Leveraging Historical Turns for Retrieval-Augmented Response Generation in Conversational Search

Anonymous ACL submission

Abstract

Conversational search enables the users to interact with the systems by multi-turns to address their complex information needs, which consist of two key components: retrieval and generation. Although retrieval has achieved significant improvement recently by understanding context-dependent queries, response generation has not been well studied. The existing methods only adapt the single-turn retrieval-augmented generation (RAG) pipeline, which overlooks the historical information (e.g., historical search results) as the conversation dives in. In this paper, we first define conversational RAG scenarios and verify the feasibility of leveraging historical turns for current turn RAG, e.g., the historical search results and the turn dependency. Then, we investigate various strategies toward a better practice for conversational RAG on three public benchmarks and demonstrate the effectiveness of integrating abundant information in historical turns. We also analyze the potential principle behind our observations, aiming to understand when and why historical information can contribute to the conversational RAG, which could facilitate the build-up of modern conversational search systems.

1 Introduction

Conversational search enables users to interact with the systems through multiple turns to address their complex information needs with two key components: retrieval and generation (Gao et al., 2022; Zamani et al., 2023; Zhu et al., 2023). The retriever first identifies the relevant passages from external resources, and then the generator further crafts an exact response based on the search results. Although existing studies on conversational retrieval have achieved significant improvements by leveraging the abundant historical information (Lin et al., 2021; Yu et al., 2021; Mo et al., 2024c), how to conduct response generation within conversational scenarios is not well-studied in the literature.

With the development of large language models (LLMs) (Ouyang et al., 2022; Chen et al., 2024b), the conversational mode becomes a common practice to generate desirable content for the users. However, most existing methods (Dinan et al., 2018; Fang et al., 2022) simply adapt the single-turn retrieval-augmented generation (RAG) (Lewis et al., 2020) pipeline even under the conversational scenario, which first reformulate the context-dependent query of current turn and leverage its associated retrieved results as the input for the response generation. Such a paradigm overlooks the information from historical turns and might result in sub-optimal performance.

Different from the single-turn scenario, the conversational interface could produce more abundant information, e.g., the historical query-response pairs, their associated retrieved results, the implicit turn dependency, etc, which might be useful for the response generation of the current turn as illustrated in Figure 1. The assumptions behind are i) some top-rank historical retrieved passages might be highly relevant to the current query (Mo et al., 2024c) serving similarly as the pseudo relevant documents in pseudo relevance feedback (Xu and Croft, 1996), but are not retrieved or top-ranked for the current turn due to the limited performance of conversational retrieval (Kim and Kim, 2022); and ii) the context-dependent query in conversation is usually ambiguous and complex even after reformulation, which requires de-noising the retrieved context by enhancement (e.g., with similar aspects contained in historical search results) (Chan et al., 2024) or diversification (e.g., with multi-aspect contained in history) (Wang et al., 2024).

However, leveraging the information from historical turns is non-trivial, due to the difficulty of modeling the lengthy and long-tailed conversation and the efficiency requirements for content generation. Besides, incorporating all historical information with respect to the current turn is infeasible since

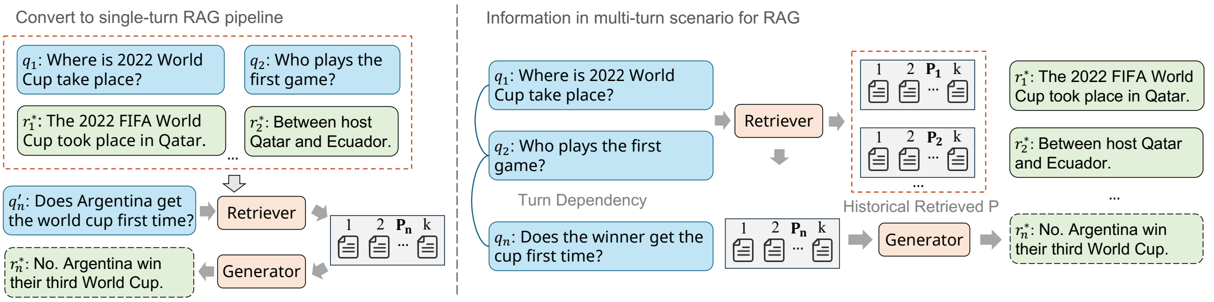

Figure 1: Illustration of the available historical information, e.g., turn dependency and historical retrieved results, within multi-turn scenarios (right), which are overlooked in existing single turn RAG pipeline (left).

it would surpass the constraints of model context input and computing resources or raise the risk of injecting additional noise. Even with the advance of long-context LLM, to exploit accurate historical information for LLM-based generation is still necessary for better performance, which is proved in existing literature (Yu et al., 2024; Xu et al., 2024) and our studies in Sec. 3.2 Thus, it is critical to investigate how the information from each historical turn can be leveraged to improve response generation performance for the current turn query in an effective and efficient way.

In this paper, we design various approaches to leverage information from historical turns for response generation in conversational search. We address the following research questions:
RQ1: How feasible to leverage historical turns to improve the generated response of the current turn?
RQ2: What is the better practice to leverage historical turns for the current turn's response generation?
RQ3: Why the information from historical turns can contribute to the generation of the current turn?

To address these inquiries, we first verify the effectiveness of leveraging oracle retrieved evidence from historical turns and the preliminary results of practical scenarios to motivate our approaches and define the task of conversational RAG. Then, we investigate different strategies to improve the task performance from different aspects, including integrating the search results, capturing and leveraging the turn dependency, and identifying the historical evidence. The principle is to decide how to leverage the information from historical turns for the current turn's generation based on the LLMs. We conduct the experiments across three conversational search benchmarks and further analyze the observation of using various historical information.

Our contributions are summarized as follows: (1) We verify the feasibility of leveraging historical turns, not limited to the query-response pairs, to improve the performance of the generated response of the current turn and define the new task/scenario of conversational RAG. (2) We investigate different strategies to leverage historical information for better performance of conversational RAG across three benchmarks. The experimental results demonstrate the effectiveness of our solutions by outperforming a set of baselines. (3) We analyze the potential principle behind our observations, aiming to understand how historical information can contribute to the conversational RAG, which could facilitate the modern conversational search system.

2 Related Work

Conversational search provides an interaction interface so that users can elaborate more complex search requirements with multi-turns, where retrieval and generation are two key components.

Conversational Retrieval. Two main methods have been developed to achieve conversational retrieval: conversational query rewriting and conversational dense retrieval. Conversational query rewriting (Voskarides et al., 2020; Wu et al., 2022; Qian and Dou, 2022; Mao et al., 2023b; Mo et al., 2023a; Ye et al., 2023; Jang et al., 2023; Mo et al., 2024a,e) aims to convert context-dependent queries into stand-alone ones, then any off-the-shelf retrievers can be applied. Another method, conversational dense retrieval (Lin et al., 2021; Kim and Kim, 2022; Mao et al., 2022b, 2023c; Jin et al., 2023; Mao et al., 2024; Mo et al., 2024b,d; Chen et al., 2024a; Cheng et al., 2024) learns to encode the user's real search intent and candidate documents into latent representations and performs dense retrieval, which leverages conversation-document relevance as supervision signals.

