KGFlow: Enhancing Large Language Model Reasoning with Multi-Graph Alignment for Knowledge-Intensive Medical Question Answering

Anonymous ACL submission

Abstract

Medical question answering (QA) is a knowledge-intensive task demanding evidence-based reasoning across diverse sources. Despite the demonstrated efficacy of augmenting large language models (LLMs) with domain-specific knowledge graphs (KGs) for QA tasks, the prevailing single-KG paradigm neglects the benefits of multi-KG complementarity. This paper introduces **KGFlow**, a novel multi-KG reasoning paradigm that mitigates the incomplete coverage and fragmented evidence of single-KG approaches through three LLM agents. Given a query, *Knowledge Explorer* extracts relevant knowledge triples in parallel from heterogeneous sources. *Graph Aligner* then semantically harmonizes these triples into global reasoning paths, which are filtered by *Contextual Pruner* to preserve the most valuable paths for inference. Extensive experiments show that KGFlow achieves promising performance improvements compared to state-of-the-art techniques on five medical QA datasets. We further observe by cases that the *plug-and-play* nature of our framework enables its application to diverse KG combinations and multilingual scenarios, resulting in broader knowledge coverage than single-KG approaches.

1 Introduction

Medical question answering (QA) aims to integrate evidence from heterogeneous sources to deliver precise answers with traceable reasoning chains. Recently, large language models (LLMs) (Yang et al., 2025) have demonstrated unprecedented capabilities in general-domain QA. However, these general-purpose LLMs may exhibit hallucinations in the medical domain, as the task necessitates precise reasoning over dynamically updated medical knowledge. Although supervised fine-tuning (SFT) provides an alternative to mitigate hallucinations in medical applications (Wu et al., 2024a), it imposes prohibitive resource requirements. Moreover, the black-box nature of LLMs makes evidence untraceable, which hinders the validation of medical answer derivations. These limitations underscore the necessity for medical QA systems that leverage structured knowledge bases to construct verifiable reasoning paths explicitly.

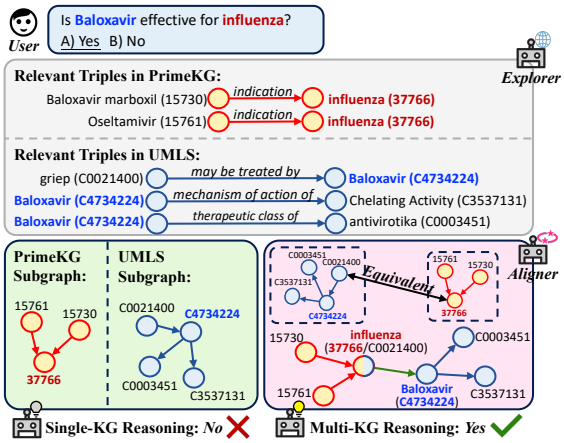

Figure 1: An example of medical QA, where the topic entities are colored. Yellow and blue triples represent relevant subgraphs retrieved from PrimeKG and UMLS, respectively. Underline denotes the correct answer.

Recognizing these challenges, a prevailing strategy is to leverage external medical knowledge to enhance the reasoning process of LLMs. Typically, retrieval-augmented generation (RAG) (Lewis et al., 2020) augments the input via embedded passage retrieval without further training, but constrains multi-hop reasoning due to evidence fragmentation. Building upon traditional RAG, GraphRAG (Edge et al., 2024) constructs knowledge graphs (KGs) from unstructured text chunks, enabling complex reasoning through graph topologies. This KG-based retrieval paradigm succeeds by providing structured knowledge to remedy the shortcomings of LLMs. Leveraging existing KGs, ToG (Sun et al., 2024) advances this paradigm by guiding LLMs to perform iterative path explo-

ration rather than executing naive retrieval. Despite the great potential of these knowledge-augmented methods, a critical issue persists due to their overdependence on a single KG, which neglects heterogeneity and complementarity across multiple KGs. Isolated KG usage inevitably suffers from incomplete information coverage. The multi-KG reasoning framework solves this problem by integrating disjointed knowledge islands through alignment technology, establishing evidence-complete paths for complex medical queries.

Consider the example in Figure 1, where *Baloxavir* and *influenza* are topic entities in the query. In this case, *Baloxavir* and *influenza* appear in the retrieved subgraphs of Unified Medical Language System (UMLS) (Bodenreider, 2004) and Precision Medicine Knowledge Graph (PrimeKG) (Chandak et al., 2023), respectively. However, neither UMLS nor PrimeKG alone contains a direct connection between the topic entities, making it difficult to establish a coherent reasoning path within a single KG. Therefore, relying on a single KG for LLMs inference will truncate reasoning paths, thus hindering accurate outcomes. Through multi-graph alignment, the entity *influenza* (ID: 37766) in PrimeKG can be mapped to the concept *grip* (CUI: $C0021400$) in UMLS, thereby forming a comprehensive reasoning path.

Motivated by the above observations, this paper proposes KGFlow, a novel multi-KG reasoning algorithm driven by three LLM agents: *Knowledge Explorer*, *Graph Aligner*, and *Contextual Pruner*. To address the incompleteness challenge of single-KG reasoning, Knowledge Explorer first extracts relevant subgraphs from heterogeneous KGs in parallel by initiating from topic entities and iteratively expanding reasoning paths via beam search. Afterwards, fragmented paths from heterogeneous KGs are assembled into a unified reasoning space through Graph Aligner, which generates semantic summaries for each entity and performs pairwise alignment to merge local paths into a global reasoning graph. Finally, Contextual Pruner converts the global reasoning chains into natural language and refines them through a filtering mechanism to retain valuable paths for answer generation. We evaluate our model against state-of-the-art (SOTA) baselines using several datasets, including MMLU-Med, MedQA-US, MedMCQA, PubMedQA, and BioASQ-Y/N. From the extensive experimental results, the proposed multi-KG reasoning strategy is demonstrated to be effective in utilizing complementary knowledge to construct coherent reasoning paths. For example, in Figure 1, KGFlow provides the accurate answer by integrating PrimeKG and UMLS, whereas single-KG or non-augmented methods yield errors due to incomplete knowledge.

To sum up, the contributions of this paper are as follows: (1) We pioneer a multi-KG reasoning paradigm that addresses the challenges of coverage incompleteness and evidence fragmentation in single-KG systems for medical QA; (2) We introduce the KGFlow framework that seamlessly integrates Knowledge Explorer, Graph Aligner, and Contextual Pruner to build coherent and traceable reasoning paths across heterogeneous KGs; (3) Extensive experiments show that our model achieves superior performance on five public datasets, which also demonstrates the effectiveness of the multi-graph alignment strategy in constructing coherent reasoning chains.

2 Related Work

2.1 LLMs for Medical Reasoning

The strong QA capabilities of general-domain LLMs, such as GPT-5 (OpenAI, 2025) and Qwen3 (Yang et al., 2025), have accelerated the development of their medical-domain counterparts. To adapt general-purpose LLMs for medical QA tasks, a straightforward approach involves SFT on domain-specific datasets (e.g., PMC-LLaMA (Wu et al., 2024a), HuatuoGPT-o1 (Chen et al., 2025)). However, SFT imposes significant temporal and computational resource demands. To reduce cost, methods like chain-of-thought (CoT) prompting leverage LLMs' intrinsic reasoning to decompose queries (Wei et al., 2022). However, their performance is constrained by the parametric knowledge, making them prone to erroneous reasoning in knowledge-intensive medical scenarios. To utilize the information from external sources, RAG (Lewis et al., 2020) is proposed to dynamically retrieve information from domain-specific corpora without additional training, mitigating hallucinations. Since KGs offer advantages in explicit, structured knowledge representation, graph-based retrieval has been studied. For instance, GraphRAG (Edge et al., 2024) and MedGraphRAG (Wu et al., 2024b) enhance complex reasoning by organizing RAG data into graphs. In addition, emerging research leverages established medical knowledge bases, such as UMLS, to augment LLMs with curated domain expertise, directly addressing their domain

knowledge deficiencies. ToG (Sun et al., 2024) introduces an LLM-KG integrating paradigm for interactive knowledge exploration, while KGARevion (Su et al., 2025) pioneers a KG-based agent that integrates the parametric knowledge of LLMs with codified knowledge in KGs. However, all these approaches rely on a singular KG, overlooking the flow of knowledge between multiple KGs.

In our framework, we propose a multi-KG reasoning algorithm that first retrieves relevant triples in parallel across heterogeneous KGs, then aligns them semantically and prunes irrelevant paths, thereby resolving the knowledge fragmentation inherent in single-KG reasoning.

