Generative History Augmentation for Context-Aware Dense Retrieval in Conversational Search

Anonymous ACL submission

Abstract

Conversational search needs to accurately understand the actual search intent in multi-turn interactions to retrieve relevant passages. Traditional conversational query rewriting methods primarily rely on manually rewritten queries. In contrast, conversational dense retrieval approaches directly utilize the entire conversation context as input, which introduces redundant noise and is further constrained by the limited availability of human-annotated supervisory signals in the dataset. To address these limitations, we propose the **G**enerative **H**istory Augmentation for Context-**A**ware **D**ense **R**etrieval (**GHADR**) system. Initially, we propose an iterative prompt refinement mechanism to leverage large language models (LLMs) to augment the conversation history and generate high-quality rewritten queries. Subsequently, we implement a semantically guided clustering algorithm to mine additional supervision signals for model training. Finally, we train a context-aware passage retriever using both the rewritten queries and the extracted signals from historical turns. Experiments on four public conversational search datasets demonstrate the effectiveness of GHADR in improving retrieval performance and reducing reliance on human-annotated signals.

1 Introduction

Conversational search enables users to engage in multi-turn interactions to satisfy their information needs by retrieving relevant passages from a collection of passages, based on the current query and its conversation history that including previous queries and responses (Kim and Kim, 2022). Unlike traditional single-turn ad-hoc retrieval, which relies primarily on keyword and phrase matching, conversational search requires modeling the whole conversation context to accurately capture the underlying search intent, as this intent may be distributed across the entire conversation history (Yu et al., 2020; Qian and Dou, 2022; (Mo et al., 2024b)). Therefore, conversational search is much more challenging than ad-hoc retrieval. Existing methods can be roughly categorized into two groups: *Conversational Query Rewriting (CQR)* and *Conversational Dense Retrieval (CDR)*, as illustrated in Figure 1.

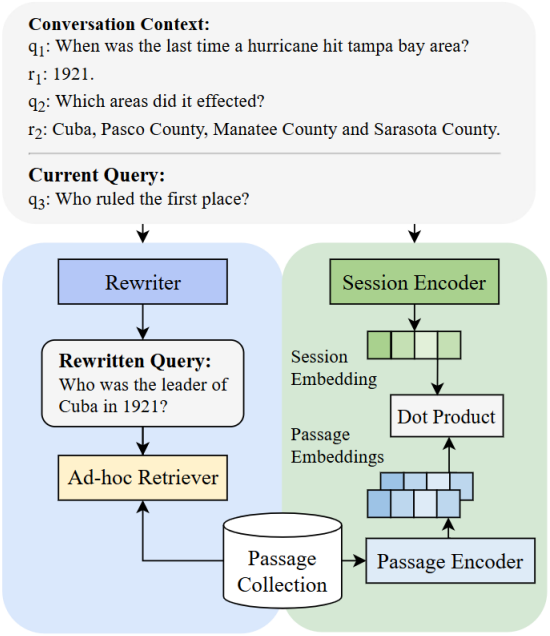

Figure 1: A conceptual illustration for the CQR and CDR.

To capture the real information needs in multi-turn conversation, CQR aims to reformulate conversational queries into stand-alone queries that can be submitted to any off-the-shelf retrievers (Vakulenko et al., 2021; Fang et al., 2022). Previous studies often fine-tune a pre-trained language model, such as T5 (Chung et al., 2024). However, these methods rely on manually rewritten queries as supervision signals to train the rewrite model, yet obtaining large-scale manually annotated data for

training remains challenging in practice. Furthermore, this rewrite-then-retrieve pipeline prevents CQR models from being directly optimized for downstream retrieval tasks (Wu et al., 2022; Mo et al., 2023a), as the two-stage process hinders end-to-end training.

In contrast, CDR leverages a pre-trained ad-hoc retriever to encode the entire conversation context and candidate passages into a unified embedding space, followed by end-to-end fine-tuning on conversational data (Yu et al., 2021; Lin et al., 2021b; Mao et al., 2022). The end-to-end CDR models can be directly optimized for better retrieval performance (Cheng et al., 2024). However, previous studies often treat the entire conversation context as input, while prior queries and responses in the conversation may be ambiguous or irrelevant to the current query. These approaches inevitably introducing noise into the training process of CDR models (Ye et al., 2023). Moreover, fine-tuning the retriever typically requires a large volume of labeled context-passage pairs. In practice, however, obtaining accurate annotations for such pairs is significantly more challenging than collecting conversational data itself.

To tackle these problems, we propose **G**enerative **H**istory Augmentation for Context-**A**ware **D**ense **R**etrieval (GHADR), a novel method that integrates the strengths of both CQR and CDR. Specifically, GHADR adopts the CQR framework to reduce ambiguities in the conversation history and reformulate the query with complete information, thereby reducing the introduction of unwanted noise. Simultaneously, it inherits the end-to-end characteristics of CDR to optimize retrieval performance in conversational search scenarios.

There are three key components in our proposed GHADR. Initially, we leverage the strong language understanding and text generation capabilities of large language models (LLMs) to resolve contextual ambiguities in conversation history, enhancing both the informativeness of historical context and the quality of the generated search query. Subsequently, based on the augmented conversation history and the rewritten query, we employ a semantic-guided clustering algorithm to mine additional supervision signals. This component effectively addresses the challenge of data scarcity in retriever training. Finally, we jointly incorporate the rewritten query and the extracted historical supervision signals into the contrastive learning framework, strengthening the retriever's implicit context modeling capabilities.