Conversational Response Generation. In the context of conversational search, the most related work for response generation is retrieval-augmented generation (RAG) (Lewis et al., 2020), which aims to craft the response for a single-turn query with retrieved external knowledge. The literature focuses on improving the response accuracy (Jiang et al., 2023; Zamani and Bendersky, 2024), detecting and reducing the hallucination (Shuster et al., 2021; Su et al., 2024), arousing the reasoning ability of LLMs (Asai et al., 2023b; Kaddour et al., 2023), fine-tuning the LLMs with external knowledge (Lin et al., 2023), etc, but not explores leveraging the abundant conversational information.

Some earlier studies (Zheng et al., 2020; Meng et al., 2020; Ren et al., 2021) have attempted to select the knowledge provided in previous dialogue content to improve the generation for current turn, but the candidate knowledge is provided in a limited pool rather than an open-retrieval setting (Qu et al., 2020), which is impractical. Although some recent studies (Pan et al., 2024; Ye et al., 2024; Roy et al., 2024) attempt to investigate the response generation within the conversational scenario, they still simply adapt the single-turn RAG pipeline by only leveraging the reformulated query of the current turn and its associated retrieved ranking list for response generation, which overlooks the historical search results and conversational turn dependency after the query rewriting. Different from them, our studies aim to figure out the principle of leveraging more historical information to improve current turn response generation, which is defined as following.

3 Motivation

3.1 Task Definition of Conversational RAG

A conversational session contains the current query q_n, and $n-1$ historical turns $\mathcal{H}_n = \{q_i, \mathcal{P}_i, r_i\}_{i=0}^{n-1}$, where $\{\mathcal{P}_i\}_{i=0}^{n-1}$ and r_i denote the search results and generated response of each preceding turns, respectively. The conversational search system require to retrieve top-k relevant passages $\mathcal{P}_n = \{p_n^t\}_{t=1}^{k}$ from a large collection $\mathcal{C}$, then apply a mechanism $\mathcal{M}$ to manipulate the input for the generator model as $\mathcal{G}_n^{\text{IN}} = \mathcal{M}(q_n, \mathcal{P}_n, \mathcal{H}_n)$. We expect the performance of the final generated response could be better by leveraging $\{\mathcal{P}_i\}_{i=0}^{n-1}$ compared with using $\mathcal{P}_n$ only, while the challenge lies in what and how to use the information in the historical turns $\mathcal{H}_n$.

3.2 Effectiveness of Historical Search Results

We first address the **RQ1** under oracle and practical settings, by comparing the quality of the generated response between using the search results from the current turn only and incorporating with the ones from historical turns based on the used LLM. Besides, we further analyze a mixed setting to evaluate the context de-noise ability. The verified experiments are conducted on three conversational search datasets, with the setup described in Sec. 5.1.

Used History t	TopiOCQA		QReCC		OR-QuAC	
Turn's evidence	F1	EM	F1	EM	F1	EM
t = 0	40.0	6.7	29.1	7.5	18.4	9.5
t = 1	42.4	7.7	30.0	9.1	19.3	10.4
t = 3	44.7	9.8	31.7	8.9	19.8	11.0
t = all	**45.3**	**10.6**	**32.1**	**9.7**	**20.2**	**12.3**

Table 1: Performance of response generation by using different historical turns' evidence in **oracle setting**.

Used top-k	TopiOCQA		QReCC		OR-QuAC	
retrieved evidence	F1	EM	F1	EM	F1	EM
cur. top-3	22.8	4.2	23.3	3.2	12.1	6.6
+ his. top-3	**32.7**	**6.9**	**27.5**	**5.1**	**14.4**	**8.1**
cur. top-10	23.3	4.5	23.7	4.1	13.6	7.3
cur. top-20	24.4	4.1	23.6	3.4	14.0	7.9

Table 2: Performance of response generation by using current turn's retrieved results with different top-k and incorporating with historical ones in **practical setting**.

Used top-k	TopiOCQA		QReCC		OR-QuAC	
retrieved evidence	F1	EM	F1	EM	F1	EM
top-5 w/. gold	29.1	4.6	28.5	9.1	17.1	9.6
top-10 w/. gold	27.6	4.4	27.3	8.5	16.8	8.7
top-15 w/. gold	27.1	3.9	26.9	8.2	16.4	8.2
top-20 w/. gold	26.2	3.8	26.4	8.1	16.2	7.9

Table 3: Performance of response generation by using different top-k of current turn and always replacing the top-1 as the gold evidence in **mixed setting**.

Oracle Setting. The oracle setting assumes the gold evidence would be retrieved by an oracle retriever for each query turn. From Table 1, we can observe that incorporating historical turn's gold evidence can improve the quality of the generated response, and using evidence from more previous turns can consistently improve it.

Practical Setting. For a practical setting, only the retrieved top-k ranking list is available rather than the gold ones. As present in Table 2, incorporating historical retrieved evidence (+ his. top-3) achieves better performance than the single-turn RAG (only cur. top-k). Although using a larger k for the current turns' retrieved ranking list might slightly improve the generation performance, it still underperforms the one with historical retrieved evidence and increases the latency cost.

Mixed Setting. For a mixed setting, we combine the previous two settings by replacing the top-1 rank position with the gold evidence and keeping the remaining retrieved evidence in the ranking list.

Symbol	Functionality
$\mathcal{I}^{\text{SRI}}$	Rank $\mathcal{P}_n \cup \{\mathcal{P}_i\}_{i=1}^{n-1}$ for query turn q_n
$\mathcal{I}^{\text{Judge}}$	Judge information needs between q_n and $\{q_i\}_{i=1}^{n-1}$
$\mathcal{I}^{\text{HEI}}$	Generate r_n with grounded passages $\mathcal{P}_n^{\text{index}}$

Table 4: Illustration of the instruction for each strategy.

The goal is to evaluate the context de-noise ability of the generator model, which can be considered as long-context LLM evaluation since the supported context length of the employed LLM is longer than sum up input information.

Table 3 shows that although the gold evidence is ranked in the top position, the generation performance keeps dropping as the k becomes larger. The results indicate that the longer input context might be challenging for the generator to deal with even though the gold evidence is included. This observation is the same as the previous studies (Xu et al., 2023; Jiang et al., 2024) that the existing LLMs supported with longer input context windows might not always deal with long context de-noise. Thus, filtering the noise within the input context before generation is necessary, especially under conversational scenarios with abundant available context.

The high impact of leveraging historical information for current turn response generation in the above observations confirms our conjecture for **RQ1**. Then, the problem is how to use the historical information (**RQ2**) and understand why they can contribute to the conversational RAG (**RQ3**).

