

# HyKGE: A Hypothesis Knowledge Graph Enhanced RAG Framework for Accurate and Reliable Medical LLMs Responses

Xinke Jiang<sup>1,2\*</sup>, Ruizhe Zhang<sup>1,2\*</sup>, Yongxin Xu<sup>1,2\*</sup>, Rihong Qiu<sup>1,2\*</sup>  
Yue Fang<sup>1,2</sup>, Zhiyuan Wang<sup>1,2</sup>, Jinyi Tang<sup>1,2</sup>, Hongxin Ding<sup>1,2</sup>  
Xu Chu<sup>1,2,4,6</sup>, Junfeng Zhao<sup>1,2,5†</sup>, Yasha Wang<sup>2,3,6†</sup>

<sup>1</sup>School of Computer Science, Peking University, Beijing, China

<sup>2</sup>Key Lab of HCST (PKU), MOE; SCS, Peking University, China

<sup>3</sup>National Engineering Research Center For Software Engineering, Peking University, China

<sup>4</sup>Center on Frontiers of Computing Studies, Peking University, Beijing, China

<sup>5</sup>Big Data Technology Research Center, Nanhu Laboratory, Jiaxing, China

<sup>6</sup>Peking University Information Technology Institute, Tianjin Binhai, China

{xinkejiang, nostradamus, xuyx, rihongqiu}@stu.pku.edu.cn, {zhaojf, wangyasha}@pku.edu.cn

## Abstract

In this paper, we investigate the retrieval-augmented generation (RAG) based on Knowledge Graphs (KGs) to improve the accuracy and reliability of Large Language Models (LLMs). Recent approaches suffer from insufficient and repetitive knowledge retrieval, tedious and time-consuming query parsing, and monotonous knowledge utilization. To this end, we develop a **Hypothesis Knowledge Graph Enhanced (HyKGE)** framework, which leverages LLMs' powerful reasoning capacity to compensate for the incompleteness of user queries, optimizes the interaction process with LLMs, and provides diverse retrieved knowledge. Specifically, HyKGE explores the zero-shot capability and the rich knowledge of LLMs with Hypothesis Outputs to extend feasible exploration directions in the KGs, as well as the carefully curated prompt to enhance the density and efficiency of LLMs' responses. Furthermore, we introduce the HO Fragment Granularity-aware Rerank Module to filter out noise while ensuring the balance between diversity and relevance in retrieved knowledge. Experiments on medical and general domain datasets two LLM turbos demonstrate the superiority of HyKGE in terms of accuracy and explainability. Code is available at <sup>1</sup>.

## 1 Introduction

**Large Language Models (LLMs)**, such as ChatGPT (OpenAI, 2022) and GPT-4 (OpenAI, 2023), have achieved remarkable progress in pivotal areas. By undergoing pre-training on massive text corpora and aligning fine-tuning to follow human instructions (Ziegler et al., 2020; Wang et al., 2023e; Ma et al., 2025; ?), they have recently demonstrated exceptional performance in

a range of downstream tasks (Kaplan et al., 2020). These achievements underscore the vast potential of LLMs in understanding and generating natural language (Vu et al., 2024), especially in the medical domain (Kraljevic et al., 2022; Yang et al., 2023c; Zhu et al., 2023; Wang et al., 2023c; Xiong et al., 2023; Wang et al., 2023a; Zhang et al., 2023a; Pal and Sankarasubbu, 2023; Bao et al., 2023). Despite the advancements of fine-tuning, they still encounter significant challenges, including the difficulty in avoiding factual inaccuracies (i.e., hallucinations and limited explainability) (Ji et al., 2023a; Cao et al., 2020; Ji et al., 2023b), data constraints (i.e. token resource limit, high training costs, and privacy concerns)<sup>2</sup>, catastrophic forgetting (Gao et al., 2024), outdated knowledge (He et al., 2022), and a lack of expertise in handling specific domains or highly specialized queries (Kandpal et al., 2023). This undermines their reliability in areas where accountability and trustworthiness are crucial and infallible in the medical area (Ji et al., 2023a; Song et al., 2024; Lai et al., 2023; Xu et al., 2023b, 2024b).

**Retrieval-Augmented Generation (RAG)**, enhances content generation by retrieving external information, reduces factual errors in knowledge-intensive tasks with the help of external knowledge and is seen as a promising solution to address incorrect answers, hallucinations, and insufficient interpretability (Izacard et al., 2022b; Asai et al., 2023b,a; Jiang et al., 2025; Xu et al., 2024a). Among the numerous external information sources (Yang et al., 2024), **knowledge graphs** (KGs), as a structured data source refined and extracted through advanced information extraction algorithms, can provide higher quality context. Compared to documents, KGs embody structured knowledge (Zhong et al., 2023; Ji et al., 2022; Xu et al., 2025, 2023a; Yang et al., 2023b; Ma et al., 2023b), providing succinct content and facilitating the analysis of intricate relationships among entities, leading to advanced inference capabilities and enabling extrapolation for efficient knowledge retrieval. They are considered by many research works

\*Equal contribution.

†Corresponding author.

<sup>1</sup><https://github.com/Artessay/HyKGE>

<sup>2</sup><https://www.youtube.com/watch?v=ahnGLM-RC1Y>

to improve the accuracy and reliability of answers provided by LLMs (Pan et al., 2024; Wen et al., 2023; Jiang et al., 2024). However, the gap between unstructured user queries of inconsistent quality and structured, high-quality KGs (Soman et al., 2023a) poses significant challenges on how to properly parse user intent for improving the robustness of retrieved knowledge (**pre-retrieval phase**) and how to handle the abundant retrieved knowledge (**post-retrieval phase**), which are detailed as:

**Challenge I: At the pre-retrieval phase, previous works suffer from how to parse user intent and retrieve reasonable knowledge based on varying-quality user query.** Some works are based on the Retrieve-Read framework, which initially obtains knowledge through dense vector retrieval according to user queries (Ma et al., 2023a; Qu et al., 2021; Yu et al., 2022). However, they are stricken with issues such as unclear expressions and lack of semantic information in the user’s original query. This misalignment between the semantic spaces of user queries and high-quality structured knowledge leads to the retrieval of knowledge that is of insufficient quality and may contain redundant information and noise (Barnett et al., 2024). In addition, the excessive redundant knowledge can lead to a waste of token resources, and the response speed of LLMs will drop sharply, which adversely damages the performance in real-world applications (Finardi et al., 2024).

**Challenge II: At the pre-retrieval phase, how to align user intent with high-quality structured knowledge while reducing interactions with LLMs remains an unresolved issue.** Some works enable LLMs to step-by-step utilize knowledge to enhance intent parsing and inference of user queries. They facilitate the acquisition of planning and reflective abilities in LLMs’ interactions with KGs through multi-round chain-of-thought requests (Sun et al., 2023; Li et al., 2023b; Wang et al., 2023b; Asai et al., 2023b; Yu et al., 2023a). However, they are constrained by the expensive time overhead of multiple interactions with LLMs and the cumulative errors in the distributed reasoning process.

**Challenge III: At the post-retrieval phase, previous studies often struggle with the dilemma of balancing the diversity and relevance of the retrieved knowledge.** Recent post-retrieval models typically apply similarity filtering or a reranking approach in response to user queries to prune retrieved results (Toro et al., 2023; Finardi et al., 2024; Cuconasu et al., 2024). However, user queries often exhibit notably monotonous properties and sparsely distributed keywords because the prevalence of natural language descriptions will tend to dilute its concentration (Breuer et al., 2023). Conversely, KGs are characterized by their inherently structured nature, resulting in a high knowledge density within retrieved results. As a consequence, pruning knowledge solely based on the user query can lead to a misalignment in knowledge density and the final result is often highly correlated yet excessively repetitive, significantly diminishing the efficacy of RAG. Therefore,

one of the primary challenges in the post-retrieval phase is to balance the trade-off between relevant knowledge and diverse ones (Breuer et al., 2023).

To cope with these challenges, we put forward the **Hypothesis Knowledge Graph Enhanced (HyKGE)** framework, a novel method based on the hypothesis output module (HOM) (Gao et al., 2022a) to explore, locate, and prune search directions for accurate and reliable LLMs responses in pre-retrieval phase and greatly preserve the relevance and diversity of search results at in post-retrieval phase. **i)** Specifically, in the **pre-retrieval phase**, our key idea is that the zero-shot capability and rich knowledge of LLMs can compensate for the incompleteness of user queries, facilitating alignment with high-quality external knowledge. For example, when facing the question “*After meals, I feel a bit of stomach reflux. What medicine should I take?*”, if retrieval is based solely on the key entity “*stomach reflux*” as illustrated in Figure 1(a), a large amount of noise will be introduced due to the broad semantics of the entity. However, if LLMs are guided to explore how to solve the problem, they will provide additional clues related to “*stomach acid*”, “*H2 receptor antagonists*” and “*proton pump inhibitors*” as illustrated in Figure 1(b), based on the knowledge acquired during their pre-training and instruction fine-tuning phases, offering exploration directions for retrieval on the KGs. **ii)** Meanwhile, HyKGE utilizes the flexibility of natural language in prompts to set constraints, enabling LLMs to provide as comprehensive information as possible when outputting hypothesis results, thereby reducing the number of interactions and improving efficiency. **iii)** In the **post-retrieval stage**, to further enhance the alignment between user queries and external knowledge inference paths, we propose a Hypothesis Output-based (HO) Fragment Granularity-aware, which utilizes multiple short snippets from the hypothesis outputs as well as the user query to rerank and filter the retrieved knowledge, greatly avoiding the filtering of diverse knowledge. It ensures fine-grained interaction and filtering while addressing the issue of imprecise matching between monotonous and sparse text (user query) with multi-element and dense text (retrieved knowledge). Through comprehensive experiments, our main contributions can be summarized as follows:

- At the pre-retrieval phase, we leverage the zero-shot capability of LLMs to obtain an exploratory and hypothesis output, transforming the incomplete and non-professional nature of user queries. Corresponding anchor entities are then identified from the hypothesis output on the KGs, providing a direction for exploration and pruning retrieval space. Simultaneously, we utilize the knowledge chains to rectify errors and illogicalities in the hypothesis outputs, mitigating hallucinations and false knowledge problems.
- At the post-retrieval stage, we propose a HO Fragment Granularity-aware rerank module to further enhance the knowledge density alignment between the

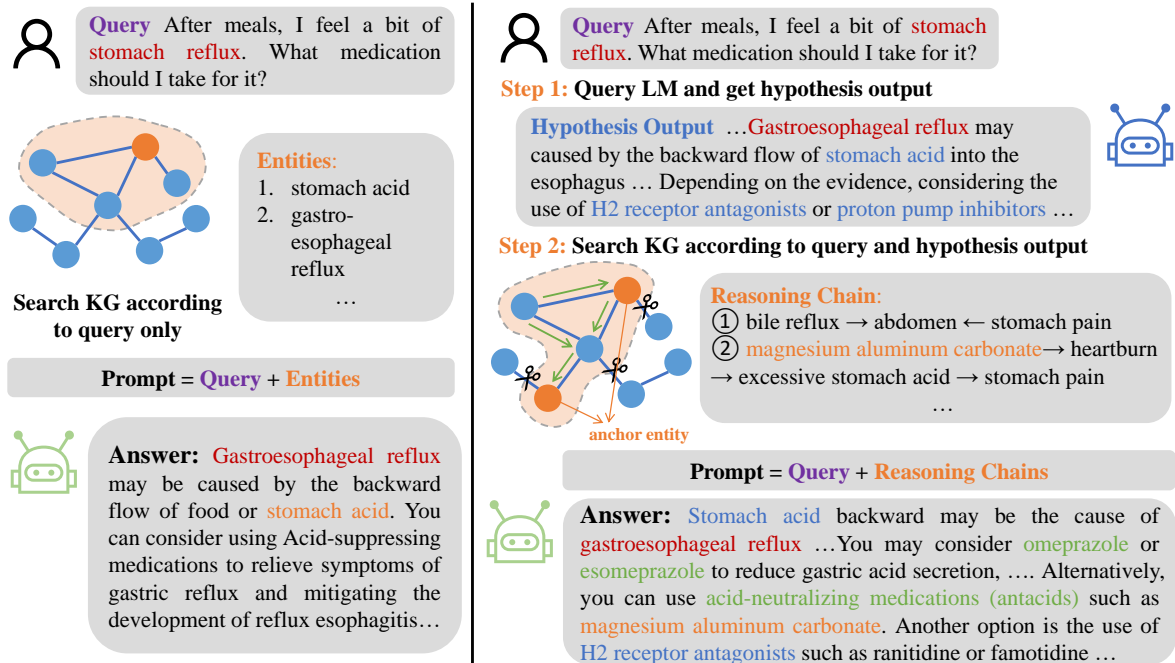


Figure 1: (a) KGRAG (Left). Basic KGRAG extracts key entities from user queries and searches for corresponding entities within KG, which are then fed into LLMs along with the query. (b) HyKGE (Right). HyKGE first queries LLMs to obtain hypothesis output and extracts entities from both the hypothesis output and the query. Then HyKGE retrieves reasoning chains between any two anchor entities and feeds the reasoning chains and query into LLMs.

retrieved reasoning chains and hypothesis outputs at a finer granularity, greatly preserving relevant yet diverse knowledge through the idea of divide-and-conquer.

- We validate the superiority of the HyKGE through various observations by experiments on two Chinese medical multiple-choice question datasets and one Chinese medical open-domain Q&A dataset with two LLM turbos. This integration of LLMs and KGs addresses key challenges in medical LLMs, notably in accuracy and explainability, and has potential applications in improving medical consultation quality, diagnosis accuracy, and expediting medical research.

