UniConv: Unifying Retrieval and Response Generation for Large Language Models in Conversations

Fengran Mo[1*], **Yifan Gao**[2], **Chuan Meng**[3*], **Xin Liu**[2], **Zhuofeng Wu**[2], **Kelong Mao**[4*]
Zhengyang Wang[2], **Pei Chen**[2], **Zheng Li**[2], **Xian Li**[2], **Bing Yin**[2], **Meng Jiang**[5]
[1]University of Montreal; [2]Amazon.com; [3]University of Amsterdam
[4]Renmin University; [5]University of Notre Dame
fengran.mo@umontreal.ca, yifangao@amazon.com, mjiang2@nd.edu

Abstract

The rapid advancement of conversational search systems revolutionizes how information is accessed by enabling the multi-turn interaction between the user and the system. Existing conversational search systems are usually built with two different models. This separation restricts the system from leveraging the intrinsic knowledge of the models simultaneously, which cannot ensure the effectiveness of retrieval benefiting the generation. The existing studies for developing unified models cannot fully address the aspects of understanding conversational context, managing retrieval independently, and generating responses. In this paper, we explore how to unify dense retrieval and response generation for large language models in conversation. We conduct joint fine-tuning with different objectives and design two mechanisms to reduce the inconsistency risks while mitigating data discrepancy. The evaluations on five conversational search datasets demonstrate that our unified model can mutually improve both tasks and outperform the existing baselines.

1 Introduction

The rapid advancement of conversational search systems revolutionizes how information is accessed by enabling the multi-turn interaction between the user and the system (Zamani et al., 2023). With the recent advances of large language models (LLMs), commercial conversational AI search engines, such as Perplexity.ai and SearchGPT[1], have been deployed with increasing attraction.

Existing conversational search systems are usually composed of two different models: a retriever and a generator (Gao et al., 2022; Mo et al., 2024b), which are trained and inferred separately. The retriever aims to identify the relevant passages by understanding conversational queries, while the generator crafts the final response for the information-seeking goal. The deployment of separate models in the whole pipeline induces the problems in two folds: *i)* The separation restricts the system from leveraging the model's intrinsic knowledge simultaneously, which raises the risk of lacking correlation with the performance of both tasks, leading to inconsistent results, i.e., the effectiveness of retrieval might not always benefit response generation (Salemi and Zamani, 2024); and *ii)* Deploying and maintaining two distinct models adds extra hardware requirements and increases maintenance expenses (Zhang et al., 2024). An intuitive solution is to develop a unified model that acts as a retriever and generator in conversational scenarios. This model is expected to mutually improve retrieval and generation performance through seamless integration with end-to-end optimization.

Recent existing studies have demonstrated the feasibility of developing LLM-based unified models in conversational question answering, involving open-domain retrieval[2] and response generation. However, these systems can only address two aspects of understanding conversational context, managing retrieval independently, or generating responses, as illustrated in Table 1. Among them, RepLLaMA (Ma et al., 2024) and E5 (Wang et al., 2024) successfully implement generative LLMs for retrieval tasks and ChatRetriever (Mao et al., 2024a) further adapt it to conversational scenarios. However, the fine-tuning for retrieval objectives leads to the collapse of generation ability in these systems. The RankRAG (Yu et al., 2024) and ChatQA (Liu et al., 2024) enable the system to exploit a more accurate input context for the generator to produce better responses to user queries. How-

[*] Work done when interned in Amazon.

[1] Perplexity.ai: https://www.perplexity.ai/, SearchGPT: https://searchgpt.com/

[2] In this paper, the retrieval denotes retrieving information from a large external collection as an open-domain setting, rather than only identifying specific pieces from the initial search results with limited top-k candidates similar to ranking.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 6936–6949
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

System	Conv.	Ret.	Gen.
RepLLaMA (Ma et al., 2024)	✗	✓	✗
E5 (Wang et al., 2024)	✗	✓	✗
ChatRetriever (Mao et al., 2024a)	✓	✓	✗
RankRAG (Yu et al., 2024)	✓	✗	✓
ChatQA (Liu et al., 2024)	✓	✗	✓
GRIT (Muennighoff et al., 2024)	✗	✓	✓
Our UniConv	✓	✓	✓

Table 1: The functionality comparison between ours and existing systems, including whether support to conversational scenario (Conv.), first-stage retrieval (Ret.), and response generation (Gen.).

ever, they should rely on an additional retriever to address retrieval needs. To develop a unified model capable of handling both retrieval and generation tasks, GRIT (Muennighoff et al., 2024) attempts to train an LLM with distinguished instructions but it is not designed for conversational applications.

To address the limitations of previous studies, we propose **UniConv**, a unified LLM-based model to investigate the feasibility of handling both retrieval and response generation in conversation. To achieve this, we inherit the training data selected by ChatRetriever (Mao et al., 2024a) to adapt LLM to serve as a powerful conversational dense retriever. To improve the response generation ability while fine-tuning dense retrieval, we design a context identification instruction mechanism as part of the learning objective. This mechanism is designed to seamlessly integrate retrieved information into the response generation process. Additionally, we identify a discrepancy in previous training data: the same data format is applied to different learning objectives, which does not align well with the distinct output requirements of retrieval and generation tasks. To mitigate this issue, we include additional well-formatted conversational search data for model fine-tuning. We conduct extensive evaluations on five widely used datasets, where UniConv demonstrates strong generalization capabilities for representing complex conversational sessions in dense retrieval, along with robust generation abilities for crafting responses. Moreover, UniConv achieves better seamless consistency between retrieval and its augmentation for response generation in terms of effectiveness and reliability compared to non-unified models.

Our contributions can be summarized as follows:

(1) We investigate the feasibility of developing a unified LLM for conversational search and propose our UniConv model for better unification.

(2) We design two mechanisms to improve the seamless consistency between conversational retrieval and its augmented response generation while addressing the issue of data discrepancy.