2.2 Entity Alignment

Entity alignment (EA) serves as the cornerstone for integrating heterogeneous KGs by identifying semantically equivalent entities across distinct KGs. Traditional EA techniques primarily rely on entity embedding comparisons, falling into two methodological categories. The first paradigm comprises translation-based methods, which model relations as translational operations between entity embeddings, as exemplified by TransE (Bordes et al., 2013). Extended variants like TransMS (Güngör et al., 2022) and TransT (Chen et al., 2021) enhance this paradigm, improving scalability at the cost of expressiveness on complex relationships and fine-grained attributes. The second paradigm employs graph neural networks (GNNs) to derive structural embeddings through neighborhood feature aggregation, capturing complex topological dependencies. Representative methods encompass GCN-Align (Mao et al., 2020) for convolutional cross-graph alignment, and Dual-AMN (Mao et al., 2021) utilizing dual attention mechanisms. Although GNN-based methods proficiently model multi-hop graph structures and excel on dense topologies, they struggle with structural heterogeneities across KGs and underutilize contextual features. Emerging research pioneers a third paradigm leveraging LLMs to unify structural and semantic information for enhanced entity alignment. Approaches like ChatEA (Jiang et al., 2024) introduce KG-structured prompting via KG-code, and LLM4E (Chen et al., 2024) establishes fully automated alignment frameworks requiring zero manual annotations. These methods demonstrate LLMs' unique capacity to harness contextual semantics while enabling complex cross-graph reasoning and reducing reliance on supervised alignment data.

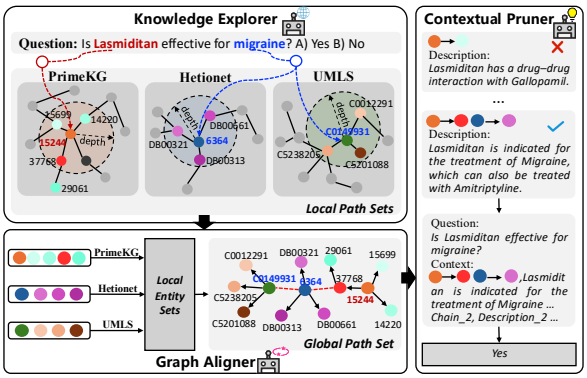

Figure 2: The overall architecture of KGFlow.

In this paper, we facilitate EA with LLMs to achieve unified modeling without retraining. Specifically, the proposed Graph Aligner agent generates entity descriptions by integrating entity names, types, and one-hop structures to resolve cross-KG semantic ambiguities.

3 Method

3.1 Problem Statement

The medical QA task takes a question q as input, which consists of a question stem q_s and a set of candidate answers q_c. The objective is to predict the correct answer $a \in q_c$ by leveraging an LLM $\mathcal{L}$ and m medical KGs $\{\mathcal{G}^i\}_{i=1}^m$. Each KG can be represented as a set of triples $\mathcal{G}^i = \{(h^i, r^i, t^i) | h^i, t^i \in \mathcal{E}^i, r^i \in \mathcal{R}^i\}$, where $\mathcal{E}^i$ and $\mathcal{R}^i$ are the entity and relation sets of the i-th KG, respectively.

3.2 Model Architecture

The overall architecture is illustrated in Figure 2. Knowledge Explorer first leverages beam search to explore from topic entities to relevant triples across heterogeneous sources. Graph Aligner then performs EA by integrating multi-dimensional information across KGs to consolidate local knowledge into global reasoning paths. Finally, Contextual Pruner retains the valuable paths and their explanations for answer generation.

3.3 Knowledge Explorer

Considering the exploration speed, we utilize m Knowledge Explorers to search for relevant triples on m KGs in parallel. For simplicity, we only introduce how a single Knowledge Explorer works. Given a question q and KG $\mathcal{G}$, Knowledge Explorer outputs a set of reasoning paths $\mathcal{P}$ where each path is an ordered sequence of triples from a relevant

subgraph $\mathcal{G}_{sub} \subset \mathcal{G}$ through *Initiation*, *Exploration*, and *Termination*.

3.3.1 Initiation

Following prior work (Sun et al., 2024), we employ the beam search strategy to search for the top-N relevant triples. At the beginning, Knowledge Explorer prompts the underlying LLM $\mathcal{L}$ to extract the topic entities $E^0 = \{e_i^0\}_{i=1}^{|E^0|}$ from the question q. Based on these topic entities, Knowledge Explorer locates seed entities in KG $\mathcal{G}$ and iteratively expands the reasoning path set $\mathcal{P}$ until the knowledge is sufficient or the maximum iteration depth D is reached. Note that the initial reasoning path set is empty, namely $\mathcal{P}^0 = \emptyset$.

3.3.2 Exploration

The exploration phase of each iteration comprises *relation exploration* and *entity exploration*. Specifically, the reasoning path set $\mathcal{P}^d$ of the d-th iteration expands iteratively from $\mathcal{P}^{d-1}$ by incorporating new triples into existing paths. To decouple relation and entity exploration, we extract the entity set $E^{d-1} = \{e_i^{d-1}\}_{i=1}^{N}$ and the relation set $R^{d-1} = \{r_j^{d-1}\}_{j=1}^{N}$ from $\mathcal{P}^{d-1}$, where E^{d-1} and R^{d-1} include all the tail entities and relations in the previous iteration, respectively. For the first iteration, E^0 comprises the topic entities, while $R^0 = \emptyset$ due to the absence of prior reasoning paths. In the d-th iteration, we initiate relation exploration via a depth of 1 beam search of width N from each tail entity in E^{d-1}, retrieving connected relations from $\mathcal{G}$ to form candidate relation set R^d_{cand}. It should be noted that relations in R^d_{cand} can be either inwards or outwards. Then, the top-N relevant relations R^d are selected from R^d_{cand} by querying the LLM $\mathcal{L}$ with the question q and the entity set E^{d-1}. Similar to the relation exploration, entity exploration executes a depth $= 1$, width $= N$ beam search over R^d relations, constructing E^d_{cand} and filtering to the top-N set E^d. Using the explored relations R^d and entities E^d, we perform one-hop expansions on every path in $\mathcal{P}^{d-1}$, incrementally growing the reasoning path set:

$$\mathcal{P}^d = \mathcal{P}^{d-1} \cup \{(h, r, t) \mid h \in E^{d-1}, r \in R^d, t \in E^d\} \quad (1)$$

3.3.3 Termination

Iteration terminates when either (1) the maximum depth D is reached, or (2) the explored knowledge is sufficient for answer generation. Specifically, we introduce an LLM-based judgment mechanism after each iteration to quantify the sufficiency of knowledge. Upon updating $\mathcal{P}^d$, we input the path set and question to the LLM, which autonomously assesses whether the explored paths are sufficient for answer generation. The exploration process terminates upon a positive judgment; otherwise, it proceeds to extend reasoning paths until reaching the maximum depth D.

3.4 Graph Aligner

The knowledge exploration stage yields m path sets $\{\mathcal{P}^i\}_{i=1}^{m}$, each originating from the KG $\mathcal{G}^i$. Graph Aligner integrates these fragmented path sets $\{\mathcal{P}^i\}_{i=1}^{m}$ into a global path set $\mathcal{P}_{\text{glob}}$ through *Summarization* and *Alignment*.

3.4.1 Summarization

The summarization stage employs the LLM to synthesize multi-dimensional entity information into natural language summaries. Initially, we aggregate all unique entities from the local path set $\mathcal{P}^i$ to construct the entity set. Subsequently, for each entity $e_j^i \in \mathcal{E}^i_{\text{loc}}$, we extract its local structure information $\mathcal{S}_j^i$ by retrieving one-hop neighbors (triples directly linked to e_j^i) from $\mathcal{G}^i$:

$$\mathcal{S}_j^i = \left\{(h, r, t) \in \mathcal{T}^i \;\middle|\; h = e_j^i \ \vee \ t = e_j^i\right\} \quad (2)$$

Given the prohibitive computational cost of maintaining exhaustive one-hop neighbors in large-scale medical KGs, we further restrict the size of $\mathcal{S}_j^i$ to contain at most N triples. Priority is given to triples within $\mathcal{P}^i$; if fewer than N exist, we uniformly sample additional triples from the entity's one-hop neighbors $\mathcal{S}_j^i$ to satisfy the bound. Finally, the name, type τ_j^i, and structure information $\mathcal{S}_j^i$ of entity e_j^i are integrated via LLM prompting to generate a natural language summary m_j^i.

3.4.2 Alignment

The alignment phase aims to link ambiguous entities with equivalent semantics across distinct medical KGs. Firstly, for a specific entity pair (e_j^i, e_s^r) from $\mathcal{G}^i$ and $\mathcal{G}^r$, we integrate four-dimensional information: (1) canonical names, (2) semantic types, (3) local structure, and (4) LLM-generated summaries to compute their equivalence scores. Secondly, the LLM computes three $[0, 10]$-scaled equivalence metrics $sim_{\mathcal{N}}, sim_{\mathcal{S}}, sim_{\mathcal{M}}$ based on name, structure, and summary similarities, respectively. Note that type information is mandated in the LLM prompt to constrain $sim_{\mathcal{N}}$ and $sim_{\mathcal{S}}$ computations, ensuring semantic consistency during

equivalence scoring. Based on these metrics, the final alignment decision is determined by comparing the sum of $sim_{\mathcal{N}}, sim_{\mathcal{S}}, sim_{\mathcal{M}}$ against a configurable threshold K, yielding a binary verdict:

$$a(e_j^i, e_s^r) = \begin{cases} 1, & \text{if } sim_{\mathcal{N}} + sim_{\mathcal{S}} + sim_{\mathcal{M}} \geq K \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

where $a(e_j^i, e_s^r) = 1$ indicates that the two entities are semantically equivalent and should be merged. Upon completing pairwise alignment across all local entity sets $\{\mathcal{E}_{\text{loc}}^i\}_{i=1}^m$, the identified equivalent entities serve as bridges to interconnect paths from different KGs. This process synthesizes a unified multi-graph reasoning network, which establishes the global reasoning path set $\mathcal{P}_{\text{glob}}$ for subsequent answer generation.