## 2 Related Work

**Retrieval-Augmented Generation.** RAG incorporates the external knowledge retrieval component via prompt engineering to achieve more factual consistency, enhancing the reliability and interpretability of LLMs' responses (Lewis et al., 2021; Xu et al., 2024c). Classic RAG methods leverage retriever models to source relevant documents from large knowledge corpora (Xu et al., 2024a), followed by reranker models that distill contents and reader models for further processing (Pan et al., 2024; Sarthi et al., 2024; Zhang et al., 2024). Despite advancements in retriever (Qu et al., 2021; Ma et al., 2023a; Kim et al., 2024) and reranker efficiency (Cheng et al., 2021; Yu et al., 2022), they still encounter difficulty in acquiring high-quality datasets for training query-document pair retrievers or limited information in

user queries which weakens their generalization capability (Gao et al., 2022b). Moreover, some researches focus on fine-tuning reader LLMs, applying instruction-tuning with retrieved knowledge or RAG API calls (Luo et al., 2023a; Izacard et al., 2022b; Asai et al., 2024; Yoran et al., 2024; Wang et al., 2024; Zhang et al., 2023b; Lin et al., 2024). However, creating such datasets is also challenging due to the need for manual label correction, which in turn, may erode LLMs generalization capabilities and cause catastrophic forgetting in routine Q&A tasks.

Beyond optimizing submodels, HyDE (Gao et al., 2022b) introduces an innovative method where instruction-following LLMs generate hypothesis documents based on user queries to enhance retriever performance, particularly in zero-shot scenarios. Other methods like CoN (Yu et al., 2023b) and CoK (Li et al., 2023c) involve LLMs in note-making and step-wise reasoning verification through customized prompts, and greatly rely on frequent interactions with LLMs. However, such an approach is excessively inefficient for deployment in real-world Q&A scenarios.

Our HyKGE, uses LLM hypothesis output for exploratory directions in KGs and corrects model errors using graph reasoning chains during pre-retrieval, and applies fine-grained alignment in post-retrieval to maintain effective, diverse knowledge, enhancing retrieval efficiently without fine-tuning or excessive interactions.

**Knowledge Graph Query-Answer.** Compared to knowledge stored in document repositories (Izacard



et al., 2022a), the knowledge contained within KGs has the advantages of being structured and inferable, rendering it a more suitable source for supplementing LLMs (Luo et al., 2023b; Jiang et al., 2023; Liu et al., 2021; Kang et al., 2023; Sen et al., 2023; Varshney et al., 2023). However, how to design a retriever to extract knowledge from KGs and how to design interaction strategies between LLM and KGs are still in the exploratory stage<sup>3</sup>. KGRAG (Soman et al., 2023a) and GraphRAG uses the user query as a reference for retrieval in KGs, which suffers from misalignment between high-quality structured knowledge and varying-quality queries. Semantic parsing methods allow LLMs to convert the question into a structural query (e.g., SPARQL), which can be executed by a query engine to derive the answers on KGs (Sun et al., 2020; Li et al., 2023a,c). However, these methods depend heavily on the quality of generated query sentences, displaying sub-par performance when confronted with intricate queries.

### 3 Preliminaries

**Definition 1. (Knowledge Graph)** Given a medical knowledge graph, denoted by  $\mathcal{KG} = (\mathcal{E}, \mathcal{R}, \mathcal{T}, \mathcal{D}, \mathcal{N})$ , where  $\mathcal{E} = \{e_1, \dots, e_N\}$  is the set of entities,  $\mathcal{R} = \{r_1, \dots, r_P\}$  is the set of relations, and  $\mathcal{T} = \{(e_{i_{head}}, r_{t_i}, e_{i_{tail}}) \mid 1 \leq i \leq T, e_{i_{head}}, e_{i_{tail}} \in \mathcal{E}, r_{t_i} \in \mathcal{R}\}$  is the set of head-relation-tail triplets (facts). Additionally,  $d_i \in \mathcal{D}$  represents the entity description of  $e_i$ , and  $\mathcal{N}_v = \{(r, u) \mid (v, r, u) \in \mathcal{T}\}$  stands for the set of neighboring relations and entities of an entity  $v$ .

**Definition 2. (Knowledge Graph Retrieval)** Knowledge Graph Retrieval (Reinanda et al., 2020) is a module that focuses on efficiently retrieving relevant information from  $\mathcal{KG}$  based on the user query  $Q$ . In KGs, information is represented as entities, relations, and attributes, forming a structured network. The goal of retrieval is to find entities or relationships that additionally supply knowledge for LLMs. Particularly, we retrieve knowledge from the matched entities  $\{e_j\}$  such as entity names, entity types, descriptions  $\{d_j\}$  and even triplets or subgraphs  $\mathcal{G}_{e_j} = (e_j, \mathcal{T}_j, d_j)$ .

## 4 Method

In this section, we detail our proposed HyKGE, and the overall framework is illustrated in Figure 2. In general, we will discuss our model from the four pipeline architectures:

- **Pre-Retrieval Phase** includes the Hypothesis Output Module (HOM) and the NER Module (NM). HOM leverages LLMs to obtain hypothesis output by exploring possible answers. Then NM extracts medical entities from HO and the user query.
- **Retrieval on Knowledge Graph** utilizes the extracted entities as anchors to search three distinct types of reasoning chains interlinking these anchors, providing relevant and logical knowledge.

- **Post-Retrieval Phase** utilizes the HO Fragment Granularity-aware rerank approach. First, the hypothesis output and the user query are segmented into discrete fragments, and subsequently, we rerank the retrieved reasoning chains based on the fragments.
- **LLM Reader** is fed with the user query and the pruned retrieved reasoning chains, organized with carefully designed prompts.

Next, we will delineate each phase in detail in the following subsections.

### 4.1 Pre-Retrieval Phase

Firstly, we let LLMs generate hypothesis outputs ( $\mathcal{HO}$ ) in response to user query  $Q$ , and then use the NER model to extract entities from both  $\mathcal{HO}$  and  $Q$ . During this process, LLMs utilize inherent medical knowledge to explore potential answers. Although  $\mathcal{HO}$  may contain factual errors or hallucinations between entities, the NER Module focuses solely on the extraction of entities while disregarding the relations, thus significantly isolating the correlation among medical entities. The subsequent graph retrieval phase (c.f. Section 4.2) searches the correct reasoning chains to discern and reintegrate the relationships between medical entities, avoiding LLMs’ shortages. The combination of HOM and NM provides us with a direction for exploration and identifies corresponding anchors in the KGs to guide subsequent graph retrieval, ensuring consistency and effectiveness in information processing.

#### 4.1.1 Hypothesis Output Module

To enhance the quality of  $\mathcal{HO}$ , due to LLMs’ robust reasoning abilities and potential as knowledge bases, we meticulously design instructions to guide LLMs in a step-by-step exploration and thoughtful consideration of problems. Prompts used in  $\mathcal{HO}$  can be referred to in Appendix A.1. Here, the prompt (a textual instruction) is denoted as  $\mathcal{P}_{\mathcal{HO}}$ , and  $Q$  to  $\mathcal{HO}$  as:

$$\mathcal{HO} = \text{LLM}(Q \mid \mathcal{P}_{\mathcal{HO}}). \quad (1)$$

Thus, in light of the powerful reasoning abilities as well as the knowledgeable medical cognition, galore medical knowledge relevant to  $Q$  is discovered.

#### 4.1.2 NER Module

Although there still remains a possibility of an inaccurate comprehension of relationships within  $\mathcal{HO}$  (i.e. hallucinations or misunderstanding between medical entities), training a discriminative model or using other general-domain LLMs for authenticity  $\mathcal{HO}$  is extremely labor-intensive and will lead to error accumulation. To tackle this issue, we extract entities instead of relationships, and utilize the completely unmistakable triplets in KGs for authenticity instead of the relations analyzed in  $\mathcal{HO}$ . As a consequence, we have trained a medical Named Entity Recognition (NER) model using the CMEEE dataset<sup>4</sup> (Zhang et al., 2022; Hongying et al.,

<sup>3</sup><https://siwei.io/talks/graph-rag-with-jerry/1>

<sup>4</sup><https://tianchi.aliyun.com/dataset/144495>

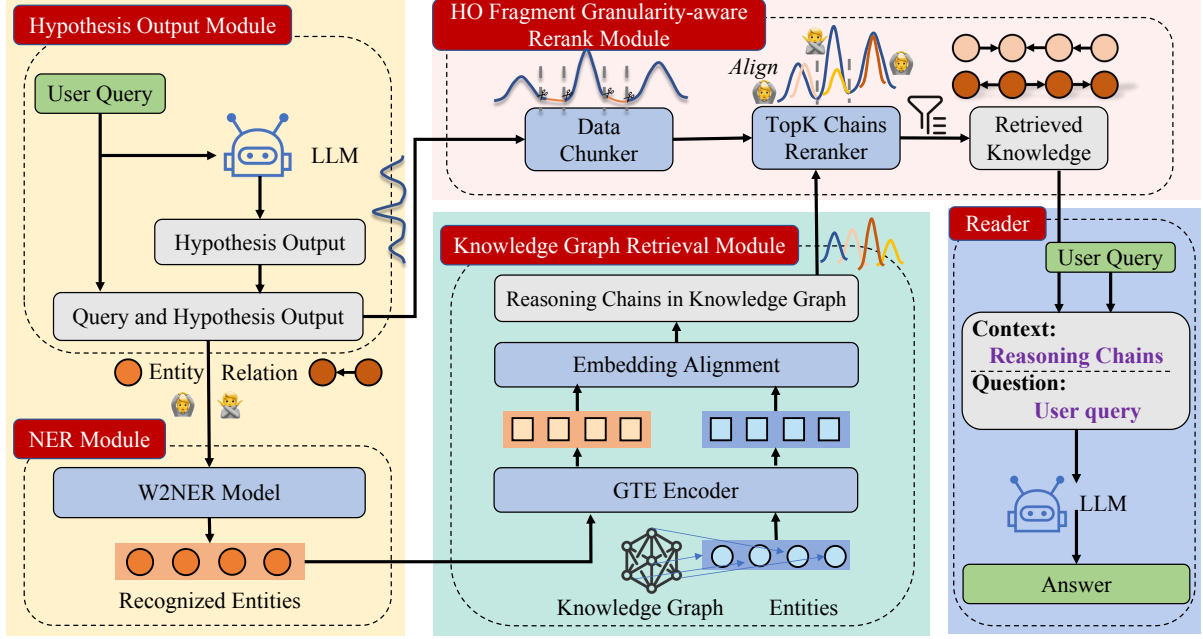


Figure 2: Overall framework. HyKGE first feeds user query  $Q$  through LLMs and obtains Hypothesis Output  $HO$ . Then through NER Module, a W2NER model is applied to recognize entities and isolate relations. Through GTE Encoder, these recognized entities are then linked with KG entities. After that, HyKGE extracts three types of relevant reasoning chains from KGs. Then, due to sparseness of  $Q$ , in HO Fragment Granularity-aware Rerank Module, HyKGE chunks  $Q$  and  $HO$  and align with reasoning chains via a TopK Chains Reranker, to eliminate noise. Finally, we organize retrieved knowledge with the user query and obtain responses through the LLM Reader.

2020). Our NER Module is built upon the W2NER model (Li et al., 2021), the state-of-the-art word-word NER model that effectively addresses three primary types of NER situations (flat, overlapped, discontinuous). This medical NER model can wonderfully extract medical entities from complex medical contexts:

$$\mathcal{U} = [u_1, \dots, u_{|\mathcal{U}|}] = \text{NER}(Q \oplus HO), \quad (2)$$

where  $\oplus$  is the concatenation function and  $u_i$  represents the corresponding extracted entity.

## 4.2 Knowledge Graph Retrieval Module

### 4.2.1 Embedding Alignment

Subsequently, we link the potential entity to  $\mathcal{KG}$  using dense retrieval methods. This process involves employing an encoding model, denoted as  $\text{enc}(\cdot)$ , to encode the potential entity  $u_i$  and entities  $\mathcal{E}$  within  $\mathcal{KG}$ . To be specific, we utilize the GTE embedding model (Li et al., 2023d) "gte\_sentence-embedding"<sup>5</sup>, which is currently the top-performing model for text vector embedding in the retrieval field. GTE Encoder follows a two-stage training process: initially using a large-scale dataset with weak supervision from text pairs, followed by fine-tuning with high-quality manually labeled data using Contrastive Learning (Le-Khac et al., 2020).

<sup>5</sup>[https://www.modelscope.cn/models/damo/nlp\\_gte\\_sentence-embedding](https://www.modelscope.cn/models/damo/nlp_gte_sentence-embedding)

Then, the inner product similarity between the embeddings of  $u_i$  and  $\mathcal{E}$  is then computed. The entity with the highest similarity, surpassing a predefined threshold  $\delta \in [0, 1]$ , is considered a match. This linkage process can be formulated as follows:

$$\begin{aligned} \text{sim}(u_i, e_j) &= \langle \text{enc}(u_i), \text{enc}(e_j) \rangle, \quad u_i \in \mathcal{U}, e_j \in \mathcal{E}, \\ u_i \leftrightarrow e_j &\text{ iff } e_j = \{\underset{e_k \in \mathcal{E}}{\text{argmax}} \text{sim}(u_i, e_k) \mid \text{sim}(u_i, e_j) > \delta\}, \end{aligned}$$

where  $\delta \in [0, 1]$  is the threshold hyper-parameter. We utilize the same encoding model  $\text{enc}(\cdot)$  to embed each medical entity, and  $\langle \text{enc}(u_i), \text{enc}(e_j) \rangle$  denotes the inner product between extracted entities and KGs entities for achieving graph entity linking. Finally, the matched entities set is denoted as  $\mathcal{E}_Q$ .

### 4.2.2 Search Reasoning Chains in KGs

Next, using matched entities, we explore reasoning chains within  $k$  hops and consolidate this knowledge along with descriptions of the head and tail entities. Considering various knowledge graph retrieval methods, we opt for utilizing reasoning chains between entities for several reasons: **i)** Reasoning Chains provide richer logical knowledge provided for LLMs to help it digest, compared to entities and entity descriptions alone. **ii)** Reasoning chains help LLM Reader understand the relationships between different entities, thereby alleviating hallucinations and error problems. **iii)** Reasoning chains act as an efficient pruning mechanism, filtering

out noise more effectively than subgraphs and saving token resources.

