KG-CQR: Leveraging Structured Relation Representations in Knowledge Graphs for Contextual Query Retrieval

Anonymous ACL submission

Abstract

The integration of knowledge graphs (KGs) with large language models (LLMs) offers significant potential to improve the retrieval phase of retrieval-augmented generation (RAG) systems. In this study, we propose KG-CQR, a novel framework for Contextual Query Retrieval (CQR) that enhances the retrieval phase by enriching the contextual representation of complex input queries using a corpus-centric KG. Unlike existing methods that primarily address corpus-level context loss, KG-CQR focuses on query enrichment through structured relation representations, extracting and completing relevant KG subgraphs to generate semantically rich query contexts. Comprising subgraph extraction, completion, and contextual generation modules, KG-CQR operates as a model-agnostic pipeline, ensuring scalability across LLMs of varying sizes without additional training. Experimental results on RAGBench and MultiHop-RAG datasets demonstrate KG-CQR's superior performance, achieving a 4-6% improvement in mAP and a 2-3% improvement in Recall@25 over strong baseline models. Furthermore, evaluations on challenging RAG tasks such as multi-hop question answering show that, by incorporating KG-CQR, the performance consistently outperforms the existing baseline in terms of retrieval effectiveness[1].

1 Introduction

Large Language Models (LLMs) have significantly advanced the field of natural language processing (NLP), particularly in understanding and generating human-like text. However, LLMs still suffer from two critical limitations: a lack of reliable factual knowledge and limited reasoning capabilities (Wang et al., 2024b). These limitations become more pronounced when LLMs are applied to domain-specific knowledge retrieval, especially in addressing queries within vertical domains (Bang et al., 2023). To address these challenges, recent research has explored the integration of Knowledge Graphs (KGs) into LLMs as a means to provide structured, accurate knowledge sources for enhanced reasoning (Pan et al., 2024). KGs, which store facts in the form of triples (i.e., head entity, relation, tail entity), offer a robust and interpretable representation of knowledge. As a result, LLM-based applications have increasingly incorporated KGs to improve performance in tasks such as question answering (Ding et al., 2024), fact-checking (Pham et al., 2025), and recommendation systems (Abu-Rasheed et al., 2024).

In the context of question answering over knowledge graphs (KGQA), current approaches can be broadly categorized into two main strategies: (i) us-

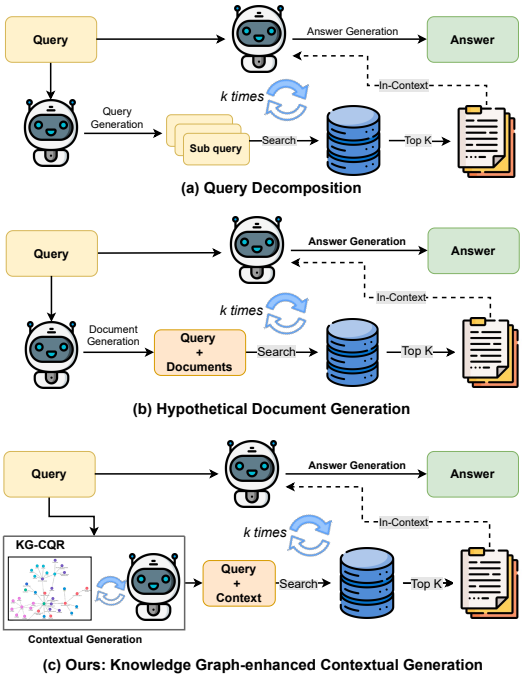

Figure 1: Overview of query expansion approaches for RAG systems: a) query decomposition; b) document generation; c) ours: KG-enhanced contextual generation

[1] https://github.com/anonymous/repo

ing LLMs to convert natural language queries into formal logical queries, which are then executed on KGs to derive answers (Nguyen et al., 2024; Wang et al., 2024a); and (ii) retrieving relevant triples from KGs and presenting them as contextual knowledge for the LLM to generate the final answer (Sarmah et al., 2024; Sun et al., 2024). Similarly, in retrieval-augmented generation (RAG) tasks, external knowledge sources, in terms of both structured (KGs) and unstructured (vectorized documents), are retrieved and incorporated into the input prompt to support answer generation by LLMs (Li et al., 2024; Edge et al., 2024). Despite these advances, the retrieval process involving KGs remains underexplored in the aforementioned approaches.

This study focuses on enhancing the retrieval process for RAG systems by integrating KG technologies to enable contextual information for the input complex queries. Specifically, the objective is to tackle a critical challenge in current systems: the embedding misalignment between sentence-level queries and document-level corpus representations (Ma et al., 2023). Accordingly, existing methods often employ large language models (LLMs) to decompose complex queries (Mao et al., 2024) (Figure 1(a)). Nonetheless, in terms of retrieval performance, this approach frequently underperforms due to insufficient contextual alignment with the corpus. Sequentially, Gao et al. (2023) proposed a new approach by generating hypothetical documents to facilitate document-document similarity comparisons (Figure 1(b)). However, this method heavily relies on underlying LLMs, introducing risks of hallucination. In terms of knowledge-grounded expansion generation, Xia et al. (2025) introduced a knowledge-aware approach that leverages both unstructured data and structured relations. Nevertheless, their reliance on predefined relation schemas between entities (e.g., title) and documents constrains the scalability and adaptability.

To overcome the aforementioned limitations, we propose KG-CQR (Knowledge Graph for Contextual Query Retrieval), a novel framework that leverages a corpus-centric KG to generate contextual information for input queries (Figure 1(c)). The key idea is to extract a relevant subgraph from the KG to semantically enrich each query. KG-CQR comprises three main modules: (i) subgraph extraction, which identifies relevant triples; (ii) subgraph completion, which infers missing triples; and (iii) contextual generation, which constructs enriched query contexts. These modules utilize a new structured representation of relations, combining textual information with KG triplets, to address the limitations of traditional entity-based scoring in KG extraction. By retrieving directly relevant data and inferring missing knowledge, KG-CQR significantly improves query contextualization. The main contributions of this work are as follows:

- We propose Contextual Query Retrieval (CQR), a novel paradigm designed to enhance the context of domain-specific queries using a predefined corpus. Our framework, KG-CQR, leverages a corpus-centric knowledge graph to improve both query understanding and retrieval effectiveness, achieving these improvements without the need for additional training.
- KG-CQR functions as a model-agnostic pipeline that employs structured relation representations to generate contextual information, ensuring adaptability and scalability across backbone LLMs with varying parameter sizes.
- We evaluate KG-CQR on recent, complex benchmark datasets specifically designed for multi-step retrieval processes in RAG systems. The experimental results demonstrate the effectiveness of KG-CQR in enhancing retrieval quality.

2 Literature Review

2.1 Query Expansion using LLM

To handle complex queries effectively, query expansion is often essential for improving the performance of the retrieval process (Azad and Deepak, 2019). Traditional approaches decompose input queries into multi-view representations to enhance retrieval accuracy (Zhang et al., 2022). Recently, with the rapid advancement LLMs, a promising direction involves query enhancement, either through prompt-based techniques leveraging LLMs (Wang et al., 2023), or by developing trainable frameworks that generate refined queries (Mao et al., 2024). These methods aim to reformulate input queries to produce more effective semantic representations (Chan et al., 2024; Chen et al., 2024). However, they still struggle to bridge the inherent gap between user queries and the knowledge corpus at the retrieval embedding level (Liu et al., 2025). Accordingly, to further improve retrieval effectiveness, especially in domain-specific applications, a deeper exploitation of contextual generation remains essential (Li et al., 2025).

2.2 Contextual Retrieval

Contextualized retrieval has been introduced to enhance retrieval performance, particularly in challenging scenarios (Morris and Rush, 2024). Recent approaches such as RAPTOR (Sarthi et al., 2024) and GraphRAG (Edge et al., 2024) employ recursive processes that combine embedding, clustering, and summarization to build hierarchical representations of documents using tree and graph structures, respectively. These hierarchical representations help improve contextual retrieval across the original corpus. In terms of query expansion through contextualization, Gao et al. (2023) proposes HyDE, a novel approach that leverages large language models (LLMs) to generate hypothetical documents conditioned on the input query. Accordingly, the query is first processed by an LLM following specific instructions to produce hypothetical documents, which are then used as pseudo-contexts for retrieval based on document-to-document similarity. However, a key limitation of HyDE lies in its dependence on LLM-powered generated content, where potential inaccuracies or hallucinations can degrade retrieval effectiveness (Zhang et al., 2024; Xia et al., 2025). Moreover, query expansion strategies must account for domain-specific context sensitivity, as the same entities may vary in meaning or relevance across different domains (Bui et al., 2021). Therefore, this study proposes a novel contextual retrieval approach, which focuses on providing contextual information for the input query, based on the structured relation of the corpus-centric KG.