3.5 Contextual Pruner

Leveraging the global path set $\mathcal{P}_{\text{glob}}$, Contextual Pruner performs *Pruning* and *Generation* stages to ensure that only the most pertinent information flows into the final path set $\mathcal{P}_{\text{final}}$ for answer generation.

3.5.1 Pruning

The multi-hop exploration and EA procedures inevitably introduce extraneous paths into $\mathcal{P}_{\text{glob}}$ that are of low relevance to the target question q. Consequently, this phase performs pruning to preserve high-value reasoning paths. Initially, we traverse all reasoning paths in $\mathcal{P}_{\text{glob}}$ originating from topic entities and nodes with in-degree 0. The set of starting nodes $\mathcal{E}_{\text{start}}$ is formally defined as:

$$\mathcal{E}_{\text{start}} = \left\{ e \in \mathcal{E}_{glob} \;\middle|\; e \in E^0 \;\vee\; \deg^-(e) = 0 \right\} \quad (4)$$

where $\mathcal{E}_{glob}$ is the set of all entities in $\mathcal{P}_{\text{glob}}$, $\deg^-(e)$ denotes the in-degree of node e. After identifying the starting nodes, the traversal process explores each reasoning path in $\mathcal{P}_{\text{glob}}$ that originates from $\mathcal{E}_{\text{start}}$. A path p_k is formally defined as an ordered sequence of triples, representing the chain of reasoning steps from the initial entity to subsequent entities via relations. Subsequently, we deploy the LLM to transform each structured reasoning path p_k into a natural language description d_k through instruction-based semantic parsing. Finally, we utilize the LLM to analyze whether a specific path in $\mathcal{P}_{\text{glob}}$ is relevant to the target question q, and add valuable paths to the final path set $\mathcal{P}_{\text{final}}$ to support answer generation.

3.5.2 Generation

Upon obtaining the final paths $\mathcal{P}_{\text{final}}$ through the pruning process, we prompt the LLM to infer the final result a_{pred} based on the target question q, structured reasoning paths $\mathcal{P}_{\text{final}}$, and descriptions of reasoning paths $\mathcal{D}_{\text{final}}$. This prune-then-generate paradigm ensures answer reliability through high-value path selection while enhancing reasoning interpretability via transparent reasoning traces. The details of the KGFlow reasoning process are described in Algorithm 1. See Appendix B for more details.

Algorithm 1 KGFlow Algorithm

Require: Question q, KGs $\{\mathcal{G}^i\}_{i=1}^m$, LLM $\mathcal{L}$, alignment threshold K, max depth D, max width N
Ensure: The predicted answer
Initialization: $E^0 \leftarrow \mathcal{L}(q, \rho_{\text{extract-ent}})$, $R^0 \leftarrow \emptyset$ ▷ Knowledge Explorer
for $i = 1$ **to** m **in parallel do**
  $\mathcal{P}^i \leftarrow \emptyset, d \leftarrow 0$
  while $d < D$ **and not** knowledge-sufficient **do**
    $R_{\text{cand}}^d \leftarrow$ 1-hop relations from E^{d-1} in $\mathcal{G}^i$
    $R^d \leftarrow \mathcal{L}$-filtered top-$N$ relations from R_{cand}^d
    $E_{\text{cand}}^d \leftarrow$ 1-hop entities from R^d in $\mathcal{G}^i$
    $E^d \leftarrow \mathcal{L}$-filtered top-$N$ entities from E_{cand}^d
    $\mathcal{P}^i \leftarrow \mathcal{P}^i \cup \{(h, r, t) \mid h \in E^{d-1}, r \in R^d, t \in E^d\}$
    $d \leftarrow d + 1$
    Check knowledge sufficiency via $\mathcal{L}$
  end while
end for
$\{\mathcal{E}_{\text{loc}}^i\}_{i=1}^m \leftarrow$ unique entities from $\{\mathcal{P}^i\}_{i=1}^m$ ▷ Graph Aligner
for $i = 1$ **to** m **do**
  for $e_j^i \in \mathcal{E}_{\text{loc}}^i$ **do**
    $m_j^i \leftarrow \mathcal{L}(\text{name: } e_j^i, \text{type: } \tau_j^i, \text{structure: } \mathcal{S}_j^i, \rho_{\text{summarize}})$
  end for
end for
for $e_j^i \in \mathcal{E}_{\text{loc}}^i, e_s^r \in \mathcal{E}_{\text{loc}}^r$ $(i \neq r)$ **do**
  Compute $sim_{\mathcal{N}}, sim_{\mathcal{S}}, sim_{\mathcal{M}}$ via $\mathcal{L}$
  if $sim_{\mathcal{N}} + sim_{\mathcal{S}} + sim_{\mathcal{M}} \geq K$ **then**
    Align e_j^i and e_s^r
  end if
end for
$\mathcal{P}_{\text{glob}} \leftarrow$ merge paths via aligned entities
Convert $\mathcal{P}_{\text{glob}}$ to descriptions ▷ Contextual Pruner
$\mathcal{P}_{\text{final}} \leftarrow \mathcal{L}$-filtered paths in $\mathcal{P}_{\text{glob}}$ relevant to q
$a_{\text{pred}} \leftarrow \mathcal{L}(\text{question: } q, \text{path: } \mathcal{P}_{\text{final}}, \text{description: } \mathcal{D}_{\text{final}}, \rho_{\text{generate}})$
return a_{pred}

4 Experiments

4.1 Experimental Settings

4.1.1 Datasets and Evaluation Metrics

We evaluate our model on five datasets: MMLU-Med (Hendrycks et al., 2021), MedQA-US (Jin et al., 2020), MedMCQA (Pal et al., 2022), PubMedQA (Jin et al., 2019), and BioASQ-Y/N (Tsatsaronis et al., 2015). To assess multilingual capability, we include two Chinese datasets: CMMLU-Med (Li et al., 2024) and CEval-Med (Huang et al., 2023). Table 8 summarizes the dataset statistics. KGFlow utilizes four KGs: PrimeKG (Chandak et al., 2023), Hetionet (Himmelstein et al., 2017), UMLS (Bodenreider, 2004), and CPubMedKG.

Following established practice (Su et al., 2025), we adopt accuracy (acc) as the primary evaluation metric and report the standard deviation (std) across runs to quantify variability and robustness.

4.1.2 Baselines

KGFlow is evaluated against three model categories: general-domain LLMs (Jiang et al., 2023; Dubey et al., 2024; Yang et al., 2025), medical-domain LLMs (Han et al., 2023; Wu et al., 2024a; Chen et al., 2025), and LLM-KG integration models (Asai et al., 2024; Yang et al., 2024; Soman et al., 2023; Xiong et al., 2024; Su et al., 2025). For general- / medical-domain LLMs, experiments cover both standard and CoT reasoning settings.

4.1.3 Implementation Details

Experiments were conducted on four NVIDIA RTX 4090 GPUs (24GB each) without quantization. The temperature was set to 0.4 for exploration and alignment, and 0 for answer generation. Based on the reasoning mechanisms of the LLMs, maximum token limits were set to 256 for standard models and 4096 for models producing intermediate reasoning traces to avoid truncation. Key hyperparameters included a beam width of $N = 5$, a maximum depth of $D = 3$, and an EA threshold of $K = 18$. All LLMs used zero-shot prompting, with KGs stored in Neo4j. Results are the mean of three independent runs (random seeds: 42, 3407, 114514) to ensure reproducibility. More details of experimental settings can be found in Appendix A.

4.2 Main Results

To evaluate our approach, we compare KGFlow with several SOTA baselines across five multi-choice medical QA datasets. As shown in Table 2, our model outperformed all other models, which improved 1.6 accuracy points on average over the baselines on MMLU-Med, BioASQ-Y/N, MedQA-US, PubMedQA, and MedMCQA datasets. In addition, we infer from Table 2 that neither domain SFT nor CoT achieves stable performance improvement. In some cases, these techniques even introduce inconsistency or marginal degradation, indicating their insufficiency in resolving knowledge-intensive medical reasoning tasks. However, the proposed KGFlow significantly boosts reasoning performance across LLMs through the multi-KG strategy, improving LLaMA3.1-8B by 3.6 points, Qwen3-8B by 2.3 points, and HuatuoGPT-o1-8B by 3.6 points in average accuracy.