As a consequence, in light of (Xu et al., 2022), we consider three possible reasoning chains from medical perspective: **i) Path** (head-to-tail) as  $\text{path}_{ij}$ , for comprehensively analyzing the triggering and causal relationships between diseases and symptoms (Mas-Ponte and Supek, 2020; Otlu et al., 2023). **ii) Co-ancestor chain** (tail-to-tail) as  $\text{chainCA}_{ij}$ , for referring similar physiological or environmental factors for better analogical diagnosis (Bergstrom et al., 2022). **iii) Co-occurrence chain** (head-to-head) as  $\text{chainCO}_{ij}$ , for better capturing the pathological characteristics and evolution of diseases (Dong et al., 2021). In general, the reasoning chain set  $\mathcal{RC}$  after the graph retrieval are as:

$$\begin{aligned} \text{path}_{ij} &= \underbrace{(e_i \rightarrow r. \rightarrow e. \rightarrow \dots \rightarrow r. \rightarrow e_j, d_i, d_j)}_{\text{within } k \text{ hops}}, \\ \text{chainCA}_{ij} &= \underbrace{(e_i \rightarrow r. \rightarrow e. \leftarrow \dots \leftarrow r. \leftarrow e_j, d_i, d_j)}_{\text{within } k \text{ hops}}, \\ \text{chainCO}_{ij} &= \underbrace{(e_i \leftarrow r. \leftarrow e. \rightarrow \dots \rightarrow r. \rightarrow e_j, d_i, d_j)}_{\text{within } k \text{ hops}}, \end{aligned}$$

where  $e_i, e_j \in \mathcal{E}_Q$ , and  $r.$  is the relation. For any entity pair in  $\mathcal{E}_Q$ , we collect reasoning chains within  $k(k \geq 2)$  hops and description of head & tail entity  $d_i, d_j$  in  $\mathcal{KG}$ .

### 4.3 Post-Retrieval Phase

Through retrieval, a large amount of reasoning chains will be collected. However, due to the considerable noise and the shortage of token resources (c.f. Challenge III in Section 1), we employ a reranker model to prune and eliminate irrelevant noise knowledge by reranking reasoning chains, leading to more efficient token resource utilization. For the reranker base model, we use the "bge\_reranker\_large"<sup>6</sup> (Xiao et al., 2023), trained through large-scale text pairs with asymmetric instruction tuning, to map text to a low-dimensional dense vector to rerank  $\text{top}K$  documents.

Moreover, due to the varying knowledge densities between queries and reasoning chains, traditional reranking based solely on  $Q$  may filter out valuable knowledge acquired through HOM, resulting in a repetitive and monotonous situation. As a consequence, we innovatively combine  $\mathcal{HO}$  and  $Q$ , rather than relying solely on user query, utilizing the richer medical knowledge contained in  $\mathcal{HO}$ . Practically, we first remove stop words from natural language and then we use the chunk method to segment  $\mathcal{HO}$  and  $Q$ :

$$\{\mathcal{C}\} = \text{Chunk}(Q \oplus \mathcal{HO}), \quad (3)$$

where  $\{\mathcal{C}\} = \{c_1, \dots, c_i, \dots, c_{|\{\mathcal{C}\}|}\}$  is the segmented fragments, with carefully selected chunk window size  $lc$  and overlap size  $oc$ . Then, we leverage a reranking model denoted as  $\text{Rerank}(\mathcal{RC}, \{\mathcal{C}\}; \text{top}K)$ , which means referring to segment set  $\mathcal{RC}$ , we select the  $\text{top}K$

reranked retrieved chains from set  $\{\mathcal{C}\}$ . Acting as a filter, the reranking model reevaluates the significance of each chain, considering various factors such as relevance, coherence, and informativeness:

$$\mathcal{RC}_{\text{prune}} = \text{Rerank}(\mathcal{RC}, \{\mathcal{C}\}; \text{top}K), \quad (4)$$

where  $|\mathcal{RC}_{\text{prune}}| = \text{top}K$ .

### 4.4 LLM Reader

Finally, we link  $\mathcal{RC}_{\text{prune}}$  with directed arrows, combined with the description of the head and tail entities, and feed the retrieved knowledge as well as user query  $Q$  to LLM Reader via prompt engineering. The prompt format  $\mathcal{P}_{\text{Reader}}$  can be referred to in Appendix A.1 and the LLM's answer can be expressed as:

$$\text{Answer} = \text{LLM}(Q, \mathcal{RC}_{\text{prune}} \mid \mathcal{P}_{\text{Reader}}). \quad (5)$$

### 4.5 Overall Algorithm

Algorithm 1 shows the overall RAG process of HyKGE. Given the knowledge graph  $\mathcal{KG}$ , HyKGE first pre-embed the entity name using  $\text{enc}(\cdot)$ , and saves the vector locally (Lines 1-3). Then, we query LLM to obtain  $\mathcal{HO}$  in response to user query  $Q$  (Line 4). After that, we extract entities from  $\mathcal{HO}$  and  $Q$  (Line 5) and match them with  $\mathcal{KG}$  (Line 6). HyKGE then retrieves the reasoning chains from  $\mathcal{KG}$  (Line 7) while filtering the noise path with HO Fragment Granularity-aware rerank module (Line 8). At last, HyKGE organizes the retrieved knowledge and query via prompt (Line 9) and queries LLM Reader to get optimized answers (Line 10).

---

**Algorithm 1** The RAG process of HyKGE.

---

**Require:** Knowledge Graph  $\mathcal{KG} = (\mathcal{E}, \mathcal{R}, \mathcal{T}, \mathcal{D}, \mathcal{N})$ , token vocabulary set  $\mathcal{V}$ , user query  $Q$ , trained NER model  $\text{NER}(\cdot)$ , trained embedded model  $\text{enc}(\cdot)$ , trained Reranking model  $\text{Rerank}(\cdot)$ , Large Language Model  $\text{LLM}(\cdot)$ , hyper-parameters  $\delta, k, \text{top}K$ .

- 1: **for**  $e_i \in \mathcal{E}$  **do**       $\triangleright$  **Embed Knowledge Graph**
  - 2:     Save  $\text{enc}(e_i)$  locally;
  - 3: **end for**
  - 4: Obtain  $\mathcal{HO}$  via LLM:  $\mathcal{HO} = \text{LLM}(Q \mid \mathcal{P}_{\mathcal{HO}})$ ;       $\triangleright$  **Hypothesis Output**
  - 5: Extract entities  $\mathcal{U}$  from  $\mathcal{HO}$  and  $Q$  via Eq.(2);       $\triangleright$  **NER Module**
  - 6: Match  $\mathcal{U}$  with  $\mathcal{E}$  via Eq.(4.2.1) and attain  $\mathcal{E}_Q$ ;       $\triangleright$  **Entity Linking**
  - 7: Retrieved reasoning chains between any two anchor entities from  $\mathcal{KG}$ ;  $\triangleright$  **Knowledge Graph Retrieval**
  - 8: Filter noise reasoning chains with HO Fragment Granularity-aware Rerank Module;       $\triangleright$  **Prune knowledge**
  - 9: Organize knowledge  $\mathcal{RC}_{\text{prune}}$  with user query  $Q$  into prompt;
  - 10: Get optimized answer of LLMs;       $\triangleright$  **LLM Reader**
- 

<sup>6</sup><https://huggingface.co/BAAI/bge-reranker-large>

## 5 Experiments

In this section, we conduct experiments on three datasets to answer the following research questions:

- **RQ1** (Section 5.2): Does HyKGE outperform the state-of-the-art Knowledge Graph RAG methods using the same database source?
- **RQ2** (Appendix A.3, Appendix A.4, Appendix A.5): Is the framework we designed effective? What impact does each module have on the overall performance?
- **RQ3** (Section 5.3, Appendix A.3): Does the retrieved knowledge we provide enhance the interpretability of LLMs answers?
- **RQ4** (Appendix A.6): How sensitive is HyKGE to hyper-parameters retrieval hop  $k$  and rerank threshold  $topK$ ?

### 5.1 Experimental Setup

**Dataset.** Our experiments are conducted on two open-source query sets: MMCU-Medical (Zeng, 2023) and CMB-Exam (Wang et al., 2023d) datasets, which are designed for multi-task Q&A and encompass single and multiple-choice questions in the medical field, and one open-domain Q&A dataset CMB-Clin (Wang et al., 2023d) which is the inaugural multi-round question-answering dataset based on real, complex medical diagnosis and treatment records. We use *CMeKG* (Clinical Medicine Knowledge Graph) (BYAMBASUREN et al., 2020), *CPubMed-KG* (Large-scale Chinese Open Medical Knowledge Graph) and *Disease-KG* (Chinese disease Knowledge Graph) as external-knowledge source. Details of datasets are summarized in Appendix A.2.

**Baselines** To fairly verify whether HyKGE can effectively enhance LLMs, we selected the following two types of general-domain large models as the base model and explored the gains brought by HyKGE: GPT 3.5 and Baichuan13B-chat (Yang et al., 2023a). We compare the HyKGE results against eight other models: **Base Model (Base)** include GPT 3.5 and Baichuan13B-chat, **Knowledge Graph Retrieval-Augmented Generation (KGRAG)** (Soman et al., 2023b,a; Sen et al., 2023), **Query Expansion (QE)** (Azad and Deepak, 2019), **CHAIN-OF-NOTE (CoN)** (Yu et al., 2023b), **Chain-of-Knowledge (CoK)** (Li et al., 2023c), **Knowledge-Augmented Language Model Verification (KALMV)** (Baek et al., 2023), **Knowledge Graph Generative Pre-Training (KG-GPT)** (Kim et al., 2023) and **Summarizing Retrievals (SuRe)** (Kim et al., 2024). Detailed description of baselines can be referred in Appendix A.2.

**Evaluation Metrics.** We select **EM**, **PCR** to evaluate the multi-task Q&A dataset, and **ACJ**, **PPL**, **ROUGE-R**, **BLEU-1**, **BLEU-4** to evaluate the open-domain Q&A dataset. The detailed descriptions of these metrics are in Appendix A.2.2.

Table 1: Analysis Comparison of RAG methods. Average Duration is computed based on GPT 3.5 turbo.

Method	External Knowledge	LLMs RAG Opt.		Avg. Time (s)
		Finetuning	Retriever	
Base	✓	✗		7.42
KGRAG	✓	✗	1	13.28
QE	✓	✗	2	18.54
CoN	✓	✓	$\geq 2$	34.33
CoK	✓	✗	$\geq 4$	45.84
KALMV	✓	✗	$\geq 4$	47.23
KG-GPT	✓	✗	5	55.19
SuRE	✓	✗	$\geq 5$	63.08
HyKGE (ours)	✓	✗	2	19.76

#### 5.1.1 Baselines

### 5.2 Performance Comparison (RQ 1)

To answer RQ1, we conduct experiments and report results of the accuracy on the MMCU-Medical, CMB-Exam and CMB-Clin datasets with two LLM turbos GPT 3.5 and Baichuan 13B-Chat, as illustrated in Table 2 and Table 3. From the reported accuracy, we can find the following observations:

**Comparison of RAG methods and Base LLMs.** Through comparison, we observe that most RAG approaches do not consistently yield effective outcomes when integrated with KGs, especially in contrast with the Base model. For instance, the KGRAG method extracts triples from KG without engaging in essential post-processing steps like reranking and filtering, thereby infusing an overabundance of noise and compromising the interpretative performance of LLMs. As for QE tasks, while traditional QE methods typically show efficacy, LLMs demonstrate a notable difficulty in comprehending instructions that necessitate the task-specific rewriting of multiple-choice questions, which, in turn, detrimentally impacts LLMs performance in such scenarios. Moreover, this effect is particularly pronounced in weaker models, such as Baichuan, where the repercussions of these deficiencies are significantly magnified. However, the improvement in CoN, CoK, KG-GPT, SuRe and HyKGE is more remarkable, because leveraging LLMs to explore or organize knowledge can assist in finding more relational knowledge and the reranking or filtering methods can highly likely remove irrelevant noise knowledge chains, and contribute to accuracy improvement.

**Comparison of HyKGE and other RAG methods.** Firstly, it is evident that our model, HyKGE, outperforms the baseline models across all metrics. For instance, the EM and PCR scores see an improvement of approximately **8.55%-28.15%** and **10.45%-33.29%** for the MMCU-Medical dataset with GPT 3.5 turbo, and the BLEU-1 and ROUGE-R scores see an improvement of approximately **40.29%-278.47%** and **7.30%-51.69%** for the CMB-Clin dataset with GPT 3.5 turbo. This highlights the effectiveness of our modules in locating valid information and filtering noises in retrieved knowledge. Although CoK, CoN, KG-GPT and SuRe have achieved commendable results, their advancements are constrained in the knowledge search space, due to their focus on continuous knowledge understanding rather



Table 2: Performance comparison (in percent) on CMB-Exam and MMCU-Medical for medical Q&A answer. **Red shading** indicates the best-performing model, while **blue** signifies the second-best in the ablation study, and **green** signifies the second-best in baselines.

LLM Turbo	LLM	GPT 3.5				Baichuan 13B-Chat			
		MMCU-Medical		CMB-Exam		MMCU-Medical		CMB-Exam	
Method	Dataset	EM	PCR	EM	PCR	EM	PCR	EM	PCR
Baselines	Base	43.52	50.55	38.40	46.76	42.20	46.09	36.91	40.95
	KGRAG	38.74	43.38	38.00	42.26	34.37	38.51	39.92	45.84
	QE	40.28	46.79	36.35	41.84	38.25	44.23	34.27	38.79
	CoN	45.74	51.15	42.45	45.65	44.98	50.65	41.37	47.58
	CoK	45.15	52.35	42.32	45.98	45.15	51.19	41.87	47.95
	KALMV	39.24	43.77	38.24	43.37	36.17	40.85	38.61	43.92
	KG-GPT	45.08	52.16	41.49	45.72	44.25	50.97	39.92	45.20
	SuRe	44.81	51.49	41.37	44.27	44.77	50.24	39.49	46.22
Ours	HyKGE	<b>49.65</b>	<b>57.82</b>	<b>45.94</b>	<b>50.63</b>	<b>49.33</b>	<b>58.12</b>	<b>45.44</b>	<b>51.25</b>
Ablation	HyKGE (w/o HO)	41.08	49.74	34.40	40.14	39.55	45.28	33.33	35.42
	HyKGE (w/o Chains)	48.15	54.53	44.60	48.27	48.65	55.45	43.40	48.81
	HyKGE (w/o Description)	48.30	54.01	44.80	48.56	48.22	55.23	43.77	49.86
	HyKGE (w/o Fragment)	47.87	54.34	42.33	47.54	47.95	53.45	44.72	49.29
	HyKGE (w/o Reranker)	46.38	52.48	41.44	48.84	43.59	46.88	40.65	46.25

Table 3: RAG relevance and answer performance comparison (mean) on CMB-Exam, MMCU-Medical, and CMB-Clin for medical Q&A answer with GPT 3.5 Turbo.

Method	Dataset	MMCU-Medical			CMB-Exam			CMB-Clin			
		ACJ	PPL	ROUGE-R	ACJ	PPL	ROUGE-R	BLEU-1	BLEU-4	PPL	ROUGE-R
Baselines	Base	/	47.42	/	/	62.54	/	4.83	6.51	10.38	23.99
	KGRAG	13.38	151.22	5.31	18.40	218.67	11.25	5.34	8.77	61.81	22.15
	QE	25.53	28.75	14.05	31.91	29.57	16.64	8.85	18.67	28.32	26.24
	CoN	19.14	29.01	16.46	14.89	27.35	17.31	12.48	25.81	17.65	31.37
	CoK	18.45	24.38	18.23	16.77	28.69	19.94	12.35	24.79	21.57	30.86
	KALMV	14.42	147.22	7.21	18.77	233.49	12.84	5.72	8.27	80.46	23.16
	KG-GPT	32.03	25.76	15.90	38.70	24.01	17.72	13.03	26.14	15.54	28.42
	SuRe	20.16	26.49	16.91	22.27	30.81	16.18	10.54	24.82	16.84	29.18
Ours	HyKGE	<b>59.57</b>	<b>12.55</b>	<b>26.89</b>	<b>71.28</b>	<b>10.14</b>	<b>32.11</b>	<b>18.28</b>	<b>30.21</b>	<b>8.56</b>	<b>33.66</b>
Ablation	HyKGE (w/o HO)	41.49	15.57	22.30	51.48	11.23	29.01	7.15	11.55	8.96	30.48
	HyKGE (w/o Fragment)	38.30	18.95	23.63	41.91	11.26	26.89	11.28	23.09	8.99	31.40

than exploration. Moreover, compared to CoK, CoN, KG-GPT and SuRe, HyKGE avoids accumulating errors in the chain of thought while acquiring and retaining more relevant yet diverse knowledge. In summary, the HyKGE exhibits superior performance over all baselines with fewer interaction times with LLMs (c.f. Table 1). Evidenced by comprehension experiments, HyKGE demonstrates HO Module’s and the HO Fragment Granularity-aware rerank module’s effectiveness compared to CoN, CoK, KG-GPT and SuRe.

### 5.3 Interpretability Analysis (RQ 3)

In this section, we concentrate on evaluating the interpretability with three metrics **ACJ**, **BLEU**, **PPL** and **ROUGE-R** as shown in Table 3 to find out whether retrieved knowledge is effective and whether it can help LLMs reduce hallucinations. Several observations can be derived from the results.

**The relevance of knowledge retrieval.** For methods that interacted with LLMs and applied noise filtering modules, such as QE, CoK, CoN, SuRE and HyKGE, we notice that they often score higher on ACJ on MMCU-Medical and CMB-Exam, and ROUGE-

R on CMB-Clin dataset, reflecting the efficacy of the LLMs’ inherent knowledge and reasoning abilities as well as the importance of removing irrelevant knowledge. Moreover, the ACJ value of KG-GPT and QE is the second-to-best as they do not alter the semantics of the user query. Therefore, the knowledge retrieved by KG-GPT and QE have higher relevance with ACJ score, compared to CoK and CoN. Furthermore, it is noticed that our proposed HyKGE surpasses baselines with a performance gain of **84.19%**-**378.71%** and **133.33%**-**345.22%** on MMCU-Medical and CMB-Exam respectively, which demonstrates our superiority in solving misaligned knowledge density between query and retrieved knowledge. The marked decline in ACJ of w/o Fragment also supports HO Fragment Granularity-aware reranker’s role in keeping relevant knowledge. The BLEU and ROUGE-R scores on CMB-Clin also demonstrate HyKGE’s superiority, indicating that HyKGE could be more appropriate for and aligned with real-life doctor consultations, proving the effectiveness of HyKGE in information retrieval.

**Can LLMs utilize retrieved knowledge to reduce hallucinations?** As for method KGRAG, it fails to per-



form well on PPL and ROUGE-R, which is attributed to the provision of overly lengthy retrieved knowledge and redundant noise, resulting in the inability of the LLMs to extract useful information from the knowledge. The performance test of baselines consistently shows that our proposed HyKGE greatly reduces hallucinations and promotes LLMs to better utilize the retrieved knowledge, with performance gain of **57.77%-95.36%** and **61.03%-185.42%** on MMCU for PPL and ROUGE-R respectively. We argue the reason that the retrieved knowledge is more relevant and diverse because of the HOM and HO Fragment Granularity-aware Reranker, and its chain structure also stimulates the reasoning ability of LLMs. Others, i.e. QE, CoN, and CoK's have been greatly reduced because their rerankers cannot retain more diverse knowledge, resulting in LLMs' answers being too singular and ROUGE-R surely being lower. Notably, our performance on the CMB-Exam test set was superior, due to richer and more detailed description of medical questions, allowing us to obtain more diverse and relevant knowledge based on *HO & Q*.

## 6 Conclusion

In this paper, we proposed HyKGE, a hypothesis knowledge graph enhanced framework for LLMs to improve accuracy and reliability. In the pre-retrieval phase, we leverage the zero-shot capability of LLMs to compensate for the incompleteness of user queries by exploring searching directions through hypothesis outputs. In the post-retrieval phase, HyKGE applies a fragment reranking module to enhance the knowledge density alignment between user queries and retrieved knowledge, preserving relevant and diverse knowledge chains. The comprehensive experiments conducted on three medical Q&A tasks with two LLMs turbo demonstrate the effectiveness of HyKGE. Nevertheless, it remains worthwhile to contemplate how to dynamically optimize fragment granularity in the post-retrieval phase—a direction that we are committed to exploring actively in the future. In addition, despite the limitations of data sources and the high computational cost of LLMs, we will explore other language or domain-specific KGs in the future to enhance the scalability and generalization of HyKGE.

## Limitations

We acknowledge two limitations in this work. First, many knowledge graphs suffer from limited coverage, and in the future, we plan to address the challenge of handling queries that contain entities not represented in the knowledge graph. Second, our current approach is tailored to knowledge graphs and data sources in a specific language. Moving forward, we aim to extend our methodology to support multiple languages and validate its effectiveness using more general benchmarks.

## Acknowledgments

This work is supported by the National Natural Science Foundation of China (No.U23A20468).

## References

- Zehmakan Ahad N., Out Charlotte, and Khelejan Sajjad Hesamipour. 2023. Why rumors spread fast in social networks, and how to stop it. In *IJCAI*.
- Akari Asai, Sewon Min, Zexuan Zhong, and Danqi Chen. 2023a. Retrieval-based language models and applications. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 6: Tutorial Abstracts)*, pages 41–46.
- Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023b. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *arXiv preprint arXiv:2310.11511*.
- Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *ICLR*.
- Hiteshwar Kumar Azad and Akshay Deepak. 2019. Query expansion techniques for information retrieval: A survey. *Information Processing & Management*, page 1698–1735.
- Jinheon Baek, Soyeong Jeong, Minki Kang, Jong C. Park, and Sung Ju Hwang. 2023. Knowledge-augmented language model verification. In *EMNLP*.
- Zhijie Bao, Wei Chen, Shengze Xiao, Kuang Ren, Jiaao Wu, Cheng Zhong, Jiajie Peng, Xuanjing Huang, and Zhongyu Wei. 2023. [Disc-medllm: Bridging general large language models and real-world medical consultation](#). *Preprint*, arXiv:2308.14346.
- Scott Barnett, Stefanus Kurniawan, Srikanth Thudumu, Zach Brannelly, and Mohamed Abdelrazek. 2024. [Seven failure points when engineering a retrieval augmented generation system](#). *Preprint*, arXiv:2401.05856.
- Noam Berger, Christian Borgs, Jennifer T. Chayes, and Amin Saberi. 2005. On the spread of viruses on the internet. In *SODA*.
- Erik N. Bergstrom, Jens Luebeck, Mia Petljak, Azhar Khandekar, Mark Barnes, Tongwu Zhang, Christopher D. Steele, Nischalan Pillay, Maria Teresa Landi, Vineet Bafna, Paul S. Mischel, Reuben S. Harris, and Ludmil B. Alexandrov. 2022. [Mapping clustered mutations in cancer reveals apobec3 mutagenesis of ccdna](#). *Nature*, 602:510–517.
- Timo Breuer, Norbert Fuhr, and Philipp Schaer. 2023. [Validating synthetic usage data in living lab environments](#). *Journal of Data and Information Quality*.
- Odmaa BYAMBASUREN, Yunfei YANG, Zhi-fang SUI, Damai DAI, Baobao CHANG, Sujian LI, and Hongying ZAN. 2020. Preliminary study on the construction of chinese medical knowledge graph. *Journal of Chinese Information Processing*.
- Meng Cao, Yue Dong, Jiapeng Wu, and Jackie Chi Kit Cheung. 2020. [Factual error correction for abstractive summarization models](#). In *Proceedings of the*

- 2020 *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6251–6258, Online. Association for Computational Linguistics.
- Hao Cheng, Yelong Shen, Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2021. [Unitedqa: A hybrid approach for open domain question answering](#). *Preprint*, arXiv:2101.00178.
- Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonello, and Fabrizio Silvestri. 2024. [The power of noise: Redefining retrieval for rag systems](#). *Preprint*, arXiv:2401.14887.
- Guiying Dong, Jianfeng Feng, Fengzhu Sun, Jingqi Chen, and Xing-Ming Zhao. 2021. A global overview of genetically interpretable multimorbidities among common diseases in the uk biobank. *Genome Medicine*, 13(110).
- Paulo Finardi, Leonardo Avila, Rodrigo Castaldoni, Pedro Gengo, Celio Larcher, Marcos Piau, Pablo Costa, and Vinicius Caridá. 2024. [The chronicles of rag: The retriever, the chunk and the generator](#). *Preprint*, arXiv:2401.07883.
- Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2022a. [Precise zero-shot dense retrieval without relevance labels](#). *Preprint*, arXiv:2212.10496.
- Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2022b. [Precise zero-shot dense retrieval without relevance labels](#). *Preprint*, arXiv:2212.10496.
- Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2024. Retrieval augmented generation for large language models: A survey.
- Hangfeng He, Hongming Zhang, and Dan Roth. 2022. [Rethinking with retrieval: Faithful large language model inference](#). *Preprint*, arXiv:2301.00303.
- Zan Hongying, Li Wenxin, Zhang Kunli, Ye Yajuan, Chang Baobao, and Sui Zhifang. 2020. Building a pediatric medical corpus: Word segmentation and named entity annotation. In *Workshop on Chinese Lexical Semantics*, pages 652–664.
- Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022a. [Unsupervised dense information retrieval with contrastive learning](#). *Preprint*, arXiv:2112.09118.
- Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2022b. [Atlas: Few-shot learning with retrieval augmented language models](#). *Preprint*, arXiv:2208.03299.
- Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and Philip S. Yu. 2022. [A survey on knowledge graphs: Representation, acquisition, and applications](#). *IEEE Transactions on Neural Networks and Learning Systems*, 33(2):494–514.
- Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023a. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.
- Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023b. [Survey of hallucination in natural language generation](#). *ACM Comput. Surv.*, 55(12).
- Jinhao Jiang, Kun Zhou, Wayne Xin Zhao, Yaliang Li, and Ji-Rong Wen. 2023. [Reasoninglm: Enabling structural subgraph reasoning in pre-trained language models for question answering over knowledge graph](#). *Preprint*, arXiv:2401.00158.
- Xinke Jiang, Yue Fang, Rihong Qiu, Haoyu Zhang, Yongxin Xu, Hao Chen, Wentao Zhang, Ruizhe Zhang, Yuchen Fang, Xu Chu, et al. 2025. Tc-rag: Turing-complete rag’s case study on medical llm systems. *ACL 2025*.
- Xinke Jiang, Rihong Qiu, Yongxin Xu, Wentao Zhang, Yichen Zhu, Ruizhe Zhang, Yuchen Fang, Xu Chu, Junfeng Zhao, and Yasha Wang. 2024. Ragraph: A general retrieval-augmented graph learning framework. *NeurIPS 2024*.
- Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *International Conference on Machine Learning*, pages 15696–15707. PMLR.
- Minki Kang, Jin Myung Kwak, Jinheon Baek, and Sung Ju Hwang. 2023. [Knowledge graph-augmented language models for knowledge-grounded dialogue generation](#). *Preprint*, arXiv:2305.18846.
- Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. [Scaling laws for neural language models](#). *Preprint*, arXiv:2001.08361.
- Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen tau Yih. 2020. [Dense passage retrieval for open-domain question answering](#). *Preprint*, arXiv:2004.04906.
- Jaehyung Kim, Sangwoo Mo Jaehyun Nam, Jongjin Park, Sang-Woo Lee, Minjoon Seo, Jung-Woo Ha, and Jinwoo Shin. 2024. Sure: Summarizing retrievals using answer candidates for open-domain qa of llms. In *ICLR*.
- Jiho Kim, Yeonsu Kwon, Yohan Jo, and Edward Choi. 2023. Kg-gpt: A general framework for reasoning on knowledge graphs using large language models. In *ACL*.
- Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR*.

- Zeljko Kraljevic, Dan Bean, Anthony Shek, Rebecca Bendayan, Harry Hemingway, Joshua Au Yeung, Alexander Deng, Alfie Baston, Jack Ross, Esther Idowu, James T Teo, and Richard J Dobson. 2022. Fore-sight – generative pretrained transformer (gpt) for modelling of patient timelines using ehrs.
- Tuan Manh Lai, ChengXiang Zhai, and Heng Ji. 2023. Keblm: Knowledge-enhanced biomedical language models. *Journal of Biomedical Informatics*, 143:104392.
- Phuc H. Le-Khac, Graham Healy, and Alan F. Smeaton. 2020. Contrastive representation learning: A framework and review. *IEEE Access*, 8:193907–193934.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Preprint*, arXiv:2005.11401.
- Jingye Li, Hao Fei, Jiang Liu, Shengqiong Wu, Meishan Zhang, Chong Teng, Donghong Ji, and Fei Li. 2021. Unified named entity recognition as word-word relation classification. *Preprint*, arXiv:2112.10070.
- Tianle Li, Xueguang Ma, Alex Zhuang, Yu Gu, Yu Su, and Wenhui Chen. 2023a. Few-shot in-context learning for knowledge base question answering. *Preprint*, arXiv:2305.01750.
- Xingxuan Li, Ruochen Zhao, Yew Ken Chia, Bosheng Ding, Shafiq Joty, Soujanya Poria, and Lidong Bing. 2023b. Chain-of-knowledge: Grounding large language models via dynamic knowledge adapting over heterogeneous sources. *Preprint*, arXiv:2305.13269.
- Xingxuan Li, Ruochen Zhao, Yew Ken Chia, Bosheng Ding, Shafiq Joty, Soujanya Poria, and Lidong Bing. 2023c. Chain-of-knowledge: Grounding large language models via dynamic knowledge adapting over heterogeneous sources. *Preprint*, arXiv:2305.13269.
- Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023d. Towards general text embeddings with multi-stage contrastive learning. *Preprint*, arXiv:2308.03281.
- Xi Victoria Lin, Xilun Chen, Mingda Chen, Weijia Shi, Maria Lomeli, Rich James, Pedro Rodriguez, Jacob Kahn, Gergely Szilvasy, Mike Lewis, Luke Zettlemoyer, and Scott Yih. 2024. Ra-dit: Retrieval-augmented dual instruction tuning. In *ICLR*.
- Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. Lost in the middle: How language models use long contexts. *Preprint*, arXiv:2307.03172.
- Ye Liu, Yao Wan, Lifang He, Hao Peng, and Philip S. Yu. 2021. Kg-bart: Knowledge graph-augmented bart for generative commonsense reasoning. *Preprint*, arXiv:2009.12677.
- Hongyin Luo, Yung-Sung Chuang, Yuan Gong, Tianhua Zhang, Yoon Kim, Xixin Wu, Danny Fox, Helen Meng, and James Glass. 2023a. Sail: Search-augmented instruction learning. *Preprint*, arXiv:2305.15225.
- Linhao Luo, Yuan-Fang Li, Gholamreza Haffari, and Shirui Pan. 2023b. Reasoning on graphs: Faithful and interpretable large language model reasoning. *Preprint*, arXiv:2310.01061.
- Kaixin Ma, Hao Cheng, Yu Zhang, Xiaodong Liu, Eric Nyberg, and Jianfeng Gao. 2023a. Chain-of-skills: A configurable model for open-domain question answering. *Preprint*, arXiv:2305.03130.
- Xinyu Ma, Yasha Wang, Xu Chu, Liantao Ma, Wen Tang, Junfeng Zhao, Ye Yuan, and Guoren Wang. 2023b. Patient health representation learning via correlational sparse prior of medical features. *IEEE Transactions on Knowledge and Data Engineering*, 35(11):11769–11783.
- Xinyu Ma, Yifeng Xu, Yang Lin, Tianlong Wang, Xu Chu, Xin Gao, Junfeng Zhao, and Yasha Wang. 2025. DRESSing up LLM: Efficient stylized question-answering via style subspace editing. In *The Thirteenth International Conference on Learning Representations*.
- David Mas-Ponte and Fran Supek. 2020. Dna mismatch repair promotes apobec3-mediated diffuse hypermutation in human cancers. *Nature Genetics*, 52:958–968.
- OpenAI. 2022. Introducing chatgpt. <https://openai.com/blog/chatgpt>.
- OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.
- Burçak Otlu, Marcos Díaz-Gay, Ian Vermes, Maria Zhivagui, Mark Barnes, and Ludmil B. Alexandrov. 2023. Topography of mutational signatures in human cancer. *Cell Reports*, 42:112930.
- Ankit Pal and Malaikannan Sankarasubbu. 2023. Gemini goes to med school: Exploring the capabilities of multimodal large language models on medical challenge problems and hallucinations.
- Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jipu Wang, and Xindong Wu. 2024. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*.
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*.
- Thomas Pouplin, Hao Sun, Samuel Holt, and Mihaela van der Schaar. 2024. Retrieval-augmented thought process as sequential decision making. *Preprint*, arXiv:2402.07812.



- Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. 2021. [Rocketqa: An optimized training approach to dense passage retrieval for open-domain question answering](#). *Preprint*, arXiv:2010.08191.
- Ridho Reinanda, Edgar Meij, and Maarten de Rijke. 2020. *IEEE*.
- Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D. Manning. 2024. Raptor: Recursive abstract processing for tree-organized retrieval. In *ICLR*.
- Priyanka Sen, Sandeep Mavadia, and Amir Saffari. 2023. [Knowledge graph-augmented language models for complex question answering](#). *Proceedings of the 1st Workshop on Natural Language Reasoning and Structured Explanations (NLRSE)*.
- Karthik Soman, Peter W Rose, John H Morris, Rabia E Akbas, Brett Smith, Braian Peetoom, Catalina Villouta-Reyes, Gabriel Ceron, Yongmei Shi, Angela Rizk-Jackson, Sharat Israni, Charlotte A Nelson, Sui Huang, and Sergio E Baranzini. 2023a. [Biomedical knowledge graph-enhanced prompt generation for large language models](#). *Preprint*, arXiv:2311.17330.
- Karthik Soman, Peter W Rose, John H Morris, Rabia E Akbas, Brett Smith, Braian Peetoom, Catalina Villouta-Reyes, Gabriel Ceron, Yongmei Shi, Angela Rizk-Jackson, et al. 2023b. Biomedical knowledge graph-enhanced prompt generation for large language models. *arXiv preprint arXiv:2311.17330*.
- Inhwa Song, Sachin R. Pendse, Neha Kumar, and Munmun De Choudhury. 2024. [The typing cure: Experiences with large language model chatbots for mental health support](#). *Preprint*, arXiv:2401.14362.
- Russell K. Standish. 2004. [Why occam’s razor](#). *Foundations of Physics Letters*, 17(3):255–266.
- Jiashuo Sun, Chengjin Xu, Lumingyuan Tang, Saizhuo Wang, Chen Lin, Yeyun Gong, Lionel M. Ni, Heung-Yeung Shum, and Jian Guo. 2023. [Think-on-graph: Deep and responsible reasoning of large language model on knowledge graph](#). *Preprint*, arXiv:2307.07697.
- Yawei Sun, Lingling Zhang, Gong Cheng, and Yuzhong Qu. 2020. [Sparqa: Skeleton-based semantic parsing for complex questions over knowledge bases](#). *Preprint*, arXiv:2003.13956.
- Sabrina Toro, Anna V Anagnostopoulos, Sue Bello, Kai Blumberg, Rhiannon Cameron, Leigh Carmody, Alexander D Diehl, Damion Dooley, William Duncan, Petra Fey, Pascale Gaudet, Nomi L Harris, Marcin Joachimiak, Leila Kiani, Tiago Lubiana, Monica C Munoz-Torres, Shawn O’Neil, David Osumi-Sutherland, Aleix Puig, Justin P Reese, Leonore Reiser, Sofia Robb, Troy Ruemping, James Seager, Eric Sid, Ray Stefanick, Magalie Weber, Valerie Wood, Melissa A Haendel, and Christopher J Mungall. 2023. [Dynamic retrieval augmented generation of ontologies using artificial intelligence \(dragon-ai\)](#). *Preprint*, arXiv:2312.10904.
- Harsh Trivedi, Niranjana Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. [Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions](#). *Preprint*, arXiv:2212.10509.
- Deeksha Varshney, Aizan Zafar, Niranshu Kumar Bhera, and Asif Ekbal. 2023. [Knowledge grounded medical dialogue generation using augmented graphs](#). *Scientific Reports*, 13.
- Minh Duc Vu, Han Wang, Zhuang Li, Jieshan Chen, Shengdong Zhao, Zhenchang Xing, and Chunyang Chen. 2024. [Gptvoicetasker: Llm-powered virtual assistant for smartphone](#). *Preprint*, arXiv:2401.14268.
- Haochun Wang, Chi Liu, Nuwa Xi, Zewen Qiang, Sendong Zhao, Bing Qin, and Ting Liu. 2023a. [Huatu: Tuning llama model with chinese medical knowledge](#).
- Keheng Wang, Feiyu Duan, Sirui Wang, Peiguang Li, Yunsen Xian, Chuantao Yin, Wenge Rong, and Zhang Xiong. 2023b. [Knowledge-driven cot: Exploring faithful reasoning in llms for knowledge-intensive question answering](#). *Preprint*, arXiv:2308.13259.
- Rongsheng Wang, Yaofei Duan, ChanTong Lam, Jiexi-ChenandJiangsheng Xu, Haoming Chen, Xiaohong Liu, PatrickCheong-Iao Pang, and Tao Tan. 2023c. [Ivygpt: Interactive chinese pathway language model in medical domain](#).
- Xidong Wang, Guiming Hardy Chen, Dingjie Song, Zhiyi Zhang, Zhihong Chen, Qingying Xiao, Feng Jiang, Jianquan Li, Xiang Wan, Benyou Wang, and Haizhou Li. 2023d. [Cmb: A comprehensive medical benchmark in chinese](#). *Preprint*, arXiv:2308.08833.
- Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023e. [Self-instruct: Aligning language models with self-generated instructions](#). *Preprint*, arXiv:2212.10560.
- Yuhao Wang, Ruiyang Ren, Junyi Li, Wayne Xin Zhao, Jing Liu, and JiRong Wen. 2024. [Rear: A relevance aware retrieval augmented framework for open-domain question answering](#).
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. [Chain-of-thought prompting elicits reasoning in large language models](#). *Preprint*, arXiv:2201.11903.
- Yilin Wen, Zifeng Wang, and Jimeng Sun. 2023. [Mindmap: Knowledge graph prompting sparks graph of thoughts in large language models](#). *Preprint*, arXiv:2308.09729.
- Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. [C-pack: Packaged resources to advance general chinese embedding](#). *Preprint*, arXiv:2309.07597.



- Honglin Xiong, Sheng Wang, Yitao Zhu, Zihao Zhao, Yuxiao Liu, Linlin Huang, Qian Wang, and Dinggang Shen. 2023. Doctorglm: Fine-tuning your chinese doctor is not a herculean task.
- Peng Xu, Wei Ping, Xianchao Wu, Lawrence McAfee, Chen Zhu, Zihan Liu, Sandeep Subramanian, Evelina Bakhturina, Mohammad Shoeybi, and Bryan Catanzaro. 2024a. Retrieval meets long context large language models. In *ICLR*.
- Tianyu Xu, Wen Hua, Jianfeng Qu, Zhixu Li, Jiajie Xu, An Liu, and Lei Zhao. 2022. [Evidence-aware document-level relation extraction](#). In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management, CIKM '22*, page 2311–2320, New York, NY, USA. Association for Computing Machinery.
- Yongxin Xu, Xu Chu, Kai Yang, Zhiyuan Wang, Peinie Zou, Hongxin Ding, Junfeng Zhao, Yasha Wang, and Bing Xie. 2023a. Seqcare: Sequential training with external medical knowledge graph for diagnosis prediction in healthcare data. In *Proceedings of the ACM Web Conference 2023*, pages 2819–2830.
- Yongxin Xu, Xinke Jiang, Xu Chu, Rihong Qiu, Yujie Feng, Hongxin Ding, Junfeng Zhao, Yasha Wang, and Bing Xie. 2025. Dearllm: Enhancing personalized healthcare via large language models-deduced feature correlations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 941–949.
- Yongxin Xu, Xinke Jiang, Xu Chu, Yuzhen Xiao, Chaohe Zhang, Hongxin Ding, Junfeng Zhao, Yasha Wang, and Bing Xie. 2024b. Protomix: Augmenting health status representation learning via prototype-based mixup. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3633–3644.
- Yongxin Xu, Kai Yang, Chaohe Zhang, Peinie Zou, Zhiyuan Wang, Hongxin Ding, Junfeng Zhao, Yasha Wang, and Bing Xie. 2023b. Vecocare: Visit sequences-clinical notes joint learning for diagnosis prediction in healthcare data. In *IJCAI*, volume 23, pages 4921–4929.
- Yongxin Xu, Ruizhe Zhang, Xinke Jiang, Yujie Feng, Yuzhen Xiao, Xinyu Ma, Runchuan Zhu, Xu Chu, Junfeng Zhao, and Yasha Wang. 2024c. Parenting: Optimizing knowledge selection of retrieval-augmented language models with parameter decoupling and tailored tuning. *arXiv preprint arXiv:2410.10360*.
- Zhichao Xu. 2023. [Context-aware decoding reduces hallucination in query-focused summarization](#). *Preprint*, arXiv:2312.14335.
- Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, Fan Yang, Fei Deng, Feng Wang, Feng Liu, Guangwei Ai, Guosheng Dong, Haizhou Zhao, Hang Xu, Haoze Sun, Hongda Zhang, Hui Liu, Jiaming Ji, Jian Xie, JunTao Dai, Kun Fang, Lei Su, Liang Song, Lifeng Liu, Liyun Ru, Luyao Ma, Mang Wang, Mickel Liu, MingAn Lin, Nuolan Nie, Peidong Guo, Ruiyang Sun, Tao Zhang, Tianpeng Li, Tianyu Li, Wei Cheng, Weipeng Chen, Xiangrong Zeng, Xiaochuan Wang, Xiaoxi Chen, Xin Men, Xin Yu, Xuehai Pan, Yanjun Shen, Yiding Wang, Yiyu Li, Youxin Jiang, Yuchen Gao, Yupeng Zhang, Zenan Zhou, and Zhiying Wu. 2023a. [Baichuan 2: Open large-scale language models](#). *Preprint*, arXiv:2309.10305.
- Kai Yang, Yongxin Xu, Peinie Zou, Hongxin Ding, Junfeng Zhao, Yasha Wang, and Bing Xie. 2023b. Kerprint: local-global knowledge graph enhanced diagnosis prediction for retrospective and prospective interpretations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 5357–5365.
- Songhua Yang, Xinke Jiang, Hanjie Zhao, Wenxuan Zeng, Hongde Liu, and Yuxiang Jia. 2024. Faima: Feature-aware in-context learning for multi-domain aspect-based sentiment analysis. In *COLING*.
- Songhua Yang, Hanjia Zhao, Senbin Zhu, Guangyu Zhou, Hongfei Xu, Yuxiang Jia, and Hongying Zan. 2023c. Zhongjing: Enhancing the chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue.
- Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2024. Making retrieval-augmented language models robust to irrelevant context. In *ICLR*.
- Donghan Yu, Chenguang Zhu, Yuwei Fang, Wenhao Yu, Shuohang Wang, Yichong Xu, Xiang Ren, Yiming Yang, and Michael Zeng. 2022. [Kgfid: Infusing knowledge graph in fusion-in-decoder for open-domain question answering](#). *Preprint*, arXiv:2110.04330.
- Wenhao Yu, Hongming Zhang, Xiaoman Pan, Kaixin Ma, Hongwei Wang, and Dong Yu. 2023a. [Chain-of-note: Enhancing robustness in retrieval-augmented language models](#). *Preprint*, arXiv:2311.09210.
- Wenhao Yu, Hongming Zhang, Xiaoman Pan, Kaixin Ma, Hongwei Wang, and Dong Yu. 2023b. [Chain-of-note: Enhancing robustness in retrieval-augmented language models](#). *Preprint*, arXiv:2311.09210.
- Hui Zeng. 2023. [Measuring massive multitask chinese understanding](#). *Preprint*, arXiv:2304.12986.
- Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Jianquan Li, Guiming Chen, Xiangbo Wu, Zhiyi Zhang, Qingying Xiao, Xiang Wan, and Benyou Wang. 2023a. Huatuogpt, towards taming language model to be a doctor.
- Ningyu Zhang, Mosha Chen, Zhen Bi, Xiaozhuan Liang, Lei Li, Xin Shang, Kangping Yin, Chuanqi Tan, Jian Xu, Fei Huang, Luo Si, Yuan Ni, Guotong Xie, Zhi-fang Sui, Baobao Chang, Hui Zong, Zheng Yuan, Linfeng Li, Jun Yan, Hongying Zan, Kunli Zhang, Buzhou Tang, and Qingcai Chen. 2022. [CBLUE: A Chinese biomedical language understanding evaluation benchmark](#). In *Proceedings of the 60th Annual*

*Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7888–7915, Dublin, Ireland. Association for Computational Linguistics.

Ruizhe Zhang, Yongxin Xu, Yuzhen Xiao, Runchuan Zhu, Xinke Jiang, Xu Chu, Junfeng Zhao, and Yasha Wang. 2024. Knowpo: Knowledge-aware preference optimization for controllable knowledge selection in retrieval-augmented language models. *AAAI 2024*.

Yichi Zhang, Zhuo Chen, Yin Fang, Lei Cheng, Yanxi Lu, Fangming Li, Wen Zhang, and Huajun Chen. 2023b. Knowledgeable preference alignment for llms in domain-specific question answering. *Preprint*, arXiv:2311.06503.

Lingfeng Zhong, Jia Wu, Qian Li, Hao Peng, and Xindong Wu. 2023. A comprehensive survey on automatic knowledge graph construction. *Preprint*, arXiv:2302.05019.

Fengbin Zhu, Wenqiang Lei, Chao Wang, Jianming Zheng, Soujanya Poria, and Tat-Seng Chua. 2021. Retrieving and reading: A comprehensive survey on open-domain question answering. *Preprint*, arXiv:2101.00774.

He Zhu, Ren Togo, Takahiro Ogawa, and Miki Haseyama. 2023. A medical domain visual question generation model via large language model.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2020. Fine-tuning language models from human preferences. *Preprint*, arXiv:1909.08593.

## A Appendix

### A.1 Design Details

#### A.1.1 Prompts Design

Prompts used in Hypothesis Output Module and LLM Reader are illustrated in Figure 3.

### A.2 Experiment Settings

#### A.2.1 Datasets

For MMCU-Medical (Zeng, 2023), the questions are from the university medical professional examination, covering the three basic medical sciences, pharmacology, nursing, pathology, clinical medicine, infectious diseases, surgery, anatomy, etc., with a total of 2,819 questions. The CMB-Exam (Wang et al., 2023d) dataset utilizes qualifying exams as a data source in the four clinical medicine specialties of physicians, nurses, medical technicians, and pharmacists, with a total of 269,359 questions. Given the extensive size of the CMB-Exam dataset, we randomly sample 4,000 questions for testing. The CMB-Clin (Wang et al., 2023d) dataset contains 74 high-quality, complex, and real patient cases with 208 medical questions.

*CMeKG* (Clinical Medicine Knowledge Graph)<sup>7</sup> (BYAMBASUREN et al., 2020), *CPubMed-KG* (Large-scale Chinese Open Medical Knowledge Graph)<sup>8</sup> and *Disease-KG* (Chinese disease Knowledge Graph)<sup>9</sup> are open-source medical KGs, which integrates extensive medical text data, including diseases, medications, symptoms and diagnostic treatment technologies. The fused KG has 1,288,721 entities and 3,569,427 relations. However, due to the lack of medical entity descriptions in its entities, we collect relevant entity knowledge from Wikipedia<sup>10</sup>, Baidu Baike<sup>11</sup>, and Medical Baike<sup>12</sup>, and store them as entity descriptions.

(1) **Base Model (Base)** servers as the model without any external knowledge, used to check the improvement effect of different RAG methods. We use GPT 3.5 and Baichuan13B-chat (Yang et al., 2023a) as base models. (2) **Knowledge Graph Retrieval-Augmented Generation (KGRAG)** (Soman et al., 2023b,a; Sen et al., 2023) uses user query as a reference to retrieve in the KGs, which is the base model of RAG on KG and has been widely applied in (Soman et al., 2023b,a; Sen et al., 2023). (3) **Query Expansion (QE)** (Azad and Deepak, 2019) reformulate the user’s initial query by adding additional terms with a similar meaning with the help of LLMs. (4) **CHAIN-OF-NOTE (CoN)** (Yu et al., 2023b) generates sequential reading notes for retrieved knowledge, enabling a thorough evaluation of their relevance to the given question and integrating

<sup>7</sup><https://cmekg.pcl.ac.cn/>, [https://github.com/king-yyf/CMeKG\\_tools](https://github.com/king-yyf/CMeKG_tools)

<sup>8</sup><https://cpubmed.openi.org.cn/graph/wiki>

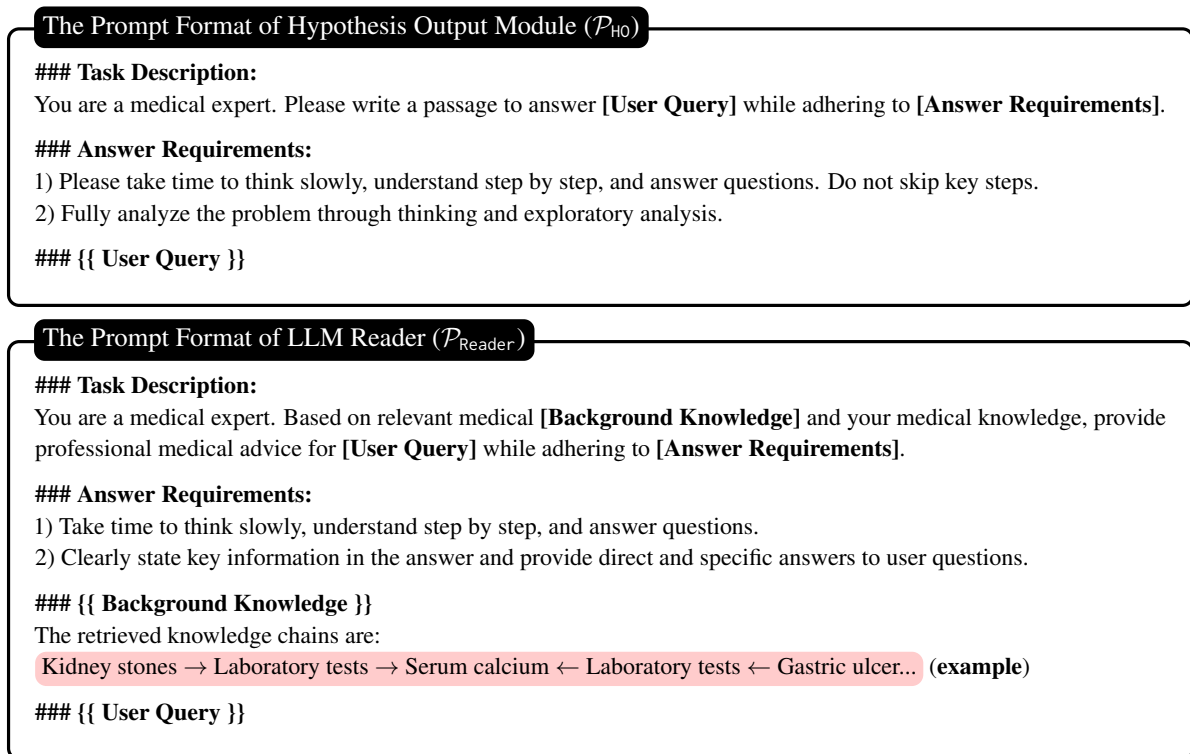
<sup>9</sup><https://github.com/nuolade/disease-kb>

<sup>10</sup><https://www.wikipedia.org/>

<sup>11</sup><https://baike.baidu.com/>

<sup>12</sup><https://www.yixue.com/>

Figure 3: The prompt formats of (Up.) Hypothesis Output Module and (Down.) LLM Reader.



these notes to formulate the final answer. (5) **Chain-of-Knowledge (CoK)** (Li et al., 2023c) utilize the power of LLMs and consists of reasoning preparation, dynamic knowledge adapting, and answer consolidation. (6) **Knowledge-Augmented Language Model Verification (KALMV)** (Baek et al., 2023) verifies the output and the knowledge of the knowledge-augmented LLMs with a separate verifier. (7) **Knowledge Graph Generative Pre-Training (KG-GPT)** (Kim et al., 2023) comprises three steps: Sentence Segmentation, Graph Retrieval, and Inference, each aimed at partitioning sentences, retrieving relevant graph components, and deriving logical conclusions. (8) **Summarizing Retrievals (SuRe)** (Kim et al., 2024) constructs summaries of the retrieved passages for each of the multiple answer candidates and confirms the most plausible answer from the candidate set by evaluating the validity and ranking of the generated summaries. Note that we follow the prompts of the baselines as stated strictly. Besides, the baselines and running time are summarized in Table 1. In RAG Options, CoN requires fine-tuning the retriever, implying a higher training overhead and the prerequisite of preparing a dataset. In addition, it is also difficult to migrate to other domain-specific KGs. In terms of LLMs interactions, QE, CoN, CoK, KALMV, KG-GPT, SuRe and HyKGE all necessitate engagement with LLMs. However, CoN, CoK, KALMV, KG-GPT and SuRe entail multiple interactions (more than twice), significantly escalating the time expenditure.

### A.2.2 Evaluation Metrics

As for the evaluation of multi-task medical choice question performance, we guide LLMs to only answer the correct answer and employ established metric **Exact Match (EM)** as suggested by prior work (Zhu et al., 2021; Karpukhin et al., 2020). For the EM score, an answer is deemed acceptable if its form corresponds to all correct answers in the provided list. For multiple-choice questions, we also calculate a **Partial Correct Rate (PCR)**. In comparison to EM, if there is a missing answer without any incorrect ones, PCR classifies it as correct. In addition, to verify the effectiveness of the retrieved knowledge, we also let LLMs output a complete analysis process. Then, we measure **Artificial Correlation Judgement (ACJ)** by inviting 20 medical experts to rate the retrieved knowledge according to the criteria of (correlation=1, relevant but useless=0, irrelevant=-1), and calculate the relevant scores for each question by sampling 100 questions from the two datasets. Moreover, we also objectively evaluated the **Perplexity (PPL)** of LLMs output. The smaller the PPL, the greater the role of retrieved knowledge in reducing LLMs' hallucinations. Moreover, we also complement our analysis with **ROUGE-Recall (ROUGE-R)** (Xu, 2023). **ROUGE-R** measures the extent to which the LLMs' responses cover the retrieved knowledge, which is crucial for ensuring comprehensive information coverage. For open-domain medical Q&A tasks, we utilize **ROUGE-R** and **Bilingual Evaluation Understudy (BLEU-1 for answer precision, BLEU-4 for answer fluency)** (Xu, 2023) to gauge the similarity

of LLMs responses to the ground-truth doctor analysis. Additionally, we employ **PPL** to assess the quality of LLMs responses. Next, we provide comprehensive descriptions and formulas for the metrics used to evaluate our models. Each metric targets a specific aspect of performance, allowing for a thorough and balanced assessment.

**Exact Match (EM).** Exact Match measures the percentage of predicted answers that exactly align with the ground truth. It is equivalent to Accuracy (ACC).

$$\text{EM} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Questions}} \quad (6)$$

**Partial Correct Rate (PCR).** For multiple-choice questions, partial correctness is permissible. If a subset of the selected answers matches the ground truth, the prediction is deemed correct.

$$\text{PCR} = \frac{\text{Number of Partially or Fully Correct Predictions}}{\text{Total Number of Questions}} \quad (7)$$

**Artificial Correlation Judgment (ACJ).** ACJ evaluates the relevance of retrieved knowledge using expert annotation. Twenty medical experts score 100 sampled questions from each dataset under three criteria:

- **Correlation = 1:** Retrieved knowledge is relevant *and* useful.
- **Relevant but Useless = 0:** Knowledge is relevant but does not help answer the question.
- **Irrelevant = -1:** Knowledge is not relevant.

The final ACJ score is the mean of the expert ratings:

$$\text{ACJ} = \frac{\text{Sum of Expert Ratings}}{\text{Number of Questions}} \quad (8)$$

**Perplexity (PPL).** Perplexity quantifies the fluency of the language model’s output by exponentiating the average negative log-likelihood of the predicted token sequence.

$$\text{PPL} = 2^{-\frac{1}{N} \sum_{i=1}^N \log_2 P(w_i)}, \quad (9)$$

where  $P(w_i)$  is the probability assigned to token  $w_i$  and  $N$  is the total number of tokens.

**ROUGE (Recall-Oriented Understudy for Gisting Evaluation).** We employ ROUGE-R (recall) to measure the overlap between generated text and reference text in terms of  $n$ -grams, longest common subsequences, or skip-bigrams.

$$\text{ROUGE-R} = \frac{\# \text{ Overlapping } n\text{-grams}}{\text{Total } \#n\text{-grams in Reference}} \quad (10)$$

where  $\#$  denotes the number.

**BLEU (Bilingual Evaluation Understudy).** BLEU assesses the precision of overlapping  $n$ -grams between generated and reference texts, incorporating a brevity penalty (BP) to discourage overly short outputs:

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^N w_n \log P_n\right), \quad (11)$$

where  $P_n$  denotes  $n$ -gram precision and  $w_n$  is typically uniform ( $w_n = 1/N$ ). The brevity penalty is defined as

$$\text{BP} = \begin{cases} 1, & \text{if } c > r, \\ \exp\left(1 - \frac{r}{c}\right), & \text{if } c \leq r, \end{cases} \quad (12)$$

with  $c$  and  $r$  representing the lengths of the generated and reference texts, respectively.

### A.2.3 Experimental Implementation.

In HyKGE,  $k = 3$ ,  $\text{top}K = 10$ ,  $\delta = 0.7$ ,  $lc = 10$ ,  $oc = 4$ . The prompts for LLMs can refer to Table 3. Moreover, for all the baselines and HyKGE, we set the maximum number of returned tokens for LLMs to 500 and the temperature to 0.6. In all baselines and HyKGE, we first use the Jieba library in Python to perform word segmentation, and then use filtered text to filter out tone words and invalid characters following “chinese\_word\_cut.txt”<sup>13</sup> to avoid errors in knowledge extraction. For a fair comparison, we apply the same W2NER, GTE and FlagEmbedding models for all baselines. Moreover, the parameters of W2NER are optimized with Adam optimizer (Kingma and Ba, 2015) with  $L_2$  regularization and dropout on high-quality medical dataset (Zhang et al., 2022; Hongying et al., 2020), the learning rate is set to 1e-3, the hidden unit is set to 1024 and weight decay is 1e-4. Similar to previous work (Soman et al., 2023a), because of the randomness of LLMs’ outputs, we repeat experiments with different random seeds five times and report the average and standard deviation results. Experimental results are statistically significant with  $p < 0.05$ . Implementations are done using the PyTorch 1.9.0 framework (Paszke et al., 2019) in Python 3.9, on an Ubuntu server equipped with 8 A100 GPU and an Intel(R) Xeon(R) CPU.

### A.3 Case Study (RQ2 and RQ3)

This case study presents a representative sample that illustrates the effectiveness of our HyKGE model using GPT-3.5 Turbo as shown in Figure 5. The color coding within the table is key to understanding the source and validity of the information and we have these observations: **i)** Compared to a brief user query, semantic spaces of  $\mathcal{HO}$  are more abundant and have a clear direction for answering, helping us better understand user intention and extract more effective entity information. Ultimately, HyKGE extracted 23 entities from  $\mathcal{HO}$  compared to only 1 from  $\mathcal{Q}$ . **ii)** Comparing the  $\mathcal{RC}$  with  $\mathcal{RC}_{\text{prune}}$ , it can be observed that the pre-filtered

<sup>13</sup>[https://github.com/RobustJay/NLP\\_Chinese\\_WordCut/blob/master/stopwords.txt](https://github.com/RobustJay/NLP_Chinese_WordCut/blob/master/stopwords.txt)



Table 4: Performance and computation time comparison (in mean  $\pm$  standard deviation) on MMCU-Medical for medical Q&A answer with GPT 3.5 Turbo.

Method / Metric	EM	PCR	Avg. Time (s)
HyKGE	<b>49.65<math>\pm</math>1.39</b>	<b>57.82<math>\pm</math>1.54</b>	<b>19.76</b>
HyKGE(+ LLM for NER)	48.17 $\pm$ 1.13	56.77 $\pm$ 1.02	26.61
HyKGE(+ LLM for Reranker)	42.72 $\pm$ 2.06	48.24 $\pm$ 1.17	32.51
HyKGE(+ LLM for Summary)	43.02 $\pm$ 3.11	46.54 $\pm$ 2.08	28.51

chains contain a large amount of irrelevant or repetitive knowledge, marked in black. After reranking, retrieved knowledge is highly non-redundant and relevant to  $\mathcal{H}\mathcal{O}$  and  $\mathcal{Q}$ , demonstrating the effectiveness of our fragment-based reranker. Ultimately, out of 125 reasoning chains, HyKGE selected  $topK = 10$  of the most valuable chains. **iii)** Note that retrieved knowledge effectively assisted LLMs in correcting errors, mitigating the issue of hallucinations. In  $\mathcal{H}\mathcal{O}$ , LLMs posited that “*calcium carbonate could not treat GERD*”; however, with the supplemental knowledge about “*calcium carbonate*” in our retrieved reasoning chains, marked in green. LLMs corrected this error in its final response. In general, this case study demonstrates HyKGE’s strong ability to generate hypotheses and validate them against a structured KG, effectively leveraging  $\mathcal{H}\mathcal{O}$  for exploring and reasoning chains for error correction. In general, the integration of these components ensures that the RAG’s outputs are not only contextually relevant but also accurate, showcasing the interpretability and potential for AI-assisted decision-making in healthcare.

#### A.4 Efficiency Analysis (RQ2)

To illustrate the effectiveness of our HyKGE module, we conducted a comparative analysis of the time overhead between HyKGE and other knowledge graph-enhanced LLM approaches, as presented in Table 1. The KGRAG method demonstrates the shortest time overhead among RAG methods, as it solely necessitates conveying the retrieved knowledge to the LLM Reader. However, when juxtaposed with QE and HyKGE, KGRAG’s performance notably lags behind, even resulting in a negative gain because of the huge noise. In contrast to QE, HyKGE incurs slightly higher time primarily due to the noise filtering process, which consumes some time. Nonetheless, the performance enhancement achieved by HyKGE outweighs this marginal increase in time overhead. Furthermore, CoN and CoK, which adopt the chain-of-thought strategy (Wei et al., 2023), entail multiple interactions with LLMs, which proves to be considerably restrictive, particularly in real-world medical Q&A scenarios where time is a critical consideration. Therefore, striking a balance between time overhead and model accuracy becomes imperative, in which regard HyKGE emerges as the most efficient and high-performing framework.

Moreover, inspired by these Chain-of-thought works (Pouplin et al., 2024; Trivedi et al., 2023), which respectively employ LLMs in different processes of RAG, we embarked on similar endeavors. Specifically, we integrated LLMs into the modules of NER, Reranker,

and summarization modules (to summarize the retrieved knowledge) (Kim et al., 2024), as shown in Table 4. However, our findings underscored that leveraging such large-parameter models for tasks amenable to smaller counterparts incurs substantial time costs with marginal benefits. For instance, incorporating LLMs into NER, aimed at enhancing entity extraction, a task that could be efficiently handled by specialized pre-trained medical NER models, not only doubling interaction time but also introducing complexities such as misinterpretation of instructions, thus impeding subsequent processing. Similarly, the utilization of LLMs in Reranker considerably strained token resources. For instance, upon retrieving the query “*I feel stomach reflux after eating. What medicine should I take?*” it generated a whopping 125 reasoning chains. However, employing LLMs to eliminate noisy knowledge from these chains resulted in decreased effectiveness. We argue that this was primarily due to the inundation of tokens, causing LLMs to lose in the middle (Liu et al., 2023), thereby impeding their ability to discern genuinely relevant knowledge from the retrieved chains and even ignore LLMs’ tasks. Consequently, LLMs employed for Reranker inadvertently filtered out valuable knowledge, yielding negative outcomes and exacerbating computational overhead. Likewise, employing LLMs for knowledge summarization encountered challenges akin to those encountered in Reranker. Although LLMs are quite effective, according to Occam’s razor principle (Standish, 2004), it is not always beneficial to use LLMs in every RAG step. Excessive reliance on LLMs can only lead to wasted time costs. In summary, because RAG involves a process of continuous trial and error (Barnett et al., 2024), we experimented with many strategies and ultimately arrived at HyKGE.

#### A.5 Ablation Study (RQ 2)

To answer RQ2, we perform ablation studies to verify the effectiveness of the critical components of HyKGE, as illustrated in Table 2. Our observation can be summarized as follows:

**In pre-retrieval phase.** When we remove the Hypothesis Output Module, results are even deteriorating than base model. This is attributed to the fact that retrieved knowledge simply based on user queries is either insufficient or futile because of lacking direction for exploration. Nevertheless, the results of w/o  $\mathcal{H}\mathcal{O}$  are still better than KGRAG and we argue the reason is the reranking of reasoning chains effectively filters out noise during the post-retrieval phase.

**In post-retrieval phase.** The removal of the Reranker leads to a noticeable decline in performance compared to HyKGE, which indicates that Reranker effectively eliminates excessive noise introduced by the retrieved knowledge, retaining only the most pertinent parts for answering the question. When we use entire  $\mathcal{H}\mathcal{O}$  and  $\mathcal{Q}$  instead of chunk ( $\mathcal{Q} \oplus \mathcal{H}\mathcal{O}$ ) to perform reranking with reasoning chains, a decline in performance is also observed. This is attributable to the misalignment between

dense retrieved knowledge and sparsely distributed keywords in  $\mathcal{H}\mathcal{O}$  and  $\mathcal{Q}$ , inducing a tendency to select more general or lengthier knowledge, thereby diminishing the HOM’s capability to supplement diverse knowledge.

Moreover, results of w/o Chains and w/o Description demonstrate that even when KG lacks certain knowledge, descriptive information or relevant knowledge chains can still enhance the answering capabilities of LLMs, which is believed to be associated with the inherent implicit knowledge within the LLMs themselves.

## A.6 Hyper-parameter Study (RQ4)

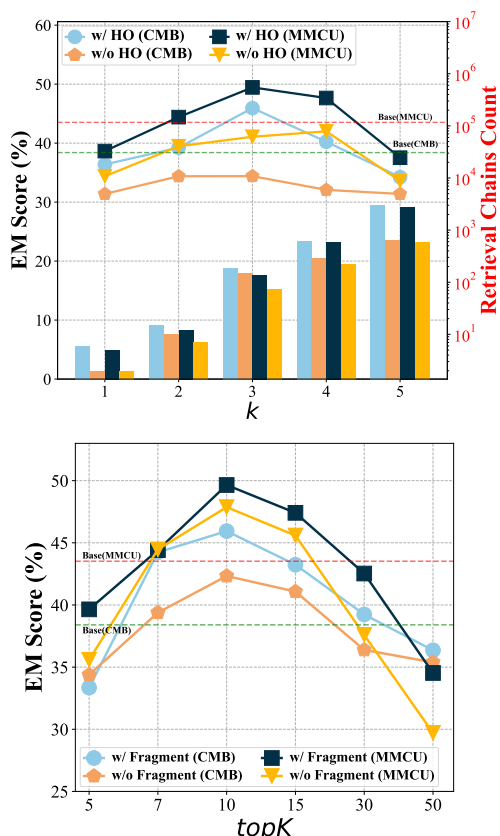


Figure 4: (Left.) Hyper-parameter study with the KG hop  $k$  on MMCU-Medical and CMB-Exam with GPT 3.5 turbo, from 1 to 5. (Right.) Hyper-parameter study with the reranker  $topK$  on MMCU-Medical and CMB-Exam with GPT 3.5 turbo, from 5 to 50.

In this part, we concentrate on evaluating the influence of different hyper-parameters on HyKGE for RQ4. Specifically, we perform a series analysis of KG hop  $k$  from the list [1, 2, 3, 4, 5] and reranker  $topK$  from the list [5, 7, 10, 15, 30, 50] to verify the sensitive:

Figure 4 (Left.) depicts EM and the number of retrieved knowledge before pruning. We observe that as  $k$  increases, the amount of knowledge retrieved explodes exponentially following a power-law distribution (Berger et al., 2005; Ahad N. et al., 2023), exceeding  $10^3$  when  $k = 5$ . However, an excessive amount of knowledge not only fails to improve EM, but also burdens LLMs with

an increased number of tokens. Concurrently, EM exhibits a trend of initial increase followed by a decrease as  $k$  increases. This phenomenon can be attributed to the fact that at lower values of  $k$ , the retrieved knowledge predominantly consists of isolated snippets of information, offering minimal utility. Conversely, with larger  $k$ , the LLMs encounter limitations in comprehending extensive reasoning chains, thereby rendering them incapable of effectively utilizing the complex and abundant retrieved information, with the performance even worse than the base model. Besides, we notice removing the HO will result in a substantial reduction in the quantity of knowledge retrieved, because of the limited diversity of knowledge obtained based solely on user query.