(3) We conduct extensive experiments to evaluate UniConv on information-seeking conversations across various settings, comparing it against several strong baselines. Its superior performance in both conversational dense retrieval and response generation highlights its remarkable effectiveness.

2 Related Work

Conversational Retrieval. Conversational retrieval aims to identify the relevant passages to satisfy users' information needs through multi-turn interaction (Meng et al., 2025; Mo et al., 2025). The main challenge lies in enabling the system to understand the real user search intents expressed in context-dependent queries. The literature outlines two main approaches to achieve the retrieval goal: *i)* conversational query rewriting (Voskarides et al., 2020; Wu et al., 2022; Mo et al., 2023a,b; Mao et al., 2023a,b; Ye et al., 2023; Jang et al., 2023; Mo et al., 2024f,a; Lai et al., 2024) that decomposes the conversational retrieval into a rewrite-then-retrieval pipeline and *ii)* conversational dense retrieval (Qu et al., 2020; Yu et al., 2021; Lin et al., 2021; Kim and Kim, 2022; Mao et al., 2022, 2023c; Jin et al., 2023; Mo et al., 2024d,e,c; Lupart et al., 2025) that directly encode the whole conversational session to perform end-to-end dense retrieval.

Conversational Response Generation. Conversational response generation aims to synthesize information from the top-retrieved passages into a single response (Meng et al., 2020b,a, 2021; Ren et al., 2021; Cheng et al., 2024; Li et al., 2024a). Different from single-turn retrieval-augmented generation (RAG) (Lewis et al., 2020; Asai et al., 2023; Mao et al., 2024b; Zhang et al., 2025), which only needs to consider the given stand-alone query with its associated retrieved results for response generation, a conversational response generator (Ye et al., 2024) requires modeling conversational turn dependency and understanding the context-depend query and search results.

LLM-based Retrieval. To explore the potential of LLMs in retrieval tasks, some existing studies (Wang et al., 2024; Ma et al., 2024) attempt to follow the observed scaling law (Kaplan et al., 2020) in search model (Ni et al., 2022) by replacing the backbone model from the small ones (e.g., BERT-base (Devlin et al., 2019) and T5-base (Raffel et al., 2020)) into the generative LLMs (e.g., Mistral (Jiang et al., 2023) and LLaMa (Touvron et al., 2023)). They keep the training paradigm similar to DPR (Karpukhin et al., 2020) using relevance judgments as supervision signals while changing the representation of queries and passages different from the ones in encoder-based models.

Unified LLMs for Retrieval and Generation. The motivation to develop unified LLMs for retrieval and generation is to attempt to mutually leverage the intrinsic knowledge from both sides to improve the model's general multi-task ability and reduce cost. To this end, Muennighoff et al. (2024) propose GRIT, to train LLMs to handle both generative and retrieval tasks by distinguishing between them through instructions, and Li et al. (2024b) design a unified framework based on generative retrieval and open-domain question answering. Then, Yu et al. (2024) propose RankRAG, which unifies the re-ranking and generation through simultaneously instructing the LLMs on context ranking and answer generation. However, they cannot address multi-turn scenarios due to a lack of conversational adaptation. In a conversational setting, Mao et al. (2024a) and Liu et al. (2024) fine-tune LLMs specifically for conversational retrieval and response generation tasks, respectively, but these adaptations do not preserve the model's ability to perform both functions concurrently. Recently, a parallel study, OneGen (Zhang et al., 2024), propose unifying the traditionally separate training approaches for generation and retrieval by incorporating retrieval tokens generated in an autoregressive manner. However, it cannot follow conversational context and independently handle retrieval tasks.

Our Goal is to develop a unified LLM-based model capable of handling both retrieval and generation in conversation, a scenario that has not been extensively explored in the existing literature.

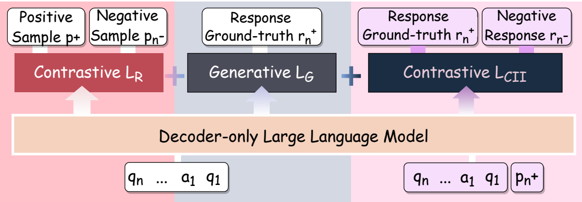

Figure 1: Overview of our UniConv. Three learning objectives are designed with various input and supervision signals for retrieval and generation in conversation.

3 Methodology

3.1 Task Formulation

The task is to establish a unified model which can handle both conversational retrieval and response generation. Formally, given a conversational session that contains $n-1$ historical turns $\mathcal{H}_n = \{q_i, r_i\}_{i=1}^{n-1}$ and current query q_n, the unified model $\mathcal{M}$ is expected to act as a retriever to identify the relevant passages $\mathcal{P}_n$ from a large collection $\mathcal{C}$ and also act as a generator to produce a response r_n on top of external knowledge $\mathcal{P}_n$ to satisfy the information needs in q_n. Thus, the unified model $\mathcal{M}$ is required to handle the multi-turn session query to retrieve the relevant passages as $\mathcal{P}_n = \mathcal{M}(q_n, \mathcal{H}_n)$, and generate the response as $r_n = \mathcal{M}(q_n, \mathcal{H}_n, \mathcal{P}_n)$. In our setting, the unified model $\mathcal{M}$ is a generative LLM with decoder-only architecture.

3.2 Generative Language Models for Conversational Search

The overview of our proposed UniConv is illustrated in Figure 1, which consists of three parts, including various learning objectives toward conversational retrieval (Sec. 3.2.1), conversational response generation (Sec. 3.2.2), and context identification instruction (Sec. 3.3). We describe each component as follows.