4 Methodology

Although the historical retrieved passages might contain useful information for current turn generation, it is infeasible to incorporate them from all previous turns as the conversation dives in, due to the efficiency concerns and the de-noising requirement within the context $\mathcal{H}_n$ from the historical turns for better generation performance. Thus, it is crucial to conduct conversational context modeling to deal with the complex requirements of information identification. To alleviate the above risks, we design three strategies based on the powerful capacity of LLMs to conduct response generation in the context of conversational scenarios from different aspects, including integrating the search results (Sec. 4.1), capturing and leveraging the dependency of historical turns (Sec. 4.2), and identifying the evidence from historical context (Sec. 4.3). The instruction and corresponding function for each strategy are summarized in Table 4. We aim to explore a better practice of leveraging historical turns for conversational RAG with LLMs and test their ability to deal with complex historical information.

4.1 Search Results Integration

Some useful information might be contained in top-ranked historical retrieved passages but fail to be retrieved or not top-ranked for the current turn. Thus, the search results from historical turns could be supplementary to the retrieved set used for the response generation of the current turn. Formally, given the retrieved passages from the current (n-th) turn $\mathcal{P}_n = \mathcal{R}(q_n, \mathcal{C})$ and each its associated historical turn $\{\mathcal{P}_i\}_{i=1}^{n-1} = \mathcal{R}(\{q_i\}_{i=1}^{n-1}, \mathcal{C})$ by a retriever $\mathcal{R}$, the designed mechanism $\mathcal{M}^{\text{SRI}}$ acts as a re-ranker to obtain the final input retrieved set by reordering the passages from $\mathcal{P}_n$ and $\{\mathcal{P}_i\}_{i=1}^{n-1}$. Concretely, the candidate passages $\mathcal{P}_n^{\text{new}}$ with new order is obtained according to the pair-wise preference $\mathcal{S}$ determined by instructing the LLMs with $\mathcal{I}^{\text{SRI}}$, where $\mathcal{P}_n^{\text{new}} \subset \mathcal{G}_n^{\text{IN}}$ as following:

$$\mathcal{P}_n^{\text{new}} = \mathcal{M}^{\text{SRI}}\left(q_n, \mathcal{P}_n, \{(q_i, \mathcal{P}_i)\}_{i=0}^{n-1}, \mathcal{I}^{\text{SRI}}\right)$$

$$\mathcal{I}^{\text{SRI}}(q_n, \{\mathcal{P}_u\}) = \begin{cases} \mathcal{S}\left(q_n, p_u^t\right), & \text{if } u = n \\ \mathcal{S}\left(q_n \oplus q_u, p_u^t\right), & \text{if } u \neq n \end{cases}$$

where $u \in [0, n]$ denotes the candidate passage p_u^t is from the top-k ranking list of u-th turn and $\{\mathcal{P}_u\} \subseteq \mathcal{P}_n \cup \{\mathcal{P}_i\}_{i=1}^{n-1}$. Different from the traditional re-ranking, we not only consider the current query q_n but also append the historical query q_u when the candidate passage p_u^t is from previous turns. The principle we use is to enhance the conversational modeling by providing explicit semantics of historical turns and avoiding topic drift when applying $\mathcal{S}$ to the candidate query-passage pairs.

4.2 Historical Turns Dependency

The previous studies (Mao et al., 2022a; Mo et al., 2023b) demonstrate that simply leveraging the information from all historical turns might inject noise to the model and adapting single-turn RAG pipeline without further conversational context modeling is sub-optimal according to our observation in Table 2. Thus, we design a strategy to first capture the dependency of historical turns and then leverage them for response generation after the retrieval procedure. This strategy can also help to reduce the latency cost of the system by maintaining only the relevant queries based on the captured turn dependency. Concretely, we instruct the LLMs with $\mathcal{I}^{\text{Judge}}$ to interactively determine whether a given query pair shares similar information needs between the current (n-th) turn with each associated historical turn and eventually obtain a binary judgment list J_n as Eq. 1, where $|J_n| = n - 1$.

$$J_n \leftarrow \mathcal{I}^{\text{Judge}}(q_n, q_i), \quad i \in [0, n-1] \qquad (1)$$

Then, we explicitly select which historical turns would be used for response generation based on the turn dependency judgment results from J with two different strategies as shown below.

$$\text{Dep}^{\text{Hard}}(\mathcal{H}_n) = \{\mathcal{H}_i^J\}, \quad J_n[i] = 1$$
$$\text{Dep}^{\text{Soft}}(\mathcal{H}_n) = \mathcal{H}_i^J \oplus \cdots \mathcal{H}_{n-1}^J, \min\{i|J_n[i] = 1\}$$

The Dep-Hard strategy retains only the turn with similar information needs, while Dep-Soft starts retention from the first turn judged as with similar information needs. The Dep-Soft allows more flexible turn dependency by maintaining transitions across consecutive turns. By applying the specific selection mechanism $\mathcal{M}^{\text{Dep}}$ on the historical information $\mathcal{H}_n$ to produce the input context based on the judgment list J_n as Eq. 2, we expect to reduce the noise contained in the input of generator $\mathcal{G}_n^{\text{IN}}$ and improve the efficiency for the response generation in inference.

$$\begin{aligned} \mathcal{G}_n^{\text{IN}} &= \mathcal{M}^{\text{Dep}}(q_n, \text{Dep}^{\text{Hard/Soft}}(J_n, \mathcal{H}_n)) \\ &= q_n \oplus \cdots \mathcal{H}_i^J \oplus \cdots \end{aligned} \qquad (2)$$

4.3 Historical Evidence Identification

Evidence identification is widely used in single-turn RAG systems (Gao et al., 2023) aiming to increase the credibility of the generation results and arouse the self-check ability of LLMs to obtain a more accurate answer (Asai et al., 2023b). Inheriting a similar idea, we design the historical evidence identification (HEI) strategy to not only provide the answer but also indicate the referenced historical retrieved passages. Specifically, by instructing the LLMs with $\mathcal{I}^{\text{HEI}}$, the output of current (n-th) turn should contain two parts of information, the response r_n for the query q_n and the grounded passages $\mathcal{P}_n^{\text{index}}$ to r_n as Eq. 3.

$$\begin{aligned} r_n &= \mathcal{M}^{\text{HEI}}\left(\mathcal{I}^{\text{HEI}}\left(q_n, \mathcal{P}_n, \mathcal{H}_n\right) | \mathcal{P}_n^{\text{index}}\right), \\ &\text{where} \quad \mathcal{P}_n^{\text{index}} \subseteq \mathcal{P}_n \cup \{\mathcal{P}_i\}_{i=1}^{n-1} \end{aligned} \qquad (3)$$

This strategy also helps us to re-examine the implicit conversational modeling ability of LLMs by explicitly obtaining their attention on historical information. In other words, we can test whether the LLMs enable to identify the useful information from each historical turn by comparing $\mathcal{P}_n^{\text{index}}$ and J_n from Eq. 1 for each turn, rather than just modeling on the historical query-response pairs.