Our contributions are summarized as follows:

- We develop an iterative prompt framework to augment conversation history, and then employ a semantic guidance method to mine additional supervision signals.
- We innovatively propose GHADR to train a context-aware conversational passage retriever by leveraging supervision signals mined from historical turns. It manages to comprehensively improve the effectiveness of conversational dense passage retrievers.
- Extensive experiments on four publicly available datasets show the effectiveness of the proposed GHADR. Our analysis reveals the complementary effects of all components within the proposed method.

2 Related Work

Conversational Dense Retrieval. CDR (Yu et al., 2021; Qian and Dou, 2022; Jeong et al., 2023; Huang et al., 2023; Mo et al., 2024d) leverages conversational search sessions to fine-tune an end-to-end, ad-hoc retriever that enables encoding sessions into embedding space for dense retrieval. Considering that the context of the entire conversation may be lengthy and contains a significant amount of noise, some studies (Lin et al., 2021b; Mao et al., 2022; Mao et al., 2024) design sophisticated context-denoising approaches for better CDR models.

While recent approaches (Mo et al., 2023b; Mo et al., 2024c) have demonstrated strong performance by leveraging actual retrieval outcomes as relevance indicators, we highlight potential deployment limitations in practical scenarios where historical ground-truth annotations are unavailable. Therefore, we propose a method that explicitly selects semantically relevant conversational turns as additional supervision signals. Furthermore, we incorporate the supervision signals derived from historical ground-truth passages to enhance the training of the CDR model.

Conversational Query Rewriting. CQR aims to enhance conversational search performance by transforming context-dependent queries into standalone ad-hoc queries (Yu et al., 2020; Vakulenko et al., 2021). To optimize query rewriting, some studies have leveraged reinforcement learning (Chen et al., 2022; Wu et al., 2022; Liu et al.,

2024) or incorporated ranking signals during model training (Qian and Dou, 2022; Mo et al., 2023a). However, these approaches rely on manually annotated rewritten queries for training CQR models, which are difficult to obtain in practice. Recently, LLMs have been demonstrated to be capable of rewriting conversational queries (Mao et al., 2023; Ye et al., 2023; Jang et al., 2024; Mo et al., 2024a), the generated queries are ideal for downstream retrieval tasks. To address these issues, we implement iterative enhancement of conversation context by prompting LLMs, which helps effectively resolve ambiguities in conversation history and reduces the need for manually rewritten queries. We then utilize the generated rewritten queries as training data to assist in training CDR models. By leveraging the strengths of CDR models in implicit context modeling, we aim to enhance the semantic correlation between the retrieval system and downstream tasks.

3 Methodology

3.1 Task Formulation

Given a new query q_k and the conversation history $\mathcal{H}_{k-1} = \{q_i, r_i\}_{i=1}^{k-1}$, where q_i and r_i denote the query and the system response to each previous turn, respectively. The i-th historical turn is denoted as (q_i, r_i, p_i^*), where p_i^* is the ground-truth passage corresponding to q_i. For given the current query q_k and the conversation history $\mathcal{H}_{k-1}$, our task is to retrieve the gold passage p_k^* from a passage collection $\mathcal{D}$.

3.2 Overview of the Methodology

In this section, we present the proposed three-stage framework GHADR, as shown in Figure 2. In the first stage (Sec. 3.3), we leverage LLMs to augment the conversation history and then prompt LLMs to rewrite the current query based on the augmented history. In the second stage (Sec. 3.4), we utilize the rewritten queries and augmented conversation history to extract additional positive and negative training sample pairs. For this purpose, we employ a semantic embedding-guided hierarchical clustering algorithm. In the third stage (Sec. 3.5), we use these additional supervision signals to train the dense passage retriever through contrastive learning, thereby improving its ability to distinguish between relevant and irrelevant historical turns.

3.3 History-Augmented Query Rewriting

Recent studies (Mao et al., 2023; Mo et al., 2024a; Jang et al., 2024) have demonstrated that open-source LLMs with language understanding and text generation capabilities can be directly applied to real-world scenarios as an effective approach for query rewriting without requiring fine-tuning. In this section, we propose an iterative prompt refinement framework for conversation history augmentation and query rewriting. Specifically, the framework comprises two core prompts: **R**ewriting-**w**ith-**R**esponse (RWR) and **R**ewriting-**a**fter-**R**ewriting (RAR).

Rewriting-with-Response for History Augmentation. In this section, we propose RWR instruction to tackle the problems of co-reference and omission. For the current k-th query q_k, the conversation history $\mathcal{H}_{k-1}$ from the first $k-1$ turns is known. As shown in Eq. 1, for any turn t in the first $k-1$ turns, we concatenate q_t, r_t and $\mathcal{H}_{t-1}$ into a prompt, where the prompt is then fed into LLMs to obtain the de-contextualized search query $q_t^{'}$.

$$q_t^{'} = \mathcal{LLM}(\mathcal{I}^{RWR} \oplus \mathcal{H}_{t-1} \oplus q_t \oplus r_t) \quad (1)$$

We obtain the augmented conversation history $\mathcal{H}_{k-1}^{'}$ by replacing all the original queries (q_t) in the first $k-1$ turns with the corresponding disambiguated queries ($q_t^{'}$).