Table 1: Statistics of the multi-choice medical QA datasets. "Avg. L": Average token counts per query. "Lang.": Language. "EN": English. "ZH": Chinese.

Lang.	Dataset	Size	Options	Avg. L	Source
EN	MMLU-Med	1,089	A/B/C/D	102	Examination
	MedQA-US	1,273	A/B/C/D	212	Examination
	MedMCQA	4,183	A/B/C/D	57	Examination
	PubMedQA	1,000	Yes/No/Maybe	334	Literature
	BioASQ-Y/N	618	Yes/No	25	Literature
ZH	CMMLU-Med	1,333	A/B/C/D	72	Examination
	CEVal-Med	375	A/B/C/D	65	Examination

4.3 Ablation Study

To evaluate the impact of each module, we progressively disable components and report the performance comparison in Table 2. Removing Graph Aligner from KGFlow (HuatuoGPT-o1-8B, w/o Graph Aligner) reduces accuracy by 0.6 points on average compared to the full model. The decline occurs because, even if multiple KGs collectively contain the necessary knowledge, the lack of an alignment mechanism prevents the formation of a complete reasoning path. The KGFlow (HuatuoGPT-o1-8B, w/o Contextual Pruner) method represents the omission of Contextual Pruner, resulting in an average decrease of 1.9 points. The performance drop arises because unpruned redundant and noisy triples pollute the reasoning paths, contaminating the supporting evidence. Furthermore, KGFlow (HuatuoGPT-o1-8B, w/o UMLS and Hetionet) surpasses SOTA LLM-KG integration methods with the same knowledge source, demonstrating the effectiveness of Knowledge Explorer. Collectively, these results validate the necessity of each module.

4.4 Resolution of the Incompleteness Problem

KGFlow overcomes the incompleteness of single knowledge sources through multi-graph integration. As shown in Table 2, combining all three KGs yields the best performance on **English** benchmarks: KGFlow (HuatuoGPT-o1-8B, PrimeKG + Hetionet + UMLS) surpasses dual- and single-KG variants by 0.4 and 1.0 points on average. This gain is consistent in **Chinese** benchmarks (Table 3), where integrating CPubMedKG and bilingual UMLS achieves SOTA accuracies of 0.822 and 0.880. The bilingual UMLS [EN&ZH] consistently outperforms its monolingual counterpart, confirming that cross-lingual alignment mitigates language-specific gaps. The plug-and-play nature of KGFlow enables a flexible combination of KGs for specific scenarios. These results demonstrate that our multi-KG paradigm enables **cross-graph**

Table 2: Performance on the MMLU-Med, BioASQ-Y/N, MedQA-US, PubMedQA, and MedMCQA datasets in comparison with the SOTA methods. "†" denotes paper-reported results; others are reproduced using officially released code. "-": results not reported.

Method	MMLU-Med	BioASQ-Y/N	MedQA-US	PubMedQA	MedMCQA	Avg.
	Acc. (std)	Acc. (std)	Acc. (std)	Acc. (std)	Acc. (std)	Acc. (std)
General-domain LLMs						
Mistral-7B (Jiang et al., 2023)	0.625(.004)	0.638(.001)	0.470(.007)	0.682(.011)	0.488(.015)	0.581(.004)
Mistral-7B (CoT)	0.630(.003)	0.650(.002)	0.474(.003)	0.688(.005)	0.545(.008)	0.597(.002)
LLaMA3.1-8B (Dubey et al., 2024)	0.665(.003)	0.725(.001)	0.599(.002)	0.756(.002)	0.556(.001)	0.660(.001)
LLaMA3.1-8B (CoT)	0.680(.002)	0.705(.002)	0.550(.004)	0.753(.006)	0.561(.005)	0.650(.002)
Qwen3-8B (Yang et al., 2025)	0.910(.001)	0.773(.012)	0.908(.004)	0.783(.003)	0.717(.005)	0.818(.003)
Qwen3-8B (CoT)	0.902(.003)	0.775(.009)	0.912(.004)	0.790(.001)	0.714(.004)	0.819(.002)
Medical-domain LLMs						
MedAlpaca-7B (Han et al., 2023)	0.614(.004)	0.498(.028)	0.404(.005)	0.355(.002)	0.306(.006)	0.435(.006)
MedAlpaca-7B (CoT)	0.607(.004)	0.499(.025)	0.400(.003)	0.352(.005)	0.311(.001)	0.434(.005)
PMC-LLaMA-7B (Wu et al., 2024a)	0.203(.011)	0.338(.017)	0.245(.004)	0.217(.012)	0.297(.008)	0.260(.005)
PMC-LLaMA-7B (CoT)	0.210(.004)	0.341(.002)	0.210(.002)	0.217(.007)	0.297(.013)	0.255(.003)
HuatuoGPT-o1-8B (Chen et al., 2025)	0.754(.003)	0.775(.004)	0.738(.002)	0.792(.002)	0.702(.001)	0.752(.001)
HuatuoGPT-o1-8B (CoT)	0.768(.006)	0.779(.009)	0.753(.004)	0.795(.003)	0.695(.002)	0.758(.003)
LLM-KG Integration Models						
Self-RAG (13B) (Asai et al., 2024)	0.506(.004)	0.646(.050)	0.410(.020)	0.347(.023)	0.376(.016)	0.457(.012)
KG-Rank (13B) (Yang et al., 2024)	0.452(.005)†	0.503(.015)†	0.362(.011)†	-	-	-
KG-RAG (8B) (Soman et al., 2023)	0.516(.005)†	0.662(.005)†	0.343(.001)†	-	-	-
MedRAG (70B) (Xiong et al., 2024)	0.567(.015)	0.712(.018)	0.484(.014)	0.404(.008)	0.437(.010)	0.521(.006)
KGARevion (8B) (Su et al., 2025)	0.734(.004)†	0.763(.001)†	0.620(.003)†	-	-	-
KGFlow Group (Ours)						
KGFlow (LLaMA3.1-8B, PrimeKG + Hetionet + UMLS)	0.729(.009)	0.782(.010)	0.625(.003)	0.769(.006)	0.577(.011)	0.696(.004)
KGFlow (Qwen3-8B, PrimeKG + Hetionet + UMLS)	**0.921(.006)**	0.816(.005)	**0.921(.006)**	0.796(.008)	**0.756(.003)**	**0.842(.003)**
KGFlow (HuatuoGPT-o1-8B, PrimeKG + Hetionet + UMLS)	0.822(.002)	0.830(.009)	0.806(.005)	**0.798(.007)**	0.716(.004)	0.794(.003)
KGFlow (HuatuoGPT-o1-8B, w/o Graph Aligner)	0.818(.003)	0.825(.002)	0.801(.003)	0.791(.005)	0.706(.004)	0.788(.002)
KGFlow (HuatuoGPT-o1-8B, w/o Contextual Pruner)	0.809(.002)	0.818(.005)	0.773(.003)	0.785(.007)	0.692(.003)	0.775(.002)
KGFlow (HuatuoGPT-o1-8B, w/o UMLS)	0.821(.003)	**0.832(.011)**	0.796(.002)	0.796(.004)	0.712(.003)	0.791(.003)
KGFlow (HuatuoGPT-o1-8B, w/o Hetionet)	0.813(.004)	0.827(.003)	0.797(.005)	0.795(.002)	0.710(.005)	0.788(.002)
KGFlow (HuatuoGPT-o1-8B, w/o PrimeKG)	0.815(.003)	0.825(.004)	0.801(.002)	0.797(.008)	0.713(.004)	0.790(.002)
KGFlow (HuatuoGPT-o1-8B, w/o UMLS and Hetionet)	0.809(.002)	0.823(.003)	0.796(.009)	0.795(.002)	0.704(.002)	0.785(.002)
KGFlow (HuatuoGPT-o1-8B, w/o UMLS and PrimeKG)	0.818(.005)	0.825(.009)	0.801(.005)	0.796(.010)	0.705(.003)	0.789(.003)
KGFlow (HuatuoGPT-o1-8B, w/o Hetionet and PrimeKG)	0.809(.008)	0.818(.012)	0.773(.007)	0.791(.005)	0.704(.004)	0.779(.004)

Table 3: Accuracy on two Chinese medical QA datasets.

Method	CMMLU-Med	CEval-Med
Llama3-8B (Dubey et al., 2024)	0.516	0.537
Qwen3-8B (Yang et al., 2025)	0.773	0.816
CHIMED-GPT (Tian et al., 2024)	0.529	0.683
CareBot-Chat (Zhao et al., 2025)	0.616	0.637
MEDCARE-14B (Liao et al., 2024)	0.716	0.781
HuatuoGPT-o1-7B-Chinese (Chen et al., 2025)	0.816	-
KGFlow (Qwen3-8B, $\text{UMLS}_{[\text{EN}]}$)	0.780	0.827
KGFlow (Qwen3-8B, $\text{CPubMedKG}_{[\text{ZH}]}$)	0.793	0.853
KGFlow (Qwen3-8B, $\text{UMLS}_{[\text{EN\&ZH}]}$)	0.793	0.837
KGFlow (Qwen3-8B, $\text{CPubMedKG}_{[\text{ZH}]}$ + $\text{UMLS}_{[\text{EN\&ZH}]}$)	**0.822**	**0.880**

Table 4: Performance of different EA methods on MED-BBK dataset.

