# Evolving Graph-Based Context Modeling for Multi-Turn Conversational Retrieval-Augmented Generation


Yiruo Cheng
Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China
chengyr@ruc.edu.cn

Hongjin Qian
Beijing Academy of Artificial Intelligence
Beijing, China
chienqhj@gmail.com

Fengran Mo
University of Montreal
Montreal, Quebec, Canada
fengran.mo@umontreal.ca

Yongkang Wu
Zhonghua Li
Qi Ye
Huawei Poisson Lab
Hangzhou, Zhejiang, China

Ji-Rong Wen
Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China
jrwen@ruc.edu.cn

Zhicheng Dou*
Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China
dou@ruc.edu.cn



## Abstract

Conversational Retrieval-Augmented Generation (RAG) systems enhance user interactions by integrating large language models (LLMs) with external knowledge retrieval. However, multi-turn conversations present significant challenges, including implicit user intent and noisy context, which hinder accurate retrieval and response generation. Existing approaches often struggle with the unstructured conversational context and fail to model explicit relations among conversational turns. Moreover, they do not leverage historically relevant passages effectively. To overcome these limitations, we propose EvoRAG, a novel framework that maintains an evolving knowledge graph aligned with the unstructured conversational context. This graph explicitly captures relations among user queries, system responses, and relevant passages across conversational turns, serving as a structured representation of the context. EvoRAG includes three key components: (1) a dual-path retrieval module for context denoising, (2) a unified knowledge integration module for query rewriting and summarization, and (3) a graph-enhanced RAG module for accurate retrieval and response generation. Experiments on four public conversational RAG datasets show that EvoRAG significantly outperforms strong baselines, particularly in handling topic shifts and long dialogue contexts.



## CCS Concepts

• **Information systems → Users and interactive retrieval**.

## Keywords

Conversational RAG; Knowledge Graph; Knowledge-Intensive Tasks



*Corresponding author.





*CIKM '25, Seoul, Republic of Korea*





ACM ISBN 979-8-4007-2040-6/2025/11
https://doi.org/10.1145/3746252.3761355

**ACM Reference Format:**
Yiruo Cheng, Hongjin Qian, Fengran Mo, Yongkang Wu, Zhonghua Li, Qi Ye, Ji-Rong Wen, and Zhicheng Dou. 2025. Evolving Graph-Based Context Modeling for Multi-Turn Conversational Retrieval-Augmented Generation. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM '25), November 10–14, 2025, Seoul, Republic of Korea.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3746252.3761355


## 1 Introduction

Conversational Retrieval-Augmented Generation (RAG) systems, e.g., ChatGPT [48], Doubao [56], and KimiChat [2], have recently demonstrated strong capability in integrating with search models to fulfill users' information needs through multi-turn interactions. By combining large language models (LLMs) with retrieved external knowledge, conversational RAG systems are able to generate accurate and context-aware conversational responses in the scenarios of multi-turn conversations.

Different from single-turn RAG [11, 16, 69], which relies solely on a complete stand-alone query, conversational RAG [38, 60, 65] requires the model to go beyond isolated queries and deal with the retrieved passages over both current and historical turns dynamically and adaptively [10, 22, 36, 53]. The challenges of the conversational RAG lie in two aspects. First, user search intent is usually implicit and context-dependent in conversation [52], which results in difficulty in context-dependent query understanding on top of preceding conversational turns. Second, as the conversation dives in, earlier turns might include both irrelevant and useful information for the current turn, resulting in either a negative or positive impact on the query understanding of the current turn [8]. Identifying the useful parts from the lengthy and redundant historical interactions is non-trivial [35]. For example, as shown in Figure 1, the real query intent is "What should we consider in terms of safety and environmental impact of the silicones, particularly with cyclic siloxanes D4 and D5?", while the conversation history contains irrelevant context like "silicones' uses in medicine and cosmetic surgery", which would hinder accurate session understanding for retrieving relevant information and generating accurate responses.

Existing studies attempt to address these challenges from two aspects. One line of studies is the conversational query rewriting

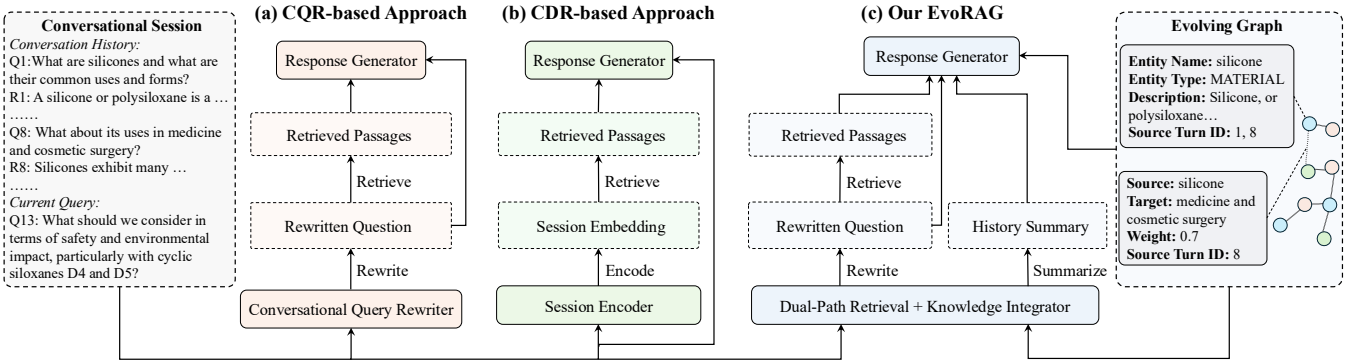


**Figure 1: A conceptual illustration for the three types of conversational RAG. Figure (a) is the CQR-based approach, figure (b) is the CDR-based approach, and figure (c) is our EvoRAG, which explicitly filters irrelevant context, integrates knowledge, and performs graph-enhanced retrieval and generation by combining conversational context with the evolving graph.**

(CQR)-based approach. As shown in Figure 1(a), this approach reformulates the current query into a stand-alone question by integrating conversation history [60, 65], and then performs the standard single-turn RAG using the rewritten question. Another line of research is the conversational dense retrieval (CDR)-based approach [29, 38]. As shown in Figure 1(b), the CDR-based approach first encodes the entire conversational session to learn latent representations for holistic retrieval [30, 68], and then leverages the complete conversational session and retrieved passages for response generation. While both approaches demonstrate certain improvements, they still face notable challenges. First, both CQR-based methods and CDR-based methods fail to explicitly model the relations among conversational turns, including topic progression, branching, or shifts [8]. Instead, they process the whole unstructured conversational session heavily relying on the model's denoising capacity [35, 41, 68] to handle internal redundancy or inconsistency. Second, neither approach adequately incorporates historical relevant passages, despite recent studies [53] showing that strategically leveraging historical relevant passages can substantially improve the response quality for subsequent conversational turns.

To address the aforementioned challenges, one promising direction is to convert the conversational context-including user queries, system responses, and relevant passages-into a structured representation. While existing studies like GraphRAG [13] have shown the value of using knowledge graphs as the structured representation to explicitly model the relations among retrieved passages, they primarily rely on static knowledge graphs. Such fixed representations lack the adaptability required for dynamic multi-turn conversations, where each interaction may introduce new queries, responses, passages, or evolving relational dependencies. We argue that an effective structured representation of the conversational context can dynamically model the relations among conversational elements, offering a coherent understanding of the dialogue flow.

Motivated by this, we propose EvoRAG, a novel framework that dynamically maintains an **evolving knowledge graph** with structured knowledge aligned with the unstructured diving conversational context when the new turn arrives. The evolving graph explicitly models the relations among user queries, system responses, and the relevant passages across conversational turns. The evolving graph's structured representation and the original conversational context's unstructured information provide complementary signals, facilitating more effective context-aware retrieval and nuanced reasoning in response generation. As shown in Figure 1(c), our EvoRAG leverages both the unstructured conversational context and the evolving graph through three key modules, thereby distinguishing it from the two existing paradigms. First, a **dual-path retrieval module** performs coarse-grained denoising for the conversation history and the corresponding evolving graph, aiming to identify relevant candidate history and candidate entities that are potentially related to the current query when diving into a new turn. Second, a **unified knowledge integration module** integrates the relevant candidates to generate a rewritten query, summarizes the conversational context, and extracts the most relevant entities for precise graph updates. Finally, a **graph-enhanced RAG module** is implemented to conduct contextual passage retrieval and response generation, with the structured knowledge representation from the evolving graph.

We evaluate EvoRAG on the four public conversational RAG datasets: TopiOCQA, QReCC, INSCIT, and CORAL. Experimental results demonstrate that EvoRAG significantly outperforms strong baselines, particularly in scenarios involving topic shifts and long conversational contexts. The superior performance implies the advantage of explicit, structured context modeling over conventional unstructured methods.

Our contributions can be summarized as follows:

(1) We propose EvoRAG, a novel framework that dynamically maintains an evolving knowledge graph to model the logical relations among user queries, system responses, and the relevant passages across conversational turns.

(2) By jointly utilizing the structured information in the evolving graph and the original unstructured context, EvoRAG captures contextual dependencies, improving both retrieval performance and generation quality in the conversations.