2.3 LLM-Powered KG Construction

One of the primary challenges in utilizing knowledge graphs (KGs) lies in their construction. Prior work relies on predefined KGs (Xia et al., 2025), which limits the flexibility and scalability of the approach. In order to automatically construct a KG, given a set of unstructured data sources (corpus), Knowledge Graph Construction (KGC) is typically framed as a structured prediction task, where models are trained to approximate target functions associated with various NLP tasks such as Named Entity Recognition (NER), Relation Extraction (RE), Entity Linking (EL), and Knowledge Graph Completion (Ye et al., 2022). However, training task-specific discriminative models often results in error propagation and limited adaptability across diverse tasks. To address these limitations, recent approaches reformulate KGC as a generative problem using sequence-to-sequence (Seq2Seq) models (Lu et al., 2022). Powered by pre-trained models such as T5 (Raffel et al., 2020), the Seq2Seq paradigm has demonstrated strong performance in multi-task training settings for KG construction. More recently, the emergence of LLMs has spurred interest in their application to KGC through zero-shot prompting techniques (Pan et al., 2024; Zhu et al., 2024). Building on this line of work, our study leverages modern open-source LLMs, e.g., LLama-3-70B, to construct knowledge graphs by parsing and categorizing entities and their relationships directly from unstructured data.

3 Methodology

3.1 Preliminary

3.1.1 Structure Relation Representation

A corpus-centric KG includes a set of triplets (structure relation) $\mathcal{T}_{KG}$, which are defined as follows:

$$\begin{aligned} KG &= \{E_{KG}, R_{KG}, \mathcal{T}_{KG}\} \\ \mathcal{T}_{KG} &= \{(u, r, v), u, v \in E_{KG}, r \in R_{KG}\} \end{aligned} \quad (1)$$

where E_{KG} is the set of entities and R_{KG} is the set of relations. Since the KG is not available for most specific domains, we follow the work in GraphRAG (Edge et al., 2024) to construct the corpus-centric KG, which includes three sequential steps: i) Ingesting specific-domain unstructured data; ii) Extracting entities and their relationships using an external LLM; iii) Mapping entities through edges (relations) that contain detailed information about their relationships.

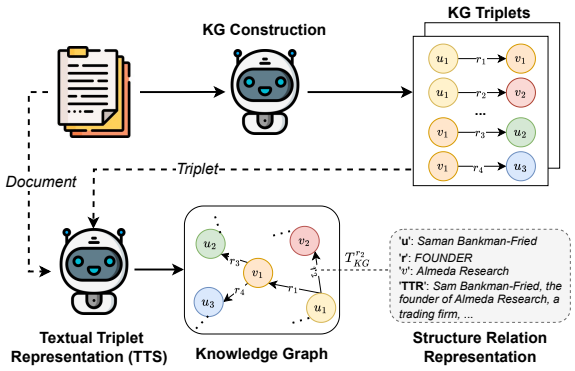

Figure 2: Construction of structured relation representations using LLM-based prompting. Detailed prompt templates are provided in Appendix A.5.

To further enhance the expressiveness of the KG, we extend each triplet $\mathcal{T}^i_{KG}$ with a textual

triplet representation (TTR). Unlike traditional approaches that rely solely on structured relational properties, our method leverages LLMs to generate rich, natural language representations of each triplet, as defined below:

$$TTR(\mathcal{T}^i_{KG}) = llm(Promt_{ttr}, D^i_d, \mathcal{T}^i_{KG}) \quad (2)$$

where $\text{llm}(Prompt_{ttr}, D^i_d, \mathcal{T}^i_{KG})$ denotes the textual description of the relation, generated by an LLM based on the instruction prompt $Prompt_{ttr}$, the corresponding triplet $\mathcal{T}^i_{KG}$, and the document $d \in D$ from which the triplet was extracted. An overview of this process is illustrated in Figure 2. In this regard, the structured relation in Equation 1 is reformulated as:

$$\mathcal{T}_{KG} = \{(u, r, v, TTR(u, r, v))\} \quad (3)$$

3.1.2 Problem Definition

The objective of the retrieval process is to extract the most relevant documents for the input query, in which the similarity score (i.e., cosine similarity) can be formulated as follows:

$$sim(q, d) =< \mathbf{v}_q, \mathbf{v_d} > \quad (4)$$

The core challenge in this process lies in ensuring that the query vector $\mathbf{v}_q$ (obtained via encoder enc_q) and the document vector $\mathbf{v}_d$ (obtained via encoder enc_d) are embedded into a shared semantic space. Traditional retrieval models typically rely on supervised learning frameworks that train encoders using query-document pairs to learn such a shared embedding space (Karpukhin et al., 2020; Santhanam et al., 2022). However, directly optimizing for query-document similarity often results in suboptimal retrieval performance, particularly when dealing with sparse or domain-specific queries. To address this limitation, we draw inspiration from the approach in (Gao et al., 2023), which shifts focus toward generating contextual embeddings for the query. Notably, instead of encoding the query directly, we enrich it with contextual information derived from the corpus-centric KG. This enriched representation is then embedded in the document space, allowing the similarity computation to align with the document-document similarity paradigm. The revised retrieval formulation is as follows:

$$\begin{aligned} \mathbf{v}_{\text{KG-CQR(q)}} &= enc_d(\text{KG-CQR(q)}) \\ sim(q, d) &=< \mathbf{v}_{\text{KG-CQR(q)}}, \mathbf{v}_d > \end{aligned} \quad (5)$$

Here, KG-CQR(q) denotes the KG-enhanced contextual information of the input query q.

3.2 KG-CQR

The overview architecture of KG-CQR is illustrated in Figure 3, which includes three main sequence components, such as subgraph extraction, subgraph completion, and contextual generation.

3.2.1 Subgraph Extraction

Given an input query q and a knowledge graph KG, the subgraph extraction module first identifies a set of relevant triples $\hat{\mathcal{T}}_{KG}$ ($\hat{\mathcal{T}}_{KG} \subset \mathcal{T}_{KG}$), based on the input query. Traditional subgraph extraction methods typically begin by identifying entities mentioned in the query q and then linking them to entities in the KG using entity linking (EL) techniques, such as using LLM prompting or specialized EL tools (Sun et al., 2024). However, these approaches often assume that the KG is complete, i.e., all factual triples relevant to the query are present in the graph, which is rarely the case in real-world scenarios (Xu et al., 2024). Furthermore, current subgraph extraction techniques predominantly rely on assessing semantic similarity at the entity or keyword level (Sun et al., 2024; Luo et al., 2024). Nevertheless, this limited granularity often fails to capture sufficient textual context, thereby reducing extraction performance, particularly when input queries involve ambiguous entities (Pham et al., 2025; Xia et al., 2025). To address these limitations, we leverage textual representations of triples (as defined in Equation 2) to measure similarity with the input query. This approach enables subgraph extraction at the sentence level, rather than relying solely on the entity level. The subgraph extraction is formalized as follows:

$$\begin{aligned} \mathbf{v}^i_r &= enc(TRR(\mathcal{T}^i_{KG})) \\ \hat{\mathcal{T}}_{KG} &= \text{argmax}_{\mathcal{T}^i_{KG} \in \mathcal{T}_{KG},\, k}\{sim(\mathbf{v}_q, \mathbf{v}^i_r)\} \end{aligned} \quad (6)$$

where $\mathbf{v}_q$ is the embedding of the input query, and k is a hyperparameter controlling the number of top-matching triples retrieved.

Sequentially, inspired by previous work for the subgraph extraction process (Sun et al., 2024), a filtering step is performed using an LLM with a task-specific prompt to remove irrelevant triples:

$$\begin{aligned} \hat{\mathcal{T}'}_{KG} = \{\mathcal{T}^i_{KG} \in \hat{\mathcal{T}}_{KG} | \\ llm(Promt_{filter}, q, \mathcal{T}^i_{KG}) = True\} \end{aligned} \quad (7)$$

Here, $Prompt_{filter}$ denotes the instruction prompt used by the LLM for the final selection. The details of $Prompt_{filter}$ are provided in Appendix A.5.