Method	Settings		Results	
	Struc.	*Name.*	Hits@1	MRR
MTransE (Chen et al., 2017)	✓		0.004	0.012
BootEA (Sun et al., 2018)	✓		0.307	0.399
AliNet (Sun et al., 2020)	✓		0.017	0.033
RDGCN (Wu et al., 2019)	✓	✓	0.306	0.365
AttrE (Trisedya et al., 2019)	✓	✓	0.194	0.279
MultiKE (Zhang et al., 2019)	✓	✓	0.213	0.289
OntoEA (Xiang et al., 2021)	✓	✓	0.517	0.604
HMEA (Wang et al., 2025)	✓	✓	0.627	0.705
Graph Aligner (ours)	✓	✓	$\mathbf{0.719}^{\uparrow 0.092}$	$\mathbf{0.768}^{\uparrow 0.063}$

and **cross-lingual** knowledge transfer, strengthening medical QA reasoning.

4.5 Resolution of the Fragmentation Problem

We evaluate our EA method on the MED-BBK-9K dataset (Zhang et al., 2020), aiming to assess its capability in addressing the evidence fragmentation challenge. We adopt a two-stage paradigm following established EA frameworks (Xiang et al., 2021), which comprises candidate entity pair generation followed by a reranking of these candidates with our proposed Graph Aligner. As shown in Table 4, our approach outperforms all baselines by 9.2 points in Hits@1 and 6.3 points in MRR. The results demonstrate that jointly modeling structural and contextual information enables more precise EA, establishing reliable connectivity across heterogeneous KGs to support knowledge-aware generation.

4.6 Impact of Hyperparameters

We analyze the impact of key hyperparameters in Figure 3, 4 and Table 5. For knowledge exploration, accuracy increases with beam width N and depth D, saturating near $N = 5$ with diminishing gains beyond $D = 3$. The EA threshold K balances precision and recall, with $K = 18$ maintaining high accuracy and sufficient aligned pairs. For reasoning-enhanced LLMs, expanding the maximum token limit substantially boosts performance, with minimal gains beyond 4096. As shown in Table 5, integrating both structured chains and natural language descriptions achieves the best per-

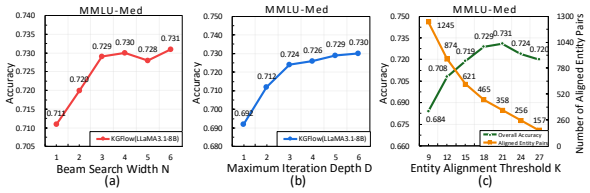

Figure 3: Performance of KGFlow (LLaMA3.1-8B, PrimeKG+Hetionet+UMLS) with different configurations. (a) Different beam search widths. (b) Different maximum iteration depths. (c) Different EA thresholds.

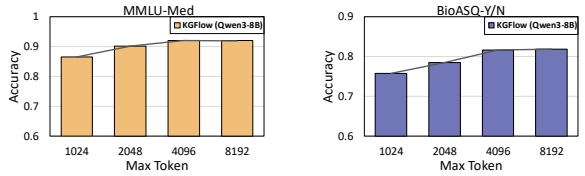

Figure 4: Performance of KGFlow (Qwen3-8B, PrimeKG+Hetionet+UMLS) with different maximum token lengths.

Table 5: Performance of KGFlow with different external knowledge input forms on the MMLU-Med and BioASQ-Y/N datasets.

Method	MMLU-Med	BioASQ-Y/N
KGFlow (Qwen3-8B, PrimeKG + Hetionet + UMLS)	**0.921**	**0.816**
w/o Description	$0.912^{\downarrow 0.009}$	$0.807^{\downarrow 0.009}$
w/o Chain	$0.920^{\downarrow 0.001}$	$0.812^{\downarrow 0.004}$

formance, confirming their complementarity. The optimal configuration collectively ensures the best performance-efficiency trade-off.

4.7 Case Study

As illustrated in Figure 5, for the query "Is SATB1 expressed in thymocytes?", KGFlow's reasoning demonstrates how multi-KG integration resolves the limitations of any single source. In UMLS, the critical triple *(Man, organism_has_gene, SATB1)* was pruned during beam search, truncating a potential 3-hop path and leaving the topic entity *thymocytes* disconnected. Conversely, PrimeKG contains the relevant triple *(thymus, anatomy_protein_present, SATB1)*, but its head entity *thymusis* is lexically divergent from the query entity *thymocytes*. It is through cross-KG entity alignment that *thymus* (PrimeKG) is mapped to *thymocytes* (UMLS). This alignment bridges the fragmented knowledge, establishing a conclusive path between *thymocytes* and *SATB1* and enabling KGFlow to generate the correct answer. This case vividly illustrates that our framework compensates for the coverage gaps and vocabulary heterogeneity of individual KGs.

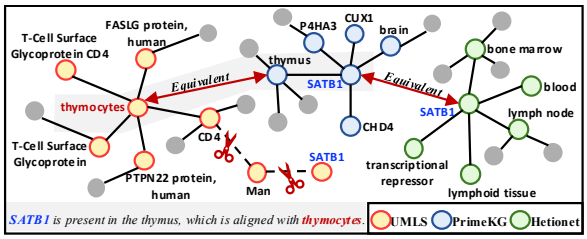

Figure 5: A case demonstrating how KGFlow resolves evidence fragmentation through cross-KG alignment. Topic entities are highlighted in color. Node colors indicate graph distance: 1-hop (colored), 2-hop (gray). 3-hop entities are not shown. The gray background indicates the final reasoning path.

4.8 Efficiency Analysis

Table 6 shows that KGFlow processes a query across three KGs in approximately 1 minute using tens of LLM calls on average, demonstrating practical viability. The framework scales efficiently despite a fivefold size disparity (Hetionet:PrimeKG:UMLS $\approx$ 1:2:5). Neo4j retrieval maintains stable latency ($\sim$10 ms/query) and low memory overhead ($<$ 100 MB), while exploration time increases sublinearly, validating the pruning strategies. This confirms that incorporating additional KGs remains computationally feasible. Moreover, operating locally on 8B open-source LLMs eliminates API fees, making KGFlow significantly more economical than proprietary or fine-tuned alternatives.

Table 6: Statistics of KGFlow computational cost.

Phase	Latency (s)	Memory (GB)	#Calls	Input Token (k)	Output Token (k)
Exploration	32.4	16.9	38.2	29.6	6.8
Summarization	6.9	16.9	8.3	7.1	1.6
Alignment	17.9	16.9	12.8	18.6	2.5
Pruning	7.6	16.9	9.6	6.2	1.7
Generation	0.8	16.9	1	0.8	0.1

5 Conclusion

This paper focuses on the knowledge incompleteness and fragmentation challenges in medical QA. Towards this, we propose KGFlow, a multi-KG reasoning model powered by Knowledge Explorer, Graph Aligner, and Contextual Pruner. Empirical results demonstrate that the multi-graph alignment strategy of our model considerably enhances the flow of knowledge while gaining high performance. In conclusion, the plug-and-play architecture of KGFlow delivers a reusable framework for multiple KGs integration, which establishes a new paradigm for next-generation medical reasoning systems.

Limitations

Despite the empirical success of KGFlow, several limitations remain. The multi-agent pipeline emphasizes comprehensive reasoning, which opens avenues for further acceleration in latency-sensitive applications. Future work will explore parallelized retrieval and alignment distillation to improve throughput. Additionally, reasoning stability depends on knowledge coverage and model verification capacity, with conflict filtering presently embedded within the pruning phase to maintain pipeline efficiency. Incorporating uncertainty-aware path selection, self-refinement mechanisms, and dynamic synchronization will further enhance robustness for evolving medical scenarios.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Olivier Bodenreider. 2004. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Res.*, 32(Database-Issue):267–270.

Antoine Bordes, Nicolas Usunier, Alberto García-Durán, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 2787–2795.

Payal Chandak, Kexin Huang, and Marinka Zitnik. 2023. Building a knowledge graph to enable precision medicine. *Scientific Data*, 10(1):67.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, and Benyou Wang. 2025. Towards medical complex reasoning with llms through medical verifiable problems. In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 14552–14573. Association for Computational Linguistics.

Muhao Chen, Yingtao Tian, Mohan Yang, and Carlo Zaniolo. 2017. Multilingual knowledge graph embeddings for cross-lingual knowledge alignment. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 1511–1517. ijcai.org.

Shengyuan Chen, Qinggang Zhang, Junnan Dong, Wen Hua, Qing Li, and Xiao Huang. 2024. Entity alignment with noisy annotations from large language models. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Xin Chen, Bin Yan, Jiawen Zhu, Dong Wang, Xiaoyun Yang, and Huchuan Lu. 2021. Transformer tracking. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 8126–8135. Computer Vision Foundation / IEEE.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 82 others. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

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

Alper Güngör, Baris Askin, Damla Alptekin Soydan, Emine Ulku Saritas, Can Baris Top, and Tolga Çukur. 2022. Transms: Transformers for super-resolution calibration in magnetic particle imaging. *IEEE Trans. Medical Imaging*, 41(12):3562–3574.