(3) We conduct extensive experiments on four public conversational RAG datasets. Experimental results show that EvoRAG consistently improves both retrieval and generation performance

and outperforms the strong baselines, especially in challenging scenarios, e.g., involving topic shifts and long-context dialogues.

## 2 Related Work

### 2.1 Conversational Search

Conversational search [40] enables users to interact with information-seeking systems through multi-turn conversations. The user's search intent is context-dependent and might involve several aspects during the interactions [42]. Thus, the main challenge for the systems is to understand the real intent accurately. Then, the relevant information could be retrieved as a response for the users. To achieve this, two main approaches have been developed in the literature.

One research line is conversational query rewriting (CQR), which converts incomplete or context-dependent queries into stand-alone rewrites based on the previous interactions. The query rewrites are then used with an off-the-shelf retriever to perform an ad-hoc search. Existing studies mainly focus on selecting the useful tokens from the conversation history [23, 28, 49, 59] or training a generative rewriter [27, 58, 67] to mimic human-annotated rewrites. However, the manually rewritten query might not necessarily achieve the best performance in the downstream tasks, as shown in previous studies [6, 61]. Besides, several studies [31, 41, 44, 61] integrate ranking signals into the query rewriting process. Recently, a series of studies [17, 32, 37, 39, 64, 66] leverage the powerful LLMs to generate query rewrites directly. These LLM-generated rewrites have shown notable improvements. Another research line, conversational dense retrieval (CDR), jointly encodes the conversation history and the current query to perform end-to-end dense retrieval [7, 30, 68]. The key challenge in training conversational dense retrievers lies in the data scarcity. To this end, data augmentation [5, 9, 21, 26, 34, 46, 47] is employed to supply the training data via various techniques. In addition, conversation history usually contains irrelevant or redundant information, which would inject unexpected noise into the retrieval phase. Thus, context denoising [33, 35, 43, 45] is effective in identifying the relevant context.

While these two approaches have demonstrated satisfactory performance in retrieval, their paradigms for concatenating the whole conversation history with the current question would result in lengthy input. Our EvoRAG, with an evolving graph for explicit context-denoising, is proposed to target this problem.

### 2.2 Graph-Based RAG

Retrieval-Augmented Generation (RAG) [14, 19, 20, 55] enhances the performance of LLMs by integrating retrieved evidence from external knowledge [18, 24, 50, 51]. Despite its effectiveness, the standard RAG pipeline faces challenges in handling complex multi-hop reasoning and multiple evidence integration [25]. To address this limitation, recent studies [4, 13, 15, 25, 54] develop graph-based RAG by integrating a knowledge graph for its strong reasoning capabilities [25], which can facilitate final response generation. For instance, GraphRAG [13] leverages LLMs to construct a knowledge graph from source documents and then generate community-level summaries for user queries. HybridRAG [54] combines text-based semantic retrieval with structured retrieval from knowledge graphs to improve the context representation. KAG [25] adopts a hierarchical knowledge graph with aligned graph structures and text blocks. Then, a hybrid reasoning engine integrates symbolic logic, semantic reasoning, and numerical computation.

However, these graph-based RAG approaches mainly rely on static knowledge graphs, which cannot directly adapt to the evolving context of multi-turn conversations. Our EvoRAG is proposed for this scenario with an evolving knowledge graph.

## 3 Preliminaries

### 3.1 Task Formulation

Conversational RAG aims to generate accurate and context-aware responses to the user's latest query within an ongoing conversation by retrieving external passages. Two sub-tasks are crucial in conversational RAG, including conversational passage retrieval and conversational response generation.

**Conversational Passage Retrieval** is the first stage to retrieve the most relevant passages from an external corpus for the current query. Formally, given the $k$-th current question $q_k$ and the corresponding conversation history $H_k = \{q_i, r_i, P_i\}_{i=1}^{k-1}$ ($q_i$, $r_i$, and $P_i$ denote the question, response and relevant passages of the $i$-th turn, respectively), the retriever R aims to identify the most relevant passages $P_k$ from the external corpus $\mathcal{P}$ that contain the evidence to answer the current query $q_k$.

**Conversational Response Generation** is the second stage to generate a contextually coherent and informative response with the retrieved passages. Formally, given the the current question $q_k$, the conversation history $H_k = \{q_i, r_i, P_i\}_{i=1}^{k-1}$, and the retrieved passages $P_k$, the generator G aims to generate a response $r_k$ to satisfy the information needs in the $k$-th question $q_k$.

### 3.2 Structure of Our Evolving Knowledge Graph

The evolving graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ serves as a structured representation of the conversational context, where the nodes $\mathcal{V}$ denote key entities extracted from semantic text, and the edges $\mathcal{E}$ denote the relations between the two entities. Both entities and relations are derived from various sources to represent the information in the conversational context, including user queries, system responses, and relevant passages.

Inspired by the GraphRAG [13], each node $v \in \mathcal{V}$ in the evolving knowledge graph is characterized by four attributes:

$$v = \Big(\text{Name}(v), \text{Type}(v), \text{Desc}(v), \text{TurnID}(v)\Big), \tag{1}$$

where $\text{Name}(v)$ refers to the entity name, $\text{Type}(v)$ is the type of the entity, $\text{Desc}(v)$ is a description of this entity, and $\text{TurnID}(v)$ is the dialogue turn(s) in which the entity was extracted.

And each edge $e \in \mathcal{E}$ is defined as:

$$e = \Big(\text{src}(e), \text{tgt}(e), w(e), \text{TurnID}(e)\Big), \tag{2}$$

where $\text{src}(e), \text{tgt}(e) \in \mathcal{V}$ are the source and target entities, respectively, $w(e)$ is the weight of the relation, and $\text{TurnID}(e)$ is the source turn ID in the conversation.

If the entities and the relations are extracted from a passage rather than extracted from a question or a response within the conversation, then the source ID should be set to the source passage ID instead of the source turn ID in the conversation, ensuring proper linkage to the original semantic text.

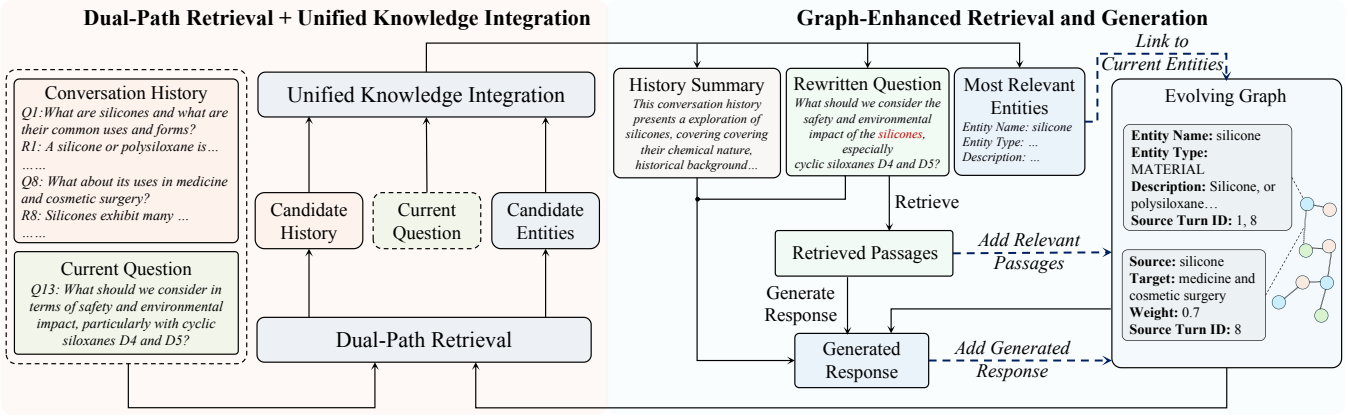


**Figure 2: The workflow of the EvoRAG, which consists of three modules: (i) Dual-Path Retrieval module for context denoising, (ii) Unified Knowledge Integration module for query rewriting and history summarization, and (iii) Graph-Enhanced Retrieval and Generation module for accurate retrieval and response generation.**

## 4 Methodology

In this section, we propose EvoRAG, a novel framework that maintains an **evolving knowledge graph** with the conversation dives in. The evolving graph explicitly models logical relations between user queries, system responses, and relevant passages across the conversational turns. By structuring the conversational context in an evolving graph, EvoRAG enables precise contextual understanding for improving retrieval performance and response quality.

### 4.1 Overview

To facilitate conversational RAG, our EvoRAG maintains an evolving knowledge graph aligned with the diving conversational context to capture the logical relations among user queries, system responses, and relevant passages.

As shown in Figure 2, our EvoRAG framework consists of three key components as follows. The detailed process is described in Algorithm 1.

(1) To construct and maintain this evolving graph efficiently, a dual-path retrieval module (§4.2) denoises the conversation history and the associated evolving graph by extracting candidate history and entities that are related to the current turn query.

(2) These denoised candidates are integrated to generate a rewritten question, a summarization of the conversation history, and the most relevant entities identified for the current turn query (§4.3). These refined outputs are used to guide subsequent retrieval, generation, and graph update.

(3) The conversational passage retrieval and response generation are executed in order by leveraging the evolving graph and the refined outputs (§4.4).