5 Experiments

The experiments are designed to answer the research questions **RQ2** and **RQ3** by evaluating our three strategies proposed in Sec. 4 and analyzing the experimental observations, respectively.

5.1 Experimental Setup

Datasets and Evaluation Metrics. We evaluate our methods on three widely-used conversational search datasets, including TopiOCQA (Adlakha et al., 2022), QReCC (Anantha et al., 2021), and OR-QuAC (Qu et al., 2020). Each of them contains the ground-truth for both passage retrieval and response generation. The statistics and more details of the datasets are provided in Appendix B.1. The evaluation metrics contain two parts with respect to the retrieval and the generation. For conversational retrieval, we deploy the NDCG@3, Recall@3, and Precision@3 for top-ranked results (Dalton et al., 2020). For evaluating response generation, we employ F1, Exact Match (EM), Accuracy (Acc.), and LLM-EM following the previous studies (Jeong et al., 2023; Asai et al., 2023b; Pan et al., 2024).

Baselines. We compare our methods with a variety of prompt-based systems that can mainly be classified into three categories. More precisely, the first group is LLM-based reasoning methods, including Zero-Shot (ZS) (AnthropicAI, 2023), Chain-of-Thought (CoT) (Wei et al., 2022), and Tree-of-Thought (ToT) (Yao et al., 2024). The second group integrates with retrieval-augmented generation (RAG) component on top of the previous systems, including vanilla RAG (Asai et al., 2023a), Self-Ask (SA) (Press et al., 2023), Reasoning and Action (ReAct) (Yao et al., 2023), and Demonstrate-Search-Predict (DSP) (Khattab et al., 2022). The third group leverages historical information to conduct response generation for the current turn, which is our target defined as conversational RAG. Among them, the Conversational Chain-of-Action (CCoA) (Pan et al., 2024) relies on a dynamic reasoning-retrieval mechanism that extracts the intent of the question and decomposes it into a reasoning chain with an updated contextual knowledge set. The Our-Base and Oracle serve as our baseline and the ideal situation of our proposed methods corresponding to the practical and oracle settings in Sec. 3.2, respectively. **Note** that the Our-Base can be considered as a long-context LLM method, which directly takes all the information as model input without specific identification/de-noising.

Dataset	Reasoning-based			with RAG				Conversational RAG					
	ZS	CoT	ToT	RAG	SA	React	DSP	CCoA	Our-Base	Our-SRI	Our-Dep	Our-HEI	Oracle
TopiOCQA	60.1	63.9	58.7	64.5	66.2	66.8	64.9	68.4	72.1	75.2	75.0	**76.3**	89.1
QReCC	52.8	54.5	50.6	55.2	54.9	56.8	58.3	60.7	63.3	64.8	63.9	**66.5**	73.2
OR-QuAC	38.5	42.8	37.7	43.9	44.6	45.3	46.8	49.7	50.0	51.7	50.4	**53.7**	64.1

Table 5: The performance comparison among different systems, including reasoning-based methods (with RAG) and conversational RAG-based ones. The reported scores are exact match via instructing LLM (LLM-EM) following (Pan et al., 2024). **Bold** and underline indicate the best and the second-best results except the `Oracle`.

Method	TopiOCQA					QReCC					OR-QuAC				
	N@3	R@3	P@3	F1	EM	N@3	R@3	P@3	F1	EM	N@3	R@3	P@3	F1	EM
Vanilla RAG	30.5	37.9	12.6	24.4	4.2	40.5	47.7	17.6	23.6	3.4	44.7	51.6	19.4	14.0	7.9
Our-Base				32.7	6.9				24.7	5.1				14.4	8.0
+ SRI-T5	33.4	40.5	13.5	32.9	7.2	48.6	56.2	20.7	27.3	7.9	50.0	**57.5**	**21.4**	16.7	10.1
+ SRI-LLM	**34.9**	**41.0**	**13.7**	**33.7**	**8.3**	**54.9**	**60.9**	**22.3**	**27.8**	**8.2**	**52.5**	57.1	20.5	17.1	**10.4**

Table 6: The performance of retrieval and response generation across three conversational search datasets. The results of retrieval are based on the ranking list of current turn, while the `SRI` strategies integrate the ranking list of historical turns with the current turn. Thus, the generation of `Our-Base` leverages the ranking list of both historical and current turns, while the others are based on the current turn only. **Bold** indicates the best result.

Implementation Details. We deploy ANCE retriever for searching relevant passages (Xiong et al., 2020) and Claude-Sonnet (AnthropicAI, 2023) as the generator to obtain the final response. The results of using the other LLMs are provided in Appendix C.3. For each of our proposed strategies to manipulate the input for generation, we still use Claude-Sonnet as the backbone model and investigate the other LLMs in ablation studies. For search results integration, we also implement monoT5 (Nogueira et al., 2020) for comparison. To make the baseline methods directly comparable, we follow the evaluation protocol from (Pan et al., 2024) to implement the compared systems with the same instruction by using our deployed LLM. More details can be found in Appedix B.2.

5.2 Results Comparison

Table 5 shows the comparison of the results between our investigated methods and previous studies on three conversational datasets in terms of response generation. First, we observe that leveraging the historical information within conversational RAG scenarios consistently outperforms both the single-turn RAG-based and the reasoning-based prompt systems. Among the conversational RAG systems, our `Base` system reports an absolute gain of 3.7%, 2.6%, and 0.3% over the previous best system `ConvCoA` on each dataset and our proposed three strategies `SRI`, `ST`, and `HEI` significantly improves the performance on top of it. The strong effectiveness is attributed to the sophisticated strategies to leverage historical turns. The impact of each strategy is provided in the following sections.

Besides, we notice that although leveraging the reasoning ability (e.g., `CoT`) and further adapting the external retrieved knowledge (e.g., `SA`, `React`, and `DSP`) can improve the performance of response generation, ignoring historical information within multi-turn scenario limits the better results, which suggests that using suitable historical search results is necessary to contribute to the current turn response generation. We also find that there is still improvement room compared with oracle results, indicating a better strategy to leverage historical information is desirable, e.g., integrating different strategies, where we leave for future exploration.

5.3 Impact of Search Results Integration

Table 6 presents the retrieval and response generation performance on three conversational search datasets. We can see that the `Our-Base` method using historical ranking list generally outperforms the `Vanilla RAG` which only uses the current turn's search result, demonstrating the necessity of leveraging historical search results. Besides, applying our search results integration (`SRI`) strategy can further improve the retrieval performance and generation quality. We find that using LLM for `SRI` performs better than deploying monoT5, implying better results could be produced by higher model capacity. The performance improvement by `SRI` strategy indicates the usefulness of the search results from historical turns for response generation.