Tianyu Han, Lisa C. Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K. Bressem. 2023. Medalpaca - an open-source collection of medical conversational AI models and training data. *CoRR*, abs/2304.08247.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Daniel Scott Himmelstein, Antoine Lizee, Christine Hessler, Leo Brueggeman, Sabrina L Chen, Dexter Hadley, Ari Green, Pouya Khankhanian, and Sergio E Baranzini. 2017. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *elife*, 6:e26726.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. In *Advances in Neural Information Processing Systems*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *CoRR*, abs/2310.06825.

Xuhui Jiang, Yinghan Shen, Zhichao Shi, Chengjin Xu, Wei Li, Zixuan Li, Jian Guo, Huawei Shen, and Yuanzhuo Wang. 2024. Unlocking the power of large language models for entity alignment. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 7566–7583. Association for Computational Linguistics.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. *CoRR*, abs/2009.13081.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 2567–2577. Association for Computational Linguistics.

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. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2024. CMMLU: measuring massive multitask language understanding in chinese. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 11260–11285. Association for Computational Linguistics.

Yusheng Liao, Shuyang Jiang, Zhe Chen, Yanfeng Wang, and Yu Wang. 2024. Medcare: Advancing medical llms through decoupling clinical alignment and knowledge aggregation. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 10562–10581. Association for Computational Linguistics.

Xin Mao, Wenting Wang, Yuanbin Wu, and Man Lan. 2021. Boosting the speed of entity alignment 10 ×: Dual attention matching network with normalized hard sample mining. In *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*, pages 821–832. ACM / IW3C2.

Xin Mao, Wenting Wang, Huimin Xu, Man Lan, and Yuanbin Wu. 2020. MRAEA: an efficient and robust entity alignment approach for cross-lingual knowledge graph. In *WSDM '20: The Thirteenth ACM International Conference on Web Search and Data Mining, Houston, TX, USA, February 3-7, 2020*, pages 420–428. ACM.

OpenAI. 2025. Gpt-5 system card. `https://cdn.openai.com/gpt-5-system-card.pdf`. Accessed: 2025-10-24.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on Health, Inference, and Learning, CHIL 2022, 7-8 April 2022, Virtual Event*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR.

Karthik Soman, Peter W. Rose, John H. Morris, Rabia E. Akbas, Brett Smith, Braian Peetoom, Catalina Villouta-Reyes, Gabriel Cerono, Yongmei Shi, Angela Rizk-Jackson, Sharat Israni, Charlotte A. Nelson, Sui Huang, and Sergio E. Baranzini. 2023. Biomedical knowledge graph-enhanced prompt generation for large language models. *CoRR*, abs/2311.17330.

Xiaorui Su, Yibo Wang, Shanghua Gao, Xiaolong Liu, Valentina Giunchiglia, Djork-Arné Clevert, and Marinka Zitnik. 2025. Kgarevion: An AI agent for knowledge-intensive biomedical QA. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Jiashuo Sun, Chengjin Xu, Lumingyuan Tang, Saizhuo Wang, Chen Lin, Yeyun Gong, Lionel M. Ni, Heung-Yeung Shum, and Jian Guo. 2024. Think-on-graph: Deep and responsible reasoning of large language model on knowledge graph. In *The Twelfth International Conference on Learning Representations,ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Zequn Sun, Wei Hu, Qingheng Zhang, and Yuzhong Qu. 2018. Bootstrapping entity alignment with knowledge graph embedding. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 4396–4402. ijcai.org.

Zequn Sun, Chengming Wang, Wei Hu, Muhao Chen, Jian Dai, Wei Zhang, and Yuzhong Qu. 2020. Knowledge graph alignment network with gated multi-hop neighborhood aggregation. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances*

in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020, pages 222–229. AAAI Press.

Yuanhe Tian, Ruyi Gan, Yan Song, Jiaxing Zhang, and Yongdong Zhang. 2024. Chimed-gpt: A chinese medical large language model with full training regime and better alignment to human preferences. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 7156–7173. Association for Computational Linguistics.

Bayu Distiawan Trisedya, Jianzhong Qi, and Rui Zhang. 2019. Entity alignment between knowledge graphs using attribute embeddings. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 297–304. AAAI Press.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R. Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, Yannis Almirantis, John Pavlopoulos, Nicolas Baskiotis, Patrick Gallinari, Thierry Artières, Axel-Cyrille Ngonga Ngomo, Norman Heino, Éric Gaussier, Liliana Barrio-Alvers, and 3 others. 2015. An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. *BMC Bioinform.*, 16:138:1–138:28.

Weiguang Wang, Lijuan Ma, Wei Cai, Haiyan Zhao, and Xia Zhang. 2025. HMEA: A hierarchical medical knowledge graph entity alignment model fusing multi-aspect information. *Artif. Intell. Medicine*, 168:103188.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Weidi Xie, and Yanfeng Wang. 2024a. Pmc-llama: toward building open-source language models for medicine. *J. Am. Medical Informatics Assoc.*, 31(9):1833–1843.

Junde Wu, Jiayuan Zhu, and Yunli Qi. 2024b. Medical graph RAG: towards safe medical large language model via graph retrieval-augmented generation. *CoRR*, abs/2408.04187.

Yuting Wu, Xiao Liu, Yansong Feng, Zheng Wang, Rui Yan, and Dongyan Zhao. 2019. Relation-aware entity alignment for heterogeneous knowledge graphs. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 5278–5284. ijcai.org.

Yuejia Xiang, Ziheng Zhang, Jiaoyan Chen, Xi Chen, Zhenxi Lin, and Yefeng Zheng. 2021. Ontoea: Ontology-guided entity alignment via joint knowledge graph embedding. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 1117–1128. Association for Computational Linguistics.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 6233–6251. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Rui Yang, Haoran Liu, Edison Marrese-Taylor, Qingcheng Zeng, Yuhe Ke, Wanxin Li, Lechao Cheng, Qingyu Chen, James Caverlee, Yutaka Matsuo, and Irene Li. 2024. Kg-rank: Enhancing large language models for medical QA with knowledge graphs and ranking techniques. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing, BioNLP@ACL 2024, Bangkok, Thailand, August 16, 2024*, pages 155–166. Association for Computational Linguistics.

Qingheng Zhang, Zequn Sun, Wei Hu, Muhao Chen, Lingbing Guo, and Yuzhong Qu. 2019. Multi-view knowledge graph embedding for entity alignment. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 5429–5435. ijcai.org.

Ziheng Zhang, Hualuo Liu, Jiaoyan Chen, Xi Chen, Bo Liu, Yuejia Xiang, and Yefeng Zheng. 2020. An industry evaluation of embedding-based entity alignment. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020 - Industry Track, Online, December 12, 2020*, pages 179–189. International Committee on Computational Linguistics.

Lulu Zhao, Weihao Zeng, Xiaofeng Shi, and Hua Zhou. 2025. Carebot: A pioneering full-process open-source medical language model. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pages 26039–26047. AAAI Press.

A Detailed Experimental Settings

A.1 Datasets

We used five English multi-choice medical QA datasets (MMLU-Med, MedQA-US, MedMCQA, PubMedQA, and BioASQ-Y/N) and two Chinese multi-choice medical QA datasets (CMMLU-Med, CEval-Med) to evaluate our proposed model. The samples in these datasets are shown in Table 8.

A.2 Knowledge Bases

KGFlow utilizes four medical knowledge bases as its primary knowledge sources: PrimeKG (Chandak et al., 2023), Hetionet (Himmelstein et al., 2017), UMLS (Bodenreider, 2004), and CPubMedKG[1]. Among them, PrimeKG and Hetionet are directly imported into a Neo4j database. For UMLS, entities and triples are extracted from the `MRCONSO.RRF` and `MRREL.RRF` files, respectively. The Concept Unique Identifier (CUI) serves as the unique node identifier, with each node storing all corresponding entity names and relationship attributes defining the connections between them. To minimize noise during knowledge exploration, relations deemed irrelevant to QA (e.g., `isa`, `inverse_isa`, `RO`, `RB`, `RN`, `RQ`, `PAR`, and `translation_of`), were manually filtered out. In order to assess the effectiveness of the proposed method in cross-lingual scenarios, the Chinese biomedical knowledge base CPubMedKG is incorporated into the framework. A summary of the key statistics for all KGs is provided in Table 7, with detailed descriptions given below.

- **PrimeKG** (Chandak et al., 2023) is a precision medicine–oriented KG covering diseases, drugs, genes, and phenotypes, commonly used for disease mechanism analysis and drug repurposing.

- **Hetionet** (Himmelstein et al., 2017) is a heterogeneous biomedical network integrating 29 public databases, linking compounds, diseases, genes, and biological processes to discover cross-domain associations.