3.2.1 Conversational Dense Retrieval

The common practice for dense retrieval fine-tuning is the paradigm of DPR (Karpukhin et al., 2020), which leverages a bi-directional encoder-only model to encode the queries and passages separately on top of a bi-encoder architecture. Then, the first sequence token [CLS] is employed as the text representation for similarity calculation. When the backbone model $\mathcal{M}$ turns into generative ones with uni-directional decoder-only architecture, e.g., LLaMA, the adaption is to form the representation $\mathcal{V}_{\text{seq}}$ using an appended end-of-sequence token

</s> to both the queries and passages (Ma et al., 2024) as $\mathcal{V}_{\text{seq}} = \mathcal{M}(\cdot)[-1]$.

To adapt the conversational scenario, the input query for each turn n is reformulated as $q'_n = q_n \circ \mathcal{H}_n$ by concatenating the context of the previous turn. Then it is vectorized with candidate passages p_n by the model $\mathcal{M}$ and calculate their similarity $\mathcal{S}\left(q'_n, p_n\right) =< \mathcal{V}_{q'_n}, \mathcal{V}_p >$ via dot product. With the established representation, contrastive learning with InfoNCE loss is used for end-to-end conversational dense retrievers optimization as

$$\mathcal{L}_{\text{R}} = -\log \frac{e^{\mathcal{S}\left(q'_n, p_n^+\right)}}{e^{\mathcal{S}\left(q'_n, p_n^+\right)} + \sum_{\mathcal{P}_n^- \in \{\mathcal{P}_N\}} e^{\mathcal{S}\left(q'_n, \mathcal{P}_n^-\right)}}$$

3.2.2 Conversational Response Generation

For information-seeking response generation in the conversation, the generator shares the same comprised query input q'_n as the retriever and is required to maintain the generation ability while learning for retrieval. To enhance the robustness of the generator, we inherit the idea of Seq2Seq (Sutskever, 2014), enabling the model to only be conditional on the representation of the input query $\mathcal{V}_{q'_n}$ rather than attention on all previous input and generated tokens. This is achieved by applying the session-masked technique in (Mao et al., 2024a) and the training objective to generate the ground-truth for turn n is shown below, where $|r_n| = T$.

$$\mathcal{L}_{\text{G}} = -\frac{1}{T}\sum_{i=1}^{T} \log p(r_n^i | r_n^1, ..., r_n^{i-1}, \mathcal{V}_{q'_n})$$

3.3 Context Identification Instruction

During the inference phase with the retrieval-augmented setting, the model input is usually the query together with the retrieved evidence serving as the main part of the instruction, where the model is expected to generate the response grounding on the useful piece of the retrieved evidence. Since the relevant passage and ground-truth response used as supervision signals are separately conducted during the training phase within the unified model, potential inconsistency risk might occur (Yu et al., 2024). To this end, we design a context identification instruction to help the model implicitly identify the useful passage during the fine-tuning, which is consistent with the input instruction format of inference. This is achieved by combining the query with the positive passage in the same sequence and using different responses as contrastive samples as

$$\mathcal{L}_{\text{CII}} = -\log \frac{e^{\mathcal{S}\left(q'_n \circ p_n^+, r_n^+\right)}}{e^{\mathcal{S}\left(q'_n \circ p_n^+, r_n^+\right)} + \sum\limits_{r_n^- \in \{r\}} e^{\mathcal{S}\left(q'_n \circ p_n^+, r_n^-\right)}}$$

3.4 Data Discrepancy Mitigation

To equip the LLMs with conversational dense retrieval capability, Mao et al. (2024a) leverage the ad-hoc search data with relevant query-passage pairs and instructional conversation with multi-turn query-response pairs to enable the model to obtain retrieval and conversational context understanding ability. In practice, their implementation utilizes each turn's response r_n^+ in the conversation dataset as the relevant passage p_n^+ and each ad-hoc query's corresponding relevant passage p_n^+ as the ground-truth r_n^+ for retrieval and generation fine-tuning, respectively. However, a unified model should have different outputs for conversational retrieval (e.g., rank-list) and generation (e.g., synthesized response), whose requirement is not exactly matched with the fine-tuned data form in existing studies and thus might lead to sub-optimal results. A more practical way is to ensure each data sample includes both the relevant passage p_n^+ and the corresponding ground-truth response r_n^+ as supervision signals for the given query turn q_n. Then, the model can learn the consistency from the various targets between retrieval and generation. Thus, we include the conversational search data (Adlakha et al., 2022) to meet this requirement to mitigate the data discrepancy issue. Another alternative is to construct synthetic data (Liu et al., 2024) with well-formed signals, which is not the focus of our paper.

3.5 Training and Inference

For the training phase, we integrate the conversational dense retrieval, retrieval-augmented response generation, and the context identification instruction tuning to form the training objective $\mathcal{L}$ of our unified model as Eq. 1, where α is a hyper-parameter to control the fine-tuning effect. For the inference phase, we use the same fine-tuned model to perform retrieval to produce a top-k rank list and generation to produce a response within zero-shot and RAG settings under conversational scenarios.

$$\mathcal{L} = \mathcal{L}_{\text{R}} + \mathcal{L}_{\text{G}} + \alpha\mathcal{L}_{\text{CII}} \tag{1}$$

4 Experiments

4.1 Experimental Setup

Evaluation Datasets and Metric. We conduct the main evaluation on four widely-used conversational search datasets, including TopiOCQA (Adlakha et al., 2022), QReCC (Anantha et al., 2021), OR-QuAC (Qu et al., 2020), and INSCIT (Wu et al., 2023). Each of them contains the gold standard for passage retrieval and response generation. Besides, FaithDial (Dziri et al., 2022), an information-seeking dialogue benchmark, and TopiOCQA are used for evaluating the reliability of the generated content via the given evidence/rationale. The statistics and details of the datasets are provided in Appendix A.1. We use NDCG@3, Recall@10, and F1 to evaluate the retrieval and generation performance to conduct a fair comparison with baselines.

Training data. We use the ad-hoc search dataset MSMARCO (Bajaj et al., 2016), the *The Question About the World* subset of the instructional conversation dataset UltraChat (Ding et al., 2023), and the whole conversational search dataset TopiOCQA for fine-tuning the unified model.