### 4.2 Dual-Path Retrieval

As the conversation dives in, the irrelevant information in terms of the current query would be maintained in the increased conversation history and the evolving graph. These noisy parts would inevitably affect the accuracy of query intent understanding [8]. To address these issues, we introduce a dual-path retrieval approach to filter out irrelevant historical contexts in conversation and entities in the evolving graph, which consists of two parallel retrieval paths: Semantic-Aware Turn Retrieval and Graph-Based Entity Extraction.

*4.2.1 Semantic-Aware Turn Retrieval.* To identify relevant turns, we compute the cosine similarity between the current query $q_k$ and each historical turn in the conversation. Each historical turn $t$ is represented as the concatenation between its question $q_t$ and response $r_t$, which is then encoded into the embedding $\mathbf{emb}_t$. The $q_k$ is encoded as $\mathbf{emb}_k$. Then, we select historical contexts $H_k^{\text{sem}}$ with cosine similarity over a pre-defined threshold $\theta$:

$$H_k^{\text{sem}} = \{(q_t, r_t) \mid \cos(\mathbf{emb}_k, \mathbf{emb}_t) > \theta\}. \tag{3}$$

For each relevant context $(q_t, r_t) \in H_k^{\text{sem}}$, we aim to identify its corresponding entities $\mathcal{V}_t$ in the evolving graph, and select the most central ones based on node degree as

$$\mathcal{V}_k^{\text{sem}} = \left\{ v_t^* \mid v_t^* = \underset{v \in \mathcal{V}_t}{\text{argmax}}\, \text{degree}(v),\ (q_t, r_t) \in H_k^{\text{sem}} \right\}. \tag{4}$$

*4.2.2 Graph-Based Entity Extraction.* The *Semantic-Aware Turn Retrieval* might not be able to capture the related entities from the logical aspect. To this end, we leverage the evolving graph to identify entities that may be logically related to the current question. Specifically, we first perform a clustering on the evolving graph $\mathcal{G}$ to obtain a set of subgraphs $\{\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_n\}$. For each subgraph $\mathcal{G}_i = (\mathcal{V}_i, \mathcal{E}_i)$, we select the most representative entity based on node degree as shown in Equation (5), resulting in the set $\mathcal{V}_k^{\text{stru}}$. With the source turn ID of the node $v \in \mathcal{V}_k^{\text{stru}}$, we can locate the corresponding historical contexts $H_k^{\text{stru}}$.

$$\mathcal{V}_k^{\text{stru}} = \left\{ v_i^* \mid v_i^* = \underset{v \in \mathcal{V}_i}{\text{argmax}}\, \text{degree}(v),\ i = 1, 2, \ldots, n \right\}. \tag{5}$$

*4.2.3 Candidate Union.* The candidate entities $\mathcal{V}_k^c$ and candidate history $H_k^c$ are obtained by merging the results from the two complementary paths: Semantic-Aware Turn Retrieval Path and the

**Algorithm 1** Evolving Graph for Multi-Turn Conversational RAG

**Procedure** EvoRAG
**Input:** the $k$-th question $q_k$, the $k$-th conversation history $H_k$, and the external corpus $\mathcal{P}$.
**Output:** the $k$-th response $r_k$, the evolving graph $\mathcal{G}$.
Initialize graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}), \mathcal{V} = \emptyset, \mathcal{E} = \emptyset$
**for** each turn $k = 1, 2, \ldots$ **do**
  $H_k^c, \mathcal{V}_k^c \leftarrow$ Dual-Path Retrieval$(H_k, q_k, \mathcal{G})$
  $\hat{q_k}, S_k, \mathcal{V}_k^* \leftarrow$ Unified Knowledge Integration$(q_k, H_k^c, V_k^c)$
    ▷ $\hat{q_k}$ is the rewritten question, $S_k$ is the history summary
      ▷ $\mathcal{V}_k^*$ is set of the most relevant entities in the graph
  $\mathcal{V}_k^{(q)} \leftarrow$ Extract Entities$(q_k)$
  $\mathcal{E}_k^{(q)} \leftarrow$ Build Edges$(\mathcal{V}_k^{(q)}, \mathcal{V}_k^*)$
  $\mathcal{V} \leftarrow \mathcal{V} \cup \mathcal{V}_k^{(q)}, \mathcal{E} \leftarrow \mathcal{E} \cup \mathcal{E}_k^{(q)}$ ▷ merge current question
  $P_k \leftarrow \mathrm{R}(\hat{q_k}, \mathcal{P})$ ▷ retrieve passages
  $r_k \leftarrow \mathrm{G}(S_k, \hat{q_k}, P_k, \mathcal{G})$ ▷ generate response
  **for** each passage $p \in P_k$ **do**
    $\mathcal{V}_p, \mathcal{E}_p \leftarrow$ Extract Entities and Edges$(p)$
    $\mathcal{V} \leftarrow \mathcal{V} \cup \mathcal{V}_p, \mathcal{E} \leftarrow \mathcal{E} \cup \mathcal{E}_p$ ▷ add relevant passages
  **end for**
  $\mathcal{V}_k^{(r)}, \mathcal{E}_k^{(r)} \leftarrow$ Extract Entities and Edges$(r_k)$
  $\mathcal{V} \leftarrow \mathcal{V} \cup \mathcal{V}_k^{(r)}, \mathcal{E} \leftarrow \mathcal{E} \cup \mathcal{E}_k^{(r)}$ ▷ append generated response
  **yield** $r_k$ ▷ output response for current turn
**end for**
**return** $\mathcal{G}$

Graph-Based Entity Extraction Path. Formally, they are defined as:

$$\mathcal{V}_k^c = \mathcal{V}_k^{\text{sem}} \cup \mathcal{V}_k^{\text{stru}}, H_k^c = H_k^{\text{sem}} \cup H_k^{\text{stru}}. \tag{6}$$

## 4.3 Unified Knowledge Integration

The candidate outputs from the dual-path retrieval might not guarantee the relevant connection between each part of the information extracted from the conversational context and the evolving graph. The redundancy might persist even after coarse denoising and thus necessitate precise subset selection to focus on the most relevant entities in the evolving graph.

To address this, we introduce a unified knowledge integration module that explicitly combines the useful information from the evolving graph and its corresponding conversational context. This module aims to leverage precisely aligned useful information from both sources to simultaneously enhance retrieval performance, improve response quality, and enable dynamic updating of the evolving graph.

First, a self-contained rewritten question $\hat{q}_k$ is generated by leveraging the relevant candidate history $H_k^c$ from conversational context and candidate entities $V_k^c$ from the evolving graph. Second, it produces a concise summary $S_k$ by integrating semantic information from both the candidate history $H_k^c$ and the candidate entities $V_k^c$, filtering out irrelevant details to retain only the most important context. Third, the most relevant entities $V_k^*$ are identified from the evolving graph for the current question, including both directly referenced entities and implicitly inferred entities. Besides, the contextual background entities provide necessary supplementary knowledge for comprehensive interpretation.

## 4.4 Graph-Enhanced Retrieval and Generation

With the output of the *Unified Knowledge Integration* module, we perform retrieval, generation, and graph updating for the current turn. The key point is to update the evolving graph to improve retrieval performance and response quality with the sophisticated graph representation. The graph evolves through three stages: Merge Current Question, Add Relevant Passages, and Append Generated Response, which is shown in Figure 3.

*4.4.1 Merge Current Question.* To effectively merge the current question to the evolving graph, we extract entities from the current question $q_k$ to obtain the current question entities $\mathcal{V}_k^{(q)}$, then link them to the most relevant entities $\mathcal{V}_k^*$ in the evolving graph. The created edges for the linking process are conducted as

$$\mathcal{E}_k^{(q)} = \left\{ (v_k, v^*, w_{\max}, t_k) \mid v_k \in \mathcal{V}_k^{(q)}, v^* \in \mathcal{V}_k^* \right\}, \tag{7}$$

where the edge weight is set to the maximum one among all edges in the graph, and $t_k$ denotes the timestamp of the $k$-th turn. Then, the graph merged with the current question $q_k$ is updated as

$$\mathcal{V} \leftarrow \mathcal{V} \cup \mathcal{V}_k^{(q)}, \mathcal{E} \leftarrow \mathcal{E} \cup \mathcal{E}_k^{(q)}. \tag{8}$$

*4.4.2 Add Relevant Passages.* After merging the current question to the evolving graph, we enrich the evolving graph with the relevant passages, in which the information is overlooked in existing literature. Specifically, a retriever R is used to retrieve the top-5 passages $P_k = \{p_1, p_2, ..., p_5\}$ from the external collection $\mathcal{P}$ with the rewritten question $\hat{q_k}$. For each passage $p \in P_k$, we independently extract its entities $\mathcal{V}_p$ and relations $\mathcal{E}_p$ using an LLM. Each entity $v \in \mathcal{V}_p$ is characterized as:

$$v = \Big( Name(v), Type(v), Desc(v), ID(p) \Big), \tag{9}$$

where $Name(v)$, $Type(v)$, $Desc(v)$, and $ID(p)$ denote the name, type, description, and source passage ID information of this entity.