Although the effectiveness of retrieval for RAG is still an open question in the community (Salemi and Zamani, 2024), our observations suggest the correlation in terms of effectiveness between retrieval and generation, i.e., improving the performance of retrieval can help promote the quality of response generation although it could be slightly in some cases. Thus, we leave the alignment between these two components, i.e., improving generation to match that of retrieval, in future work. The results based on the other evaluation metric can be found in Appendix C.1.

5.4 Impact of Historical Turns Dependency

Effectiveness and Efficiency. Table 7 shows the results of leveraging the turn dependency judgment information in terms of effectiveness and efficiency. We can observe that both strategies can reduce the latency cost, i.e., lower average input token (Avg. $|T|$) per turn for the generator model. The `Dep-soft` that maintains more implicit turn dependency can obtain better effectiveness and even outperform `Our-Base` by eliminating the noise, while the `Dep-hard` might result in a degradation due to the information loss. The results imply that the judgment for turn dependency from LLM might not exactly be accurate but can still help in achieving effectiveness and efficiency trade-off. This trade-off is important in the RAG task as the context-denoising requirement, especially the input context is usually much longer in conversational scenarios as the corresponding analysis provided in Appendix C.2. Overall, although the implicit transition among historical turns is still hard to capture, it is still a crucial factor for performance improvement with appropriate adaption.

Method	TopiOCQA		QReCC		OR-QuAC	
	F1	Avg. $\lvert T\rvert$	F1	Avg. $\lvert T\rvert$	F1	Avg. $\lvert T\rvert$
Our-Base	32.7	8327	24.7	9434	14.4	8503
+ Dep-hard	23.9	1137	21.0	5438	15.7	4992
+ Dep-soft	33.2	3579	26.1	6004	15.9	5650

Table 7: The performance of the generated response with turn dependency judgments information based on two different strategies. The Avg. $|T|$ denotes the average number of input tokens (prompt) for the generator.

Usefulness of Historical Turns. Table 8 shows the statistical results of how historical turns contribute to the current turn from the aspect of the search intent. By applying the `Dep-soft` on top of the `Base` method, we can see that more than 30% of query turns have performance improvement, indicating the effectiveness of our strategy again. Besides, we aim to understand whether the improvement is attributed to the historical turns with similar information needs that served as context-denoising (De-noise) or the not similar ones that act to improve search results diversification for current turn (Diversify). This is calculated by the percentage of whether a historical turn is judged as sharing similar information needs with the current turn as Eq. 1 among the improved turns. Then, we can see that context-denoising contributes much more than search results diversification on TopiOCQA, which contains frequent topic-switch phenomena within the conversation. This trend is alleviated in QReCC and OR-QuAC, whose conversations are mostly around the same topic.

Dataset	Improve	De-noise	Diversify
TopiOCQA	47.42	73.69	26.31
QReCC	32.70	58.56	41.43
OR-QuAC	37.80	56.94	43.06

Table 8: The statistical results (percentage) of how historical turns contribute to the current turn response generation and the attribution of the improvement with respect to context-noising or search results diversification.

5.5 Impact of Historical Evidence Identification

Method	TopiOCQA		QReCC		OR-QuAC	
	Acc.	HD%	Acc.	HD%	Acc.	HD%
Our-Base	33.0	-	5.1	-	8.0	-
w/. HEI	38.8	60.2	11.8	78.7	14.3	80.4

Table 9: The performance of applying historical evidence identification (HEI) and analysis of how LLMs can capture the historical dependency (HD%) by calculating the ratio of grounding on historical passages to total grounding evidence for each turn.

Table 9 describes the impact of applying the historical evidence identification (`HEI`) strategy and explicitly reflects to what extent the LLMs can capture the historical dependency (HD%) by calculating the ratio of grounding on historical passages to total grounding evidence for each turn. We observe a significant improvement of accuracy after applying the `HEI` strategy across all datasets, which might be attributed to explicitly guiding the LLM to pay attention to the retrieved passages from historical turns, that might be homogeneous and can enhance the similar information contained in the current turn. For the historical dependency analy-

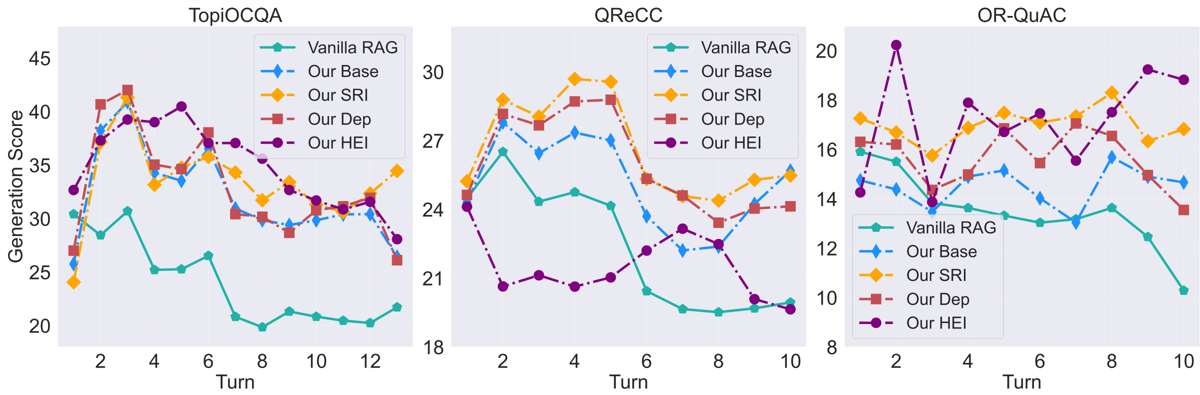

Figure 2: Impact of historical turns at the generation score with different strategies on three datasets.

sis, we find that the LLMs could leverage historical information to contribute to the current turn generation, especially when the conversation is around the same topic (QReCC and OR-QuAC). Further exploration could guide the LLMs to identify useful information in historical turns.

5.6 Impact of Statistic of Ranking List

In this analysis, we aim to understand why historical retrieved results can contribute to current turn response generation from the aspect of the information in the ranking list rather than the final performance of conversational RAG. The statistical results are shown in Table 10.

The hit information of `Bottom` and `History` denote the gold evidence of the current turn is ranked between 3 to 100 in the ranking list and occurs in any previous turns' top-3 results, respectively. These two statistics show the correlation between the position of gold evidence and the ranking list of both current turn and historical turns, where lower `Bottom Hit` and higher `History Hit` are beneficial for leveraging historical turns for response generation. Besides, the `Supplement` denotes the overlap of the passages retrieved beyond the top-3 of the current turn and the ones ranked at the top-3 of any historical turns, which means the historical top retrieved results might supply the evidence used for response generation. We observe that applying our `SRI` strategy can reduce the `Bottom Hit` and enhance the `History Hit` and `Supplement`, which implies the utility of historical search results.

5.7 Impact of Historical Turns

In this experiment, we study the impact of the historical turns for various strategies. We use their per-turn F1 score for generation evaluation. As shown in Figure 2, the performance of vanilla RAG keeps dropping as the conversation diving in and consistently underperforms our proposed methods.