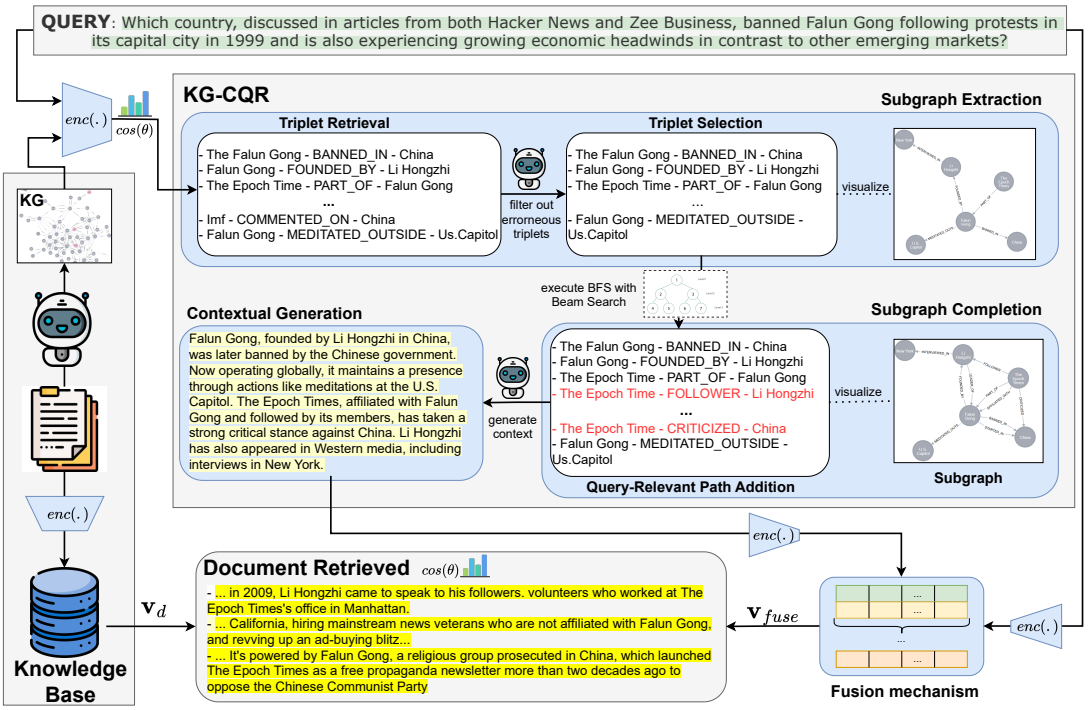

Figure 3: An illustration of KG-CQR for the retrieval process, which includes three main components: Subgraph Extraction, Subgraph Completion, and Contextual Generation

3.2.2 Subgraph Completion

The initial subgraph $\hat{\mathcal{T}'}_{KG}$ is extracted based on semantic similarity, typically resulting in a limited set of triplets that may lack sufficient contextual information. The goal of the subgraph completion function is to enrich this subgraph by incorporating additional triplets from the structure relation of KG ($\mathcal{T}_{KG}$) that form semantically meaningful paths between entities in $\hat{\mathcal{T}'}_{KG}$. Relevance is assessed by aggregating the semantic similarities between the input query and triplet textual representations along these paths. The subgraph completion proceeds through the following steps (Algorithm 1):

- Step 1: Extract entities from the initial subgraph $\hat{\mathcal{T}'}_{KG}$.
- Step 2: Apply Beam Search, a heuristic-guided variant of Breadth-First Search (BFS), to identify the top-n candidate paths.
- Step 3: Filter out paths that contain nodes not present in the initial subgraph $\hat{\mathcal{T}'}_{KG}$.
- Step 4: Select the top-K highest-scoring unique triplets, with K defaulting to 20.
- Step 5: Construct the completed subgraph $\hat{\mathcal{T}''}_{KG}$ by merging the initial subgraph $\hat{\mathcal{T}'}_{KG}$ with the selected triplets.

Notably, to reduce computational complexity in Step 2, instead of executing the naive BFS traversal, a limited number of nodes are expanded, guided by a heuristic function (BFSBeam). This function computes semantic similarity between the input query and aggregates the relevance scores of the TTRs along each path, which is illustrated in more detail in the Appendix A.3.

3.2.3 Contextual Generation

The objective of the retrieval process is to identify the most relevant documents for a given input query by computing similarity scores, typically using cosine similarity between their vector representations, which is formally defined as:

$$KG\text{-}CQR(q) = llm(Prompt_g, \hat{\mathcal{T}''}_{KG}) \quad (8)$$

where $Prompt_g$ represents the generation instruction prompt, as detailed in Appendix A.5. The enriched subgraph $\hat{\mathcal{T}''}_{KG}$ serves as contextual input to the LLM, facilitating the generation of a contextually enriched query representation. This reformulated query can then be encoded within the same embedding space as the corpus documents, enabling effective retrieval.

3.3 Retrieval Fusion Function

The input query and its synthetic contextual information are embedded using a fusion encoder-based

Algorithm 1 Query-Relevant Path Addition for Subgraph Completion

Require: $\mathcal{T}_{KG}$, $\hat{\mathcal{T}}'_{KG}$, q, top K, max-path L
Ensure: Subgraph $\hat{\mathcal{T}}''_{KG}$

$E_p \leftarrow \{u, v \mid \{u, r, v, \text{TTR}\} \in \hat{\mathcal{T}}'_{KG}\}$
Load Embedding mode: enc
$v_q \leftarrow q \neq \emptyset ?\text{enc}(q) : \text{None}$
$T_{\text{set}} \leftarrow \{\{u, r, v\} \mid \{u, r, v, \text{TTR}\} \in \hat{\mathcal{T}}'_{KG}\}$
$P \leftarrow \bigcup_{(e_i,e_j)\in E_p} \text{BFSBeam}(\mathcal{T}_{KG}, e_i, e_j, T_{\text{set}}, L)$
if $P = \emptyset$ **then**
    return $\hat{\mathcal{T}}'_{KG}$
end if
$S \leftarrow \emptyset$
for $p \in P$ **do**
    if $\{v_p \leftarrow \text{enc}(\text{TTR}) \mid \{u, r, v, \text{TTR}\} \in p\}$ **then**
        $s \leftarrow v_q \neq \text{None} ?\text{Mean}(\cos(v_p, v_q)) : 0$
        $S \leftarrow S \cup \{(p, s)\}$
    end if
end for
Sort S by score descending
$C \leftarrow \emptyset$
for $(p, s) \in S$ until $|C| \geq K$ do **do**
    if $\{u, r, v\} \in p \wedge \{u, r, v\} \notin T_{\text{set}}$ **then**
        $C \leftarrow C \cup \{u, r, v\}$
    end if
end for
$\hat{\mathcal{T}}''_{KG} \leftarrow \hat{\mathcal{T}}'_{KG} \cup C$
return $\hat{\mathcal{T}}''_{KG}$

approach. This technique enables the retrieval system to go beyond superficial query-document matching by leveraging the interaction between the query and its enriched context, resulting in more accurate and semantically relevant retrieval outcomes (Bruch et al., 2024). In this work, we adopt a weighted-sum fusion mechanism to compute the final query representation, defined as:

$$\mathbf{v}_{fuse(q)} = \alpha \cdot \mathbf{v}_q + (1 - \alpha) \cdot \mathbf{v}_{KG-CQR(q)} \quad (9)$$

This fusion mechanism proves especially effective in complex, multi-turn, or context-sensitive retrieval scenarios, where conventional query enhancement or decomposition methods often fall short. Consequently, the objective function in Equation 5 can be reformulated as:

$$\begin{aligned} sim(q, d) &= sim(\text{KG-CQR}(q), d) \\ &=< \mathbf{v}_{fuse(q)}, \mathbf{v}_d > \end{aligned} \quad (10)$$

4 Experiment

4.1 Experimental Setup

Baseline: We evaluate our method using three baseline models that encompass diverse document retrieval strategies: (i) BM25 (Robertson and Zaragoza, 2009), a classical sparse retrieval model; (ii) DPR (Karpukhin et al., 2020), a dense retrieval approach based on a dual-encoder architecture that independently encodes queries and passages, optimizing their embeddings via contrastive loss; and (iii) BGE (Xiao et al., 2024), which combines dense, sparse, and multi-vector retrieval using a self-knowledge distillation framework. To thoroughly assess the impact of KG-CQR on retrieval performance, we compare the integration of KG-CQR against the integration of HyDE (Gao et al., 2023) with the same baselines.

Benchmark Datasets: We evaluate our method on two recent and widely used benchmark datasets: (i) RAGBench (Friel et al., 2024), which spans five distinct industry-specific domains. We use its test set comprising approximately 11,000 instances for retrieval evaluation; and (ii) Multihop-RAG (Tang and Yang, 2024), which includes a knowledge base, a large set of multi-hop queries, corresponding ground-truth answers, and supporting evidence, totaling 2,556 queries for evaluation. For each dataset, the corresponding KG is constructed in three steps, as outlined in Section 3.1.1, using the LLama-3.3-70B model.