Rewriting-after-Rewriting for Query Rewriting. After obtaining the augmented history $\mathcal{H}_{k-1}^{'}$ of the first $k-1$ turns of conversation, we introduce the RAR instruction to generate a well-informed, context-independent rewritten query using LLMs. For the k-th query with conversation history, as formalized in Eq. 2, we concatenate the current query q_k and the augmented conversation history $\mathcal{H}_{k-1}^{'}$ into a prompt. This composite input is subsequently fed into LLMs to generate the final rewritten query q_k^*.

$$q_k^* = \mathcal{LLM}(\mathcal{I}^{RAR} \oplus \mathcal{H}_{k-1}^{'} \oplus q_k) \quad (2)$$

In a multi-turn conversation scenario, for the initial conversation turn where no historical context exists, the original query itself is directly used as the rewritten query i.e., $q_1^* = q_1$.

Through this iterative refinement process, our prompt framework effectively addresses the question ambiguity problem in conversation context, thereby improving both the accuracy and context

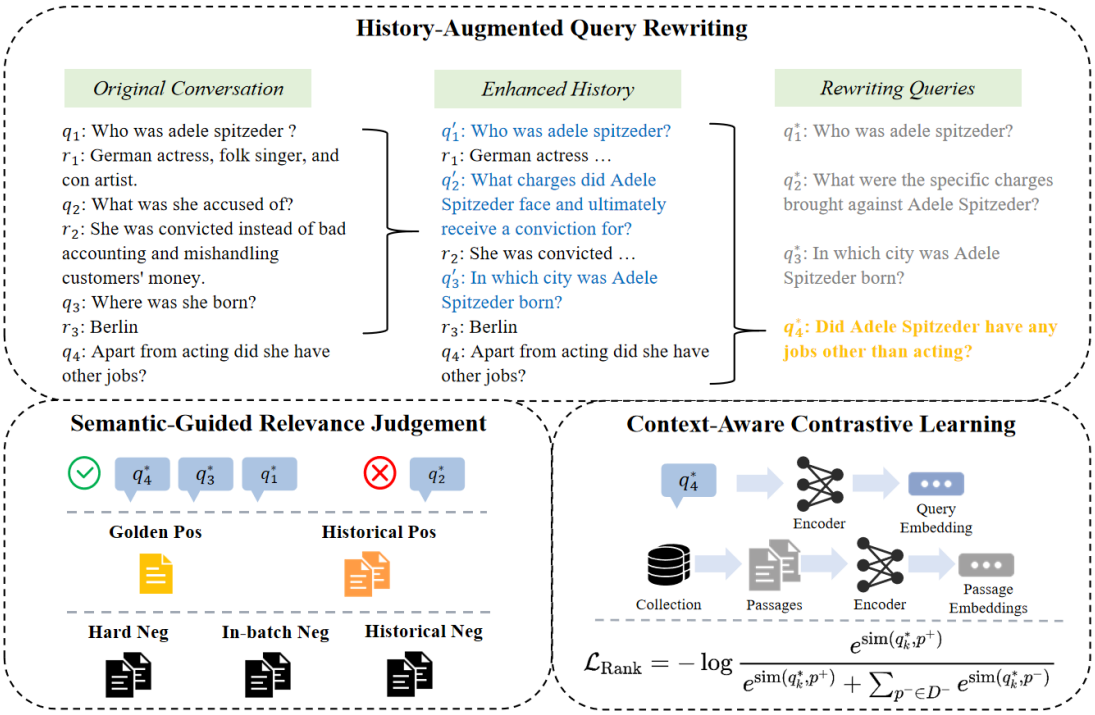

Figure 2: Overview of GHADR. The first step (top) involves leveraging LLMs to augment the conversation history and reformulate the current query. In the second step (bottom left), the reformulated queries is encoded into semantic embeddings. Subsequently, a relevance judgment is conducted between the current query and the historical turns, enabling the extraction of positive and negative samples as supervision signals. The third step (bottom right) trains a dense passage retriever through contrastive learning, incorporating the additional supervision signals.

relevance of query rewriting. Although conceptually straightforward, the experimental results presented in Sec. 5.2 demonstrate that our proposed prompt framework achieves performance comparable to several existing baselines. Notably, it surpasses the manually rewritten queries included in the QReCC dataset under specific settings. The precise prompts employed, along with representative examples for each case, are provided in Appendix A.

3.4 Semantic-Guided Relevance Judgement

We agree with Kim and Kim (2022) and Mao et al. (2022) that determining whether a historical turn is relevant to the current query is one of the crucial parts of the conversational modeling process. To leverage the full conversation context, we propose a semantic-guided approach for identifying relevant historical turns in relation to the current query. Specifically, after obtaining the augmented conversation history H'_{k-1} and rewritten query q^*_k, we employ an embedding model to encode both historical queries and the current query into semantic embeddings. Subsequently, we compute pairwise cosine similarity scores between these embeddings to construct a similarity matrix, which is then transformed into a distance matrix for clustering purposes. The agglomerative clustering (Ackermann et al., 2012) algorithm is applied to group semantically coherent queries, leveraging their hierarchical relationships. This clustering algorithm builds a hierarchy of clusters through a bottom-up approach, where each data point starts as its own cluster, and pairs of clusters are merged at each iteration based on their similarity until a desired cluster structure is formed.