Then, each relation $e \in \mathcal{E}_p$ is defined as a similar way:

$$e = \Big( v_s, v_t, w_e, ID(p) \Big), \tag{10}$$

where $v_s, v_t \in \mathcal{V}_p$ represent the source and target entities respectively, $w_e$ assigns a weight to the relation, and $ID(p)$ indicates its source passage ID. The extracted elements of passages are then added to the evolving graph:

$$\mathcal{V} \leftarrow \mathcal{V} \cup \bigcup_{p \in P_k} \mathcal{V}_p, \mathcal{E} \leftarrow \mathcal{E} \cup \bigcup_{p \in P_k} \mathcal{E}_p. \tag{11}$$

*4.4.3 Generate and Append Response.* To generate the conversational response $r_k$, we employ a structured reasoning process over the evolving graph $\mathcal{G}$. Specifically, we treat the current question entities $\mathcal{V}_k^{(q)}$ in the $\mathcal{G}$ as the starting nodes. From the starting nodes, we perform a breadth-first search traversal up to $m$ hops to extract a relevant subgraph:

$$\mathcal{G}_k^{(sub)} = \mathrm{BFS}(\mathcal{G}, \mathcal{V}_k^{(q)}, m), \tag{12}$$

where $\mathcal{G}_k^{(sub)} \subseteq \mathcal{G}$ captures the nodes and edges that are structurally related to the current turn. We identify the passages that appear most frequently as sources in the subgraph, expanding our evidence pool to include both the top-5 retrieved passages $P_k$ and

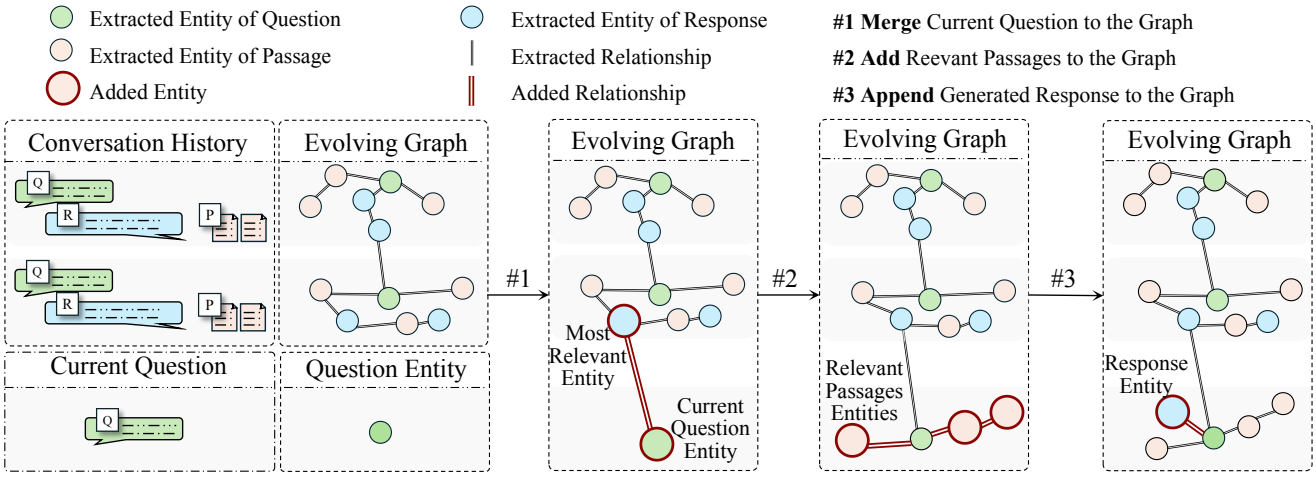


**Figure 3: The example to illustrate the update of the evolving graph with diving into a new turn.**

historically relevant passages that may not have been retrieved in the current turn but are connected through the graph structure:

$$P_k^* = P_k \cup \left\{ p \mid p \text{ is the frequent sources in } \mathcal{G}_k^{(sub)} \right\}. \tag{13}$$

Eventually, the final response $r_k$ is generated by integrating information from the history summary $S_k$, the rewritten question $\hat{q}_k$, and the expanded passage set $P_k^*$:

$$r_k = \mathrm{G}(S_k, \hat{q}_k, P_k^*), \tag{14}$$

where G denotes the generator. New entities $\mathcal{V}_k^{(r)}$ and relations $\mathcal{E}_k^{(r)}$ are subsequently extracted from the response $r_k$ and then appended into the evolving graph as follows:

$$\mathcal{V} \leftarrow \mathcal{V} \cup \mathcal{V}_k^{(r)}, \mathcal{E} \leftarrow \mathcal{E} \cup \mathcal{E}_k^{(r)}. \tag{15}$$

# 5 Experimental Setup

## 5.1 Datasets

We evaluate EvoRAG on four public conversational RAG datasets: (1) TopiOCQA [1]: an open-domain conversational dataset from Wikipedia, featuring multi-turn information-seeking questions with topic shifts; (2) QReCC [3]: a conversational question answering dataset with human rewrites and a diverse web-sourced passage corpus; (3) INSCIT [62]: a mixed-initiative information-seeking dialogue dataset with human-generated interactions; (4) CORAL [8]: a large-scale Wikipedia-based dataset for evaluating conversational RAG systems in realistic multi-turn settings.

## 5.2 Evaluation Metrics

Our evaluation follows task-specific metrics: (1) for conversational passage retrieval, we use standard retrieval metrics including MRR, NDCG@3, and Recall@10, consistent with the previous work [35]; (2) for conversational response generation, we report F1, ROUGE-1, ROUGE-L, following previous work [29, 60].

## 5.3 Baselines

We compare EvoRAG with seven baselines, including four CQR-based approaches that rewrite the current question before single-turn RAG, and three CDR-based approaches that incorporate both conversation history and the current question for retrieval, then feed all retrieved passages and full history to the generator. The CQR-based baselines include: T5QR [27], a T5-based conversational question rewriting model fine-tuned on human rewrites; ConvGQR [41], a hybrid framework combining query rewriting and expansion; LLM4CS [32], which uses GPT-3.5 to generate query rewrites; and CHIQ [39], a two-step LLM-driven method that enhances rewriting through improved history utilization. The CDR-based baselines consist of: ConvDR [68], a conversational dense retrieval method with few-shot learning via a teacher-student framework; InstructorR [21], which leverages LLMs to generate soft supervision signals for training conversational dense retrievers without labeled data; and ChatQA [29], a two-stage instruction fine-tuned model for conversational RAG.

Additionally, we include two reference settings: Vanilla, which uses raw concatenation of conversation history and current question without rewriting, and Human, which uses golden human rewrites for retrieval and generation.

## 5.4 Implementation Details

We use ANCE [63] as the retriever to retrieve relevant passages, use Qwen2.5-7B-Instruct [57] as the backbone to integrate knowledge, and use Llama-3.1-8B-Instruct [12] as the generator to generate conversational responses. To construct the conversation history graph and the passage graph, we directly leverage Qwen2.5-72B-Instruct [57] to extract entities and relations from the conversation history and the passage.

To ensure a fair comparison, for the CQR-based approaches, we only use query rewriting without additional query expansion. As for InstructorR, since their source code is not publicly available and their experimental setup closely aligns with ours, we directly adopt

**Table 1: Retrieval results of EvoRAG and baselines. The best and second-best results are in bold and underlined. The symbol † signifies that our model achieves superior results among baselines in a statistically significant manner (t-test, $p$-value < 0.05). RI-H is the relative improvement over the Human annotation.**

| Method | TopiOCQA | | | QReCC | | | INSCIT | | | CORAL | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | MRR | NDCG@3 | R@10 | MRR | NDCG@3 | R@10 | MRR | NDCG@3 | R@10 | MRR | NDCG@3 | R@10 |
| T5QR | 23.0 | 22.2 | 37.6 | 34.5 | 31.8 | 53.1 | 37.8 | 27.8 | 45.6 | 38.9 | 31.4 | 45.7 |
| ConvGQR | 23.4 | 22.5 | 39.8 | 36.4 | 33.5 | 56.6 | 38.5 | 28.5 | 45.0 | 40.3 | 32.6 | 47.1 |
| LLM4CS | 28.2 | 27.4 | 47.7 | 40.9 | 37.9 | 62.2 | 39.0 | 28.9 | 47.3 | 41.5 | 33.7 | 49.5 |
| CHIQ-FT | 30.0 | 28.9 | **51.0** | 36.9 | 34.0 | 57.6 | N/A | N/A | N/A | N/A | N/A | N/A |
| ConvDR | 27.2 | 26.4 | 43.5 | 38.5 | 35.7 | 58.2 | 39.2 | 28.9 | 47.3 | 40.0 | 32.2 | 49.6 |
| InstructorR | 25.3 | 23.7 | 45.1 | **43.5** | **40.5** | **66.7** | N/A | N/A | N/A | N/A | N/A | N/A |
| EvoRAG | **31.0** | **29.9** | 50.8 | 40.0 | 37.0 | 60.5 | **41.0**$^{\dagger}$ | **30.6**$^{\dagger}$ | **48.7**$^{\dagger}$ | **42.1** | **34.3** | **50.1** |
| Vanilla | 10.3 | 9.1 | 19.1 | 42.5 | 39.8 | 62.6 | 20.4 | 13.6 | 26.2 | 31.4 | 23.9 | 44.1 |
| Human | N/A | N/A | N/A | 38.4 | 35.6 | 58.6 | N/A | N/A | N/A | 43.2 | 35.1 | 51.9 |
| RI-H | N/A | N/A | N/A | +4.2% | +3.9% | +3.2% | N/A | N/A | N/A | -2.5% | -2.3% | -3.5% |