Category	TopiOCQA	QReCC	OR-QuAC
Bottom Hit ↓	35.08	37.55	30.52
+ SRI	33.43	35.20	28.00
History Hit ↑	16.80	48.77	65.56
+ SRI	21.01	52.63	68.65
Supplement	74.14	87.50	85.21
+ SRI	75.62	87.54	86.13

Table 10: The three statistical results (percentage) of the ranking lists with or without applying `SRI` strategies.

This is because as the conversation becomes longer, the difficulty of context modeling increases, while the vanilla RAG lacks a specific design for leveraging historical information except for the query-response pairs. Among our approaches, different strategies can keep an even trend as the historical turns increase due to the effectiveness of leveraging historical search results and the information of turn dependency modeling. Besides, various performances are observed on the different datasets and the three specific designed strategies are generally better than the base method.

6 Conclusion and Future Work

In this paper, we provide the first exploration of leveraging historical turns for retrieval-augmented generation (RAG) in conversational search by defining and verifying the new task scenario, conversational RAG. We investigate various strategies to incorporate historical information to improve the performance of response generation. Experimental results on three conversational search benchmarks demonstrate the effectiveness of our methods by comparing them with existing systems. The thorough analysis of our approaches and observations reveals the potential principle and effectiveness of historical turns. Future work could explore training-based methods by generating available supervision signals to guide the generator to leverage useful information in historical turns.

Limitations

Our study demonstrates the feasibility and effectiveness of leveraging historical information, e.g., historical search results and turn dependency, for response generation in conversational search, while the potential limitations are three aspects that can be explored in future studies.

First, in addition to using the query-response pairs in existing studies, we attempt to leverage the historical search results and the turn dependency for response generation, while more historical information can be explored, e.g., the query type of historical turn (Bolotova et al., 2022), the user thought modeling (Mao et al., 2023a), the historical user feedback (Owoicho et al., 2023), etc.

Second, it is important to understand the potential principle for the effectiveness of historical turns, where we inherit the evaluation metric from previous studies and provide empirical and quantitative analysis to explore it. However, the existing evaluation metric on the generation tasks might not reflect all aspects as it is an open question in the research community, especially in the context of RAG (Salemi and Zamani, 2024). Thus, conducting more qualitative analysis and necessary human evaluation might be helpful for interpretability and can provide concise guidance for understanding what kind of historical information is useful within different scenarios (Wu et al., 2023).

Third, we develop three LLM-based training-free strategies to leverage various historical information as initial exploration. More sophisticated methods can be designed to achieve better performance, such as a combined or multi-step strategy to cover or reason (Yue et al., 2024) different aspects of historical information, a mechanism to identify/evaluate useful parts of the context in historical turns, and a training-based method with suitable supervision signal to guide the existing LLMs, which might not be optimized toward historical information leverage (Yu et al., 2024).

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, Sewon Min, Zexuan Zhong, and Danqi Chen. 2023a. Tutorial proposal: Retrieval-based language models and applications. In *The 61st Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, page 41.

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

Valeriia Bolotova, Vladislav Blinov, Falk Scholer, W Bruce Croft, and Mark Sanderson. 2022. A non-factoid question-answering taxonomy. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1196–1207.

Chi-Min Chan, Chunpu Xu, Ruibin Yuan, Hongyin Luo, Wei Xue, Yike Guo, and Jie Fu. 2024. Rq-rag: Learning to refine queries for retrieval augmented generation. *arXiv preprint arXiv:2404.00610*.

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

Pei Chen, Shuai Zhang, and Boran Han. 2024b. CoMM: Collaborative multi-agent, multi-reasoning-path prompting for complex problem solving. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1720–1738, Mexico City, Mexico. Association for Computational Linguistics.

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

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*, pages 2174–2184.

Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. 2020. Trec cast 2019: The conversational assistance track overview. *arXiv preprint arXiv:2003.13624*.

Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. 2021. Cast 2020: The conversational assistance track overview. In *In Proceedings of TREC*.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. In *International Conference on Learning Representations*.

Hung-Chieh Fang, Kuo-Han Hung, Chen-Wei Huang, and Yun-Nung Chen. 2022. Open-domain conversational question answering with historical answers. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 319–326.

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

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488.

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*.

Soyeong Jeong, Jinheon Baek, Sung Ju Hwang, and Jong C Park. 2023. Phrase retrieval for open domain conversational question answering with conversational dependency modeling via contrastive learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6019–6031.

Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992.

Ziyan Jiang, Xueguang Ma, and Wenhu Chen. 2024. Longrag: Enhancing retrieval-augmented generation with long-context llms. *arXiv preprint arXiv:2406.15319*.

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.

Jean Kaddour, Joshua Harris, Maximilian Mozes, Herbie Bradley, Roberta Raileanu, and Robert McHardy. 2023. Challenges and applications of large language models. *arXiv preprint arXiv:2307.10169*.

Omar Khattab, Keshav Santhanam, Xiang Lisa Li, David Hall, Percy Liang, Christopher Potts, and Matei Zaharia. 2022. Demonstrate-search-predict: Composing retrieval and language models for knowledge-intensive nlp. *arXiv preprint arXiv:2212.14024*.

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.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

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.

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.

Xi Victoria Lin, Xilun Chen, Mingda Chen, Weijia Shi, Maria Lomeli, Richard James, Pedro Rodriguez, Jacob Kahn, Gergely Szilvasy, Mike Lewis, et al. 2023. Ra-dit: Retrieval-augmented dual instruction tuning. In *The Twelfth International Conference on Learning Representations*.

Kelong Mao, Chenlong Deng, Haonan Chen, Fengran Mo, Zheng Liu, Tetsuya Sakai, and Zhicheng Dou. 2024. Chatretriever: Adapting large language models for generalized and robust conversational dense retrieval. *arXiv preprint arXiv:2404.13556*.

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, and Hongjin Qian. 2022a. Curriculum contrastive context denoising for few-shot conversational dense retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 176–186.

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*, pages 2935–2946.

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, Weiwei Sun, Zhaochun Ren, Zhaopeng Tu, and Maarten de Rijke. 2020. 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.

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. *arXiv preprint arXiv:2406.05013*.

Fengran Mo, Kelong Mao, Yutao Zhu, Yihong Wu, Kaiyu Huang, and Jian-Yun Nie. 2023a. Convgqr: Generative query reformulation for conversational search. *arXiv preprint arXiv:2305.15645*.

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. 2024b. Aligning query representation with rewritten query and relevance judgments in conversational search. *arXiv preprint arXiv:2407.20189*.

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

Fengran Mo, Bole Yi, Kelong Mao, Chen Qu, Kaiyu Huang, and Jian-Yun Nie. 2024d. 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. 2024e. How to leverage personal textual knowledge for personalized conversational information retrieval. *arXiv preprint arXiv:2407.16192*.

Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. 2020. Document ranking with a pretrained sequence-to-sequence model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 708–718.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Paul Owoicho, Ivan Sekulić, Mohammad Aliannejadi, Jeffrey Dalton, and Fabio Crestani. 2023. Exploiting simulated user feedback for conversational search: Ranking, rewriting, and beyond. *arXiv preprint arXiv:2304.13874*.

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*.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A Smith, and Mike Lewis. 2023. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5687–5711.

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*, pages 4725–4737.

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.

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.

Nirmal Roy, Leonardo FR Ribeiro, Rexhina Blloshmi, and Kevin Small. 2024. Learning when to retrieve, what to rewrite, and how to respond in conversational qa. *arXiv preprint arXiv:2409.15515*.

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.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803.

Weihang Su, Yichen Tang, Qingyao Ai, Zhijing Wu, and Yiqun Liu. 2024. Dragin: Dynamic retrieval augmented generation based on the real-time information needs of large language models. *arXiv preprint arXiv:2403.10081*.

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

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.

Shuting Wang, Xin Xu, Mang Wang, Weipeng Chen, Yutao Zhu, and Zhicheng Dou. 2024. Richrag: Crafting rich responses for multi-faceted queries in retrieval-augmented generation. *arXiv preprint arXiv:2406.12566*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

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*.

Jinxi Xu and W. Bruce Croft. 1996. Query expansion using local and global document analysis. In *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Peng Xu, Wei Ping, Xianchao Wu, Lawrence McAfee, Chen Zhu, Zihan Liu, Sandeep Subramanian, Evelina Bakhturina, Mohammad Shoeybi, and Bryan Catanzaro. 2023. Retrieval meets long context large language models. In *The Twelfth International Conference on Learning Representations*.

Peng Xu, Wei Ping, Xianchao Wu, Lawrence McAfee, Chen Zhu, Zihan Liu, Sandeep Subramanian, Evelina Bakhturina, Mohammad Shoeybi, and Bryan Catanzaro. 2024. Retrieval meets long context large language models. In *The Twelfth International Conference on Learning Representations*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2024. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.

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*.

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*.

Zhenrui Yue, Honglei Zhuang, Aijun Bai, Kai Hui, Rolf Jagerman, Hansi Zeng, Zhen Qin, Dong Wang, Xuanhui Wang, and Michael Bendersky. 2024. Inference scaling for long-context retrieval augmented generation. *arXiv preprint arXiv:2410.04343*.

Hamed Zamani and Michael Bendersky. 2024. Stochastic rag: End-to-end retrieval-augmented generation through expected utility maximization. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2641–2646.

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.

Chujie Zheng, Yunbo Cao, Daxin Jiang, and Minlie Huang. 2020. Difference-aware knowledge selection for knowledge-grounded conversation generation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 115–125.

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

Appendix

A Prompt

In this section, we list the prompts that we have carefully designed for different proposed strategies, as well as the prompts used for response generation.

A.1 Search Results Integration

Search Results Integration

[System Prompt]: You are an intelligent ranker. For an information-seeking dialog, please rank the retrieved passages from each utterance according to the utility to help you answer the current question.

[User Prompt]: For an information-seeking dialogue, I will provide you with {top-k * (turn - 1)} historical retrieved passages with their associated queries and {top-k} passages for current query, each indicated by number identifier []. Rank the passages based on their helpfulness to answer the current query: $\{q_i\}$

A.2 Historical Evidence Identification

Historical Evidence Identification

[System Prompt]: This is a chat between a user and an artificial intelligent assistant. The assistant gives helpful, detailed, and polite answers to the user's questions based on the retrieved evidence from historical turns and current turn. The assistant should also indicate when the answer cannot be found in the context and then answer based on its own knowledge.

[User Prompt]: Historical Question 1: $\{q_1\}$. Response 1: $\{r_1\}$. Evidence id: [pid_1, pid_2, ..., pid_k]. Evidences: [psg_1, psg_2, ..., psg_k]
Historical Question 2: $\{q_2\}$. Response 2: $\{r_2\}$. Evidence id: [pid_1, pid_2, ..., pid_k]. Evidences: [psg_1, psg_2, ..., psg_k]
...
Historical Question n-1: $\{q_{n-1}\}$. Response n-1: $\{r_{n-1}\}$ Evidence id: [pid_1, pid_2, ..., pid_k]. Evidences: [psg_1, psg_2, ..., psg_k].
Current Evidences: [psg_1, psg_2, ..., psg_k]. Current Question: $\{q_i\}$. Please give a complete answer to the question. Cite the evidences that supports your answer within brackets []. The output format should be:
Answer:
Citation: []
Current Question: $\{q_i\}$.
Never ask for clarification, say you don't understand or explain the reason. Go ahead!

A.3 Historical Turns Dependency Judgment

Historical Turns Dependency Judgment

[System Prompt]: You are an expert evaluator. For an information-seeking dialog, given the Current Question, please select all the Historical Questions that contain similar information needs. The output should be a list that only contains the index of selected questions.

[User Prompt]:
Historical Question 1: $\{q_1\}$. Response 1: $\{r_1\}$
Historical Question 2: $\{q_2\}$. Response 2: $\{r_2\}$
...
Historical Question n-1: $\{q_{n-1}\}$. Response n-1: $\{r_{n-1}\}$
Current Question: $\{q_i\}$.
Now, you should give me the selected Historical Questions. The output format should be a list that only contain the index of selected questions, e.g., [1, 3, 5]. Never ask for clarification or say you don't understand the selection. Go ahead!

A.4 Conversational Retrieval-Augmented Response Generation

Conversational Retrieval-Augmented Response Generation

[System Prompt]: You are an intelligent generator. For an information-seeking dialog, please help generate the answer that can fully address the user's information needs based on the historical conversation and retrieved evidence.

[User Prompt]: Historical Question 1: $\{q_1\}$. Evidences: [psg_1, psg_2, ..., psg_k]. Response 1: $\{r_1\}$.
Historical Question 2: $\{q_2\}$. Evidences: [psg_1, psg_2, ..., psg_k]. Response 2: $\{r_2\}$.
...
Historical Question n-1: $\{q_{n-1}\}$. Evidences: [psg_1, psg_2, ..., psg_k]. Response n-1: $\{r_{n-1}\}$.
Current Evidences: [psg_1, psg_2, ..., psg_k]. Current Question: $\{q_i\}$.
Now, you should give me the answer of the **Current Question** under the conversation and **Evidences**. The output format should always be:
Answer:
Note that you should always try to answer it concisely.
Never ask for clarification, say you don't understand or explain the reason. Go ahead!

B Experimental Setup

B.1 Datasets Details

	TopiOCQA	QReCC	OR-QuAC
#Conv.	205	2,775	771
#Turns(Qry.)	2,514	16,451	5,571
#Collection	25M	54M	11M
#Avg. Qry.	12.9	5.3	7.2
#Min Qry.	5	2	4
#Max Qry.	25	12	12
#Avg. Psg	9.0	1.6	1.0

Table 11: Statistics of conversational search datasets.