Baselines. We compare our methods with various conversational retrieval and response generation baselines. For the retrieval phase, we compete with the most effective conversational dense retrieval (CDR) systems based on small language models (SLMs), including ConvDR (Yu et al., 2021), Conv-ANCE (Mao et al., 2023c), and QRACDR (Mo et al., 2024c) and most recently LLM-based approaches, including RepLLaMA (Ma et al., 2024), E5 (Wang et al., 2024), (Conv-)GRIT (Muennighoff et al., 2024), and ChatRetriever (Mao et al., 2024a). The GRIT is the only system that can handle both retrieval and generation tasks, and its variant Conv-GRIT is fine-tuned for conversation on the same setting as ours. Besides, the compared baselines also contain the methods based on conversational query rewriting (CQR) on top of LLMs, including the ones without fine-tuning (LLM-Aided (Ye et al., 2023), LLM4CS (Mao et al., 2023a), and CHIQ (Mo et al., 2024a)) and with fine-tuning (RETPO (Yoon et al., 2024)).

In the response generation phase, we conduct the comparison under zero-shot and RAG settings. For the zero-shot manner, we include GRIT and its variants Conv-GRIT and three powerful LLMs: Mistral, Claude (AnthropicAI, 2023), and ChatGPT (OpenAI). For the RAG setting, to make the results comparable, we employ Mistral-2-7B-chat as the generator with two typical conversational dense retrievers Conv-ANCE and Chatretriever, and keep the Conv-GRIT on the same workflow as our UniConv, i.e., using the same model for both tasks. More details about the baseline methods are described in Appendix A.2.

Implementation Details. We initialize UniConv with Mistral-2-7B-chat, which can be also applied on top of any generative models. We train it on eight 40G A100 GPUs using LoRA (Hu et al., 2022) with a maximum input sequence length of 1024 for query and 384 for passages and responses. The training process involves one epoch with a learning rate of 1e-4, a gradient accumulation of 4 steps, a batch size of 32, and in-batch negatives per sample. The α for loss balance is set to 0.5. During the inference stage, we deploy Faiss (Johnson et al., 2019) for the dense retrieval, set the maximum length as 128, and use top-10 retrieved passages for the response generation. For baseline comparisons, we adhere to the implementation settings specified in their original papers. The LLM-based CQR and the SLM-based CDR methods are based on ANCE dense retriever (Xiong et al., 2020).

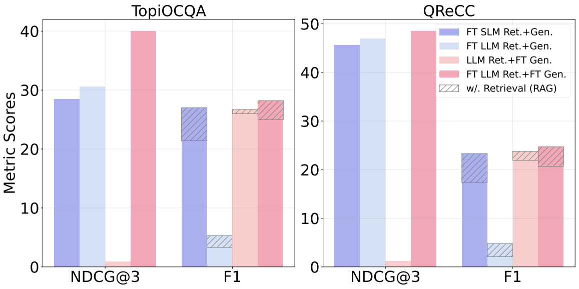

Figure 2: The performance of different systems to fine-tune language models for conversational retrieval and response generation with various settings.

4.2 Feasibility of Unifying Conversational Retrieval and Generation

We first examine the feasibility of unifying conversational retrieval and generation to verify whether jointly fine-tuning can maintain the model's generation ability while grasping retrieval capacity. The results for conversational retrieval and response generation on two datasets are shown in Figure 2, which includes four different systems: *i)* fine-tuning SLM for retrieval and using the original

Category	System	TopiOCQA		QReCC		OR-QuAC		INSCIT	
		NDCG@3	R@10	NDCG@3	R@10	NDCG@3	R@10	NDCG@3	R@10
CQR	LLM-based Conversational Query Rewriter (7B) + Ad-hoc Dense Retriever (110M)								
	LLM-Aided	-	-	41.3	65.6	-	-	-	-
	LLM4CS	26.7	43.3	42.1	66.4	-	-	-	-
	RETPO (w./ FT)	28.9	49.6	41.1	66.7	-	-	-	-
	CHIQ	32.2	53.0	44.6	70.8	-	-	-	-
CDR	SLM-based Encoder-only Dense Retriever (110M)								
	ConvDR	26.4	43.5	35.7	58.2	-	-	-	-
	Conv-ANCE	28.5	52.6	45.6	71.5	35.5	55.6	24.5	38.2
	QRACDR	36.5	57.1	49.1	74.8	40.8	60.4	30.0	43.6
	LLM-based Decoder-only Dense Retriever (7B)								
	RepLLaMA	15.0	27.2	31.8	20.4	-	-	-	-
	E5	16.9	28.7	32.9	21.1	-	-	-	-
	GRIT	17.3	30.9	33.5	23.6	-	-	-	-
	Conv-GRIT	36.0	54.2	48.3	69.7	-	-	-	-
	ChatRetriever	40.1	63.7	**52.5**	**75.8**	41.9	58.9	35.1	50.8
	UniConv (Ours)	**42.6**†	**67.4**†	47.6	68.9	**43.5**†	**63.0**†	**36.2**†	**54.2**†

Table 2: Performance of different systems for conversational retrieval on four datasets. † denotes significant improvements with t-test at $p < 0.05$ over each of the compared CDR systems. **Bold** indicates the best results.

LLM as the generator; *ii)* fine-tuning LLM for retrieval only and *iii)* for response generation only; *iv)* fine-tuning LLM for both tasks.

By comparing systems ii), iii), and iv), we observe that only fine-tuning a single part on the backbone LLM for retrieval or response generation hurts another ability. However, jointly fine-tuning the model with the objective functions for both tasks can obtain a unified model. Besides, the LLM-based retriever performs better than the SLM-based one, indicating the potential for conversational search performance with an LLM.