The clustering algorithm partitions historical ground-truth passages into two disjoint groups:

$$\mathcal{P}^+_h = \{p^*_i\}_{i=1}, \quad \mathcal{P}^-_h = \{p^*_j\}_{j=1} \tag{3}$$

Specifically, the $\mathcal{P}^+_h$ set consists of relevant passages where each passage corresponds to a historical query clustered with the current query q^*_k. Conversely, the $\mathcal{P}^-_h$ set comprises irrelevant passages, each associated with historical queries that belong to clusters different from that of the current query q^*_k.

3.5 Training Dense Retriever in GHADR

Contrastive learning is a prevalent choice for training dense passage retriever in recent studies (Kim

and Kim, 2022; Mao et al., 2024). The dense passage retriever uses the encoders E_P and E_Q to map passages and queries to embedding space, respectively. The passage embeddings can be offline computed and indexed. The similarity between a query and a passage can be compute via dot product: $\text{sim}(q, p) = E_Q(q)^T \cdot E_P(p)$.

In this work, we train E_Q using contrastive learning, with the following positive and negative samples employed throughout the training process:

- p_k^*: The ground-truth passage corresponding to the current query q_k.
- $\mathcal{P}_h^+$: Historical passages from previous turns deemed relevant to the current query, based on relevance judgments from Sec. 3.4.
- $\mathcal{P}_h^-$: Conversely, historical passages from previous conversation turns deemed irrelevant to the current query.
- $\mathcal{P}_b^-$: In-batch negatives sampled from other data instances within the same training batch.
- $\mathcal{P}_r^-$: These are retrieved passages that serve as hard negatives. They can be obtained by using the top-ranked passages retrieved for q_k by an off-the-shelf retriever after excluding p_k^* if it is present (Karpukhin et al., 2020; Mo et al., 2024c). In this work, we adopt a sparse retriever (BM25) to obtain the hard negatives.

Given the variability in the number of positive and negative samples mined from previous historical turns across different queries, we implement a randomized sampling strategy. This approach systematically selects one historical pseudo-positive sample, one historical hard-negative sample, and the top-ranked retrieved hard-negative sample per training instance. In formal terms, we formulate the final training positive and negative samples as Eq. 4.

$$\begin{aligned} \mathcal{P}_k^+ &= \{p_k^*\} \cup \mathcal{P}_h^+ \\ \mathcal{P}_k^- &= \mathcal{P}_h^- \cup \mathcal{P}_b^- \cup \mathcal{P}_r^- \end{aligned} \tag{4}$$

The contrastive learning loss for the DPR is defined in Eq. 5, where $p^+ \in \mathcal{P}_k^+$ and $p^- \in \mathcal{P}_k^-$.

$$\mathcal{L} = -\log \frac{e^{\text{sim}(q_k^*, p^+)}}{e^{\text{sim}(q_k^*, p^+)} + \sum_{p^-} e^{\text{sim}(q_k^*, p^-)}} \tag{5}$$

Dataset	Split	#Conv.	#Turns(Qry.)
TopiOCQA	Train	3,509	45,450
	Test	205	2,514
QReCC	Train	10,823	63,501
	Test	2,775	16,451
CAsT-19	Test	50	479
CAsT-20	Test	25	208

Table 1: Statistics of conversational search datasets.

4 Experimental Setup

4.1 Datasets and Metrics

Following previous studies (Yu et al., 2021; Jang et al., 2024), four widely-used conversation datasets are used for our experiments. TopiOCQA (Adlakha et al., 2022) contains complex topic-switch phenomena within each conversational session. QReCC (Anantha et al., 2021) focuses on the query rewriting problem, most queries in a conversational session are on the same topic. In addition, we evaluate two CAsT datasets (Dalton et al., 2020a; Dalton et al., 2020b) which are used solely as test sets, to further validate the zero-shot ability of our method, e.g., when CDR models are trained on QReCC and tested on CAsTs. The statistics of the datasets are provided in Table 1.

For an adequate comparison with previous studies, we evaluate the retrieval results using the `pytrec_eval` (Van Gysel and de Rijke, 2018) tool to calculate three standard evaluation metrics: MRR, NDCG@3, and Recall@10.

4.2 Implementation details

For the large language models, we use Qwen2.5-7B (Qwen et al., 2025) to perform history augmentation and query rewriting. For Sec. 3.4, we encode queries into semantic embeddings using gte-Qwen2-7B-Instruct (Li et al., 2023), and generate relevance judgments using the agglomerative clustering algorithm implemented in the scikit-learn (Pedregosa et al., 2011) library.

We adopt ANCE (Xiong et al., 2021) as the backone model for conversation dense passage retrievers training. To train GHADR, we use the AdamW optimizer with a learning rate of 1e-5, set the batch size to 32, and train the model for 10 epochs. Following previous works (Yu et al., 2021; Mo et al., 2024c), we only update the parameters of the query encoder and the passage encoder remains frozen during training. The dense retrieval are performed

using FAISS (Johnson et al., 2019).