4.2 Main Results

Table 1 presents the evaluation results of the retrieval process on both datasets. Retrieval accuracy is evaluated using standard metrics, including mean Average Precision (mAP) and Recall@k, where $k \in \{5, 10, 25\}$. The reported results use $\alpha = 0.7$ (Equation 9), which was found to yield the best performance (the selection of this value is further discussed in Appendix A.2.2). From the results, we draw the following observations:

i) **Retrieval Performance**: KG-CQR significantly improves retrieval performance across various retrieval backbones. On the RAGBench dataset, for example, KG-CQR + BGE achieves the best performance overall, with an mAP of 0.542 and Recall@25 of 0.675, outperforming both the baseline models and the HyDE-enhanced variants. On the more challenging MultiHop-RAG dataset, KG-CQR + BM25 achieves the highest recall metrics (e.g., Recall@25 = 0.532), demonstrating KG-

	RAGBench				MultiHop-RAG			
Model	mAP	Recall@5	Recall@10	Recall@25	mAP	Recall@5	Recall@10	Recall@25
BM25	0.329	0.337	0.399	0.462	0.241	0.261	0.353	0.486
DPR	0.276	0.286	0.348	0.425	0.099	0.125	0.183	0.284
BGE	0.521	0.51	0.589	0.657	0.227	0.251	0.357	0.52
HyDE + DPR	0.286	0.293	0.354	0.426	0.099	0.125	0.183	0.284
HyDE + BGE	0.516	0.507	0.586	0.638	0.232	0.256	0.363	0.524
KG-CQR + BM25	0.398	0.398	0.454	0.514	**0.250**	**0.267**	**0.372**	**0.532**
KG-CQR + DPR	0.316	0.319	0.384	0.462	0.129	0.157	0.224	0.34
KG-CQR + BGE	**0.542**	**0.529**	**0.61**	**0.675**	0.24	0.261	0.371	0.525

Table 1: Retrieval performances on RAGBench and MultiHop-RAG datasets, using LLama-3.3-70B as the backbone

	RAGBench				MultiHop-RAG			
Backbone	mAP	Recall@5	Recall@10	Recall@25	mAP	Recall@5	Recall@10	Recall@25
LLama-3.2-3B	0.537	0.524	0.604	0.672	0.230	0.251	0.359	0.520
LLama-3.1-8B	0.538	0.526	0.606	0.672	0.235	0.255	0.370	0.522
LLama-3.3-70B	0.542	0.529	0.61	0.675	0.24	0.261	0.371	0.525

Table 2: The performance of KG-CQR across various parameter sizes of the backbone LLMs

CQR's capability to enhance retrieval accuracy over traditional methods.

ii) **Contextual Accuracy**: The relatively weaker performance of HyDE compared to its baselines suggests potential drawbacks of relying heavily on LLM-generated synthetic queries. HyDE introduces a straightforward approach for enhancing context, nonetheless, its effectiveness appears sensitive to the contextual reliability of the generated content. This underscores the limitations of insufficiently grounded synthetic information in retrieval tasks.

iii) **Diverse Benchmarks**: Although models like BGE perform well on relatively straightforward datasets such as RAGBench, more complex datasets like MultiHop-RAG demand advanced reasoning capabilities. KG-CQR demonstrates robustness in such settings by effectively handling multi-hop reasoning and maintaining strong performance. These results highlight the importance of retrieval frameworks that integrate contextual understanding and structured knowledge to perform consistently across diverse and complex benchmarks.

4.3 Detailed Analysis

4.3.1 Impact of LLM Backbone

Table 2 illustrates the retrieval performance of KG-CQR when paired with different sizes of language models, using BGE as the underlying retrieval method. Specifically, the LLama-3.3-70B model achieves the highest performance across nearly all metrics, however, the performance differences between the 8B and 70B variants are relatively modest, suggesting diminishing returns as model size increases. These findings indicate that while larger models do offer performance advantages, KG-CQR remains effective even with relatively smaller backbones such as LLama-3.2-3B and LLama-3.1-8B. This highlights KG-CQR's practicality for resource-constrained environments, offering a favorable trade-off between retrieval performance and computational cost.

4.3.2 Ablation Study

Table 3 presents an ablation study evaluating the contribution of two core components of KG-CQR: the Textual Triplet Representation (TTR) for extracting subgraph and the Subgraph Completion (Sub.Comp.). As reported results, removing TTR (Equation 2) leads to the most pronounced drop in performance (e.g., Recall@25 decreases from 0.675 to 0.641), underscoring the importance of

Method	Recall@5	Recall@10	Recall@25
W/O TTR	0.486	0.572	0.641
W/O Sub.Comp.	0.525	0.605	0.671
KG-CQR	0.529	0.61	0.675

Table 3: Ablation studies on main components of KG-CQR on RAGBench Datasets

TTR in accurately extracting relevant subgraphs that preserve semantic alignment with the query. This confirms that converting structured KG information into textual form plays a critical role in aligning the knowledge with the retrieval task. Similarly, omitting the Subgraph Completion module also results in a notable performance degradation, though less severe than removing TTR. This suggests that while the initial subgraph extraction is vital, enriching the subgraph context via completion further improves the model's ability to retrieve relevant documents.

4.3.3 Multi-Step Retrieval for RAG Results

We evaluate the effectiveness of KG-CQR in multi-step reasoning RAG tasks by integrating its retrieval outputs into the IRCoT framework (Trivedi et al., 2023). Experiments were conducted with three LLMs of varying sizes to assess the generalizability of KG-CQR. The results highlight the significance of both components in enhancing retrieval performance. We randomly sampled 500 examples

Model	Retrieval	F1	#Iter	#Score
LLama-3.2-3B	BM25	0.372	3.293	3.122
LLama-3.1-8B		0.393	2.748	3.424
LLama-3.3-70B		0.431	1.912	3.449
LLama-3.2-3B	KG-CQR +BM25	0.407	2.714	3.317
LLama-3.1-8B		0.41	1.834	3.603
LLama-3.3-70B		0.443	1.393	3.826

Table 4: Multi-step reasoning RAG performance

from the RAGBench test set and evaluated results using F1, GPT-Score ($\#Score$) (Fu et al., 2024), and average reasoning steps ($\#Iter$). GPT-Score was computed using GPT-4o through the OpenAI API (based on its performance on the Judge LLM leaderboard[2]). As shown in Table 4, several key insights can be drawn: i) **KG-CQR significantly enhances retrieval quality**: By enriching the input query with semantically and contextually relevant information, KG-CQR consistently improves performance across all model sizes compared to using BM25 alone (the results with BGE method are illustrated in Appendix A.2.1); ii) **Improved contextualization leads to fewer iterations**: KG-CQR enables models to perform reasoning with fewer iterations, likely because the contextual queries better reflect the underlying multi-hop intent. This reduces the need for redundant or corrective reasoning steps during generation. iii) **Cross-model scalability**: Gains are observed across the size of LLMs, highlighting the flexibility of KG-CQR.

[2]https://huggingface.co/spaces/AtlaAI/judge-arena, accessed by May 18th, 2025

4.3.4 Retrieval Latency

Figure 4 compares the relative retrieval latency of the baseline HyDE with three KG-CQR variants: i) **KG-CQR w Naive-BFS)**: use basic BFS algorithm for subgraph completion; ii) **KG-CQR w/o Sub.Comp.**: removes the subgraph completion module entirely; iii) **KG-CQR(ours)**: utilizes heuristic-guided Beam Search for more efficient subgraph completion. The analysis confirms that

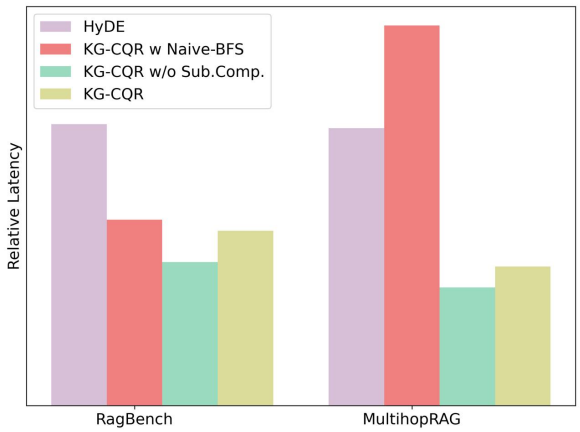

Figure 4: Retrieval latency performance

the proposed KG-CQR with Beam Search strikes an optimal balance between retrieval efficiency and reasoning capability. While KG-CQR without subgraph completion is the fastest, KG-CQR with Beam Search provides a more scalable and semantically expressive alternative with only modest additional cost. In contrast, HyDE and naive BFS approaches incur higher latency, making them less favorable for real-time or large-scale applications.

5 Conclusion

This study introduces KG-CQR, a novel framework for contextual query retrieval. By incorporating a corpus-centric KG, our approach bridges the semantic gap between user queries and the target corpus. Recognizing that real-world KGs are often incomplete, we propose an effective method for extracting subgraph triplets, which are then used to generate contextual information for the input query. Experiments on widely used benchmark datasets demonstrate the effectiveness and potential of the proposed approach.

Limitations

Although KG-CQR demonstrates promising results, several limitations warrant consideration for future improvements:

KG Construction Challenges: The construction of the corpus-centric knowledge graph relies heavily on external LLMs, such as LLama-3.3-70B, for entity and relation extraction. This process is susceptible to errors in named entity recognition (NER), relation extraction (RE), and entity linking (EL), which can propagate through the pipeline and affect the quality of the extracted subgraph. In domains with sparse or noisy unstructured data, the resulting KG may lack completeness or accuracy, limiting the effectiveness of KG-CQR.

Scalability of Subgraph Extraction: The subgraph extraction process, while effective, can be computationally intensive for large-scale knowledge graphs with millions of triples. The semantic similarity computation at the sentence level, using textual triplet representations (TTRs), increases computational overhead, potentially limiting scalability in real-time retrieval systems or resource-constrained environments.

Limited Evaluation Scope: The evaluation of KG-CQR was limited to two benchmark datasets—RAGBench and MultiHopRAG—which, although diverse, may not fully capture the breadth of real-world retrieval scenarios. To better assess the generalizability of the framework, further evaluation on additional datasets, especially those involving cross-lingual tasks or domain-specific knowledge, is warranted.

References

Hasan Abu-Rasheed, Christian Weber, and Madjid Fathi. 2024. Knowledge graphs as context sources for llm-based explanations of learning recommendations. In *IEEE Global Engineering Education Conference, EDUCON 2024, Kos Island, Greece, May 8-11, 2024*, pages 1–5. IEEE.

Hiteshwar Kumar Azad and Akshay Deepak. 2019. Query expansion techniques for information retrieval: A survey. *Inf. Process. Manag.*, 56(5):1698–1735.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, IJCNLP 2023 -Volume 1: Long Papers, Nusa Dua, Bali, November 1 - 4, 2023*, pages 675–718. Association for Computational Linguistics.

Sebastian Bruch, Siyu Gai, and Amir Ingber. 2024. An analysis of fusion functions for hybrid retrieval. *ACM Trans. Inf. Syst.*, 42(1):20:1–20:35.

Manh-Ha Bui, Toan Tran, Anh Tran, and Dinh Q. Phung. 2021. Exploiting domain-specific features to enhance domain generalization. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 21189–21201.

Chi-Min Chan, Chunpu Xu, Ruibin Yuan, Hongyin Luo, Wei Xue, Yike Guo, and Jie Fu. 2024. RQ-RAG: learning to refine queries for retrieval augmented generation. *CoRR*, abs/2404.00610.

Xinran Chen, Xuanang Chen, Ben He, Tengfei Wen, and Le Sun. 2024. Analyze, generate and refine: Query expansion with llms for zero-shot open-domain QA. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 11908–11922. Association for Computational Linguistics.

Wentao Ding, Jinmao Li, Liangchuan Luo, and Yuzhong Qu. 2024. Enhancing complex question answering over knowledge graphs through evidence pattern retrieval. In *Proceedings of the ACM on Web Conference 2024, WWW 2024, Singapore, May 13-17, 2024*, pages 2106–2115. ACM.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. 2024. From local to global: A graph RAG approach to query-focused summarization. *CoRR*, abs/2404.16130.

Robert Friel, Masha Belyi, and Atindriyo Sanyal. 2024. Ragbench: Explainable benchmark for retrieval-augmented generation systems. *CoRR*, abs/2407.11005.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2024. Gptscore: Evaluate as you desire. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 6556–6576. Association for Computational Linguistics.

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

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick S. H. 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 2020, Online, November 16-20, 2020*, pages 6769–6781. Association for Computational Linguistics.

Xiaoxi Li, Jiajie Jin, Yujia Zhou, Yuyao Zhang, Peitian Zhang, Yutao Zhu, and Zhicheng Dou. 2025. From matching to generation: A survey on generative information retrieval. *ACM Trans. Inf. Syst.* Just Accepted.

Xingxuan Li, Ruochen Zhao, Yew Ken Chia, Bosheng Ding, Shafiq Joty, Soujanya Poria, and Lidong Bing. 2024. Chain-of-knowledge: Grounding large language models via dynamic knowledge adapting over heterogeneous sources. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Yanming Liu, Xinyue Peng, Jiannan Cao, Shi Bo, Yanxin Shen, Xuhong Zhang, Sheng Cheng, Xun Wang, Jianwei Yin, and Tianyu Du. 2025. Bridging context gaps: Leveraging coreference resolution for long contextual understanding. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, Apr 24-28, 2025*. OpenReview.net.

Yaojie Lu, Qing Liu, Dai Dai, Xinyan Xiao, Hongyu Lin, Xianpei Han, Le Sun, and Hua Wu. 2022. Unified structure generation for universal information extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 5755–5772. Association for Computational Linguistics.

Linhao Luo, Yuan-Fang Li, Gholamreza Haffari, and Shirui Pan. 2024. Reasoning on graphs: Faithful and interpretable large language model reasoning. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. 2023. Query rewriting in retrieval-augmented large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 5303–5315. Association for Computational Linguistics.

Shengyu Mao, Yong Jiang, Boli Chen, Xiao Li, Peng Wang, Xinyu Wang, Pengjun Xie, Fei Huang, Huajun Chen, and Ningyu Zhang. 2024. Rafe: Ranking feedback improves query rewriting for RAG. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 884–901. Association for Computational Linguistics.

John X. Morris and Alexander M. Rush. 2024. Contextual document embeddings. *CoRR*, abs/2410.02525.

Thi Nguyen, Linhao Luo, Fatemeh Shiri, Dinh Phung, Yuan-Fang Li, Thuy-Trang Vu, and Gholamreza Haffari. 2024. Direct evaluation of chain-of-thought in multi-hop reasoning with knowledge graphs. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 2862–2883. Association for Computational Linguistics.

Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2024. Unifying large language models and knowledge graphs: A roadmap. *IEEE Trans. Knowl. Data Eng.*, 36(7):3580–3599.

Hoang Pham, Thanh-Do Nguyen, and Khac-Hoai Nam Bui. 2025. Verify-in-the-graph: Entity disambiguation enhancement for complex claim verification with interactive graph representation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5181–5197, Albuquerque, New Mexico. Association for Computational Linguistics.

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

Stephen E. Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2022. Colbertv2: Effective and efficient retrieval via lightweight late interaction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 3715–3734. Association for Computational Linguistics.

Bhaskarjit Sarmah, Benika Hall, Rohan Rao, Sunil Patel, Stefano Pasquali, and Dhagash Mehta. 2024. Hybridrag: Integrating knowledge graphs and vector retrieval augmented generation for efficient information extraction. *CoRR*, abs/2408.04948.

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D. Manning. 2024. RAPTOR: recursive abstractive processing for tree-organized retrieval. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Jiashuo Sun, Chengjin Xu, Lumingyuan Tang, Saizhuo Wang, Chen Lin, Yeyun Gong, Lionel M. Ni, Heung-Yeung Shum, and Jian Guo. 2024. Think-on-graph: Deep and responsible reasoning of large language model on knowledge graph. In *The Twelfth International Conference on Learning Representations,*

ICLR 2024, Vienna, Austria, May 7-11, 2024. OpenReview.net.

Yixuan Tang and Yi Yang. 2024. Multihop-rag: Benchmarking retrieval-augmented generation for multi-hop queries. *CoRR*, abs/2401.15391.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 10014–10037. Association for Computational Linguistics.

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

Yuqi Wang, Boran Jiang, Yi Luo, Dawei He, Peng Cheng, and Liangcai Gao. 2024a. Reasoning on efficient knowledge paths: Knowledge graph guides large language model for domain question answering. *CoRR*, abs/2404.10384.

Yuxia Wang, Minghan Wang, Muhammad Arslan Manzoor, Fei Liu, Georgi Georgiev, Rocktim Jyoti Das, and Preslav Nakov. 2024b. Factuality of large language models: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 19519–19529. Association for Computational Linguistics.

Yu Xia, Junda Wu, Sungchul Kim, Tong Yu, Ryan A. Rossi, Haoliang Wang, and Julian McAuley. 2025. Knowledge-aware query expansion with large language models for textual and relational retrieval. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4275–4286, Albuquerque, New Mexico. Association for Computational Linguistics.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-pack: Packed resources for general chinese embeddings. 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*, pages 641–649. ACM.

Yao Xu, Shizhu He, Jiabei Chen, Zihao Wang, Yangqiu Song, Hanghang Tong, Guang Liu, Jun Zhao, and Kang Liu. 2024. Generate-on-graph: Treat LLM as both agent and KG for incomplete knowledge graph question answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 18410–18430. Association for Computational Linguistics.

Hongbin Ye, Ningyu Zhang, Hui Chen, and Huajun Chen. 2022. Generative knowledge graph construction: A review. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 1–17. Association for Computational Linguistics.

Le Zhang, Yihong Wu, Qian Yang, and Jian-Yun Nie. 2024. Exploring the best practices of query expansion with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 1872–1883. Association for Computational Linguistics.

Shunyu Zhang, Yaobo Liang, Ming Gong, Daxin Jiang, and Nan Duan. 2022. Multi-view document representation learning for open-domain dense retrieval. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 5990–6000. Association for Computational Linguistics.

Yuqi Zhu, Xiaohan Wang, Jing Chen, Shuofei Qiao, Yixin Ou, Yunzhi Yao, Shumin Deng, Huajun Chen, and Ningyu Zhang. 2024. Llms for knowledge graph construction and reasoning: recent capabilities and future opportunities. *World Wide Web (WWW)*, 27(5):58.

A Appendix

A.1 GPT-score Criteria

Following the work in (Fu et al., 2024), we define the GPT-Score with three criteria for the measurement as follows:

- **Correctness**: alignment of the generated answer with the reference answer
- **Faithfulness**: whether the generated answer remains true to the given context
- **Relevance**: how well the retrieved context and the generated answer address the query

A.2 Comprehensive Experimential Results

A.2.1 Multi-Step Retrieval for RAG with BGE

Building on the earlier analysis (Table 4), Table 5 presents results for multi-step reasoning RAG performance using BGE as the retrieval baseline, along with KG-CQR. The key observations are as follows: i) **Dense retrieval outperforms sparse retrieval across all model sizes**: BGE consistently outperforms BM25 in terms of F1 score and GPT-Score,

Model	Retrieval	F1	#Iter	#Score
LLama-3.2-3B	BGE	0.411	2.665	3.242
LLama-3.1-8B		0.434	2.272	3.528
LLama-3.3-70B		0.448	1.48	3.576
LLama-3.2-3B	KG-CQR +BGE	0.432	2.378	3.317
LLama-3.1-8B		0.438	1.812	3.532
LLama-3.3-70B		0.452	1.23	3.878

Table 5: Performance of multi-step reasoning RAG with BGE

which demonstrates that dense retrieval via BGE retrieves more semantically relevant contexts than BM25, supporting more accurate and efficient reasoning; ii) **KG-CQR improves both BM25 and BGE retrieval**: Adding KG-CQR on top of both BM25 and BGE enhances performance by enriching the query with context-relevant knowledge. Although the improvement margin is narrower in the BGE setting, KG-CQR still consistently enhances performance, highlighting its generality across retrieval methods.

A.2.2 Fusion Embeddings Experiments

Table 6 show the comprehensive evaluation on the value of α to fuse the input query and context embeddings (Equation 9). As results, setting $\alpha = 0.7$ consistently yields the best overall performance.

A.2.3 Full Retrieval Results across Backbones

Table 7 and Table 8 demonstrate the full experimental results across various backbones, including LLama-3.2-3B and LLama-3.1-8B, respectively. Similar to the results on LLama-3.3-70B, the KG-CQR + BGE backbone at α = 0.7 yields the best performance for both models, in which LLama-3.1-8B shows slight improvements over LLama-3.2-3B, particularly in MultiHop-RAG tasks.

A.3 BFS with Beam Search Algorithm

Algorithm 2 presents the pseudocode for the BFS with Beam Search. Given the hyperparameter Beam width (e.g., equal to 3), the algorithm explores explicit paths (triplets) that represent meaningful connections between entities within the given subgraph.

Algorithm 2 BFS Algorithm with Beam Search

function BFSBEAM($\mathcal{T}_{KG}$, e_s, e_t, T_{set}, L)
  $Q \leftarrow \text{Queue}(\{\langle e_s, \emptyset \rangle\})$; $P \leftarrow \emptyset$
  Load Embedding mode: enc
  $v_q \leftarrow q \neq \emptyset ? \text{enc}(q) : \text{None}$
  $S \leftarrow \emptyset$
  $W \leftarrow 3$ ▷ Beam width for Beam Search
  while $Q \neq \emptyset$ **do**
    $(\text{node}, p) \leftarrow Q.\text{dequeue}()$
    if $|p| > L$ **then**
      continue
    end if
    if node $= e_t$ **and** $p \neq \emptyset$ **then**
      $P \leftarrow P \cup \{p\}$
    end if
    if node $= e_t$ **then**
      continue
    end if
    for $\{u, r, v, \text{TTR}\} \in \mathcal{T}_{KG}(u, v, r)$ **where** $u =$ node **do**
      if $v \notin p.\text{entities}$ **then**
        $v_p \leftarrow \text{enc}(\text{TTR})$
        $s \leftarrow v_q \neq \text{None} ? \text{Mean}(\cos(v_p, v_q)) : 0$
        $p_{\text{new}} \leftarrow p \cup \{u, r, v\}$
        $S \leftarrow S \cup \{(p_{\text{new}}, s)\}$
      end if
    end for
    Sort S by score descending
    for $(p_{\text{new}}, s) \in S$ **take top** W **do**
      $u \leftarrow p_{\text{new}}.\text{last\_node}$
      $Q.\text{enqueue}((u, p_{\text{new}}))$
    end for
    $S \leftarrow \emptyset$
  end while
  return P
end function

A.4 Error Analysis with Examples

To better understand the behavior of the KG-CQR, we performed a qualitative error analysis on six representative multi-hop queries from the MultiHop-RAG dataset with three corrected retrievals (Table 9) and three with incorrect retrievals (Table 10). We compared the outputs of KG-CQR against those of HyDE and the human-annotated Ground Truth.

Based on the results in Table 9, there are several assumptions as follows: i) KG-CQR demonstrates strong performance in disambiguating entities. For instance, in the query "Did one of CBS's performers create a scandal?", KG-CQR retrieves documents specifically related to the mentioned performer and event. This shows that incorporating knowledge graph information improves precision by retrieving documents more closely aligned with the query context; ii) In time-sensitive queries like

	RAGBench				MultiHop-RAG			
Backbone	mAP	Recall@5	Recall@10	Recall@25	mAP	Recall@5	Recall@10	Recall@25
				$\alpha = 0.3$				
KG-CQR + DPR	0.32	0.325	0.385	0.462	0.143	0.169	0.239	0.354
KG-CQR + BGE	0.528	0.513	0.596	0.664	0.224	0.247	0.35	0.499
				$\alpha = 0.5$				
KG-CQR + DPR	0.323	0.327	0.391	0.469	0.14	0.165	0.237	0.351
KG-CQR + BGE	0.539	0.527	0.609	**0.676**	0.235	0.253	0.364	0.515
				$\alpha = 0.7$				
KG-CQR + DPR	0.316	0.319	0.384	0.462	0.129	0.157	0.224	0.34
KG-CQR + BGE	**0.542**	**0.529**	**0.61**	0.675	**0.24**	**0.261**	**0.371**	**0.525**

Table 6: Fusion embeddings with different values of α (Equation 9)

	RAGBench				MultiHop-RAG			
Backbone	mAP	Recall@5	Recall@10	Recall@25	mAP	Recall@5	Recall@10	Recall@25
KG-CQR + BM25	0.386	0.388	0.446	0.507	**0.236**	**0.253**	0.359	0.520
				$\alpha = 0.3$				
KG-CQR + DPR	0.312	0.319	0.382	0.458	0.129	0.152	0.221	0.337
KG-CQR + BGE	0.517	0.505	0.588	0.661	0.203	0.225	0.332	0.481
				$\alpha = 0.5$				
KG-CQR + DPR	0.319	0.327	0.388	0.465	0.132	0.156	0.225	0.341
KG-CQR + BGE	0.531	0.520	0.602	0.669	0.219	0.239	0.350	0.507
				$\alpha = 0.7$				
KG-CQR + DPR	0.313	0.319	0.384	0.460	0.125	0.151	0.219	0.335
KG-CQR + BGE	**0.537**	**0.524**	**0.604**	**0.672**	0.230	0.251	**0.366**	**0.522**

Table 7: Full experimental results of LLama-3.2-3B with different setting of α (Equation9)

	RAGBench				MultiHop-RAG			
Backbone	mAP	Recall@5	Recall@10	Recall@25	mAP	Recall@5	Recall@10	Recall@25
KG-CQR + BM25	0.391	0.391	0.448	0.505	0.236	0.251	0.357	0.515
				$\alpha = 0.3$				
KG-CQR + DPR	0.325	0.329	0.391	0.462	0.138	0.166	0.234	0.352
KG-CQR + BGE	0.523	0.509	0.591	0.659	0.216	0.237	0.341	0.489
				$\alpha = 0.5$				
KG-CQR + DPR	0.327	0.330	0.394	0.467	0.136	0.162	0.233	0.351
KG-CQR + BGE	0.535	0.522	0.603	0.669	0.227	0.247	0.359	0.510
				$\alpha = 0.7$				
KG-CQR + DPR	0.318	0.322	0.387	0.462	0.127	0.151	0.220	0.338
KG-CQR + BGE	**0.538**	**0.526**	**0.606**	**0.672**	**0.236**	**0.255**	**0.370**	**0.522**

Table 8: Full experimental results of LLama-3.1-8B with different setting of α (Equation9)

"Which events occurred in Week 12?", KG-CQR accurately retrieves temporally relevant content, whereas HyDE often returns general or loosely connected documents. This suggests that KG signals enhance temporal grounding in multi-hop retrieval tasks; iii) For bridge-type queries that require chaining multiple pieces of information (e.g., "Does the article from Wendy refer to the same city?"), KG-CQR performs well by retrieving documents that correctly capture the intermediate and final entities. This indicates improved multi-hop coherence over baseline methods.

Despite these strengths, the proposed KG-CQR shows notable limitations in the following areas

(Table 10): i) **Contextual Drift and Irrelevant Retrievals**: KG-CQR struggles with queries requiring fine-grained temporal reasoning, comparative analysis, or interpretation of subjective content. These limitations stem from insufficient temporal representation and the lack of deep semantic modeling needed to capture nuanced relationships and contrasting viewpoints; ii) **Limited Multi-hop Coherence**: For queries requiring reasoning across multiple documents, KG-CQR sometimes retrieved disconnected evidence, failing to form a complete answer path.

A.5 Prompt Template

For better reproducibility, we present all prompt templates in the appendix. Below is a quick reference list outlining the prompt templates and their usages:

- Figure 5: Prompt the task instruction for KG construction.
- Figure 6: Prompt the task instruction for textual triplet representation.
- Figure 7: Prompt the task instruction for filtering irrelevant triplets.
- Figure 8: Prompt the task instruction for contextual generation.

Query q	HyDE@5	KG-CQR@5	Ground Truth
Did the CBSSports.com article report Kenneth Walker III remaining healthy and uninjured during a game, similarly to how the Sporting News article reports injuries for Tee Higgins, Noah Brown, Treylon Burks, and Kadarius Toney preventing their participation in Week 12?	**D1:** Meanwhile, corner CJ Henderson (concussion) was a full participant on Friday and carries no designation heading into the weekend...' **D2:** He left Week 2 after suffering a concussion and was absent in Week 3; then was not part of the game plan much in Week 4 (7.7% target share against Philadelphia)...' **D3:** He was ruled questionable to return. NFL Media reported on Monday that Kupp suffered a low ankle sprain and will be evaluated going forward...' **D4:** Geno Smith's struggles complicate their fantasy prospects, as well...' **D5:** Head coach Ron Rivera called the injury "significant" earlier this week...'	**D1:** Geno Smith's struggles complicate their fantasy prospects, as well...' **D2:** Meanwhile, corner CJ Henderson (concussion) was a full participant on Friday and carries no designation heading into the weekend...' **D3:** When asked if that means for Thursday's matchup against San Francisco, the coach said, "I would think so."...' **D4:** Week 11 of the 2023 NFL season has provided plenty of drama, from the Bears hanging with the Lions to the Giants getting a rebound from emergency quarterback Tommy DeVito...' **D5:** Miami ruled him questionable to return with a knee injury, and while he later returned to the sidelines from a locker-room visit, he was replaced on the field indefinitely by Raheem Mostert...'	**D1:** When asked if that means for Thursday's matchup against San Francisco, the coach said, "I would think so."...' **D2:** Geno Smith's struggles complicate their fantasy prospects, as well...' **D3:** Walker's struggled under center, Tillman is set up for a high-usage day against the Rams with Amari Cooper (ribs) banged up...' **D4:** Miami ruled him questionable to return with a knee injury, and while he later returned to the sidelines from a locker-room visit, he was replaced on the field indefinitely by Raheem Mostert...'
Does the article from Wired suggest that Sony headphones do not offer the best value in their class during the Walmart Cyber Monday Deals, while the article from Music Business Worldwide indicates that Artists are seeking deals that offer more control and better economics, or do both articles suggest a common trend in seeking value and control in their respective fields?	**D1:** Black Friday is often a boon for deals on headphones and earbuds, and this year is no different...' **D2:** Engadget has been testing and reviewing consumer tech since 2004. Our stories may include affiliate links; if you buy something through a link, we may earn a commission...' **D3:** But in both the case of Universal Music Group and Warner Music Group, they're – currently anyway – not the biggest megastars on either company's books...' **D4:** This is one of the few sales we've seen all year, which makes their very high asking price a lot more palatable...' **D5:** Nothing is more frustrating than buying a new pair of headphones, an OLED TV, or a backpack just to find out that you could have gotten it for a lot cheaper somewhere else...'	**D1:** They're light on extras like noise canceling but at this price, they're a great investment as your go-to workout companions...' **D2:** Black Friday is often a boon for deals on headphones and earbuds, and this year is no different...' **D3:** Luckily they've already gotten a discount, which makes it easier to land their class-leading noise canceling, great sound, and luxuriously comfy design that's loaded with modern features...' **D4:** Engadget has been testing and reviewing consumer tech since 2004. Our stories may include affiliate links; if you buy something through a link, we may earn a commission. Read more about how we evaluate products...' **D1:** This is one of the few sales we've seen all year, which makes their very high asking price a lot more palatable...'	**D1:** Spanish and Latin artists have much more options to develop their audiences and monetize their music at each stage of their career...' **D2:** They're light on extras like noise canceling but at this price, they're a great investment as your go-to workout companions...'
Which company, covered by both TechCrunch and The Verge, is not only claimed to have developed an AI model with superior architecture that rivals GPT-4 but also has been accused of altering the internet's appearance and harming news publishers' bottom lines through anticompetitive practices?	**D1:** Hey, folks, welcome to Week in Review (WiR), TechCrunch's regular newsletter that recaps the past few days in tech. AI stole the headlines once again, with tech giants from Google to X (formerly Twitter) heading off against OpenAI for chatbot supremacy...' **D2:** And on a company level, Meta is doing all it can to encourage collaboration and "openness," recently partnering with Hugging Face to launch a new startup accelerator designed to spur adoption of open source AI models...' **D3:** Google, OpenAI and Microsoft, a close OpenAI partner and investor, have been among the chief critics of Meta's open source AI approach, arguing that it's potentially dangerous and disinformation-encouraging...' **D4:** By 2020, the Knowledge Graph had grown to 500 billion facts about 5 billion entities. But much of the "collective intelligence" that Google tapped into was content "misappropriated from Publishers," the complaint alleges...' **D5:** The lawsuit reiterates this concern, claiming that Google's recent advances in AI-based search were implemented with "the goal of discouraging end-users from visiting the websites of Class members who are part of the digital news and publishing line of commerce."...'	**D1:** A new class action lawsuit filed this week in the U.S. District Court in D.C. accuses Google and parent company Alphabet of anticompetitive behavior in violation of U.S. antitrust law, the Sherman Act, and others, on behalf of news publishers...' **D2:** This week, Google took the wraps off of Gemini, its new flagship generative AI model meant to power a range of products and services including Bard, Google's ChatGPT competitor...' **D3:** The lawsuit reiterates this concern, claiming that Google's recent advances in AI-based search were implemented with "the goal of discouraging end-users from visiting the websites of Class members who are part of the digital news and publishing line of commerce."...' **D4:** By 2020, the Knowledge Graph had grown to 500 billion facts about 5 billion entities. But much of the "collective intelligence" that Google tapped into was content "misappropriated from Publishers," the complaint alleges...' **D5:** Hey, folks, welcome to Week in Review (WiR), TechCrunch's regular newsletter that recaps the past few days in tech. AI stole the headlines once again, with tech giants from Google to X (formerly Twitter) heading off against OpenAI for chatbot supremacy...'	**D1:** "I used to do those types of tactics, so I couldn't hate on anybody personally," she said. "If people have a problem with Google's results, they have to ask themselves, is it the fault of the SEOs?" she asked...' **D2:** A new class action lawsuit filed this week in the U.S. District Court in D.C. accuses Google and parent company Alphabet of anticompetitive behavior in violation of U.S. antitrust law, the Sherman Act, and others, on behalf of news publishers...' **D3:** Sure, she called herself a "thought leader," and yes, sure, she had changed her last name to improve her personal branding by more closely associating herself with her grandmother's uncle, the artist Man Ray...' **D4:** This week, Google took the wraps off of Gemini, its new flagship generative AI model meant to power a range of products and services including Bard, Google's ChatGPT competitor...'

Table 9: Examples of KG-CQR with correctly retrieved documents. Blue texts are corrected retrieved documents

Query q	HyDE@5	KG-CQR@5	Ground Truth
Has the approach of Sportsbooks in adjusting betting lines and odds, as reported by Sporting News after October 4, 2023, and before November 1, 2023, remained consistent?	**D1:** For instance, when examining odds for the next Super Bowl champion released shortly after the previous Super Bowl, these odds are based mostly on the recently concluded season... **D2:** They are basing their odds on past performance and expected future accomplishments, as well as the quality of the team around the top candidates for the award. Thus, the odds are quite favorable... **D3:** When such information becomes public, sportsbooks may adjust the odds accordingly. Professional Bettors: Large wagers from sharp bettors or professional gamblers can cause the lines to shift... **D4:** The past few weeks of the 2023 NFL season have reminded us that no matter how smooth you sail to start the voyage, choppy waters will surely come at some point... **D5:** Let's say the Chiefs win by exactly three, a distinct possibility since a single field goal decides most NFL games...	**D1:** For instance, when examining odds for the next Super Bowl champion released shortly after the previous Super Bowl, these odds are based mostly on the recently concluded season... **D2:** When such information becomes public, sportsbooks may adjust the odds accordingly. Professional Bettors: Large wagers from sharp bettors or professional gamblers can cause the lines to shift... **D3:** They are basing their odds on past performance and expected future accomplishments, as well as the quality of the team around the top candidates for the award... **D4:** The past few weeks of the 2023 NFL season have reminded us that no matter how smooth you sail to start the voyage, choppy waters will surely come at some point. We started the first six weeks with a best bets winning percentage of well over ... **D5:** Do point spread odds change? Yes, point spread odds can change, and these shifts are commonly referred to as "line movement."	**D1:** It's important to note that in PGA and other golf tournaments, there are usually many players, so the odds can be much higher than in head-to-head sports matchups, given the broader field of competition... **D2:** BetMGM Sportsbook: As one of the most recognizable names in the gambling industry, BetMGM knows how to attract and keep customers with competitive odds for all bet types, including futures bets and the NBA Rookie of the Year... **D3:** When the lines are first released for NBA ROTY honors, the season hasn't even started yet, so there are no statistics, trends, or player news... **D4:** Does overtime count in my moneyline bet? Yes, in most sports and with most sportsbooks (including new betting sites), overtime (or any extra time or tiebreakers) does count in a moneyline bet.
Does the TechCrunch article on generative AI in the enterprise suggest that CIOs are more cautious in their AI adoption strategy compared to the belief of business leaders mentioned in another TechCrunch article, who think AI will be essential for all businesses within five years?	**D1:** To hear the hype from vendors, you would think that enterprise buyers are all in when it comes to generative AI. But like any newer technology, large companies tend to move cautiously... **D2:** I'd venture to guess more exposure for its burgeoning generative AI platform... **D3:** Expect more moves like that from 2024's OpenAI as the caution and academic reserve that the previous board exerted gives way to an unseemly lust for markets and customers... **D4:** Google, OpenAI and Microsoft, a close OpenAI partner and investor, have been among the chief critics of Meta's open source AI approach, arguing that it's potentially dangerous and disinformation-encouraging... **D5:** The NMPA's submission, dated October 30, 2023, pulls no punches. It starts off by stressing that its membership – US music publishers major and independent – are "not opposed" to AI...	**D1:** To hear the hype from vendors, you would think that enterprise buyers are all in when it comes to generative AI... **D2:** I'd venture to guess more exposure for its burgeoning generative AI platform. IBM's most recent earnings were boosted by enterprises' interest in generative AI, but the company has stiff competition in Microsoft and OpenAI... **D4:** Expect more moves like that from 2024's OpenAI as the caution and academic reserve that the previous board exerted gives way to an unseemly lust for markets and customers... **D4:** The NMPA's submission, dated October 30, 2023, pulls no punches. It starts off by stressing that its membership – US music publishers major and independent – are "not opposed" to AI...' **D5:** Google On generative AI, Google's report discusses "recent progress in large-scale AI models" which it suggests...	**D1:** "So we've been doing this whole push for AI over the last maybe six or nine months and we're at the point right now where we're building specific use cases for each different team and function within the firm."... **D2:** Third, the application is only as sophisticated as the data that it is fed. Proprietary data is necessary for specific and relevant insights and to ensure others cannot replicate the final product... **D3:** That's going to take setting up some structure and organization around how this gets implemented over time, says Jim Rowan, principal at Deloitte, who is working with clients around how to build generative AI across companies in an organized fashion...
Does 'The Independent - Life and Style' article suggesting Prince William's emotional state regarding Princess Diana's death align with the same publication's depiction of the events leading up to her death in 'The Crown season six'?	**D1:** He is not located, but later walks back to the house on his own accord, drenched in rain. "14 hours, that poor boy was gone," the Queen later says... **D2:** The show also features the pair's death in a car crash in Paris. As the new season arrives, and fans wonder what in The Crown is based in reality, here's everything you need to know... **D3:** She then poses for them in her swimsuit, but complains in a later episode that they can "never relax" with the press "constantly" around... **D4:** After staying several days on Mohamed Al Fayed's yacht, the boys return home to London where their father, the then-Prince of Wales, accompanies them to Balmoral Castle to vacation with the rest of the royal family in Scotland... **D5:** During the interview, the outlet noted that Smith said his wife's memoir "kind of woke him up" and that he has now realised she is more...	**D1:** The show also features the pair's death in a car crash in Paris. As the new season arrives, and fans wonder what in The Crown is based in reality, here's everything you need to know... **D2:** He is not located, but later walks back to the house on his own accord, drenched in rain. "14 hours, that poor boy was gone," the Queen later says... **D3:** After staying several days on Mohamed Al Fayed's yacht, the boys return home to London where their father, the then-Prince of Wales, accompanies them to Balmoral Castle to vacation with the rest of the royal family in Scotland... **D4:** She then poses for them in her swimsuit, but complains in a later episode that they can "never relax" with the press "constantly" around... **D5:** Asks the Queen if she'd received the invitation to Camilla's 50th birthday, to which she says she has, but cannot attend as she's in Derbyshire...	**D1:** Stay ahead of the trend in fashion and beyond with our free weekly Lifestyle Edit newsletter Stay ahead of the trend in fashion and beyond with our free weekly Lifestyle Edit newsletter Please enter a valid email address... **D2:** However, at the inquest into the death in 2007, the jury were shown CCTV footage of him purchasing an engagement ring worth £11,600 in a jewellers across the square from the Ritz on the afternoon of the crash...

Table 10: Examples of KG-CQR with incorrectly retrieved documents. Red texts indicate notable limitations of KG-CQR in several areas, such as contextual drift or limited complex multi-hop coherence

Instruction Prompt for LLM-Powered Graph Construction

You are a top-tier algorithm designed for extracting information in " "structured formats to build a knowledge graph. Your task is to identify " "the entities and relations requested with the user prompt from a given " "text.

You must generate the output in a JSON format containing a list " 'with JSON objects. Each object should have the keys: "head", ' '"head_type", "relation", "tail", and "tail_type". The "head" ' "key must contain the text of the extracted entity with one of the types " "from the provided list in the user prompt. The "head_type" key must contain the type of the extracted head entity, which must be one of the types from {node_labels_str}.

if node_labels else " ", The "relation" key must contain the type of relation between the "head" ' and the "tail", which must be one of the relations from {rel_types_str}.'

if rel_types else "", The "tail" key must represent the text of an extracted entity which is the tail of the relation, and the "tail_type" key must contain the type of the tail entity from {node_labels_str}.

if node_labels else "", "Attempt to extract as many entities and relations as you can. Maintain " "Entity Consistency: When extracting entities, it's vital to ensure " 'consistency.

If an entity, such as "John Doe", is mentioned multiple ' "times in the text but is referred to by different names or pronouns " '(e.g., "Joe", "he"), always use the most complete identifier for ' "that entity.

The knowledge graph should be coherent and easily " "understandable, so maintaining consistency in entity references is " "crucial.",

"IMPORTANT NOTES:\n- Don't add any explanation and text.",

Figure 5: Prompt the task instruction for KG construction

Instruction Prompt for LLM-Powered Textual Triplet Representation (Equation 2)

You are an expert in extracting information from text, your task is to find pieces of information that mention the relationship of two objects in the relationship and synthesize them into one paragraph.
The summary paragraph must be written in English. The response MUST be the summary only without any explanation.

Passage: {}

Triplets: {}

Figure 6: Prompt the task instruction for textual triplet representation.

Instruction Prompt for Filtering Triplets (Equation 7)

You are a grader assessing relevance of a list of retrieved passages to a user question. The goal is to filter out erroneous retrievals.
Return only the passage index whether the passage is relevant to the question. Provide the output as a JSON with passage index seperated by a comma and no premable or explanation.

Here is the list of retrieved text: {text}

Here is the user question: {question}

Figure 7: Prompt the task instruction for filtering irrelevant triplets

Instruction Prompt for Contextual Generation (Equation 8)

You are a helpful assistant responsible for generating a comprehensive summary of the data provided below. Given the list of triplets that may relation with each other. Please write a Concise summary of triplets that aim to provide a contextual information. The output just generate a concise summary without any explanation.

Please note that if the provided triplets are contradictory, please resolve the contradictions and provide a single, coherent summary (no need Here is part)

Input Triplets: {triplets}

Figure 8: Prompt the task instruction for contextual representation