Then we investigate the RAG setting by incorporating the corresponding retrieved passages for the response generation, except applying the search results from system ii) to system iii), since these two systems cannot handle both tasks. We can see the improvement from RAG is higher in system iv) with a unified model compared with system iii) with a separated one. These results confirm the feasibility of developing a unified model for conversational search. In the following sections, we conduct experiments to investigate our approaches.

4.3 Results of Conversational Retrieval

Table 2 shows the conversational retrieval results on four datasets and the comparison with existing systems, where we have the following observations:

(1) Our proposed UniConv outperforms the baseline methods on most of the datasets, including the previous unified model (Conv-GRIT), the state-of-the-art conversational dense retriever (QRACDR and ChatRetriever), and conversational query rewriter (LLM4CS and CHIQ), which demonstrates that the superior dense retrieval ability of our developed system by arousing the LLM capacity with specific fine-tuning.

(2) The state-of-the-art CDR systems, either SLM-based (QRACDR) or LLM-based (ChatRetriever and UniConv) consistently perform better than the LLM-based CQR systems, which indicates the end-to-end optimization can achieve better performance compared with the rewrite-then-retrieval paradigm (Elgohary et al., 2019).

(3) The LLM-based retrievers (RepLLaMA, E5) do not always behave much more powerfully than SLM-based ones for conversational retrieval, although they are considered with strong foundational multi-turn context understanding capacity. This might be attributed to the possible reason that they are fine-tuned solely on templated instructions, which fail to handle complex and diverse conversational information-seeking scenarios via fully leveraging the generalization capabilities of LLMs. Thus, it is still necessary and important to conduct conversational dense retrieval fine-tuning when employing LLM as a backbone model.

4.4 Results of Conversational Response Generation

Table 3 shows the results of conversational response generation on four datasets with two different settings and the comparison among existing systems. In the zero-shot scenario, our UniConv does not perform as well as the current state-of-the-art LLM. This suggests that joint fine-tuning to enhance retrieval capabilities may negatively impact direct response generation performance based on parametric knowledge, due to modifications to the model parameters. In the RAG setting, where responses are generated based on retrieved passages, we observe that our UniConv outperforms the compared systems with separate retrievers and generators. This indicates that the unified framework may better leverage intrinsic consistency and shared knowledge, mutually enhancing the performance of both retrieval and generation.

System	TopiOCQA	QReCC	OR-QuAC	INSCIT
w/o retrieval (Zero-shot)				
Mistral	26.6	24.3	17.4	23.1
Claude	27.2	25.0	17.5	27.0
ChatGPT	28.5	25.5	17.8	24.4
GRIT	27.5	25.2	17.0	23.6
Conv-GRIT	26.0	23.7	14.5	23.0
UniConv	26.7	21.2	12.6	23.8
w/. retrieval (RAG)				
Conv-ANCE + Mis.	27.2	25.9	17.0	24.8
ChatRetriever + Mis.	28.3	**26.3**	17.3	30.3
Conv-GRIT	28.8	26.0	-	-
UniConv	**29.6**	26.2	**17.8**	**33.2**

Table 3: Performance of different systems for conversational response generation. For RAG, we use Mistral-7B-chat as the generator to make the results comparable, except for the Conv-GRIT with the same workflow as our UniConv. **Bold** indicates the best results.

4.5 Ablation Studies

We conduct ablation studies for conversational retrieval and response generation to study the effects of our proposed two mechanisms, a context identification instruction (CII) mechanism to improve consistency when leveraging the retrieved information for response generation within the same model and a data discrepancy mitigation (DDM) mechanism to induce well-formatted training data with desirable supervision signals. The results are reported in Table 4 and Table 5, respectively.

Ablation	TopiOCQA	QReCC	OR-QuAC	INSCIT
Our UniConv	42.6	46.6	43.5	36.2
w/o CII	45.5	49.7	47.6	40.0
w/o DDM	41.5	45.4	41.1	35.2

Table 4: Results of ablation studies for conversational retrieval on four datasets with NDCG@3 score.

Conversational Retrieval. Table 4 shows an interesting phenomenon that incorporating the CII mechanism would hurt the retrieval performance of our UniConv, while it is helpful for the response generation as shown in Table 5. This might be because the changed input query form as $q'_n \circ p_n^+$ inevitably influences the contextualized embedding obtained via the learning objective of retrieval $\mathcal{L}_R$, leading to the possible confusion within the model due to the training is conducted simultaneously. A potential solution is to perform fine-tuning for conversational retrieval $\mathcal{L}_R$ and CII $\mathcal{L}_{\text{CII}}$ separately into a two-stage process, which can be explored in future studies. Furthermore, removing the DDM mechanism leads to performance degradation, indicating that utilizing well-structured conversational search data with distinct ground-truths for the retrieval and generation stages during joint fine-tuning can enhance previously sub-optimal results.

Conversational Response Generation. Table 5 shows that removing any mechanism leads to performance degradation for both zero-shot and RAG settings. These observations validate the effectiveness of the added components in enhancing model performance by addressing inconsistencies between retrieval and generation within the unified model. The improvements vary across datasets, suggesting that the effectiveness of the added mechanisms may depend on the structure and distribution of the data. Additionally, an obvious gap remains compared to using gold evidence for generation, indicating the potential for further improvement in better integrating retrieved information with the generation process.

