Algorithm 1 Counting Topics with Confidence

Require: a conversation $\{t_1, t_2, \ldots, t_n\}$, a topic model $f(\cdot)$
  Initialize $topicCounts$ as an empty list
  Initialize $topics$ as an empty list
  for i in n **do**
    $P \leftarrow f(\{t_1, \ldots, t_i\})$
    $P \leftarrow P \setminus topics$
    $P \leftarrow$ SORT(P, DESCENDING)
    if $i == 1$ **then**
      APPEND 1 TO $topicCounts$
      APPEND ARGMAX(P) TO $topics$
      $confidence \leftarrow P[0] - P[1]$
    else
      $confidence' \leftarrow P[0] - P[1]$
      if $confidence' \geq confidence$ **then**
        APPEND $topicCounts[i-1] + 1$ TO $topicCounts$
        APPEND ARGMAX(P) TO $topics$
        $confidence \leftarrow confidence'$
      else
        APPEND $topicCounts[i-1]$ TO $topicCounts$
      end if
    end if
  end for
  return $topicCounts$

E Examples of Generated Conversations

In this section, we present two examples of the full generated data of a turn by the LLM in Figure 5 and Figure 6. We only show the data generated by the LLM and the example contexts augmented by rule-based strategies (token masking, turn masking, and reordering based on the dependency graph generated by LLM) can be found in Figure 2.

QReCC (Conversation #1, Turn #3)	
Original	*Query 1*: What can you tell me about Gary Cherone? *Response 1*: Gary Francis Caine Cherone is an American rock singer and songwriter, known for his work as the lead vocalist of Extreme and for his short stint for Van Halen. *Query 2*: Did Gary sing well? *Response 2*: Yes, Gary Cherone is also known for his work as the lead vocalist of the Boston rock group Extreme. *Query 3*: What significant fact can you tell me about Gary that you liked?
Entity Replaced	*Query 1*: What can you tell me about Axl Rose? *Response 1*: Axl Rose is an American singer, songwriter, and musician, known for his work as the lead vocalist of Guns N' Roses and for his powerful, wide-ranging voice., *Query 2*: Did Axl sing well? *Response 2*: Yes, Axl Rose is also known for his powerful, emotive voice and his ability to convey the intensity of his songs. *Query 3*: What significant fact can you tell me about Axl that you liked?
Necessary Turns	*Turn 1*
Noisy Turn	*Query*: Did Gary Cherone ever collaborate with other famous musicians? *Response*: Yes, Gary Cherone has collaborated with many artists throughout his career, including his time with Extreme and Van Halen. In fact, he wrote and recorded a song with the late, great Freddie Mercury, called "Love is the Answer".
Paraphrased	*Query 1*: Can you provide information about Gary Cherone? *Response 1*: Gary Francis Caine Cherone is a renowned American rock singer and songwriter who gained fame as the lead vocalist of Extreme and for his brief tenure as the lead vocalist of Van Halen. *Query 2*: Did Gary perform well vocally? *Response 2*: Yes, Gary Cherone is a talented singer and songwriter who has made significant contributions to the music industry as the lead vocalist of Extreme. *Query 3*: What significant fact can you tell me about Gary that you liked?
Intent Shifted	*Query 1*: Can you tell me more about Gary Cherone's influence on the rock scene? *Response 1*: Absolutely, Gary Cherone is known for his unique vocal style and his ability to blend different musical elements, which has had a significant impact on the rock music industry. *Query 2*: How did he differ from other rock singers of his time? *Response 2*: Gary Cherone's distinctive vocal style, which combines elements of hard rock, funk, and R&B, has set him apart from other rock singers of his time. He has been an inspiration to many young rock singers and has helped to shape the sound of contemporary rock music. *Query 3*: What are some of his most notable achievements in the industry?

Figure 5: An example to show the generated data of the LLM for a turn in QReCC.

TopiOCQA (Conversation #150, Turn #3)	
Original	*Query 1*: "Call me by your name" what is it about? *Response 1*: It is about the relationship that grows between Elio and Oliver after Oliver comes to live with Elio's family in Italy for academic purposes. *Query 2*: Is it a movie? *Response 2*: Yes, based on a book of the same name. *Query 3*: Who is the author of the book?
Entity Replaced	*Query 1*: "A Quiet Place" what is it about? *Response 1*: It centers on the Abbott family struggling to survive in silence while hiding from creatures that hunt by sound. *Query 2*: Is it a movie? *Response 2*: Yes, adapted from an original screenplay. *Query 3*: Who penned the screenplay?
Necessary Turns	*Turn 1, Turn 2*
Noisy Turn	*Query:* Did you know that the film was shot in various locations in Italy, including the Amalfi Coast and Rome? *Response*: Yes, the film's cinematography highlights the beauty of Italy's landscape, which adds to the romantic atmosphere of the story.
Paraphrased	*Query 1*: What is the context of this story "call me by your name"? *Response 1*: The story revolves around the emotional connection that develops between Elio and Oliver when Oliver stays with Elio's family in Italy for academic purposes. *Query 2*: Is this a film? *Response 2*: Yes, it is an adaptation of the book of the same name. *Query 3*: Who is the author of the book?
Intent Shifted	*Query 1*: ”Call me by your name" who directed it? *Response 1*: Luca Guadagnino directed the movie. *Query 2*: Does it feature original music? *Response 2*: Yes, it includes original music by Sufjan Stevens. *Query 3*: Who composed the film's score?

Figure 6: An example to show the generated data of the LLM for a turn in TopiOCQA.