The statistics of each dataset are presented in Table 11, including TopiOCQA, QReCC, and OR-QuAC. The TopiOCQA has the longest conversation and the average turn is 12.9, while the other two are relatively shorter. Besides, one of the main differences among them is the average number of passages per conversation, which are 9.0, 1.6, and 1.0 with respect to TopiOCQA, QReCC, and OR-QuAC. It reveals that in QReCC and OR-QuAC, most conversations only involve one topic/passage.
TopiOCQA is a relatively new published dataset featured with topic switching. As the conversation goes on, the topics may switch to related topics, a phenomenon commonly observed in information-seeking search sessions. Therefore, TopiOCQA requires the ability of accurate context modeling more, since the previous turns may not be directly related to the current turn.
QReCC is conducted specifically for conversational open-domain QA. The questions were sourced from the QuAC (Choi et al., 2018), NQ (Kwiatkowski et al., 2019), and CAsT datasets (Dalton et al., 2020, 2021). The annotators were required to give answers using a web search engine.
OR-QuAC is transformed from the conversational MRC dataset QuAC (Choi et al., 2018). The questions in a conversation are sourced from the same section in Wikipedia, and the answers are extractive, i.e., exact text spans in passages.

B.2 Implementation Details

We implement the retrieval evaluation metrics from the `pytrec_eval` tool (Van Gysel and de Rijke, 2018) based on Faiss (Johnson et al., 2019) libraries. The lengths of the query, concatenated input, and passage are truncated to 32, 512, and 384 tokens, respectively. For response generation, the evaluation is implemented following the code released by Qu et al. (2020), Asai et al. (2023b), and Pan et al. (2024). Besides, we limit the generation length as 128 with temperature 1. All the experiments are conducted on an NVIDIA A100 GPU

C Additional Experiment Results

C.1 Main Results with Other Metrics

Evaluating generation results is still an open question in the research community, since the existing metric, e.g., F1, EM, Rouge, and BLEU, can only reflect the quality of generated response from a specific aspect. The evaluation could be more difficult when the ground-truth answer is free-form rather than exact spans. Thus, Table 12 reports the additional evaluation metric for both retrieval and generation in terms of Table 6 as references.

C.2 Efficiency Analysis

In this section, we supply the impact of context for efficiency analysis. Figure 3 shows the average input token numbers for the generator model per turn. We can see that when applying our `Base` which leverages historical search results with a basic strategy, the average input token numbers increase quickly as the conversation goes on. Although it improves the performance as shown in Figure 2, the efficiency reduces. However, our designed `Dep-soft` strategy can reduce the inference cost and keep good effectiveness at the meanwhile. This experimental analysis suggests a better effectiveness and efficiency trade-off strategy is important and desirable.

C.3 Impact of Different Backbone Models

In this section, we evaluate the generalization ability of our proposed methods based on different types of LLMs, including open-source ones with various sizes and commercial ones. Table 13 shows the performance of various LLMs based on whether leveraging historical information, i.e., using top-20 retrieved passages of only current turn as `vanilla RAG` or our `Base` method with top-3 retrieved passages of each historical turn and current turn.We find that our method outperforms the vanilla RAG paradigm across various backbone models, indicating the effectiveness of historical information in improving conversation RAG tasks and the robustness of our strategies.

Method	TopiOCQA			QReCC			OR-QuAC		
	MRR	Rouge	BLEU	MRR	Rouge	BLEU	MRR	Rouge	BLEU
Vanilla RAG	31.5	21.7	7.2	43.7	21.1	7.3	47.7	12.9	4.4
Our-Base		29.2	12.0		24.1	9.6		13.5	4.7
+ SRI-T5	32.4	29.4	12.2	51.1	24.4	9.9	52.9	15.5	6.2
+ SRI-LLM	34.5	29.8	12.5	54.2	24.7	10.0	54.0	15.7	6.3

Table 12: The performance of conversational retrieval and response generation across three conversational search datasets with other metrics based on Table 6

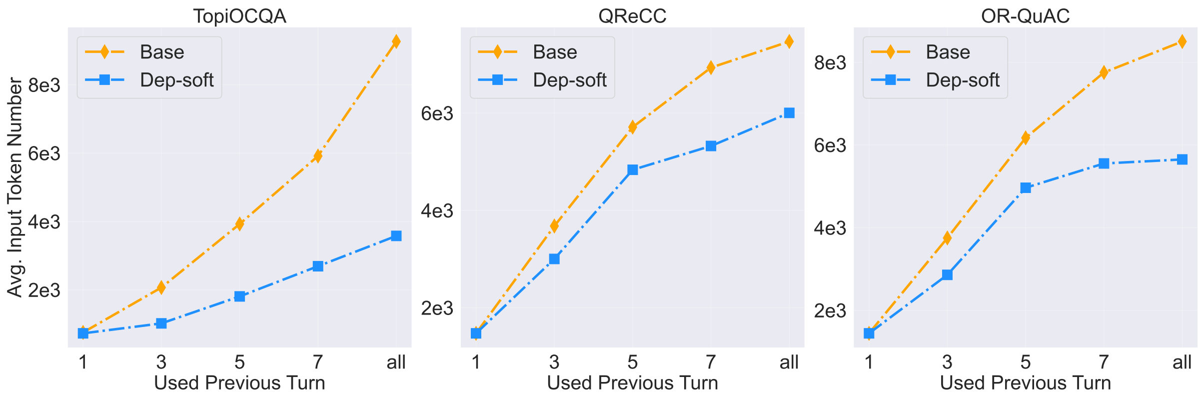

Figure 3: Impact of historical turns at the generation efficiency based on the average input token numbers with two strategies on three datasets.

$\mathcal{LLM}$	TopiOCQA		QReCC		OR-QuAC	
	F1	EM	F1	EM	F1	EM
w/o Historical Search Results (Vanilla RAG)						
Mistral-2-7B	27.0	3.1	22.4	3.9	14.2	4.7
Mixtral-8x7B	27.8	3.7	24.0	4.1	14.3	5.6
ChatGPT-3.5	28.6	4.5	25.3	4.4	14.5	7.7
Claude-Sonnet	24.4	4.1	23.6	3.4	14.0	7.9
w/. Historical Search Results (Ours)						
Mistral-2-7B	28.0	4.4	23.0	4.3	14.5	4.8
Mixtral-8x7B	29.6	5.7	25.8	4.7	14.7	6.0
ChatGPT-3.5	32.1	6.3	26.9	5.0	15.6	8.7
Claude-Sonnet	32.7	6.9	27.5	5.1	14.4	8.1

Table 13: Response quality with using various LLMs, where the generation is performed either only on top of the retrieved evidence of current turn (Vanilla RAG) or also leveraging historical search results (Our-Base).