4.6 Impact of Generated Response Reliability

In this section, we investigate whether the unified model can produce a more accurate and reliable response than the system with the separated models. We use the variants of UniConv without adding the $\mathcal{L}_{\text{CII}}$ term in Eq. 1 as the generator and employ ChatRetriever as the retriever for the RAG

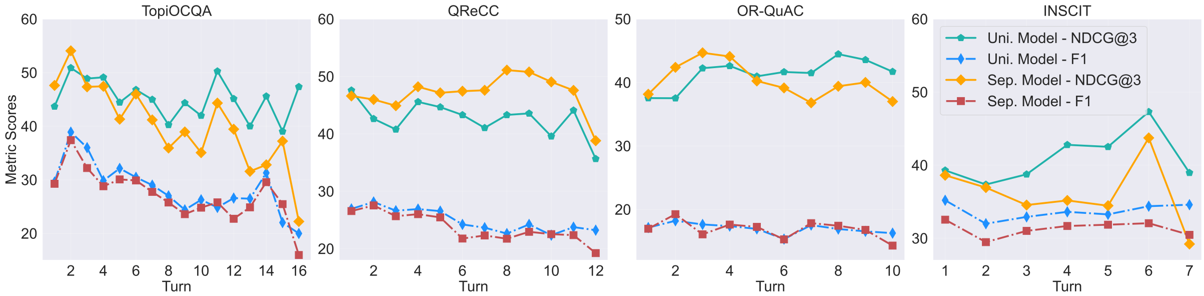

Figure 3: The performance of unified (Uni.) and separated (Sep.) models on dense retrieval (NDCG@3) and response generation (F1) with different conversation turns on four different datasets.

System	TopiOCQA	QReCC	OR-QuAC	INSCIT
w/o retrieval (Zero-shot)				
Our UniConv	26.7	21.2	12.6	23.8
w/o CII	25.2	20.6	12.4	23.0
w/o DDM	24.8	20.8	12.3	23.7
w/. retrieval (RAG)				
Our UniConv	29.6	26.2	17.8	33.2
w/o CII	29.4	26.0	17.3	31.4
w/o DDM	29.1	24.7	16.8	25.3
For Reference (Optimal retrieved results)				
w/. gold	41.1	26.9	23.3	34.6

Table 5: Results of ablation studies for conversational response generation two settings with F1 scores.

setting within the separated system while deploying the full UniConv as the unified system. We evaluate both systems on the TopiOCQA and FaithDial datasets, measuring similarity using F1 and BERT scores to assess the accuracy of the generated response r' compared to the ground-truth response r. Faithfulness is evaluated by comparing the generated responses against the evidence or rationale E provided by the datasets. Since FaithDial does not include a retrieval collection, we utilize the same database as TopiOCQA for this purpose.

The results presented in Table 6, show that the unified system consistently enhances the accuracy of generated responses across both datasets in two settings. For faithfulness, the RAG setting further improves the unified system's performance, whereas a performance drop is observed for the separated system in TopiOCQA. These observations suggest that developing the system as a unified model can improve reliability to a certain extent.

4.7 Impact of Conversational Context

We examine the impact of conversational context (multi-turn conversations) on retrieval and response generation tasks for systems with unified and separated models. The evaluation is based on per-turn performance, with the implementation for both systems consistent with the setup described in Sec. 4.6. As shown in Figure 3, the unified model consistently outperforms the separated model on both tasks as the conversation progresses, except for the retrieval task on QReCC. This observation highlights the unified model's more robust ability to understand conversations and maintain better consistency between retrieved results and its augmented generation, even in longer conversations.

4.8 Impact of History-Aware Ability

We analyze the history-aware ability of the developed model by incorporating the top-3 search results from each historical turn for the current turn's response generation, since the existing studies (Pan et al., 2024; Ye et al., 2024) demonstrate that useful information should be contained in history. To ensure a fair comparison, we use the same search results from our UniConv for both systems, varying only the generators as the previous sections. The results shown in Table 7 indicate the better performance of the unified model, which suggests the implicit de-noising capacity could be enhanced via the jointly fine-tuning. This observation also implies that more advantages are still to be discovered within the unified framework.

5 Conclusion and Future Work

In this paper, we present UniConv, a unified LLM-based model capable of handling both dense retrieval and response generation in complex conversational scenarios. We propose two mechanisms to seamlessly integrate retrieved information into response generation and address data discrepancy issues during joint fine-tuning. Experimental results on five conversational search datasets demonstrate the superior performance and enhanced reliability of UniConv. For future studies, developing a unified system for a broader range of complex con-

System	TopiOCQA				FaithDial			
	F1 (r',r)	Bert (r',r)	F1 (r', E)	Bert (r', E)	F1 (r',r)	Bert (r',r)	F1 (r', E)	Bert (r', E)
Separated	23.8 (↑ 2.9)	86.0 (↓ 0.3)	25.5 (↑ 2.6)	87.0 (↓ 1.1)	11.4 (↑ 0.7)	85.0 (↑ 10.4)	10.9 (↑ 3.9)	84.3 (↑ 9.5)
Unified	26.7 (↑ 2.9)	86.5 (↑ 0.5)	25.1 (↑ 6.9)	87.4 (↑ 0.7)	11.6 (↑ 0.9)	85.5 (↑ 9.8)	12.1 (↑ 4.2)	87.4 (↑ 7.8)

Table 6: The performance comparison on two datasets between the system with separated models for retrieval and generation and the unified ones. The evaluation is conducted between the generated response r' with the ground-truth response r and the evidence E. Arrows denote the change in the results by incorporating RAG.

System	TopiOCQA	QReCC	OR-QuAC	INSCIT
w/. historical top-3 search results				
Separated	30.3	25.3	17.2	32.0
Unified	31.1	26.6	18.3	33.5

Table 7: The response generation performance comparison by investigating the history-aware ability of different types of systems with F1 scores.

versational search scenarios is valuable, including product search, item recommendation, proactive retrieval, etc. Besides, it is important to continue improving the consistency between retrieval and generation and conduct specific training based on large-scale synthetic data.

Limitations

Despite our comprehensive studies, some potential limitations can be addressed in the future:

Efficiency. The used backbone model with 7B size LLM is larger than the previous SLM-based CDR systems, which raises efficiency concerns. Nevertheless, on the one hand, the LLM-based retriever with superior search performance reduces the requirement for extensive passage re-ranking. In real-world applications, this could help reduce the initial higher costs by ultimately decreasing the overall time required for ranking. On the other hand, the multi-task ability of UniConv makes the cost worthwhile compared with the retrieval-only systems in existing studies. This is also a promising research direction that integrates more embedding-based tasks into the instruction-based generation framework in conversation. Besides, exploring the possibility of distilling UniConv into a more efficient, smaller model is desirable.