We use Pyserini (Lin et al., 2021a) to implement sparse retrieval (BM25). Following previous work, we set BM25 parameters as $k1 = 0.9, b = 0.4$ and $k1 = 0.82, b = 0.68$ for TopiOCQA and QReCC, respectively. We conduct experiments on a single NVIDIA A100 40G GPU.

4.3 Baselines

To validate the effectiveness of our approach, we compared it with advanced baseline methods. To ensure a fair comparison, all selected baselines are evaluated on dense passage retrievers.

ConvGQR (Mo et al., 2023a) reformulates better conversational queries by combining two T5-based models for query rewrite and query expansion.

LLM4CS (Mao et al., 2023) presents a simple yet effective prompt framework to leverage LLMs as a text-based search intent interpreter.

IterCQR (Jang et al., 2024) iteratively trains the conversational query rewriting model by directly leveraging information retrieval signals as a reward.

CHIQ (Mo et al., 2024a) leverages the capabilities of LLMs to resolve ambiguities in the conversation history before query rewriting.

ConvDR (Yu et al., 2021) fine-tunes an ad-hoc search dense retriever to learn the latent representation of the reformulated query.

SDRConv (Kim and Kim, 2022) performs conversational dense retrieval on conversational search data with additionally mined hard negatives.

InstructoR (Jin et al., 2023) uses LLMs to estimate the relevance score between session and passages to guide the training of dense retriever.

HAConvDR (Mo et al., 2024c) fine-tunes the ANCE model on context-denoising reformulated query and additional signals from historical turns.

ConvSDG (Mo et al., 2024d) employs LLMs to generate synthetic training data, which is subsequently used for fine-tuning dense retrievers.

5 Results and Analysis

5.1 Main Results

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

(1) We find that our GHADR consistently outperforms all compared baselines across three metrics on both datasets. On the TopiOCQA dataset, GHADR improves MRR by 4.0% and Recall@10 by 9.2% over the second-best method. On the QReCC dataset, GHADR's Recall@10 reaches 73.0%, which is close to HAConvDR's performance, but the MRR and NDCG@3 metrics are higher, exceeding those of HAConvDR. We attribute the performance advantages of GHADR to the following two aspects. First, GHADR integrates the query rewriting capability of CQR and the passage-level context modeling capability of CDR, enabling it to effectively capture intent changes and incorporate multi-turn conversation information in dynamic conversational scenarios. Second, GHADR optimizes the negative sampling strategy during training, enhancing the model's ability to distinguish contextually relevant passages.

(2) We observe that the CDR approaches overall outperform the CQR approaches on the QReCC dataset, which focuses on query rewriting. This phenomenon suggests that in QReCC scenarios that require handling contextual dependencies, the implicit context modeling technique is able to more consistently capture key information in the conversation history, leading to better performance in ranking and recall metrics. On the contrary, on the TopiOCQA dataset, the CQR and CDR approaches do not show a significant difference in performance and this phenomenon suggests that explicit query rewriting techniques are also effective in capturing dynamically changing user intent in TopiOCQA scenarios with frequent topic shifts.

5.2 Impact of LLMs with different parameter scales

To explore the impact of LLMs on the generative history augmentation strategy proposed in Sec. 3.3, we conduct experiments on open-source LLMs with different parameter scales. We perform CQR with Qwen2.5 series LLMs, using the rewritten query as input to sparse retrieval. Table 3 presents the sparse retrieval results of our proposed prompt framework.

We observe that Qwen2.5-72B achieves the highest performance, with the MRR score improves by 26.7% on TopiOCQA and 9.7% on QReCC compared to Qwen2.5-7B. This indicates that models with larger parameter scales generally outperform those with smaller scales, a finding consistent with the scaling laws of LLMs.

The improvement is greater for TopiOCQA, indicating that conversational scenarios with more topic shifts are more challenging and require LLMs with larger parameter scales to capture topic shifts

Category	Method	TopiOCQA			QReCC		
		MRR	NDCG@3	Recall@10	MRR	NDCG@3	Recall@10
CQR	Human-Rewrite	-	-	-	38.4	35.6	58.6
CQR	ConvGQR	25.6	24.3	41.8	42.0	39.1	63.5
CQR	LLM4CS	27.7	26.7	43.3	44.8	42.1	66.4
CQR	IterCQR	26.3	25.1	42.6	42.9	40.2	65.5
CQR	CHIQ-FT	30.0	<u>28.9</u>	<u>51.0</u>	36.9	34.0	57.6
CDR	ConvDR	27.2	26.4	43.5	38.5	35.7	58.2
CDR	SDRConv	26.1	25.4	44.4	47.3	43.6	69.8
CDR	InstructoR	25.3	23.7	45.1	43.5	40.5	66.7
CDR	ConvSDG	21.4	19.9	37.8	-	-	-
CDR	HAConvDR	<u>30.1</u>	28.5	50.8	<u>48.5</u>	<u>45.6</u>	<u>72.4</u>
CDR	**GHADR (Ours)**	**31.3**	**29.3**	**55.7**	**50.0**	**46.5**	**73.0**

Table 2: Performance of different retrieval methods on TopiOCQA and QReCC, all use dense passage retrievers. Only the QReCC dataset has manually rewritten queries. All compared models are initialized from ANCE. **Bold** and <u>underline</u> indicate the best and the second-best results, respectively.