**Table 2: Generation results of EvoRAG and baselines. The best and second-best results are in bold and underlined. The symbol † signifies that our model achieves superior results among baselines in a statistically significant manner (t-test, $p$-value < 0.05). RI-H is the relative improvement over the Human annotation.**

| Method | TopiOCQA | | | QReCC | | | INSCIT | | | CORAL | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | F1 | ROUGE-L | ROUGE-1 | F1 | ROUGE-L | ROUGE-1 | F1 | ROUGE-L | ROUGE-1 | F1 | ROUGE-L | ROUGE-1 |
| T5QR | 22.8 | 21.1 | 22.1 | 21.0 | 21.1 | 23.4 | 25.5 | 23.9 | 26.9 | 24.9 | 20.9 | 23.5 |
| ConvGQR | 21.1 | 19.8 | 20.8 | 22.0 | 22.0 | 24.4 | 25.6 | 24.0 | 26.9 | 25.1 | 20.8 | 23.5 |
| LLM4CS | 24.1 | 22.3 | 23.3 | 22.0 | 22.1 | 24.6 | 25.6 | 24.0 | 27.0 | **25.6** | **21.2** | 23.9 |
| ConvDR | 24.3 | 19.2 | 20.1 | 23.9 | 23.2 | **25.9** | **27.3** | **25.0** | **28.5** | 24.4 | 20.7 | 23.5 |
| ChatQA | 18.1 | 17.2 | 18.4 | 23.7 | 22.3 | 24.9 | 25.7 | 23.7 | 26.9 | 20.3 | 18.6 | 20.9 |
| EvoRAG | **26.8**$^{\dagger}$ | **24.7**$^{\dagger}$ | **25.8**$^{\dagger}$ | **24.0** | **23.3** | **25.9** | 26.8 | **25.0** | 28.3 | 25.1 | **21.2** | **24.0** |
| Vanilla | 20.6 | 19.2 | 20.1 | 24.9 | 24.2 | 26.8 | 23.6 | 21.9 | 24.9 | 22.8 | 20.0 | 22.6 |
| Human | N/A | N/A | N/A | 22.9 | 22.8 | 25.2 | N/A | N/A | N/A | 26.8 | 22.0 | 24.9 |
| RI-H | N/A | N/A | N/A | +4.8% | +2.2% | +2.8% | N/A | N/A | N/A | -6.3% | -3.6% | -3.6% |

their reported results on TopiOCQA and QReCC for comparison. For baselines without their own response generator, we use the same generator as the generator in our EvoRAG framework to ensure comparability. For evaluation, we use golden responses and golden passages from the dialogue history to ensure consistency. Code is released at https://github.com/Ariya12138/EvoRAG.

## 6 Experimental Results and Analysis

In this section, we evaluate the performance of our EvoRAG on four public conversational RAG datasets and provide a detailed analysis.

### 6.1 Main Results

*6.1.1 Retrieval Performance.* Table 1 presents the retrieval results of EvoRAG and baselines across four public conversational RAG datasets. We can make the following observations:

(1) Our EvoRAG demonstrates strong retrieval performance across all datasets, achieving the best results on TopiOCQA, INSCIT, and CORAL. Our results deliver significant improvements over existing baselines, with relative gains of 3.5%, 5.9%, and 1.8% in NDCG@3 on TopiOCQA, INSCIT, and CORAL, respectively when compared to the second-best performing baseline using the same ANCE retriever. On QReCC, although slightly behind InstructorR and LLM4CS, EvoRAG still outperforms other baselines, highlighting its robustness across diverse conversational settings. The superior performance of our EvoRAG can be attributed to two key architectural innovations: (i) The dual-path retrieval module combines both semantic and structural awareness to comprehensively filter noise in conversation history; (ii) The evolving graph provides a structured representation of the conversational context that effectively captures relations among conversational turns, which is particularly valuable in scenarios involving topic shifts and long conversational contexts.

(2) The performance gap between CDR and CQR methods further highlights EvoRAG's adaptability to diverse conversational scenarios. CDR baselines like InstructorR perform well on QReCC, where conversations focus on a single topic and history noise is

**Table 3: Ablation studies for EvoRAG. The best and second best results are in bold and underlined. R-1 is short for ROUGE-1.**

| Method | TopiOCQA | | | QReCC | | | INSCIT | | | CORAL | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | F1 | ROUGE-L | R-1 | F1 | ROUGE-L | R-1 | F1 | ROUGE-L | R-1 | F1 | ROUGE-L | R-1 |
| EvoRAG | **26.8** | **24.7** | **25.8** | **24.0** | **23.3** | **25.9** | **26.8** | 25.0 | **28.3** | 25.1 | 21.2 | 24.0 |
| *w/o* Expanding Passages | 26.3 | 24.4 | 25.5 | 22.4 | 22.2 | 24.7 | 26.2 | 24.5 | 27.6 | 25.5 | 21.4 | **24.4** |
| *w/o* History Summary | 26.4 | 24.6 | 25.6 | 23.4 | **23.3** | 25.8 | **26.8** | **25.3** | **28.3** | **25.9** | **21.6** | **24.4** |

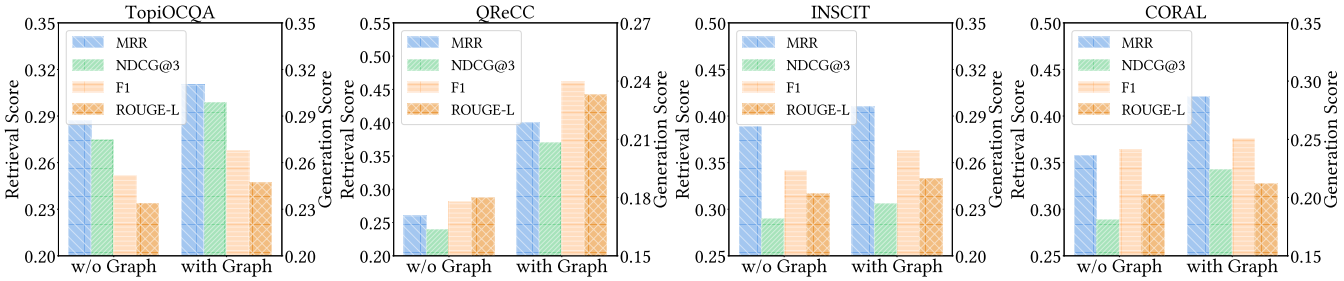


**Figure 4: Experimental results for the impact of the graph.**

minimal. However, their reliance on raw history concatenation becomes a limitation in multi-topic datasets, where irrelevant turns introduce noise. Conversely, while CQR methods can mitigate noise through query rewriting, they often lose valuable context information. EvoRAG overcomes both limitations through its dual-path retrieval and evolving graph, achieving strong performance in both single-topic and complex multi-topic conversations.

*6.1.2 Generated Response Quality.* Table 2 presents the generation results of EvoRAG and baselines across four public conversational RAG datasets. We have the following observations:

(1) EvoRAG consistently achieves strong generation performance across all datasets, which aligns well with its retrieval performance. Overall, better retrieval quality tends to translate into better generation results, with relative gains of 10.3% and 0.4% in F1 on TopiOCQA and QReCC, respectively when compared to the second-best performing baseline. However, on the INSCIT and CORAL datasets, despite EvoRAG achieving superior retrieval performance, its generation results are comparable to the second-best performing baseline. This discrepancy may be attributed to limitations in the rewritten questions and the summary of the conversation history, which could hinder the generator from fully capturing the user's intent and thereby constrain the quality of the generated responses.

(2) Compared to the CQR-based approach, we observe that the CDR-based approach demonstrates unexpectedly strong performance when the generator receives both the retrieved passages along with the conversation history and the current question. For instance, while ConvDR shows inferior retrieval performance on datasets such as TopiOCQA, INSCIT, and CORAL, it achieves notably high generation quality, with F1 scores of 24.3, 27.3, and 24.4, respectively. We attribute this to its ability to leverage the conversation history, which helps the generator better understand the user's current query intent and generate more coherent responses.

## 6.2 Further Analysis

*6.2.1 Ablation Study.* To better understand the contributions of each component within EvoRAG, we conduct a comprehensive ablation study on the four public conversational RAG datasets. From the ablation results of the Table 3, we can see: (1) Removing the passage expansion module, which leverages the evolving graph to incorporate historically relevant passages, results in F1 drops of 0.5, 1.6, and 0.6 on TopiOCQA, QReCC, and INSCIT, respectively. These declines underscore the value of integrating useful passages from earlier turns for maintaining topical continuity and improving response quality. Interestingly, on CORAL, the absence of passage expansion slightly improves F1 (from 25.1 to 25.5), likely because its conversations are more self-contained, reducing the need for historical passage expansion. (2) Removing the history summary, which synthesizes candidate history and candidate entities, leads to notable F1 drops of 0.4 and 0.6 on TopiOCQA and QReCC, respectively. This highlights the history summary's dual function: maintaining dialogue flow for coherence and preserving essential factual information for generating accurate responses. In contrast, for INSCIT and CORAL, incorporating the history summary offers limited or even negative effects. This discrepancy is likely attributable to the lower quality of the generated summaries, which may introduce noise or distort relevant context, thereby impeding rather than aiding downstream response generation.

*6.2.2 Impact of the Graph.* Figure 4 illustrates the impact of the evolving knowledge graph on the performance of EvoRAG across four datasets. The graph-based approach demonstrates significant improvements in both retrieval and generation metrics, highlighting the importance of the evolving graph in modeling conversational context. In the retrieval task, the graph notably enhances performance by effectively tracking the dialogue flow and identifying topically relevant information, especially in scenarios where topic drift and history noise are prominent. In generation tasks, the

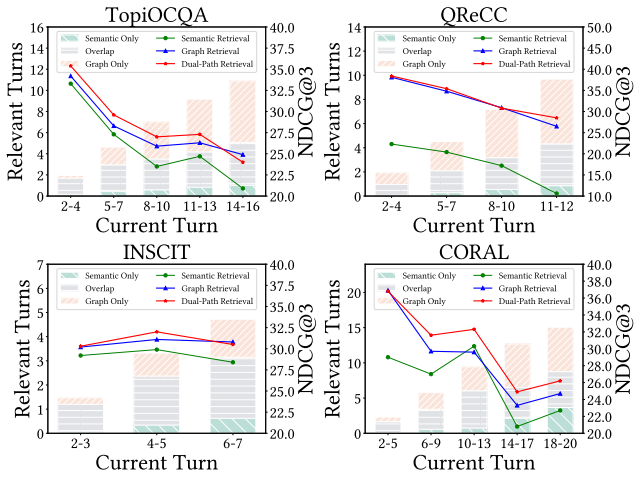


**Figure 5: Analysis of Dual-Path Retrieval. The histogram shows the relevant history turns retrieved by each path, and the curve shows retrieval performance using those turns.**

graph contributes to better generation performance by enabling the model to focus on relevant passages, ensuring coherent and contextually grounded responses. The graph's ability to denoise history and dynamically expand evidence makes it particularly effective in long-context conversations, addressing the limitations of unstructured conversational context.

*6.2.3 Analysis of the Dual-Path Retrieval.* Our analysis of the Dual-Path Retrieval, as shown in Figure 5, reveals distinct yet complementary patterns in conversational context modeling.

In the earlier conversation turns, both the Semantic-Aware Turn Retrieval and Graph-Based Entity Extraction paths show strong alignment in selecting relevant historical turns. This convergence occurs because the brief conversation history contains clear semantic cues without significant topic shifts, allowing both methods to effectively identify dependencies. Once the conversation reaches its later stages, the two paths begin to diverge noticeably. As the conversation flows, accumulated noise and topic shifts create ambiguity that challenges purely semantic retrieval. While semantic similarity becomes less reliable, the graph-based path maintains robustness by tracking relations among conversational turns through the evolving graph, prioritizing logically central nodes to preserve important information despite topic shifts. The findings highlight the complementary nature of semantic retrieval and graph retrieval: the former ensures immediate relevance, while the latter maintains cross-turn consistency. Their integration allows EvoRAG to outperform any single retrieval path, demonstrating the advantage of hybrid context modeling in conversational RAG.

*6.2.4 Analysis of the Expanded Passages.* Conversational context exhibits a key property: there exists a significant overlap between the golden passages of the current turn and those from previous historical turns. Table 4 clearly illustrates this, showing overlap rates of 22.0%, 38.6%, 23.8%, and 9.6% across the four datasets. By exploiting this characteristic, our graph-based passage expansion mechanism is able to recover relevant passages from the earlier turns that the retriever often misses. Our results show that in 11.8%, 13.7%, 3.1%, and 1.5% of turns in TopiOCQA, QReCC, INSCIT, and CORAL, respectively, the evolving graph successfully helps find historically significant passages that are missed by the current turn retrieval but align perfectly with the current query intent. These findings strongly demonstrate the necessity of structured modeling of conversational context to fully capture and utilize the relevant historical information. By maintaining and leveraging the evolving graph, EvoRAG addresses two key challenges simultaneously: it reduces the risk of missing persistently relevant passages while enhancing the coherence of generated responses through the structured representation of the conversational context.

**Table 4: Analysis of expanded passages. # Queries denotes test turns. # Qrels denotes turns with golden passages. # Repeated Qrels denotes turns with overlapping golden passages from historical turns. # Successful denotes turns where graph-based passage expansion recovers missed golden passages**

| Dataset | # Queries | # Qrels | # Overlap Qrels | # Successful |
|---|---|---|---|---|
| TopiOCQA | 2514 | 2514 | 554 (22.0%) | 297 (11.8%) |
| QReCC | 16451 | 8209 | 3166 (38.6%) | 1126 (13.7%) |
| INSCIT | 2767 | 2706 | 644 (23.8%) | 85 (3.1%) |
| CORAL | 6480 | 3778 | 362 (9.6%) | 58 (1.5%) |

## 7 Conclusion and Future Work

In this paper, we introduce EvoRAG, a novel framework that addresses the challenges of multi-turn conversational RAG by dynamically maintaining an evolving knowledge graph to model the logical relations among user queries, system responses, and relevant passages across conversational turns. EvoRAG's three core components—dual-path retrieval, unified knowledge integration, and graph-enhanced retrieval and generation—collectively enable precise context denoising, accurate retrieval, and coherent response generation. The success of EvoRAG highlights the importance of structured representation in conversational RAG, providing both a scalable architecture and an interpretable approach for improving retrieval performance and response quality. Future work could explore further optimizations for real-time graph updates and broader applications in dynamic dialogue systems.

## Acknowledgments

This work was supported by Beijing Natural Science Foundation No. L233008, Beijing Municipal Science and Technology Project No. Z231100010323009, National Science and Technology Major Project No. 2022ZD0120103, National Natural Science Foundation of China No. 62272467. The work was partially done at the Beijing Key Laboratory of Research on Large Models and Intelligent Governance.

## GenAI Usage Disclosure

In this work, generative AI tools are used selectively to enhance productivity while ensuring research integrity. Simple coding tasks are automated with generative AI tools to improve efficiency. All contributions of generative AI tools remain under our full oversight.

## References


[1] 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 (2022), 468–483.

[2] Moonshot AI. 2023. Kimi chat. https://kimi.moonshot.cn/

[3] Raviteja Anantha, Svitlana Vakulenko, Zhucheng Tu, Shayne Longpre, Stephen Pulman, and Srinivas Chappidi. 2021. Open-Domain Question Answering Goes Conversational via Question Rewriting. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tür, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (Eds.). Association for Computational Linguistics, 520–534. doi:10.18653/V1/2021.NAACL-MAIN.44

[4] Rong-Ching Chang and Jiawei Zhang. 2024. CommunityKG-RAG: Leveraging Community Structures in Knowledge Graphs for Advanced Retrieval-Augmented Generation in Fact-Checking. *CoRR* abs/2408.08535 (2024). arXiv:2408.08535 doi:10.48550/ARXIV.2408.08535

[5] Haonan Chen, Zhicheng Dou, Kelong Mao, Jiongnan Liu, and Ziliang Zhao. 2024. Generalizing Conversational Dense Retrieval via LLM-Cognition Data Augmentation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, 2700–2718. https://aclanthology.org/2024.acl-long.149

[6] Zhiyu Chen, Jie Zhao, Anjie Fang, Besnik Fetahu, Oleg Rokhlenko, and Shervin Malmasi. 2022. Reinforced Question Rewriting for Conversational Question Answering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: EMNLP 2022 - Industry Track, Abu Dhabi, UAE, December 7 - 11, 2022*, Yunyao Li and Angeliki Lazaridou (Eds.). Association for Computational Linguistics, 357–370. doi:10.18653/V1/2022.EMNLP-INDUSTRY.36

[7] Yiruo Cheng, Kelong Mao, and Zhicheng Dou. 2024. Interpreting Conversational Dense Retrieval by Rewriting-Enhanced Inversion of Session Embedding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, 2879–2893. doi:10.18653/V1/2024.ACL-LONG.159

[8] 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. *CoRR* abs/2410.23090 (2024). arXiv:2410.23090 doi:10.48550/ARXIV.2410.23090

[9] Zhuyun Dai, Arun Tejasvi Chaganty, Vincent Y. Zhao, Aida Amini, Qazi Mamunur Rashid, Mike Green, and Kelvin Guu. 2022. Dialog Inpainting: Turning Documents into Dialogs. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA (Proceedings of Machine Learning Research, Vol. 162)*, Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (Eds.). PMLR, 4558–4586. https://proceedings.mlr.press/v162/dai22a.html

[10] Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of Wikipedia: Knowledge-Powered Conversational Agents. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net. https://openreview.net/forum?id=r1l73iRqKm

[11] Guanting Dong, Yutao Zhu, Chenghao Zhang, Zechen Wang, Ji-Rong Wen, and Zhicheng Dou. 2025. Understand What LLM Needs: Dual Preference Alignment for Retrieval-Augmented Generation. In *Proceedings of the ACM on Web Conference 2025, WWW 2025, Sydney, NSW, Australia, 28 April 2025- 2 May 2025*, Guodong Long, Michale Blumestein, Yi Chang, Liane Lewin-Eytan, Zi Helen Huang, and Elad Yom-Tov (Eds.). ACM, 4206–4225. doi:10.1145/3696410.3714717

[12] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. 2024. The Llama 3 Herd of Models. *CoRR* abs/2407.21783 (2024). arXiv:2407.21783 doi:10.48550/ARXIV.2407.21783

[13] Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. 2024. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130* (2024).

[14] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. 2023. Retrieval-Augmented Generation for Large Language Models: A Survey. *CoRR* abs/2312.10997 (2023). arXiv:2312.10997 doi:10.48550/ARXIV.2312.10997

[15] Bernal Jimenez Gutierrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. 2024. HippoRAG: Neurobiologically Inspired Long-Term Memory for Large Language Models. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (Eds.). http://papers.nips.cc/paper_files/paper/2024/hash/6ddc001d07ca4f319af96a3024f6dbd1-Abstract-Conference.html

[16] Yizheng Huang and Jimmy Huang. 2024. A Survey on Retrieval-Augmented Text Generation for Large Language Models. *CoRR* abs/2404.10981 (2024). arXiv:2404.10981 doi:10.48550/ARXIV.2404.10981

[17] Yunah Jang, Kang-il Lee, Hyunkyung Bae, Seungpil Won, Hwanhee Lee, and Kyomin Jung. 2023. IterCQR: Iterative Conversational Query Reformulation without Human Supervision. *CoRR* abs/2311.09820 (2023). arXiv:2311.09820 doi:10.48550/ARXIV.2311.09820

[18] Jiajie Jin, Xiaoxi Li, Guanting Dong, Yuyao Zhang, Yutao Zhu, Yongkang Wu, Zhonghua Li, Ye Qi, and Zhicheng Dou. 2025. Hierarchical Document Refinement for Long-context Retrieval-augmented Generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, 3502–3520. https://aclanthology.org/2025.acl-long.176/

[19] Jiajie Jin, Yutao Zhu, Zhicheng Dou, Guanting Dong, Xinyu Yang, Chenghao Zhang, Tong Zhao, Zhao Yang, and Ji-Rong Wen. 2025. FlashRAG: A Modular Toolkit for Efficient Retrieval-Augmented Generation Research. In *Companion Proceedings of the ACM on Web Conference 2025, WWW 2025, Sydney, NSW, Australia, 28 April 2025 - 2 May 2025*, Guodong Long, Michale Blumestein, Yi Chang, Liane Lewin-Eytan, Zi Helen Huang, and Elad Yom-Tov (Eds.). ACM, 737–740. doi:10.1145/3701716.3715313

[20] Jiajie Jin, Yutao Zhu, Yujia Zhou, and Zhicheng Dou. 2024. BIDER: Bridging Knowledge Inconsistency for Efficient Retrieval-Augmented LLMs via Key Supporting Evidence. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, 750–761. https://aclanthology.org/2024.findings-acl.42

[21] Zhuoran Jin, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. 2023. InstructoR: Instructing Unsupervised Conversational Dense Retrieval with Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, 6649–6675. doi:10.18653/V1/2023.FINDINGS-EMNLP.443

[22] Byeongchang Kim, Jaewoo Ahn, and Gunhee Kim. 2020. Sequential Latent Knowledge Selection for Knowledge-Grounded Dialogue. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net. https://openreview.net/forum?id=Hke0K1HKwr

[23] Vaibhav Kumar and Jamie Callan. 2020. Making Information Seeking Easier: An Improved Pipeline for Conversational Search. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020 (Findings of ACL, Vol. EMNLP 2020)*, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, 3971–3980. doi:10.18653/V1/2020.FINDINGS-EMNLP.354

[24] Xiaoxi Li, Jiajie Jin, Yujia Zhou, Yongkang Wu, Zhonghua Li, Ye Qi, and Zhicheng Dou. 2025. RetroLLM: Empowering Large Language Models to Retrieve Fine-grained Evidence within Generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, 16754–16779. https://aclanthology.org/2025.acl-long.819/

[25] Lei Liang, Mengshu Sun, Zhengke Gui, Zhongshu Zhu, Zhouyu Jiang, Ling Zhong, Yuan Qu, Peilong Zhao, Zhongpu Bo, Jin Yang, Huaidong Xiong, Lin Yuan, Jun Xu, Zaoyang Wang, Zhiqiang Zhang, Wen Zhang, Huajun Chen, Wenguang Chen, and Jun Zhou. 2024. KAG: Boosting LLMs in Professional Domains via Knowledge Augmented Generation. *CoRR* abs/2409.13731 (2024). arXiv:2409.13731 doi:10.48550/ARXIV.2409.13731

[26] 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, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, 1004–1015. doi:10.18653/V1/2021.EMNLP-MAIN.77

[27] Sheng-Chieh Lin, Jheng-Hong Yang, Rodrigo Frassetto Nogueira, Ming-Feng Tsai, Chuan-Ju Wang, and Jimmy Lin. 2020. Conversational Question Reformulation via Sequence-to-Sequence Architectures and Pretrained Language Models. *CoRR* abs/2004.01909 (2020). arXiv:2004.01909 https://arxiv.org/abs/2004.01909

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

[29] Zihan Liu, Wei Ping, Rajarshi Roy, Peng Xu, Chankyu Lee, Mohammad Shoeybi, and Bryan Catanzaro. 2024. ChatQA: Surpassing GPT-4 on Conversational QA and RAG. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (Eds.). http://papers.nips.cc/paper_files/paper/2024/hash/1c0d54ebd0a6e58c4eca7d591e374b9d-Abstract-Conference.html

[30] 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. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, 1227–1240. https://aclanthology.org/2024.emnlp-main.71

[31] Kelong Mao, Zhicheng Dou, Bang Liu, Hongjin Qian, Fengran Mo, Xiangli Wu, Xiaohua Cheng, and Zhao Cao. 2023. Search-Oriented Conversational Query Editing. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, 4160–4172. doi:10.18653/V1/2023.FINDINGS-ACL.256

[32] Kelong Mao, Zhicheng Dou, Fengran Mo, Jiewen Hou, Haonan Chen, and Hongjin Qian. 2023. Large Language Models Know Your Contextual Search Intent: A Prompting Framework for Conversational Search. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, 1211–1225. doi:10.18653/V1/2023.FINDINGS-EMNLP.86

[33] Kelong Mao, Zhicheng Dou, and Hongjin Qian. 2022. Curriculum Contrastive Context Denoising for Few-shot Conversational Dense Retrieval. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, Enrique Amigó, Pablo Castells, Julio Gonzalo, Ben Carterette, J. Shane Culpepper, and Gabriella Kazai (Eds.). ACM, 176–186. doi:10.1145/3477495.3531961

[34] 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, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, 2935–2946. doi:10.18653/V1/2022.EMNLP-MAIN.190

[35] Kelong Mao, Hongjin Qian, Fengran Mo, Zhicheng Dou, Bang Liu, Xiaohua Cheng, and Zhao Cao. 2023. Learning Denoised and Interpretable Session Representation for Conversational Search. In *Proceedings of the ACM Web Conference 2023, WWW 2023, Austin, TX, USA, 30 April 2023 - 4 May 2023*, Ying Ding, Jie Tang, Juan F. Sequeda, Lora Aroyo, Carlos Castillo, and Geert-Jan Houben (Eds.). ACM, 3193–3202. doi:10.1145/3543507.3583265

[36] 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, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, Jimmy X. Huang, Yi Chang, Xueqi Cheng, Jaap Kamps, Vanessa Murdock, Ji-Rong Wen, and Yiqun Liu (Eds.). ACM, 1151–1160. doi:10.1145/3397271.3401097

[37] 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*. 64–74.

[38] Fengran Mo, Yifan Gao, Chuan Meng, Xin Liu, Zhuofeng Wu, Kelong Mao, Zhengyang Wang, Pei Chen, Zheng Li, Xian Li, Bing Yin, and Meng Jiang. 2025. UniConv: Unifying Retrieval and Response Generation for Large Language Models in Conversations. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, 6936–6949. https://aclanthology.org/2025.acl-long.344/

[39] Fengran Mo, Abbas Ghaddar, Kelong Mao, Mehdi Rezagholizadeh, Boxing Chen, Qun Liu, and Jian-Yun Nie. 2024. CHIQ: Contextual History Enhancement for Improving Query Rewriting in Conversational Search. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, 2253–2268. https://aclanthology.org/2024.emnlp-main.135

[40] Fengran Mo, Kelong Mao, Ziliang Zhao, Hongjin Qian, Haonan Chen, Yiruo Cheng, Xiaoxi Li, Yutao Zhu, Zhicheng Dou, and Jian-Yun Nie. 2024. A Survey of Conversational Search. *arXiv preprint arXiv:2410.15576* (2024).

[41] Fengran Mo, Kelong Mao, Yutao Zhu, Yihong Wu, Kaiyu Huang, and Jian-Yun Nie. 2023. ConvGQR: Generative Query Reformulation for Conversational Search. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, 4998–5012. doi:10.18653/V1/2023.ACL-LONG.274

[42] 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, SIGIR 2025, Padua, Italy, July 13-18, 2025*, Nicola Ferro, Maria Maistro, Gabriella Pasi, Omar Alonso, Andrew Trotman, and Suzan Verberne (Eds.). ACM, 4094–4097. doi:10.1145/3726302.3731686

[43] Fengran Mo, Jian-Yun Nie, Kaiyu Huang, Kelong Mao, Yutao Zhu, Peng Li, and Yang Liu. 2023. Learning to Relate to Previous Turns in Conversational Search. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2023, Long Beach, CA, USA, August 6-10, 2023*, Ambuj K. Singh, Yizhou Sun, Leman Akoglu, Dimitrios Gunopulos, Xifeng Yan, Ravi Kumar, Fatma Ozcan, and Jieping Ye (Eds.). ACM, 1722–1732. doi:10.1145/3580305.3599411

[44] Fengran Mo, Chen Qu, Kelong Mao, Yihong Wu, Zhan Su, Kaiyu Huang, and Jian-Yun Nie. 2024. Aligning Query Representation with Rewritten Query and Relevance Judgments in Conversational Search. *CoRR* abs/2407.20189 (2024). arXiv:2407.20189 doi:10.48550/ARXIV.2407.20189

[45] Fengran Mo, Chen Qu, Kelong Mao, Tianyu Zhu, Zhan Su, Kaiyu Huang, and Jian-Yun Nie. 2024. History-Aware Conversational Dense Retrieval. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, 13366–13378. https://aclanthology.org/2024.findings-acl.792

[46] Fengran Mo, Bole Yi, Kelong Mao, Chen Qu, Kaiyu Huang, and Jian-Yun Nie. 2024. ConvSDG: Session Data Generation for Conversational Search. In *Companion Proceedings of the ACM on Web Conference 2024, WWW 2024, Singapore, Singapore, May 13-17, 2024*, Tat-Seng Chua, Chong-Wah Ngo, Roy Ka-Wei Lee, Ravi Kumar, and Hady W. Lauw (Eds.). ACM, 1634–1642. doi:10.1145/3589335.3651940

[47] Fengran Mo, Jinghan Zhang, Yuchen Hui, Jia Ao Sun, Zhichao Xu, Zhan Su, and Jian-Yun Nie. 2025. ConvMix: A Mixed-Criteria Data Augmentation Framework for Conversational Dense Retrieval. *arXiv preprint arXiv:2508.04001* (2025).

[48] OpenAI. 2022. OpenAI: Introducing ChatGPT. https://openai.com/index/chatgpt/

[49] Hongjin Qian and Zhicheng Dou. 2022. Explicit Query Rewriting for Conversational Dense Retrieval. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, 4725–4737. doi:10.18653/V1/2022.EMNLP-MAIN.311

[50] Hongjin Qian, Zheng Liu, Chao Gao, Yankai Wang, Defu Lian, and Zhicheng Dou. 2025. HawkBench: Investigating Resilience of RAG Methods on Stratified Information-Seeking Tasks. *CoRR* abs/2502.13465 (2025). arXiv:2502.13465 doi:10.48550/ARXIV.2502.13465

[51] Hongjin Qian, Zheng Liu, Peitian Zhang, Kelong Mao, Defu Lian, Zhicheng Dou, and Tiejun Huang. 2025. MemoRAG: Boosting Long Context Processing with Global Memory-Enhanced Retrieval Augmentation. In *Proceedings of the ACM on Web Conference 2025, WWW 2025, Sydney, NSW, Australia, 28 April 2025- 2 May 2025*, Guodong Long, Michale Blumestein, Yi Chang, Liane Lewin-Eytan, Zi Helen Huang, and Elad Yom-Tov (Eds.). ACM, 2366–2377. doi:10.1145/3696410.3714805

[52] Filip Radlinski and Nick Craswell. 2017. A Theoretical Framework for Conversational Search. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval, CHIIR 2017, Oslo, Norway, March 7-11, 2017*, Ragnar Nordlie, Nils Pharo, Luanne Freund, Birger Larsen, and Dan Russel (Eds.). ACM, 117–126. doi:10.1145/3020165.3020183

[53] Nirmal Roy, Leonardo F. R. Ribeiro, Rexhina Blloshmi, and Kevin Small. 2024. Learning When to Retrieve, What to Rewrite, and How to Respond in Conversational QA. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, 10604–10625. https://aclanthology.org/2024.findings-emnlp.622

[54] Bhaskarjit Sarmah, Dhagash Mehta, Benika Hall, Rohan Rao, Sunil Patel, and Stefano Pasquali. 2024. HybridRAG: Integrating Knowledge Graphs and Vector

Retrieval Augmented Generation for Efficient Information Extraction. In *Proceedings of the 5th ACM International Conference on AI in Finance, ICAIF 2024, Brooklyn, NY, USA, November 14-17, 2024*. ACM, 608–616. doi:10.1145/3677052.3698671

[55] Jiejun Tan, Zhicheng Dou, Wen Wang, Mang Wang, Weipeng Chen, and Ji-Rong Wen. 2025. HtmlRAG: HTML is Better Than Plain Text for Modeling Retrieved Knowledge in RAG Systems. In *Proceedings of the ACM on Web Conference 2025, WWW 2025, Sydney, NSW, Australia, 28 April 2025- 2 May 2025*, Guodong Long, Michale Blumestein, Yi Chang, Liane Lewin-Eytan, Zi Helen Huang, and Elad Yom-Tov (Eds.). ACM, 1733–1746. doi:10.1145/3696410.3714546

[56] ByteDance Doubao Team. 2023. Doubao. https://www.doubao.com/chat/

[57] Qwen Team. 2024. Qwen2.5: A Party of Foundation Models. https://qwenlm.github.io/blog/qwen2.5/

[58] Svitlana Vakulenko, Shayne Longpre, Zhucheng Tu, and Raviteja Anantha. 2021. Question Rewriting for Conversational Question Answering. In *WSDM '21, The Fourteenth ACM International Conference on Web Search and Data Mining, Virtual Event, Israel, March 8-12, 2021*, Liane Lewin-Eytan, David Carmel, Elad Yom-Tov, Eugene Agichtein, and Evgeniy Gabrilovich (Eds.). ACM, 355–363. doi:10.1145/3437963.3441748

[59] Nikos Voskarides, Dan Li, Pengjie Ren, Evangelos Kanoulas, and Maarten de Rijke. 2020. Query Resolution for Conversational Search with Limited Supervision. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, Jimmy X. Huang, Yi Chang, Xueqi Cheng, Jaap Kamps, Vanessa Murdock, Ji-Rong Wen, and Yiqun Liu (Eds.). ACM, 921–930. doi:10.1145/3397271.3401130

[60] Yujing Wang, Hainan Zhang, Liang Pang, Binghui Guo, Hongwei Zheng, and Zhiming Zheng. 2025. MaFeRw: Query Rewriting with Multi-Aspect Feedbacks for Retrieval-Augmented Large Language Models. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, Toby Walsh, Julie Shah, and Zico Kolter (Eds.). AAAI Press, 25434–25442. doi:10.1609/AAAI.V39I24.34732

[61] Zeqiu Wu, Yi Luan, Hannah Rashkin, David Reitter, Hannaneh Hajishirzi, Mari Ostendorf, and Gaurav Singh Tomar. 2022. CONQRR: Conversational Query Rewriting for Retrieval with Reinforcement Learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, 10000–10014. doi:10.18653/V1/2022.EMNLP-MAIN.679

[62] 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 (2023), 453–468. doi:10.1162/TACL_A_00559

[63] Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. 2021. Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net. https://openreview.net/forum?id=zeFrfgyZln

[64] Fanghua Ye, Meng Fang, Shenghui Li, and Emine Yilmaz. 2023. Enhancing Conversational Search: Large Language Model-Aided Informative Query Rewriting. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, 5985–6006. doi:10.18653/V1/2023.FINDINGS-EMNLP.398

[65] 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. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024*, Grace Hui Yang, Hongning Wang, Sam Han, Claudia Hauff, Guido Zuccon, and Yi Zhang (Eds.). ACM, 2301–2305. doi:10.1145/3626772.3657980

[66] Chanwoong Yoon, Gangwoo Kim, Byeongguk Jeon, Sungdong Kim, Yohan Jo, and Jaewoo Kang. 2024. Ask Optimal Questions: Aligning Large Language Models with Retriever's Preference in Conversational Search. *CoRR* abs/2402.11827 (2024). arXiv:2402.11827 doi:10.48550/ARXIV.2402.11827

[67] Shi Yu, Jiahua Liu, Jingqin Yang, Chenyan Xiong, Paul N. Bennett, Jianfeng Gao, and Zhiyuan Liu. 2020. Few-Shot Generative Conversational Query Rewriting. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, Jimmy X. Huang, Yi Chang, Xueqi Cheng, Jaap Kamps, Vanessa Murdock, Ji-Rong Wen, and Yiqun Liu (Eds.). ACM, 1933–1936. doi:10.1145/3397271.3401323

[68] Shi Yu, Zhenghao Liu, Chenyan Xiong, Tao Feng, and Zhiyuan Liu. 2021. Few-Shot Conversational Dense Retrieval. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, Fernando Diaz, Chirag Shah, Torsten Suel, Pablo Castells, Rosie Jones, and Tetsuya Sakai (Eds.). ACM, 829–838. doi:10.1145/3404835.3462856

[69] 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. *CoRR* abs/2308.07107 (2023). arXiv:2308.07107 doi:10.48550/ARXIV.2308.07107