Broader Experimental Configuration. We only leverage the fixed hyper-parameters for model setup and ratio to mix training data. Though we obtain strong performance, the exploration within broader experimental configurations could lead to better performance. Besides, adapting our methods to other types or sizes of backbone models and using full-model fine-tuning rather than LoRA might bring additional observations and results.

Robust Evaluation for Generation. How to evaluate the generated content is still an open question for the research community. Our evaluation is conducted on a single metric following the previous studies (Mao et al., 2024a; Liu et al., 2024) for a fair comparison, which might not reflect the quality of different aspects of the generated response. Leveraging more comprehensive evaluation metrics or incorporating another LLM as an evaluator might help us to observe more insights about improving the consistency between retrieval and generation.

References

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

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

AnthropicAI. 2023. Introducing claude.

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

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, et al. 2016. Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*.

Yiruo Cheng, Kelong Mao, Ziliang Zhao, Guanting Dong, Hongjin Qian, Yongkang Wu, Tetsuya

Sakai, Ji-Rong Wen, and Zhicheng Dou. 2024. Coral: Benchmarking multi-turn conversational retrieval-augmentation generation. *arXiv preprint arXiv:2410.23090*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

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

Nouha Dziri, Ehsan Kamalloo, Sivan Milton, Osmar Zaiane, Mo Yu, Edoardo M Ponti, and Siva Reddy. 2022. Faithdial: A faithful benchmark for information-seeking dialogue. *Transactions of the Association for Computational Linguistics*, 10:1473–1490.

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

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

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

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

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

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

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

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

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781.

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

Yilong Lai, Jialong Wu, Congzhi Zhang, Haowen Sun, and Deyu Zhou. 2024. Adacqr: Enhancing query reformulation for conversational search via sparse and dense retrieval alignment. In *COLING*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Ming Li, Pei Chen, Chenguang Wang, Hongyu Zhao, Yijun Liang, Yupeng Hou, Fuxiao Liu, and Tianyi Zhou. 2024a. Mosaic-it: Free compositional data augmentation improves instruction tuning. *arXiv preprint arXiv:2405.13326*.

Xiaoxi Li, Yujia Zhou, and Zhicheng Dou. 2024b. Unigen: A unified generative framework for retrieval and question answering with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 8688–8696.

Sheng-Chieh Lin, Jheng-Hong Yang, and Jimmy Lin. 2021. Contextualized query embeddings for conversational search. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1004–1015.

Zihan Liu, Wei Ping, Rajarshi Roy, Peng Xu, Mohammad Shoeybi, and Bryan Catanzaro. 2024. Chatqa: Building gpt-4 level conversational qa models. *arXiv preprint arXiv:2401.10225*.

Simon Lupart, Mohammad Aliannejadi, and Evangelos Kanoulas. 2025. Disco meets llms: A unified approach for sparse retrieval and contextual distillation in conversational search. In *SIGIR*.

Xueguang Ma, Liang Wang, Nan Yang, Furu Wei, and Jimmy Lin. 2024. Fine-tuning llama for multi-stage text retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2421–2425.

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

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

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

Kelong Mao, Zhicheng Dou, Hongjin Qian, Fengran Mo, Xiaohua Cheng, and Zhao Cao. 2022. Convtrans: Transforming web search sessions for conversational dense retrieval. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2935–2946.

Kelong Mao, Zheng Liu, Hongjin Qian, Fengran Mo, Chenlong Deng, and Zhicheng Dou. 2024b. Ragstudio: Towards in-domain adaptation of retrieval augmented generation through self-alignment. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 725–735.

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*, pages 3193–3202.

Chuan Meng, Pengjie Ren, Zhumin Chen, Christof Monz, Jun Ma, and Maarten de Rijke. 2020a. RefNet: A reference-aware network for background based conversation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8496–8503.

Chuan Meng, Pengjie Ren, Zhumin Chen, Zhaochun Ren, Tengxiao Xi, and Maarten de Rijke. 2021. Initiative-aware self-supervised learning for knowledge-grounded conversations. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 522–532.

Chuan Meng, Pengjie Ren, Zhumin Chen, Weiwei Sun, Zhaochun Ren, Zhaopeng Tu, and Maarten de Rijke. 2020b. DukeNet: A dual knowledge interaction network for knowledge-grounded conversation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1151–1160.

Chuan Meng, Francesco Tonolini, Fengran Mo, Nikolaos Aletras, Emine Yilmaz, and Gabriella Kazai. 2025. Bridging the gap: From ad-hoc to proactive search in conversations. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

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

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

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

Fengran Mo, Chuan Meng, Mohammad Aliannejadi, and Jian-Yun Nie. 2025. Conversational search: From fundamentals to frontiers in the llm era. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Fengran Mo, Jian-Yun Nie, Kaiyu Huang, Kelong Mao, Yutao Zhu, Peng Li, and Yang Liu. 2023b. Learning to relate to previous turns in conversational search. In *29th ACM SIGKDD Conference On Knowledge Discover and Data Mining (SIGKDD)*.

Fengran Mo, Chen Qu, Kelong Mao, Yihong Wu, Zhan Su, Kaiyu Huang, and Jian-Yun Nie. 2024c. Aligning query representation with rewritten query and relevance judgments in conversational search. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 1700–1710.

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

Fengran Mo, Bole Yi, Kelong Mao, Chen Qu, Kaiyu Huang, and Jian-Yun Nie. 2024e. Convsdg: Session data generation for conversational search. In *Companion Proceedings of the ACM on Web Conference 2024*, pages 1634–1642.

Fengran Mo, Longxiang Zhao, Kaiyu Huang, Yue Dong, Degen Huang, and Jian-Yun Nie. 2024f. How to leverage personal textual knowledge for personalized conversational information retrieval. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 3954–3958.