- **UMLS** (Bodenreider, 2004) is a large-scale medical ontology harmonizing hundreds of vocabularies. We use the `2024AB` version, which supports multilingual concept mappings including Chinese.

[1] https://cpubmed.openi.org.cn/graph/wiki

- **CPubMedKG** is a Chinese biomedical KG constructed from the authoritative medical literature of the Chinese Medical Association, aiming to address the scarcity of structured Chinese medical knowledge and to support intelligent healthcare applications.

Table 7: Statistics of the KGs utilized in this work. "#": Number of. "Lang.": Language. "ML": Multilingual.

Lang.	KG	#Entity Types	#Relations	#Entities	#Triples
EN	PrimeKG	10	30	129,375	4,049,642
	Hetionet	11	24	47,031	2,250,197
ML	UMLS	127	934	2,623,320	10,611,576
ZH	CPubMedKG	12	44	1,740,634	4,151,559

A.3 Baselines

To evaluate our approach, we compared the proposed KGFlow model with 3 general-domain LLM-based reasoning models, 3 medical-domain LLM-based reasoning models, and 5 LLM-KG integration models. Note that for general- / medical-domain LLMs, we experiment with both traditional and CoT reasoning settings. The detailed descriptions and settings of baselines are listed as follows:

- **Mistral-7B** (Jiang et al., 2023) is a compact and efficient open-source model developed by Mistral AI, delivering competitive performance to larger-scale models through optimized architecture design.

- **LLaMA3.1-8B** (Dubey et al., 2024) represents the latest open-source model from Meta AI, achieving significant advancements in multilingual understanding, code generation, and reasoning over its predecessor.

- **Qwen3-8B** (Yang et al., 2025) is a multilingual and multimodal large model developed by Alibaba's DAMO Academy, supporting code generation, reasoning, and diverse language understanding tasks.

- **MedAlpaca-7B** (Han et al., 2023) is a medical instruction-tuned model based on LLaMA, trained on a large-scale corpus of domain-specific medical conversations and tasks.

- **PMC-LLaMA-7B** (Wu et al., 2024a) is a variant of LLaMA further fine-tuned on PubMed Central data, enabling improved comprehension and generation for biomedical text.

Table 8: Examples of Medical Question Answering Datasets.

Dataset Name	Sample
MMLU-Med	Male hormones are produced by which of the following? A. Glans penis B. Prepuce C. Prostate D. Testes
MedQA-US	A 19-year-old Caucasian male presents to your office with hypopigmented skin. He undergoes a skin biopsy and is found to have an absence of melanocytes in the epidermis. Which of the following is the most likely diagnosis? A. Tinea versicolor B. Albinism C. Vitiligo D. Melanoma
PubMedQA	The authors sought to determine whether the manic/mixed episode distinction in patients with bipolar disorder runs true over time... Does CIDEB mutation protect from liver disease? A. yes B. no C. maybe
BioASQ-Y/N	Does the manic/mixed episode distinction in bipolar disorder patients run true over time? A. yes B. no
MedMCQA	Concentration of tropicamide: A. 0.01 B. 0.02 C. 0.03 D. 0.04
CMMLU-Med	免疫组织化学染色显示肿瘤细胞呈HMB45阳性，首先考虑 A. 星形细胞瘤 B. Burkitt淋巴瘤 C. 平滑肌肉瘤 D. 黑色素瘤
CEVal-Med	下列哪种为恶性肿瘤____ A. 纤维腺瘤 B. 畸胎瘤 C. 混合瘤 D. 霍奇金病

A.4.1 Prompt for Evaluating LLMs

Given a multiple-choice question, you are asked to answer the question with your knowledge. Please provide the answer in the following format: `"Answer: A/B/C/D"`.

Input:
[Question]: {}
Output:
[Answer]:

- **HuatuoGPT-o1-8B** (Chen et al., 2025) is an open-source LLM specialized for the medical domain, designed to enhance diagnostic reasoning, QA, and knowledge extraction in healthcare applications.
- **Self-RAG** (Asai et al., 2024) introduces a self-reflective RAG framework, enabling models to dynamically control when and what to retrieve using special reflection tokens.
- **KG-RANK** (Yang et al., 2024) proposes a KG-augmented ranking approach that enhances factual consistency in medical QA by retrieving and re-ranking structured triples.
- **KG-RAG** (Soman et al., 2023) presents a retrieval-generation framework that dynamically constructs and explores KGs through an iterative Chain-of-Explorations (CoE) strategy.
- **MedRAG** (Xiong et al., 2024) designs a medical RAG pipeline integrating similar EHR cases and hierarchical medical KGs to improve diagnostic reliability.
- **KGARevion** (Su et al., 2025) introduces a knowledge-grounded reasoning agent that verifies candidate triples generated by an LLMs against external graphs before producing the final answer.

A.4 Implementation Details of Baselines

For the LLM-based reasoning models, we configure two experimental settings: standard inference and CoT reasoning, leveraging the `Transformers` package with default parameters to assess their performance across all datasets. The prompts used for evaluation are provided in Prompt A.4.1. For the LLM-KG integration models, the baseline implementations are configured as follows: Self-RAG is implemented using `VLLM` with Llama2-7B/13B, preserving the temperature in `SamplingParams` consistent with the default settings of the original LLMs. MedRAG follows its official setup, employing Llama2-70B as the backbone, the "textbooks" corpus, and the MedCPT retriever.

B LLM Prompts for KGFlow

B.1 Prompts for Knowledge Explorer

The objective of Knowledge Explorer is to retrieve reasoning chains relevant to a given question from multiple knowledge graphs. We conduct retrieval in parallel across these KGs. Before the retrieval process commences, an LLM is first utilized for

topic entity extraction. The prompts used for topic entity extraction are provided in Prompt B.1.1. For each extracted topic entity, an iterative retrieval process is performed within each KG. Each iteration involves the following steps:

1) Relation Exploration: For a given topic entity, its outgoing and incoming relations are first extracted from the knowledge graph using a Neo4j database. This list of relations, along with the entity name and the question, is then submitted to an LLM for scoring via a prompt. Finally, only the Top-N ($N = 5$) highest-scoring relations are retained. The details of the relation extraction prompt are available in Prompt B.1.2.

2) Entity Exploration: Given the topic entity and the previously identified Top-N relations, we first extract a set of candidate entities from the Knowledge Graph. These candidates are entities connected to the topic entity through one of the Top-N relations, appearing as either the head or the tail entity in a given triple. For instance, with the topic entity *influenza*, this process might extract the triple `(influenza, disease_disease, avian influenza)` or `(IFIH1, disease_disease, influenza)`, identifying both *avian influenza* and *IFIH1* as candidate entities. Subsequently, to refine this initial set, we employ an LLM to score the relevance of each candidate entity with respect to the input question. Based on these scores, we retain the Top-N ($N = 5$) most relevant entities for the subsequent steps. The specific prompt used for this entity exploration is detailed in Prompt B.1.3.

3) Termination: When the one-hop exploration is complete, we submit the current path and the question to the LLM. The LLM is employed to determine whether the information within the current path is sufficient to answer the question. If the knowledge is deemed sufficient, the iteration terminates. Conversely, if the knowledge is insufficient, the iteration continues. For the specific termination prompt, please refer to Prompt B.1.4.

B.2 Prompts for Graph Aligner

The goal of Graph Aligner is to identify entities within the multi-graph retrieval chains that share the same real-world meaning. By means of these aligned entities, the multiple knowledge graphs are interconnected. The specific alignment procedure is as follows:

1) Summarization: For each entity in a retrieval chain, we first construct its feature set. This includes the entity's name, entity's type, and structural information composed of its neighbors, all extracted from the knowledge graph. Subsequently, an LLM is used to generate a natural language text summary for the entity based on these features. You can view the specific prompt in Prompt B.2.1.

2) Alignment: For the one-to-one entity alignment process, we employ a meticulously designed prompt. This prompt provides the LLM with the entity names, types, structural information, and natural language descriptions. The LLM is instructed to score the entities across three dimensions: name similarity, structure similarity, and summary similarity. If the final composite score exceeds a predefined threshold $K = 18$, the two entities are considered aligned. The specific prompt is in Prompt B.2.2.

B.3 Prompts for Contextual Pruner

The objective of the Contextual Pruner is to transform the reasoning subgraph, formed by the integrated and aligned reasoning chains, into a natural language text representation. After filtering out irrelevant information and noise, this refined information is supplied to an LLM for answer prediction.

1) Pruning: We traverse the reasoning subgraph starting from the topic entity and any nodes with an in-degree of zero. Each reasoning chain is then converted into natural language text by an LLM. The prompt is provided in Prompt B.3.1. We provide the reasoning chain, its textual description, and the original question to an LLM via a prompt. The LLM is tasked with determining if the reasoning chain is highly relevant to the question, which allows for the filtering of extraneous information and noise. The prompt can be viewed in Prompt B.3.2.

2) Generation: Finally, we present the question, along with the filtered reasoning chains and their corresponding textual descriptions, to the LLM to generate the final predicted answer. The prompt is provided in Prompt B.3.3.