Figure 4 (Right.) depicts EM with different reranking thresholds. Similar to Figure 4 (Left.), as  $topK$  increases, the trends demonstrate that overwhelming reasoning chains will hamper LLMs’ ability for comprehension. Meanwhile, it is obvious that HyKGE w/o Fragment always underperforms on EM as analyzed in Section A.5.

## A.7 Quantifying Knowledge Density

Quantifying the knowledge density of user queries poses significant challenges in natural language processing and understanding. Knowledge density refers to the amount of useful information conveyed in a given statement, which can vary widely depending on how the information is presented.

For example, consider the following two statements:

1. “Doctor, my name is Jack. Recently, after having breakfast, I often experience stomach pain.”
2. “Having breakfast causes stomach pain.”

The first statement exhibits a lower knowledge density, as it includes extraneous information that does not directly contribute to the understanding of the medical issue at hand. In contrast, the second statement is concise and directly conveys the relationship between breakfast and stomach pain, demonstrating a higher knowledge density.

This disparity illustrates the challenges involved in quantifying knowledge density. The lack of fine-grained annotations in many datasets complicates efforts to develop reliable metrics for measuring knowledge density. While it is possible to qualitatively assess the information content of various statements, the absence of standardized methods or clear benchmarks makes it difficult to achieve accurate and consistent quantification.

Consequently, addressing the misalignment between the knowledge density of user queries and reasoning chains remains an area of active research. Future work may involve exploring new annotation strategies or developing automated methods to evaluate and quantify knowledge density effectively.

## A.8 Explanation and Ablation Results for the Proposed Reasoning Chains

Biomedical knowledge graphs are characterized by dense adjacency relationships, which often result in sub-

Figure 5: Case study. We demonstrate the User Query  $\mathcal{Q}$ , Hypothesis Output  $\mathcal{HO}$ , Retrieved Reasoning Chains  $\mathcal{RC}$  and Pruned Reasoning chains  $\mathcal{RC}_{\text{prune}}$  of HyKGE using GPT-3.5 Turbo to verify the interpretability and effectiveness of HyKGE. **Red shed** signifies that the knowledge or answer is derived from evidence in  $\mathcal{Q}$ , **blue shed** indicates that the evidence originates from  $\mathcal{HO}$ , and **green shed** represents a corrected answer with the help of  $\mathcal{KG}$ , despite initially being false in  $\mathcal{HO}$ .

### A case study of HyKGE pipeline

#### ### User Query $\mathcal{Q}$ :

I feel **stomach reflux** after eating. What medicine should I take?

#### ### Hypothesis Output $\mathcal{HO}$ :

**Stomach reflux** can be a symptom caused by **gastroesophageal reflux**. 1. Do you have symptoms related to **indigestion**, **excess stomach acid**, **gastric ulcers**, **bile reflux**, or **stomach pain**? ... When experiencing mild gastroesophageal reflux, the following plans can be considered:- Avoid overeating - Avoid spicy, stimulating foods, and alcohol - Maintain a light diet, eat more vegetables. Under appropriate evidence, consider taking **H2 receptor antagonists** or **proton pump inhibitors**. **It is not recommended to take calcium carbonate** or **aluminum hydroxide medications**, **as calcium carbonate can lead to kidney stones, and the aluminum element in aluminum hydroxide is toxic and may cause allergic reactions or discomfort.**

#### ### Retrieved Reasoning Chains $\mathcal{RC}$ :

Kidney stones → Laboratory tests → Serum calcium ← Laboratory tests ← Gastric ulcer. Kidney stones → Clinical symptoms and signs → Fever ← Complications ← Esophageal cancer.  
**Excess gastric acid** → Affected area → Abdomen ← Affected area ← **Stomach pain**. Niflumic acid → Adverse reaction → Discomfort ← Adverse reaction ← Calcium carbonate. Gastrointestinal bleeding → Related diseases → Gastric ulcer. **Gastric ulcer** → Treatment → **Aluminum hydroxide**.  
**Calcium carbonate** → Indication → **Excess gastric acid**. **Aluminum hydroxide** → Indication → **Gastric reflux**.  
**Niflumic acid** → Third level classification → Others ← Affected area ← **Postprandial food reflux**.  
**Duodenogastric reflux and bile reflux gastritis** → Treatment → **Calcium carbonate**. More else ... [125 reasoning chains in total.]

#### ### Pruned Reasoning chains $\mathcal{RC}_{\text{prune}}$ :

**Gastric ulcer** → Treatment → **Aluminum hydroxide**. **Calcium carbonate** → Indication → **Excess gastric acid**.  
**Aluminum hydroxide** → Indication → **Gastric reflux**.  
**Niflumic acid** → Third level classification → Others ← Affected area ← **Postprandial food reflux**.  
**Duodenogastric reflux and bile reflux gastritis** → Treatment → **Calcium carbonate**.  
**Stomach reflux** → Related disease → **Excess gastric acid**. **Caved-S** → Indication → **Excess gastric acid**.  
**Weile tablets** → Indication → **Excess gastric acid**. More else. . . [10 reasoning chains in total.]

#### ### Answer:

**Stomach reflux** can be a symptom caused by **gastroesophageal reflux disease (GERD)**, and this condition could potentially lead to reflux esophagitis. Consider the following medical recommendations:  
**Acid-suppressing medications (Proton Pump Inhibitors)**: You may consider using omeprazole or esomeprazole to effectively reduce gastric acid secretion, aiding in relieving symptoms of gastric reflux and mitigating the development of reflux esophagitis. Alternatively, you can use **acid-neutralizing medications (antacids) such as magnesium aluminum carbonate (Rennie), calcium carbonate, and aluminum hydroxide**. Another option is the use of **H2 receptor antagonists such as ranitidine or famotidine**. Additionally, medications like **Caved-S or Weile tablets** are also suitable for mild to moderate symptoms of gastric reflux. Gastric protective medications: For short-term relief of stomach pain, **you may also consider using gastric mucosal protectants such as sucralfate**. It helps form a protective mucous layer, reducing irritation to the stomach mucosa. Diet and lifestyle adjustments: First, . . . it is advisable to consult a healthcare professional for evaluation and diagnosis as soon as possible.

graphs that contain a significant number of noisy and irrelevant entities. These noisy subgraphs can degrade the performance of downstream tasks by introducing unnecessary complexity and distracting from relevant information. The results presented in Table 1 (HyKGE w/o Chains) validate this observation, underscoring the necessity of incorporating reasoning chains to reduce noise and enhance retrieval quality.

**Rationale Behind the Selection of the Three Reasoning Chains.** The inclusion of the three reasoning chains—Path, Co-ancestor, and Co-occurrence—is driven by their unique ability to address the challenges associated with biomedical reasoning tasks. Each chain pattern offers specific advantages that collectively enhance the retrieval process:

- **Path (head-to-tail):** This reasoning chain captures direct causal or triggering relationships between diseases and symptoms. For instance, it identifies connections such as "infection" leading to "fever," providing clear and logical insights into causal medical phenomena.
- **Co-ancestor Chain (tail-to-tail):** This chain identifies shared physiological or environmental factors between entities, facilitating analogical reasoning and supporting diagnostic processes. For example, "high blood pressure" and "diabetes" might share a common factor like "obesity," revealing underlying relationships that aid in understanding co-morbidities.
- **Co-occurrence Chain (head-to-head):** This reasoning chain captures pathological characteristics and disease progression by linking entities that frequently co-occur in similar contexts. For instance, it establishes connections between "tumor" and "metastasis," providing insights into disease progression and associations.

**Benefits of the Selected Reasoning Chains.** These reasoning chains collectively serve as a structured and semantically rich mechanism for organizing biomedical knowledge. Specifically, they provide several benefits: - They facilitate easier interpretation and inference by large language models (LLMs), enabling more accurate and relevant retrievals. - By focusing on logical and contextual relationships, they function as an efficient pruning mechanism, significantly reducing noise in subgraphs and improving the quality of the retrieved knowledge.

**Experimental Validation.** To further support our choice of reasoning chains, we conducted an ablation study comparing the effectiveness of each reasoning chain pattern (Path, Co-ancestor, Co-occurrence) both individually and in combination. The ablation results, along with those detailed in Table 1, confirm that the integration of these three chains yields the best performance by leveraging their complementary strengths.

By incorporating these reasoning chains, our method effectively addresses the inherent challenges of biomedical knowledge graphs while ensuring more precise and semantically meaningful knowledge retrieval for downstream tasks.

## A.9 Evaluation of Relevance, Coherence, and Informativeness by the Reranking Model

We clarify the design and rationale behind the HO Fragment Granularity-aware Rerank Module and explain how it achieves a balance between diversity and relevance in retrieval.

**Core Objective of the Module.** The HO Fragment Granularity-aware Rerank Module is specifically designed to address the challenges of semantic misalignment and noise interference in knowledge retrieval while balancing relevance and diversity. This is accomplished through two primary steps:

**1. Chunk Splitting: Enhancing Relevance and Reducing Noise.** Traditional retrieval methods often treat the user query  $Q$  as a single entity and match it directly with reasoning chains (RCs) in the knowledge graph. However, user queries frequently contain ambiguous semantics or irrelevant information, leading to semantic misalignment and the introduction of noise or redundancy. To mitigate this, we implement chunk splitting for both the user query  $Q$  and the hypothesis output (HO): - Each query  $Q$  and HO is divided into  $K$  meaningful chunks, which allows us to remove redundant or low-information parts (for example, phrases like "My name is..." or overly descriptive background details). - This chunking process ensures that each segment of the query aligns more precisely with the reasoning chains in the knowledge graph, significantly enhancing relevance and coherence while filtering out noise.

**2. Similarity Matrix: Ensuring Diversity and Information Completeness.** During the reranking process, we construct a similarity comparison matrix to balance diversity and relevance. This matrix compares the  $K$  chunks with  $M$  reasoning chains (RCs): - The similarity matrix, with dimensions  $K \times M$ , represents the semantic similarity between each query chunk and each reasoning chain. - Each value in the matrix quantifies how well a specific chunk aligns with a particular reasoning chain.

We utilize this similarity information to rank all reasoning chains and select the top  $K$  chains as the final retrieval results. This approach offers several benefits: **1. Enhancing Diversity:** By incorporating similarity scores from multiple chunks, the method avoids over-reliance on a single chunk, ensuring that the selected reasoning chains cover a broader semantic scope. **2. Preserving Information Completeness:** The selection of the top  $K$  reasoning chains not only guarantees high-quality outputs but also retains important task-relevant information, thus preventing the loss of valuable knowledge due to excessive simplification.



In summary, the HO Fragment Granularity-aware Rerank Module effectively evaluates the relevance, coherence, and informativeness of each retrieved reasoning chain through a structured process that emphasizes meaningful segmentation of queries and robust similarity analysis.

#### **A.10 Retrieving Reasonable Knowledge Based on Varying Quality User Queries**

The quality of user queries can vary significantly, and retrieving reasonable knowledge accounts for two main scenarios:

**1. Incomplete or Implicit Semantics in User Queries.** User queries often contain implicit meanings that can lead to incomplete understanding. For example, a query like “I have a fever and my throat is uncomfortable” implicitly suggests potential causes such as “viral infection” or “bacterial infection,” which are not explicitly mentioned. When performing retrieval-augmented generation (RAG) directly on such queries without leveraging hypothesis outputs, the retrieved entities and knowledge may be insufficient to address the underlying medical context.

To overcome this challenge, we utilize the powerful contextual understanding capabilities of large language models (LLMs) to generate hypothesis outputs that fill in the missing semantics. For instance, the model can explicitly identify the possibility of infections based on the symptoms provided. This strategy ensures that the retrieval process is both comprehensive and aligned with the intended meaning of the user’s query.

**2. Noisy User Queries with Redundant Medical Entities.** Consider a user query such as “I am Jack; after having breakfast this morning, I felt nausea, stomach pain, dizziness, and weakness, and then I decided to rest but later developed a headache.” This query contains multiple medical entities—“nausea,” “stomach pain,” “dizziness,” “weakness,” and “headache”—which, while all related to medical conditions, may include irrelevant or redundant information that distracts from the main concern: the “stomach pain” following breakfast.

Directly retrieving knowledge from the knowledge graph (KG) based on such a query would likely include unnecessary entities like “headache” or “weakness,” leading to an overabundance of noisy or irrelevant knowledge. To address this, we again leverage hypothesis outputs from LLMs to refine the query, focusing on the most critical medical entities that directly relate to the task. In this example, entities such as “breakfast” and “stomach pain” would be emphasized, while irrelevant or less pertinent entities would be filtered out. This refinement enhances the precision of the KG retrieval process, ensuring that the retrieved knowledge is relevant and useful.

By applying these strategies, we ensure that our approach is robust to varying query quality and capable of retrieving meaningful knowledge while minimizing redundancy and noise in the results.

#### **A.11 Training LLMs with Knowledge from Knowledge Graphs**

Indeed, in the medical domain, the performance of the base model significantly impacts the quality of the hypothesis outputs (HO). Our experiments consistently demonstrate that larger models with stronger comprehension abilities produce higher-quality HOs. This finding provides a solid basis to believe that training LLMs on medical knowledge—particularly high-quality medical KGs—could greatly enhance the effectiveness of the pre-retrieval phase.

Incorporating KG knowledge directly into an LLM’s parametric memory has the potential to bypass certain challenges encountered during the retrieval stage, such as noise and misalignment when interacting with knowledge graphs. As you suggested, this approach could offer a promising alternative, especially in domains where the growth of high-quality KGs is relatively slow or static.

To explore this idea further, we propose leveraging our high-quality KG to fine-tune or train LLMs, enabling the models to internalize this structured medical knowledge. By doing so, we can potentially enhance the model’s ability to generate contextually relevant and accurate outputs based on the internalized knowledge.