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

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

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

Zhenyu Pan, Haozheng Luo, Manling Li, and Han Liu. 2024. Conv-coa: Improving open-domain question answering in large language models via conversational chain-of-action. *arXiv preprint arXiv:2405.17822*.

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

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

Pengjie Ren, Zhumin Chen, Zhaochun Ren, Evangelos Kanoulas, Christof Monz, and Maarten De Rijke. 2021. Conversations with search engines: Serp-based conversational response generation. *ACM Transactions on Information Systems (TOIS)*, 39(4):1–29.

Alireza Salemi and Hamed Zamani. 2024. Evaluating retrieval quality in retrieval-augmented generation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2395–2400.

I Sutskever. 2014. Sequence to sequence learning with neural networks. *arXiv preprint arXiv:1409.3215*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

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

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

Zeqiu Wu, Yi Luan, Hannah Rashkin, David Reitter, 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)*.

Zeqiu Wu, Ryu Parish, Hao Cheng, Sewon Min, Prithviraj Ammanabrolu, Mari Ostendorf, and Hannaneh Hajishirzi. 2023. Inscit: Information-seeking conversations with mixed-initiative interactions. *Transactions of the Association for Computational Linguistics*, 11:453–468.

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

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

Linhao Ye, Zhikai Lei, Jianghao Yin, Qin Chen, Jie Zhou, and Liang He. 2024. Boosting conversational question answering with fine-grained retrieval-augmentation and self-check. *arXiv preprint arXiv:2403.18243*.

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

Shi Yu, Zhenghao Liu, Chenyan Xiong, Tao Feng, and Zhiyuan Liu. 2021. Few-shot conversational dense retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 829–838.

Yue Yu, Wei Ping, Zihan Liu, Boxin Wang, Jiaxuan You, Chao Zhang, Mohammad Shoeybi, and Bryan Catanzaro. 2024. Rankrag: Unifying context ranking with retrieval-augmented generation in llms. *arXiv preprint arXiv:2407.02485*.

Hamed Zamani, Johanne R Trippas, Jeff Dalton, Filip Radlinski, et al. 2023. Conversational information seeking. *Foundations and Trends® in Information Retrieval*, 17(3-4):244–456.

Jinghan Zhang, Xiting Wang, Fengran Mo, Yeyang Zhou, Wanfu Gao, and Kunpeng Liu. 2025. Entropy-based exploration conduction for multi-step reasoning. *arXiv preprint arXiv:2503.15848*.

Jintian Zhang, Cheng Peng, Mengshu Sun, Xiang Chen, Lei Liang, Zhiqiang Zhang, Jun Zhou, Huajun Chen, and Ningyu Zhang. 2024. Onegen: Efficient one-pass unified generation and retrieval for llms. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4088–4119.

Appendix

A Experimental Setup

A.1 Datasets Details

	TopiOCQA	QReCC	OR-QuAC	INSCIT	FaithDial
#Conv.	205	2,775	771	469	791
#Turns(Qry)	2,514	16,451	5,571	2,767	3,539
#Collection	25M	54M	11M	49M	-
#Avg. Qry	12.9	5.3	7.2	5.9	4.5
#Min Qry	5	2	4	2	4
#Max Qry	25	12	12	7	5
#Avg. Psg	9.0	1.6	1.0	1.6	-

Table 8: Statistics of five used datasets.

The statistics of each dataset are presented in Table 8. The first four datasets are used for the retrieval and response generation evaluation while the FaithDial does not provide the collection for retrieval so it is used for reliability evaluation only.

A.2 Baseline Details

We provide a more detailed introduction to the following baselines used for comparison:

LLM-Aided (Ye et al., 2023): An informative conversational query rewriting by directly prompting ChatGPT-3.5 as both query rewriters and rewrite editors twice to incorporate all the desirable properties for producing the final rewritten queries.

LLM4CS (Mao et al., 2023a): A state-of-the-art LLM-based prompting method for conversational query rewriting. We implement it with full aggregation by calling LLMs five times for query and response generation but without the chain-of-thought (CoT) content because of the efficient annotation consideration in practical scenarios.

RETPO (Yoon et al., 2024): A retriever preference adapted query rewriting method that fine-tunes LLaMA-2-7B-Chat as a query rewrite model with an external query rewrite dataset generated by GPT-4.

CHIQ (Mo et al., 2024a): A state-of-the-art method leverages the basic NLP capabilities of LLMs to enhance the quality of contextual history for improving the query rewriting performance.

ConvDR (Yu et al., 2021): A conversational dense retrieval method that uses knowledge distillation to learn the session embeddings with relevance judgments from the human-rewritten queries based on the ANCE model.

Conv-ANCE (Mao et al., 2023c): A conversational dense retrieval method that leverages ANCE fine-tuned on conversational search data only using the retrieval loss term in Eq. 1.

QRACDR (Mo et al., 2024c): A state-of-the-art SLM-based query representation alignment conversational dense retrieval method by incorporating relevance judgments and rewritten query annotation as supervision signals for retriever fine-tuning.

RepLLaMA (Ma et al., 2024): A large ad-hoc dense retriever fine-tuned on top of the LLaMA-7B-Chat model on the MSMARCO dataset.

E5 (Wang et al., 2024): A large ad-hoc retriever fine-tuned on top of Mistral-7B model on the combination of synthetic dataset generated by ChatGPT-3.5 and MSMARCO.

CharRetriever (Mao et al., 2024a): A state-of-the-art LLM-based conversational dense retriever with better robustness via a novel contrastive session-masked instruction tuning approach.

GRIT (Muennighoff et al., 2024): A first proposed unified model to handle both retrieval and generation tasks by incorporating vanilla instruction tuning and using different training data for its contrastive learning and instruction tuning.

Conv-GRIT (Muennighoff et al., 2024): A variant of GRIT fine-tuned on the conversational data with the same setting as our UniConv model for fair comparisons.