$\mathcal{LLM}$	TopiOCQA		
	MRR	N@3	R@10
Qwen2.5-7B	21.0	19.4	36.3
Qwen2.5-14B	21.6	19.8	37.2
Qwen2.5-32B	23.8	22.1	41.0
Qwen2.5-72B	**26.6**	**25.1**	**43.8**
$\mathcal{LLM}$	QReCC		
	MRR	N@3	R@10
Human-Rewritten	39.7	36.2	62.5
Qwen2.5-7B	39.0	35.9	60.0
Qwen2.5-14B	38.9	35.6	59.5
Qwen2.5-32B	39.9	36.8	61.3
Qwen2.5-72B	**42.8**	**39.8**	**64.1**

Table 3: Sparse (BM25) retrieval results for systems using various LLMs. Only the QReCC dataset has manually rewritten queries.

	TopiOCQA		QReCC	
	MRR	N@3	MRR	N@3
GHADR	**31.3**	**29.3**	**50.0**	**46.5**
w/o RWR	26.2	24.4	49.6	46.2
w/o RAR	25.6	23.5	49.0	45.5
w/o his pos.	29.6	27.8	48.7	45.3
w/o his neg.	28.8	26.7	47.8	44.5

Table 4: Ablation study of different components.

between conversational turns for higher-quality rewritten queries. On QReCC, the sparse retrieval results of rewritten queries with small-scale LLMs show minimal differences from those of manually rewritten queries, indicating that our proposed CQR strategy is effective across LLMs of varying scales. Furthermore, these results validate that the queries rewritten by our History-Augmented Query Rewriting component can functionally substitute manually rewritten queries, and they are subsequently employed in the components described in Sec. 3.4 and Sec. 3.5.

5.3 Ablation Study

In this section, we conduct an ablation study on both TopiOCQA and QReCC datasets to investigate the impact of different components in our GHADR. The results are shown in Table 4, and we observe that removing any component leads to performance degradation.

On the TopiOCQA dataset, there is a significant degradation in performance after removing RWR or RAR. In addition, there are significant domain differences in the contributions of RWR and RAR. For instance, removing RAR leads to a substantial decrease in metric scores for TopiOCQA, while for QReCC, the reduction is marginal. This phenomenon can be attributed to the distinct conversation characteristics of the two datasets: conversations in TopiOCQA involve more complex topic shifts than those in QReCC, which requires more complex query rewriting techniques to adapt to historical context.

On both datasets, removing historical negatives leads to more pronounced performance degradation compared with removing historical positives. This demonstrates that negative sampling is more critical than positive sampling in our approach, thereby emphasizing the model's need for noise suppres-

	CAsT-19		CAsT-20	
	MRR	N@3	MRR	N@3
ConvGQR	61.0	34.6	35.1	<u>24.3</u>
InstructoR	61.2	**46.6**	43.7	**29.6**
ConvSDG	60.6	35.3	36.5	24.2
GHADR (Ours)	**61.5**	<u>35.4</u>	**44.1**	19.0

Table 5: Retrieval performance of the zero-shot setting on CAsT-19 and CAsT-20. **Bold** and <u>underline</u> indicate the best and the second-best results, respectively.

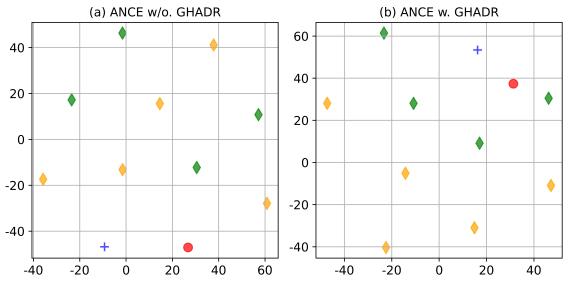

Figure 3: T-SNE visualization of query and passage embeddings based on two DPR models without and with HAConvDR training. The markers with red, blue, green and orange color represent query, gold passage, his.pos. and his.neg. respectively.

sion over positive sample expansion. Notably, the complete GHADR model achieves optimal performance on both datasets. Although the contribution of each component varies across datasets, the results suggest that the components have complementary effects, working collectively to enhance the model's overall effectiveness.

5.4 Zero-shot Analysis

The zero-shot evaluation is conducted on two CAsT datasets to assess the generalization capability of GHADR. We first train a dense passage retriever on the QReCC training set and then directly evaluate it on the CAsT test sets. As presented in Table 5, our observations are as follows:

(1) GHADR performs outstandingly in the MRR metric under zero-shot settings, demonstrating its strong cross-domain generalization ability and indicating its ability to accurately locate relevant passages. This indicates that the method can effectively leverage pre-trained knowledge to achieve accurate retrieval in unseen target domains without requiring domain-specific annotations. This feature makes it more flexible and applicable for practical applications.

(2) The NDCG@3 metric reflects the recall performance by measuring the proportion of relevant passages within the top three retrieval results. On the CAsT-19 dataset, GHADR surpasses most baselines in NDCG@3, except for InstructoR. In contrast, on the CAsT-20 dataset, the NDCG@3 score of GHADR is much lower than that of other baselines. This performance discrepancy suggests that GHADR maintains high accuracy but experiences a decline in recall when faced with distributional shifts or more complex queries. The trade-off between accuracy and recall may stem from the model's excessive focus on optimizing semantic alignment while neglecting the coverage of multiple relevant passages.

5.5 Qualitative Analysis

To provide deeper insights into our approach, we conduct a qualitative analysis by visualizing an example within the embedding space, as illustrated in Figure 3. This figure provides a t-SNE visualization (van der Maaten and Hinton, 2008) comparing the ANCE dense retriever with and without GHADR training. In contrast to the vanilla ANCE model, which fails to distinguish the gold passage from the ground-truth passages in the previous historical turns, the ANCE trained with our GHADR demonstrates significantly improved ability to differentiate relevant passages from distractors. The concrete example of this case analysis is presented in Appendix B.

6 Conclusion

In this study, we propose GHADR, a framework comprising three core components that combine the advantages of CQR and CDR approaches. The History-Augmented Query Rewriting component iteratively enhances the quality of conversation history, thereby improving the performance of query rewriting. Specifically, the Semantic-Guided Relevance Judgement component and the Context-Aware Contrastive Learning component are designed to train a dense passage retriever using context-denoised queries and additional supervision signals mined from historical turns. Comprehensive experimental evaluations on four public datasets demonstrate the effectiveness, applicability and generalizability of GHADR in handling complex multi-turn conversation.

Limitations

We identify two potential limitations of our work. First, the use of LLM-based query rewriters is inevitably subject to the inherent limitations of LLMs. In this study, our experiments are limited to the Qwen2.5 family of open-source LLMs, excluding other open-source and commercial closed-source LLMs. This is primarily due to computational and financial constraints.

Second, when employing the semantic-guided strategy to mine historical supervision signals for training dense passage retrievers, our current implementation relies solely on hierarchical clustering techniques. Note that the mined positive samples are derived from ground truth passages of historical queries that are semantically similar to the current query. However, these passages may deviate from the actual relevance of the current query. Therefore, future research should explore more effective strategies for supervision signal mining.

Ethical Statement

We conduct experiments with publicly available datasets and open-source LLMs. Our approach augments the conversation history and rewrites queries based on previous conversation history. Since these operations are dependent on the historical context of the conversation, if there are biases or inappropriate statements in the original conversation context, the results generated by our method may also contain similar biases or inappropriate statements.

References

Marcel R. Ackermann, Johannes Blömer, Daniel Kuntze, and Christian Sohler. 2012. Analysis of agglomerative clustering. *Algorithmica*, 69(1):184–215.

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, Online. Association for Computational Linguistics.

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: Industry Track*, pages 357–370, Abu Dhabi, UAE. Association for Computational Linguistics.

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)*, pages 2879–2893, Bangkok, Thailand. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tai, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, and 16 others. 2024. Scaling instruction-finetuned language models. *J. Mach. Learn. Res.*, 25(1).

Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. 2020a. Cast 2020: The conversational assistance track overview. In *Text Retrieval Conference*.

Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. 2020b. Trec cast 2019: The conversational assistance track overview. *ArXiv*, abs/2003.13624.

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, Online only. Association for Computational Linguistics.

Chao-Wei Huang, Chen-Yu Hsu, Tsu-Yuan Hsu, Chen-An Li, and Yun-Nung Chen. 2023. CONVERSER: Few-shot conversational dense retrieval with synthetic data generation. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 381–387, Prague, Czechia. Association for Computational Linguistics.

Yunah Jang, Kang-il Lee, Hyunkyung Bae, Hwanhee Lee, and Kyomin Jung. 2024. IterCQR: Iterative conversational query reformulation with retrieval guidance. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8121–8138, Mexico City, Mexico. Association for Computational Linguistics.

Soyeong Jeong, Jinheon Baek, Sung Ju Hwang, and Jong 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, Toronto, Canada. Association for Computational Linguistics.

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

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

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

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, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *Preprint*, arXiv:2308.03281.

Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021a. Pyserini: A python toolkit for reproducible information retrieval research with sparse and dense representations. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 2356–2362, New York, NY, USA. Association for Computing Machinery.

Sheng-Chieh Lin, Jheng-Hong Yang, and Jimmy Lin. 2021b. Contextualized query embeddings for conversational search. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1004–1015, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Lihui Liu, Blaine Hill, Boxin Du, Fei Wang, and Hanghang Tong. 2024. Conversational question answering with language models generated reformulations over knowledge graph. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 839–850, Bangkok, Thailand. Association for Computational Linguistics.

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*, pages 1227–1240, Miami, Florida, USA. Association for Computational Linguistics.

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*, pages 1211–1225, Singapore. Association for Computational Linguistics.

Kelong Mao, Zhicheng Dou, and Hongjin Qian. 2022. 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*, SIGIR '22, page 176–186, New York, NY, USA. Association for Computing Machinery.

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

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

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

Fengran Mo, Jian-Yun Nie, Kaiyu Huang, Kelong Mao, Yutao Zhu, Peng Li, and Yang Liu. 2023b. Learning to relate to previous turns in conversational search. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, page 1722–1732, New York, NY, USA. Association for Computing Machinery.

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

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 Web Conference 2024*, WWW '24, page 1634–1642, New York, NY, USA. Association for Computing Machinery.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in

Python. *Journal of Machine Learning Research*, 12:2825–2830.

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, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Svitlana Vakulenko, Shayne Longpre, Zhucheng Tu, and Raviteja Anantha. 2021. Question rewriting for conversational question answering. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, WSDM '21, page 355–363, New York, NY, USA. Association for Computing Machinery.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.

Christophe Van Gysel and Maarten de Rijke. 2018. Pytrec_eval: An extremely fast python interface to trec_eval. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, page 873–876, New York, NY, USA. Association for Computing Machinery.

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*, pages 10000–10014, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2021. Approximate nearest neighbor negative contrastive learning for dense text retrieval. 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, Singapore. Association for Computational Linguistics.

Shi Yu, Jiahua Liu, Jingqin Yang, Chenyan Xiong, Paul 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 '20, page 1933–1936, New York, NY, USA. Association for Computing Machinery.

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*, SIGIR '21, page 829–838, New York, NY, USA. Association for Computing Machinery.

A Prompt Examples

In Table 6, we list the prompts designed to enhance the conversation history, as well as the prompts used to rewrite the current query.

B Qualitative Example

A qualitative example corresponding to the T-SNE visualization in Sec. 5.5 is presented in Table 7.

TopiOCQA #Session 23

Rewriting with Response (RWR)

For an information-seeking dialog, please help reformulate the question into rewrite that can fully express the user's information needs without the need of context.

YOUR TASK (only questions and responses may be given):
Context:
Question: Who was adele spitzeder?
Response: German actress, folk singer, and con artist.

Current Question: What was she accused of?
Current Response: She was convicted instead of bad accounting and mishandling customers' money.

Now, you should give me the rewrite of the **Current Question** under the **Context** and the **Current Response**. Note that you should always try to rewrite it. Never ask for clarification or say you don't understand it in the generated rewrite. The output format should always be Rewrite: $Rewrite.

Model Output: What charges did Adele Spitzeder face and ultimately receive a conviction for?

Rewriting after Rewriting (RAR)

For an information-seeking dialog, please help reformulate the question into rewrite that can fully express the user's information needs without the need of context.

YOUR TASK (only questions and responses may be given):
Context:
Question: Who was adele spitzeder?
Response: German actress, folk singer, and con artist.
Question: What charges did Adele Spitzeder face and ultimately receive a conviction for?
Response: She was convicted instead of bad accounting and mishandling customers' money.

Current Question: Where was she born?

Now, you should give me the rewrite of the **Current Question** under the **Context**. Note that you should always try to rewrite it. Never ask for clarification or say you don't understand it in the generated rewrite. The output format should always be Rewrite: $Rewrite.

Model Output: In which city was Adele Spitzeder born?

Table 6: The prompts for History Augmentation and Query Rewriting. Blue denotes enhanced historical queries. Orange denotes the current user's query and its rewritten output.

Conversation (TopiOCQA #Session 179_10)

q_1: When was the last time a hurricane hit tampa bay area?

q_1^*: When was the last time a hurricane hit tampa bay area?

r_1: 1921

q_2: Which areas did it effected? (relevant)

q_2^*: What areas were impacted by the hurricane that occurred in 1921 in the Tampa Bay area?

r_2: Cuba, Pasco County, Manatee County and Sarasota County.

q_3: Who ruled the first place? (relevant)

q_3^*: Who was the leader of Cuba in 1921?

r_3: Communist Party of Cuba

q_4: Which ethnic groups immigrated here? (relevant)

q_4^*: What ethnic groups migrated to the area affected by the 1921 hurricane in the Tampa Bay area?

r_4: Afro-Cubans

q_5: What is one of the principles of its communist party?

q_5^*: What is a core principle of the Communist Party of Cuba?

r_5: It entails democratic and open discussion of policy issues within the party, followed by...

q_6: When was it founded?

q_6^*: What is the founding date of the Communist Party of Cuba?

r_6: 3 October 1965

q_7: What arrangements were made before the 1921 hurricane? (relevant)

q_7^*: What arrangements were issued before the 1921 hurricane that affected the Tampa Bay area?

r_7: Storm warnings were issued eastward from mouth of the Mississippi ...

q_8: Which states surround borders the river you just mentioned?

q_8^*: What are the states that border the Mississippi River?

r_8: States of Minnesota, Wisconsin, Iowa, Illinois, Missouri and Kentucky.

q_9: What is the significance of its name?

q_9^*: What is the origin or meaning behind the name of Mississippi?

r_9: The word Mississippi itself comes from the French rendering of the Anishinaabe ...

Current Query

q_{10}: Which species can be found here?

q_{10}^*: What types of wildlife inhabit the area affected by the 1921 hurricane in the Tampa Bay area?

Gold Passage

Mississippi River Other fauna In addition to fish, several species of turtles (such as snapping, musk, mud, map, cooter, painted and softshell turtles), American alligator, aquatic amphibians (such as hellbender, mudpuppy, three-toed amphiuma and lesser siren), and cambarid crayfish (such as the red swamp crayfish) are native to the Mississippi basin.

Table 7: An example for case study in GHADR. q_i^* indicates the rewritten query based on augmented history. A historical query with relevant indicates that the query is relevant to the current turn.