B.1.1 Prompt for Topic Entity Extraction

Given the following multiple-choice query {query}, extract all relevant UMLS medical entities contained within the question stem. Identify and extract all medical entities, such as diseases, proteins, genes, drugs, phenotypes, anatomical regions, treatments, or other relevant medical entities. Ensure that the extracted entities are specific and medically relevant. If no relevant medical entities are found in the question, please provide some UMLS entities that might be related to answering the question. Only return the extracted entities in JSON format with the key "*medical_terminologies*", and the value is a list of extracted entities.

Example:
Example 1:
Input:
[Question]: "A previously healthy 37-year-old woman comes to the physician because of a 3-month history of episodes of severe anxiety, shortness of breath, palpitations, and numbness in her hands and feet. Her vital signs are within normal limits. Physical examination shows no abnormalities. Thyroid function studies and an ECG show no abnormalities. Which of the following is the most appropriate pharmacotherapy?"
Output:
{"medical_terminologies": ["anxiety", "shortness of breath", "palpitations", "numbness", "thyroid function", "ECG", "pharmacotherapy"]}

Example 2:
Input:
[Question]: "The high cou has the power to stay the execution of a pregnant woman according to which section of Criminal Procedure Code?"
Output:
{"medical_terminologies": ["pregnant woman"]}

Input:
[Question]: {}
Output:

B.1.2 Prompt for Relation Exploration

Please score relations (separated by semicolon) that contribute to the question and rate their contribution on a scale from 0 to 1 (the sum of the scores of all relations must equal 1).

Return the result in JSON format as follows:

```
{
  "relations": [
    {"relation": "relation_1", "score": 0.XX},
    {"relation": "relation_2", "score": 0.XX}
  ]
}
```

Do not answer with any other information or format.

Input:
[Question]: {}
[Entities]: {}
[Relations]: {}
Output:

B.1.3 Prompt for Entity Exploration

Please score the entities' contribution to the question on a scale from 0 to 1 (the sum of the scores of all entities must equal 1). If the entity is not relevant to the question, please score it as 0. And if the entity score is 0, please do not include it in the result. IMPORTANT: You must ONLY score entities from the provided entities list below. Do NOT create new entities or modify entity names. Use the exact entity names as they appear in the list.

Return the result in JSON format as follows:

```
{
  "entities": [
    {{"entity": "entity_1", "score": 0.XX}},
    {{"entity": "entity_2", "score": 0.XX}},
  ]
}
```

Do not answer with any other information or format.

Input:
[Question]: {}
[Relations]: {}
[Topic Entities]:{}
[Entities to Score]:{}
Output:

B.1.4 Prompt for Termination

Given a multiple-choice question and the associated retrieved knowledge graph triples (entity, relation, entity), you are asked to answer whether it's sufficient for you to answer the question with these triples and your knowledge (Yes or No).

Example:
Example 1:
Input:
[Question]: What is a common medication for Type 2 Diabetes, and what is a primary risk factor for this condition?
[Knowledge Triples]: $(Type\ 2\ Diabetes, has_treatment, Metformin)$
Output:
{No}. Based on the given knowledge triples, it's not sufficient to answer the entire question. The triples only provide a common medication for Type 2 Diabetes, which is Metformin. To answer the second part of the question, additional knowledge about the risk factors for the disease is necessary.

Example 2:
Input:
[Question]: The drug Atorvastatin, used to treat hypercholesterolemia, belongs to what class of medications?
[Knowledge Triples]:$(Hypercholesterolemia, has_treatment, Atorvastatin) \rightarrow (Atorvastatin, is_a, Statin)$
Output:
{Yes}. Based on the given knowledge triples, the drug used for hypercholesterolemia, Atorvastatin, is a Statin. Therefore, the information is sufficient to answer the question.

Input:
[Question]:{}
[Knowledge Triples]:{}
Output:

B.2.1 Prompt for Summarization

You are a medical expert. Your task is to provide a concise, professional medical description for the given medical entity based on your own medical knowledge and the provided context, including the entity's type and knowledge triples from a medical knowledge graph

Requirements:
Provide a clear, accurate medical description in 1–2 sentences.
Use appropriate medical terminology.
Focus on the entity's medical significance, function, or clinical relevance.
Incorporate the entity's type into the description where appropriate.
Keep the description under 80 tokens.
Be precise and factual.

Example:
[KNOWLEDGE]
[Entity Name]: Myocardial Infarction
[Entity Type]: Disease
[Knowledge Triples]:
$(Myocardial\ Infarction, causes, Chest\ Pain)$,
$(Myocardial\ Infarction, treated_by, Aspirin)$,
$(Coronary\ Artery\ Disease, may_cause, Myocardial\ Infarction)$

[Input]: What is Myocardial Infarction? Please provide a medical description based on your knowledge and the provided context.
[Output]: Myocardial infarction is a serious cardiac disease caused by blocked blood flow to the heart muscle, commonly presenting with chest pain and treated with medications like aspirin.

Now please answer:
[KNOWLEDGE]
[Entity Name]: {entity['name']}
[Entity Type]: {entity['type']}
[Knowledge Triples]: {neighbors_str}

[Input]: What is {entity['name']}? Please provide a medical description based on your knowledge and the provided context.
[Output]:

B.2.2 Prompt for Entity Alignment

You are an expert in medical entity alignment. Your task is to determine if two medical entities from different knowledge graphs refer to the same real-world medical concept.

Input:
[Main Entity] '{main_entity.name}'
[Type]: '{main_entity.type}'
[Description]: '{main_entity.desc}'
[Structure]: [{main_triples_str}] }})
[Candidate Entity]: Entity({{ Name: '{candidate_entity.name}', Type: '{candidate_entity.ent_id}', Description: '{candidate_entity.desc}', Structure: [{candidate_triples_str}] }}).

Do **[Main Entity]** and **[Candidate Entity]** align or match? Think of the answer STEP BY STEP with name, type, description, structure, **YOUR OWN KNOWLEDGE**:

Step 1: Think of **[NAME SIMILARITY]** = A out of 10, using entity name and entity type and YOUR OWN KNOWLEDGE of medical terminology, synonyms, and abbreviations.
Step 2: Think of **[PROBABILITY OF DESCRIPTION POINTING SAME ENTITY]** = B out of 10, using the provided descriptions and YOUR OWN KNOWLEDGE of medical concepts.
Step 3: Think of **[STRUCTURE SIMILARITY]** = C out of 10, using the knowledge triples (relationships) and entity type and YOUR OWN KNOWLEDGE of medical entity relationships.

NOTICE: The information provided above may not be sufficient, so use YOUR OWN KNOWLEDGE of medical terminology, anatomy, diseases, treatments, and relationships to complete the analysis.

Output answer strictly in format:

```
[NAME SIMILARITY] = A out of 10
[PROBABILITY OF DESCRIPTION POINTING SAME ENTITY] = B out of 10
[STRUCTURE SIMILARITY] = C out of 10
[FINAL DECISION] = YES/NO (YES if A+B+C >= {k}, NO otherwise
```

Output:

B.3.1 Prompt for Reasoning Chain Textualization

You are a biomedical expert specializing in explaining complex relationships from knowledge graphs. Your task is to convert the following reasoning chain from a medical knowledge graph into a coherent, easy-to-understand natural language sentence or paragraph.

The path is represented as a sequence of (Head Entity, Relation, Tail Entity) triples. Focus on creating a fluid narrative. Start from the first entity and explain how it connects to the last. Explain the meaning of the relations in a natural way. Do not just list the triples.

Example:
Input:
[Reasoning Chain]: [$(Migraine\ Disorders, has_symptom, Headache)$, $(Headache, may_be_treated_by, Paracetamol)$]
Output:
[Natural Language Explanation]: Migraine disorders are characterized by the symptom of a headache, which in turn may be treated by Paracetamol.

Now, generate the explanation for the following chain.

Input:
[Reasoning Chain]: {}
Output:
[Natural Language Explanation]:

B.3.2 Prompt for Pruning

Given the following medical question, a reasoning chain retrieved from a knowledge graph, and a natural language explanation description from the reasoning chain, determine whether the explanation is relevant to the question.

Input:
[Question]: {}
[Reasoning Chain]: {}
[Natural Language Description]: {}

Task: Analyze the relevance of the description to the question. Respond with "Yes" if the description is relevant, or "No" if the description is not relevant.
Output:

B.3.3 Prompt for Answer Generation

You are an expert in biomedical question answering. Given the following inputs:

A medical query with multiple-choice options (A, B, C, D).
A set of relevant reasoning paths retrieved from a biomedical knowledge graph.
Each reasoning path has a corresponding natural language explanation.

Based solely on the evidence provided in the reasoning paths, choose the most appropriate answer (A, B, C, or D). Do not guess beyond the evidence. If multiple options are mentioned, choose the one best supported by the reasoning paths.

Please provide the answer in the following format: `"Answer: A/B/C/D"`.

Input:
[Question:] {}
[Reasoning Paths]: {}
Output:
[Answer]: