medIKAL: Integrating Knowledge Graphs as Assistants of LLMs for Enhanced Clinical Diagnosis on EMRs

Mingyi Jia[1], Junwen Duan[1]*, Yan Song[2], Jianxin Wang[1],
[1]Hunan Provincial Key Lab on Bioinformatics,
School of Computer Science and Engineering, Central South University
[2]University of Science and Technology of China,
{jiamingyi, jwduan}@csu.edu.cn, clksong@gmial.com, jxwang@mail.csu.edu.cn

Abstract

Electronic Medical Records (EMRs), while integral to modern healthcare, present challenges for clinical reasoning and diagnosis due to their complexity and information redundancy. To address this, we proposed medIKAL (**I**ntegrating **K**nowledge Graphs as **A**ssistants of **LLM**s), a framework that combines Large Language Models (LLMs) with knowledge graphs (KGs) to enhance diagnostic capabilities. medIKAL assigns weighted importance to entities in medical records based on their type, enabling precise localization of candidate diseases within KGs. It innovatively employs a residual network-like approach, allowing initial diagnosis by LLMs to be merged into KG search results. Through a path-based reranking algorithm and a fill-in-the-blank style prompt template, it further refined the diagnostic process. We validated medIKAL's effectiveness through extensive experiments on a newly introduced open-sourced Chinese EMR dataset, demonstrating its potential to improve clinical diagnosis in real-world settings. The code and dataset are publicly available at https://github.com/CSU-NLP-Group/mediKAL.

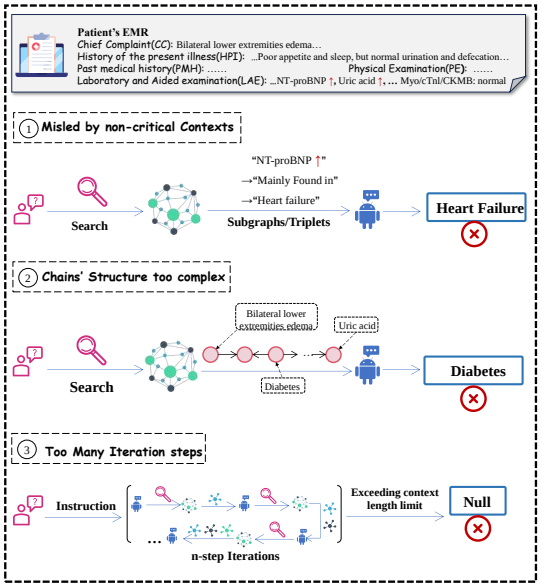

Figure 1: Limitations of existing methods using KG-augmented LLMs for application to EMR diagnostic tasks. ① use subgraphs/triplets to augment context.② use reasoning chains to augment context. ③ use the iteration-based approach to involve LLMs in KG searching and reasoning.

1 Introduction

Electronic Medical Records (EMRs) are the digitized record of a patient's medical and health information and play an important role in the modern healthcare system. However, due to their complexity and information redundancy, clinical diagnosis based on EMRs extremely requires specialized medical knowledge and clinical experience. This demand has led to the development of automated methods to assist and support clinical diagnosis and decision-making.

Recently, large language models (LLMs) have demonstrated great potential in various medical domains (Lee et al., 2023; Lee, 2023; Ayers et al., 2023; Nayak et al., 2023). But directly applying LLMs to the medical field still has raised concerns about the generation of erroneous knowledge and hallucinations because of their lack of specific medical knowledge (Bernstein et al., 2023). Training LLMs in the medical domain requires a lot of high-quality data, and the best-performing LLMs available are often closed-source, making further training difficult (Achiam et al., 2023). Furthermore, considering that knowledge in the medical field is constantly being updated and iterated, for already trained LLMs, updating their parameters can only be done through retraining, which is extremely time-consuming and expensive (Baek et al., 2023b).

As a classic form of large-scale structured knowl-

* Corresponding Author. Email: jwduan@csu.edu.cn.

Proceedings of the 31st International Conference on Computational Linguistics, pages 9278–9298
January 19–24, 2025.

edge base, knowledge graphs (KGs) can provide explicit knowledge representation and interpretable reasoning paths and can be continually modified for correction or update. Therefore, KGs become an ideal complement to LLMs (Pan et al., 2024a; Yan et al., 2024a). However, existing works on "LLM⊕KG" cannot be directly applied to EMR diagnosis tasks, mainly due to the following reasons: (1) Existing approaches rely on entity recognition in the input text to locate corresponding information in KGs, but they do not differentiate the contributions of different types of entities during searching on KGs. (2) They typically treat triplets or subgraphs obtained from KGs as direct context inputs or simply convert them into natural language, which can easily lead to the problem of exceeding the input length limit and hard to understand for LLMs when encountering complex structures and informative contexts. (3) It was found that when adopting RAG paradigm, LLMs tend to overly rely on the provided context and fail to fully utilize their internal knowledge, making it easy to be misled by incorrect knowledge (Baek et al., 2023a).

In this paper, we propose an effective framework called medIKAL (**I**ntegrating **K**nowledge Graphs as **A**ssistants of **LLM**s). Specifically, unlike other conventional approaches, we assign different weights to entities in EMRs based on their type, which enables us to more precisely localize possible candidate diseases in KGs. Meanwhile, in order to prevent the results from relying too much on KGs, we drew inspiration from the idea of residual networks to allow LLMs to first diagnose without relying on external knowledge, and then merge the diagnosis results with the search results of the knowledge graph. Subsequently, we propose a path-based rerank algorithm to rank candidate diseases. Finally, we designed a special fill-in-the-blank style prompt template to help LLMs to better inference and error correction.

In summary, our contributions can be abbreviated as: (1) We raised the problem of a shortage of high-quality open-source Chinese electronic medical record data and we introduced an open-sourced Chinese EMR dataset. (2) We proposed an effective method that allows LLMs to handle information-dense and highly redundant EMRs to make correct diagnoses. (3) We conducted extensive experiments on our collected EMR dataset to demonstrate the effectiveness of our medIKAL framework.

2 Related Work

2.1 Clinical Diagnosis and Prediction on EMRs

Electronic medical records (EMRs) provide detailed medical information about patients, including symptoms, medical history, test results, and treatment records, and are widely used in patient care, clinical diagnosis, and treatment (Xu et al., 2024). Prior research has extensively focused on designing deep learning models for EMR data, addressing downstream tasks such as disease diagnosis and risk assessment (Gao et al., 2020; Xu et al., 2022; Wang et al., 2023b).

LLMs have demonstrated impressive performance in various medical tasks, including disease diagnosis and prediction in EMRs. Researchers have explored multiple approaches: (Jiang et al., 2023a) used LLMs and biomedical knowledge graphs to construct patient-specific knowledge graphs, processed with a Bidirectional Attention-enhanced Graph Neural Network (BAT GNN); RAM-EHR (Xu et al., 2024) transformed multiple knowledge sources into text format, utilizing retrieval-enhanced and consistency-regularized co-training; DR.KNOWS (Gao et al., 2023) combined a knowledge graph built with the Unified Medical Language System (UMLS) and a clinical diagnostic reasoning-based graph model for improved diagnosis accuracy and interpretability; REALM (Zhu et al., 2024) integrated clinical notes and multivariate time-series data using LLMs and RAG technology, with an adaptive multimodal fusion network. Most studies focus on English EMR datasets like MIMIC-III (Johnson et al., 2016), which primarily contains ICU data and may not suffice for modeling mild cases, rehabilitation, or routine treatments. Research on Chinese EMR datasets remains limited.

2.2 Knowledge Graphs Augmented LLMs

Knowledge graphs have advantages in dynamic, explicit, structured knowledge representation and storage, and easy addition, deletion, modification, and querying (Pan et al., 2024b), which has led to increasing interest among researchers in exploring the integration of knowledge graphs with large language models. One typical paradigm is to incorporate knowledge graph triplets into the training data during the training phase and obtain their embedding representations through graph neural network modules (Zhang et al., 2019; Sun et al.,

2021; Li et al., 2023; Huang et al., 2024). However, LLMs often have a large-scale requirement for pre-training corpora, making it difficult and costly to find or create knowledge graphs of a matching scale (Wen et al., 2023). More importantly, combining knowledge graphs with LLMs through embedding can result in the loss of their original advantages, such as interpretability of reasoning and efficiency of knowledge updates.

In recent studies, researchers have attempted to integrate KGs with LLMs through prompts (Wen et al., 2023; Wu et al., 2024; Yang et al., 2024; Wang et al., 2023a). They typically identify entities in the input text and locate the corresponding triplets or subgraphs in the KG, which are then transformed into natural language (Wen et al., 2023), entity sets (Wu et al., 2024), or reorganized triplets (Yang et al., 2024), etc., and concatenated with the input prompts to provide additional knowledge to LLMs. Another approach is to use an iterative strategy where LLMs act as agents to explore and reason step-by-step on the KG until it obtains sufficient knowledge or reaches the maximum number of iterations (Sun et al., 2023; Jin et al., 2024). However, this approach is more suitable for shorter questions. In scenarios with longer contexts, larger knowledge graph scales, and more complex structures, it can result in excessive interactions with LLMs and the inability to find the correct paths in the knowledge graph.

3 Method

3.1 EMR Summarisation and Direct Diagnosis via LLMs

Considering that the EMRs contain a large amount of redundant information, direct use is easy to cause interference in the diagnostic process. So we first designed a series of questions to prompt the LLM to summarize the key information in the EMR, such as patient symptoms, medical history, medication usage, medical visits, etc. Detailed prompt templates are shown in Table 11 and 12 in Appendix F. This process can be represented as:

$$\mathcal{M} = \mathrm{LLM}([\mathrm{Prompt}_{\mathrm{sum}}, \mathcal{M}_{orig}]) \quad (1)$$

where $\mathcal{M}_{orig}$ represents the original input medical record, $\mathcal{M}$ represents the medical record after decomposition and summarization, and $\mathrm{Prompt}_{\mathrm{sum}}$ is the textual prompt.

Based on the decomposed and summarized medical record, we allow the LLM to rely on its internal knowledge for preliminary diagnosis and obtain a set of potential diseases $\mathcal{D}_{\mathrm{LLM}}$. This process can be represented as:

$$\mathcal{D}_{\mathrm{LLM}} = \mathrm{LLM}([\mathrm{Prompt}_{\mathrm{diag}}, \mathcal{M}]) \quad (2)$$

where $\mathrm{Prompt}_{\mathrm{diag}}$ denotes the textual instruction used to guide the LLM in performing preliminary diagnosis and providing predicted diseases (see Table 13 in Appendix F).

3.2 Candidate Disease Localization and Reranking via KG

3.2.1 Entity Recognition and Matching

Before the knowledge graph search process, we perform entity recognition on the summarized EMR $\mathcal{M}$ using a pre-trained NER model. This process can be represented as:

$$\mathcal{E}_{\mathcal{M}} = e_1, e_2, \ldots, e_{|E|} = \mathrm{NER}(\mathcal{M}) \quad (3)$$

Where the entity set extracted from the EMR is denoted as $\mathcal{E}_{\mathcal{M}}$, and NER denotes the pre-trained NER model.

Then for every $e_i \in \mathcal{E}_{\mathcal{M}}$, we link it to the corresponding node in the knowledge graph $\mathcal{G}$ using dense retrieval methods. Specifically, given an entity $e_i \in \mathcal{E}_{\mathcal{M}}$, we use an encoding model to get the embedding of e_i, and calculate the similarity score between e_i and each entity node u_j in $\mathcal{G}$'s entity node set $\mathcal{E}_{\mathcal{G}}$, and the entity node with the highest similarity score is considered as a match. This process can be formulated as follows:

$$\hat{u}_i = \arg\max_{u_j \in \mathcal{E}_{\mathcal{G}}} \mathrm{sim}(\mathrm{enc}(e_i), \mathrm{enc}(u_j)), \quad (4)$$

Where enc denotes the encoding model, and $\hat{u}_i$ denotes the matched entity node. Finally, the set of matched entities is denoted as $\mathcal{E}_{\mathcal{Q}}$.

3.2.2 Candidate Disease Localization Based on Entity-Type Weights

Most of the previous work using KG to augment LLMs has not made a strict distinction between entity types when using entities for the knowledge graph search process. However, in the EMR, different types of entities are supposed to contribute differently to the diagnosis of a disease. For example, the association between a patient's current symptoms and the disease is more direct and closer.

So in this paper, we propose an entity type-driven method for candidate disease localization and filtering. For every entity $e_i \in \mathcal{E}_{\mathcal{Q}}$, we assign a

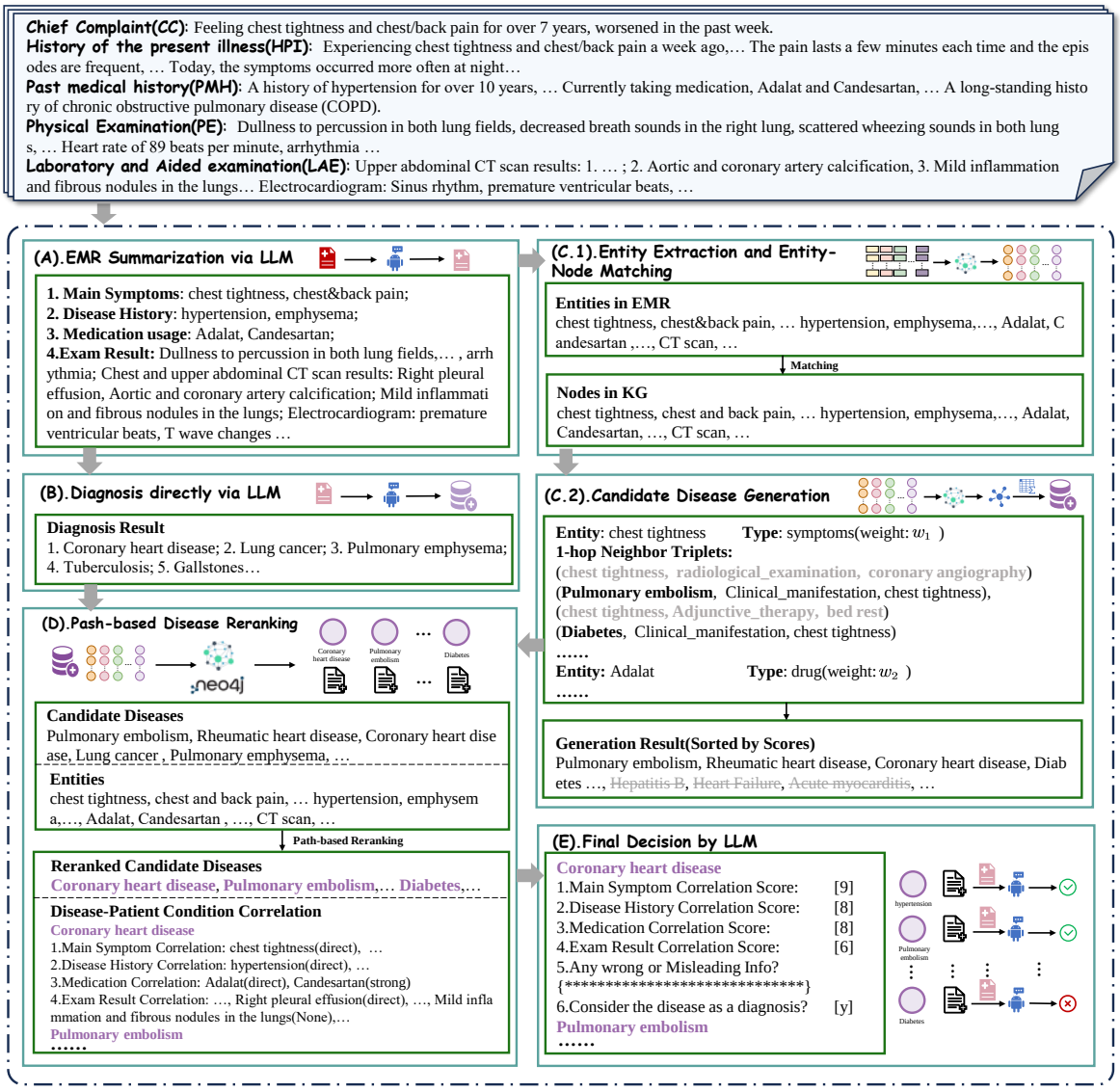

Figure 2: The overall workflow of medIKAL. It contains three main modules, namely: **Module 1.** preprocess before KG search (A, B, and C.1); **Module 2.** Candidate Disease Localization and Reranking via KG (C.2 and D); **Module 3.** Collaborative Reasoning for LLMs and KG (E).

contribution weight w_{t_i} according to its entity type t_i. Then we search for disease nodes in the 1-hop neighbors of e_i in $\mathcal{G}$ and obtain the set of disease nodes $\mathcal{D}_i$, where the score of each disease in $\mathcal{D}_i$ will be increased by w_{t_i}. The algorithm description of the above process can be found in Algorithm 1 in Appendix B. After getting the potential disease set $\mathcal{D}_\mathcal{G}$ generated by the KG search process, we merge $\mathcal{D}_\mathcal{G}$ with the potential disease set $\mathcal{D}_{\text{LLM}}$ obtained through LLMs in Section 3.1, resulting in a candidate disease set $\mathcal{D}_{can} = \mathcal{D}_{\text{LLM}} \cup \mathcal{D}_\mathcal{G}$. Here we have drawn inspiration from the idea of residual networks (He et al., 2016). We hope to make more use of the LLM's internal knowledge in this way, rather than relying solely on the knowledge graph for searching for correct diagnoses.

3.2.3 Candidate Disease Reranking Based on Paths.

In actual clinical diagnosis, doctors usually make a diagnosis based on a series of information such as the patient's symptoms, medical history, examination results, etc. Therefore, a correct diagnosis should be correlated with most of the patient information. In order to model this correlation, we propose a path-based reranking algorithm. Specifically, we define $\text{dist}(\mathcal{D}_i, e_j)$ to denote the shortest path distance between disease $\mathcal{D}_i$ and entity $e_j \in \mathcal{E}_\mathcal{Q}$ on $\mathcal{G}$. Diseases with closer total dis-

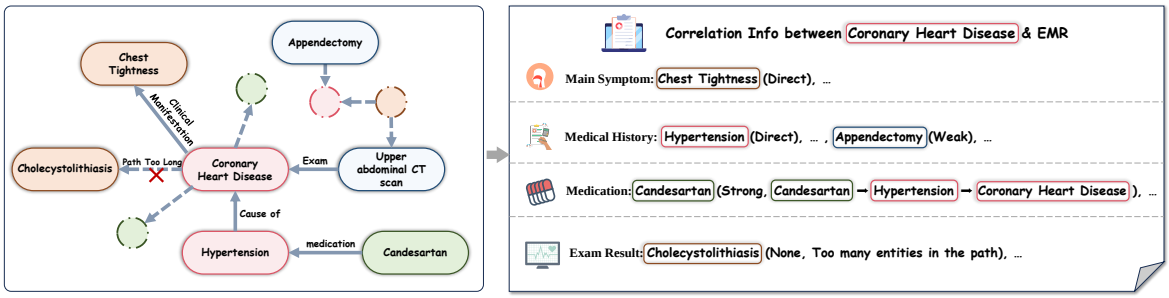

Figure 3: An illustration of how to combine reranking process with the knowledge construction process.

tances to the entity set $\mathcal{E}_Q$ are considered to have a stronger association with the patient's information, making them more likely to be the correct diagnostic results. The specific process of path-based reranking can be found in Algorithm 2.

3.3 Collaborative Reasoning between LLMs and KG Knowledge

After completing the search and reranking process based on the knowledge graph, we reconstructed the search results to provide additional contextual information for LLMs for collaborative reasoning.

3.3.1 Reconstruction of KG Knowledge

EMRs are different from conventional medical QA tasks. Even though we have previously summarized them, they are still information-dense and complex-context structures, so the retrieved KG knowledge will also become extensive. If we still follow previous work and directly input triplets or knowledge chain paths as context knowledge, it would lead to overly chaotic structures that LLMs can hardly understand, which increases the possibilities of hallucination. Therefore, in this paper, we propose a way to reconstruct knowledge graph information. For each candidate disease $\mathcal{D}_i \in \mathcal{D}_{rerank}$, we classify and organize the information related to $\mathcal{D}_i$ according to several aspects like the correlations between $\mathcal{D}_i$ and the patient's main symptoms, or between $\mathcal{D}_i$ and the patient's medical history, etc. An example illustration is shown in Figure 3.

In this way, we transform the information of paths and entities retrieved from the knowledge graph into a semi-structured representation of knowledge, which maximizes the manifestation of the association between each candidate disease and the content of the medical record, enabling LLMs to make more intuitive judgments and analyses. Moreover, since the association between the majority of entities and diseases has already been established during the processing of Section 3.2.2 and Section 3.2.3, the knowledge reconstruction process does not require re-searching $\mathcal{G}$, avoiding additional time consumption.

3.3.2 Clinical Reasoning and Diagnosis Based on Fill-in-the-Blank Prompt Templates

Based on the reconstructed knowledge described above, we designed a special prompt template in a fill-in-the-blank style to make the reasoning paths of LLMs more rational. We guide LLMs to quantitatively evaluate the degree of correlation between a specific disease $\mathcal{D}_i$ and the aspects mentioned above, giving a score ranging from 0 to 10 (the higher the score, the higher the degree of correlation) for each aspect, and then calculate a total score. If the total score is higher than a pre-defined threshold θ, we consider the current candidate disease $\mathcal{D}_i$ as one of the final diagnostic results. Additionally, to ensure the self-consistency of LLMs, we also check the consistency between this total score and the prediction made by LLMs. If they are inconsistent, we will check the original prediction $\mathcal{D}_{\text{LLM}}$ to decide whether to drop $\mathcal{D}_i$. The specific prompt template can be found in Table 14 in Appendix F.

4 Experiments

4.1 Experimental Setup

4.1.1 Datasets

CMEMR Dataset Construction: Considering the current lack of high-quality and widely covered EMR datasets in the Chinese community, we construct a dataset CMEMR (Chinese Multi-department Electronic Medical Records) collected from a Chinese medical website[1]. We filtered the collected electronic medical records, exclud-

[1] https://bingli.iiyi.com/

ing those with existing problems or missing key information. The details of the dataset can be seen in Table 5 in Appendix A. In order to ensure the correctness and usability of the collected medical records, we randomly sampled a batch of medical records in each department and consulted the corresponding department experts, mainly focusing on the correctness of the diagnosis results (i.e., the labels of our task). A complete data example is provided in Figue 5 and 6.

In addition, to further validate our proposed method, we selected the following three datasets as supplements: (1) **CMB-Clin** (Wang et al., 2023c): The CMB-Clin dataset contains 74 high-quality, complex, real EMRs, each of which will contain several medical QA pairs. To be consistent with our approach, we simplify the task of this dataset to a pure disease diagnosis task. (2) **GMD** (Liu et al., 2022): The GMD dataset was constructed based on EMRs. Each sample in the dataset contains a target disease along with its explicit and implicit symptom information. (3) **CMD** (Yan et al., 2023): The CMD dataset is a follow-up to the GMD dataset. Its format is the same as the GMD dataset, and also sourced from EMRs. The only difference is that CMD contains a more variety of diseases and symptoms.

4.1.2 Baselines

We compared our proposed medIKAL with three series of baseline methods: LLM-only, LLM⊕KG, and LLM⊗KG (Sun et al., 2023). Details of the baseline methods are provided in Appendix D.

LLM-only: They do not rely on external knowledge and only use the LLMs' internal knowledge for reasoning, including CoT (Wei et al., 2022), ToT (Yao et al., 2024), and Sc-CoT (Wang et al., 2022)).

LLM⊕KG: We selected four representative works, namely MindMap (Wen et al., 2023), ICP (Wu et al., 2024), HyKGE (Jiang et al., 2023b), , and KG-rank (Yang et al., 2024), all of which are aimed at medical question-answering and reasoning tasks, so we believe they are highly relevant to our work in this paper.

LLM⊗KG: This is the concept proposed by (Sun et al., 2023). It enables LLMs to participate in the search and reasoning process on KGs, check whether the current knowledge is sufficient to answer the question, and make decisions for the subsequent search process iteratively. We selected ToG (Sun et al., 2023) and Graph Chain-of-Thought (Jin et al., 2024) as baselines.

4.1.3 Evaluation metric

To enhance the scientific rigor and effectiveness of the evaluation, particularly in identifying disease diagnoses, following (Fan et al., 2024), we adopted the International Classification of Diseases (ICD-10) (Percy et al., 1990) as the authoritative source and link standardized disease terminologies with natural language based diagnostic results. Initially, we extract disease entities from the diagnostic results and the label in the EMR. Then we implement a fuzzy matching process with a predefined threshold of 0.5 to link these disease entities with ICD-10 terminology, building two normalized disease sets $S_{\hat{\mathcal{D}}}$ and $S_{\mathcal{R}}$. Finally we use these two sets to calculate the Precision, Recall and F1-score metrics. More details are shown in Appendix E.

4.1.4 Implementation Details

For the backbone model, we choose Qwen models with different parameter scales ([7B, 14B, 72B]). In all experiments, we set do_sample to false for consistent responses. For the knowledge graph, we choose the CPubMed-KG. For the NER model mentioned in section 3.2.1, we choose the RaNER (Wang et al., 2021) model released by Tongyi-Laboratory. For the Entity-node matching process in section 3.2.1, we choose the CoROM (Long et al., 2022) model as our embedding model. More implementation details are listed in Appendix C.

4.2 Experimental Results

4.2.1 Overall Performance

The main experimental results on CMEMR dataset are shown in Table 1. From the results, we can draw the following analysis:

(1) Our method significantly outperforms other baselines using **LLM⊕KG** paradigm on CMEMR dataset, which demonstrated the effectiveness of our method on EMR-diagnosis task.

(2) The methods using **LLM⊗KG** (i.e., ToG (Sun et al., 2023) and Graph-CoT (Jin et al., 2024)) perform poorly on EMR-diagnosis Tasks, since they are designed for short multi-hop QA task. The iteration steps and the complexity of beam search increase greatly as the amount of context and the size of KG increase, which makes it easily reach the upper limit of the number of iterative steps without collecting enough information, or exceeding the input length limit of LLMs.

	Methods	Qwen-7b-chat			Qwen-14b-chat			Qwen-72b-chat		
		R	P	F1	R	P	F1	R	P	F1
I	Direct	41.07	31.23	35.48	42.98	32.50	37.01	45.12	34.45	39.06
II	CoT	41.24	31.06	35.43	42.58	31.67	36.32	46.01	33.19	38.56
	ToT	39.25	31.77	35.11	43.19	32.56	37.12	45.45	34.87	39.46
	SC-CoT	41.99	31.69	36.12	42.34	32.90	37.40	45.49	34.59	39.29
III	MindMap	41.42	32.30	36.29	43.59	**33.81**	38.08	45.14	35.62	39.81
	KG-Rank	39.13	28.61	33.05	41.34	31.45	35.72	44.79	32.95	37.96
	ICP	40.13	30.67	34.76	41.58	30.23	35.00	44.00	32.38	37.30
	HyKGE	42.05	32.42	36.61	43.76	33.45	37.91	45.91	34.30	39.26
	ToG	38.78	26.94	31.79	39.09	27.31	32.15	40.39	27.81	32.93
	Graph-CoT	35.90	24.01	28.77	38.67	25.11	30.44	39.68	27.48	32.47
	Ours	**42.16**	**32.86**	**36.93**	**43.96**	33.65	**38.12**	**46.43**	**35.72**	**40.37**

Table 1: Experimental results on CMEMR dataset with different scale of backbone models. The best results are highlighted in bold.

	Methods	CMB-Clin			GMD			CMD		
		R	P	F1	R	P	F1	R	P	F1
I	Direct	40.35	26.77	32.18	42.01	21.03	28.02	50.26	25.11	33.48
II	CoT	40.66	27.23	32.62	42.44	21.30	28.36	51.02	25.49	33.99
	ToT	39.94	25.90	31.42	41.68	20.80	27.75	49.39	24.48	32.73
	SC-CoT	41.10	26.31	32.08	**42.73**	21.37	**28.49**	51.14	25.57	34.09
III	MindMap	39.26	**29.24**	33.51	41.44	21.18	28.03	49.75	25.62	33.82
	KG-Rank	41.70	27.12	32.86	38.16	19.54	25.84	47.91	23.92	31.90
	ICP	40.27	25.54	31.25	39.38	19.63	26.20	46.26	23.15	30.85
	HyKGE	41.53	28.21	**33.59**	40.33	21.36	27.92	48.67	24.35	32.45
	ToG	35.41	19.18	24.88	41.76	20.85	27.81	50.73	25.24	33.70
	Graph-CoT	36.35	20.66	26.07	38.13	19.06	25.54	49.07	24.51	32.69
	Ours	**41.89**	27.68	33.33	42.37	**21.43**	28.46	**51.26**	**25.74**	**34.27**

Table 2: Experimental results on CMB-Clin, GMD, and CMD datasets using Qwen-7B-chat. The best results are highlighted in bold.

(3) As we expected, the performance of medIKAL improves with the scale of backbone models due to the increase of models' reasoning and instruction-following ability. Considering the plug-and-play and train-free nature of our method, it can be flexibly deployed to backbone models of different sizes depending on the needs of different scenarios.

We also tested our method on three additional datasets and the experimental results are shown in Table 2. Our method performs stably on the CMB-Clin dataset, whose data format is also standard EMRs. On the GMD and CMD datasets, there is a slight degradation in the performance of our method. This is because although GMDs and CMDs are also constructed using EMRs, they contain too little patient information (only symptoms), which can easily localize to other related diseases on the knowledge graph leading to errors.

4.2.2 In-depth Analysis

How do different knowledge graph augmented prompts affect medIKAL's performance? In order to verify our proposed special prompt template's superiority, we compare it with several knowledge graph-augmented prompt templates, including entities (Wu et al., 2024), relevant triplets (Yang et al., 2024), natural language, reasoning chains (Jiang et al., 2023b), and mindmap (Wen et al., 2023). The experimental results are shown in Table 3. According to the results, using relevant entities is very ineffective as it does not utilize the relational information contained in the knowledge graph at all. For the reasoning chains and mindmap, due to the information-

Methods	R	P	F1
Relevant Entities	39.22	28.74	33.17
Natural Language	39.88	28.92	33.52
Relevant Triples	40.26	29.61	34.12
Reasoning Chains	40.97	31.16	35.39
MindMap	41.10	31.41	35.60
FBP(ours)	**42.16**	**32.86**	**36.93**

Table 3: Performances of medIKAL using different knowledge graph-augmented prompt templates on CMEMR dataset. Note that we kept all the rest parts of the medIKAL and only replaced the final "fill-in-the-blanks" prompts (FBP) with other methods to conduct this experiment.

intensive nature of EMR data, they can easily form overly large and complex-structure prompt contexts, making it difficult for LLMs (especially models with small parameters) to reason.

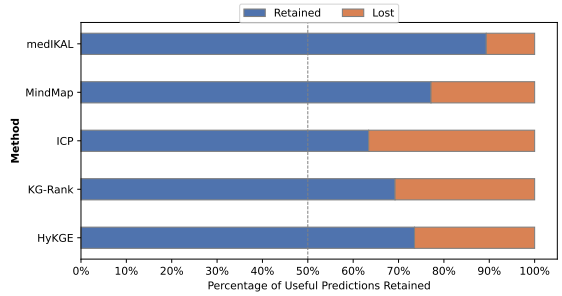

Figure 4: Evaluation results for medIKAL and other baseline methods' capabilities of utilizing LLM's internal knowledge. "Retained" denotes that the useful diagnoses from LLM's original predictions are kept as final results, and "Lost" denotes the opposite.

Does medIKAL integrate KG and LLMs better compared with other baselines? The problem with most of the existing work based on knowledge graphs is that LLMs can be overly dependent on the information obtained from KG and fail to use their own knowledge. Therefore, we counted the proportion of useful predictions in the original predictions of LLMs retained by medIKAL and other baseline methods. From the experimental results in Figure 4, medIKAL is able to minimize LLM's over-reliance on KG's knowledge and retains the majority of useful predictions compared to other baselines.

4.2.3 Ablation Study

We conduct the following ablation studies to demonstrate the importance of different modules in medIKAL.

Method	R	P	F1
medIKAL	**42.16**	**32.86**	**36.93**
w/o SUM	41.56	32.37	36.39
w/o ETW	41.19	29.88	34.63
w/o PR	41.91	32.44	36.57
w/o RI	40.16	30.32	34.55

Table 4: Ablation study results on CMEMR dataset. w/o indicates removal of the corresponding module. "SUM" denotes "summarization". "ETW" denotes "**E**ntity **T**ype **W**eight". "PR" denotes "**P**ath-based **R**eranking". "RI" denotes "**R**esnet-like **I**ntegration".

(a).w/o SUM (summarization): Remove the summarization step when pre-processing medical records and i nstead use the raw content directly.
(b).w/o ETW (Entity-Type Weight): Remove the entity-type weight when performing entity-based candidate disease searches, with all entities contributing equal weights.
(c).w/o PR (Path-based Reranking): Remove the reranking process for candidate diseases.
(d).w/o RI (Resnet-like Integration): Do not integrate the LLM's direct diagnosis result into the candidate disease.

The results in Table 4 show that both removing the "SUM" module and the "ETW" settings can seriously interfere with the performance, as the former leads to the introduction of a lot of redundant information in the original EMRs, while the latter leads to unimportant entities overly influencing the results. Removing the "RI" module would result in results that are entirely dependent on the KG search process, while the internal knowledge of LLMs is almost completely unused, thus causing a severe performance decrease.

4.2.4 Case Study

We show representative case studies in Figure 7 in the Appendix to demonstrate the effectiveness of our proposed medIKAL. From Figure 7, we can find that medIKAL can not only complement (Figure 7-(a)) and correct (Figure 7-(d)) the predictions of LLMs using KG, but also effectively guide LLMs to analyze and reason (Figure 7-(b)). Besides, the cross-validation approach through quantitative assessment and model judgment can also effectively improve the fault tolerance for LLMs' hallucination(Figure 7-(c)).

4.2.5 Error Analysis

We have analyzed the errors of our medIKAL framework on the CMEMR dataset. Through a systematic manual review, these errors are mainly categorized into three distinct classes, which are detailed in Table 10 in the Appendix.

Overlapping Disease Characteristics This type of error can be divided into two cases: (1) For distinct diseases with significant feature overlap (e.g., symptoms, affected sites, diagnostic methods, medications), LLMs may misdiagnose during preliminary diagnosis based on patient information. Similarly, feature overlap during the KG-retrieval stage often leads to diagnostic failures. (2) For a disease d and its subtypes (e.g., "anemia" and "iron-deficiency anemia"), nodes related to subtype diseases in the KG often have similar or stronger associations with the parent disease d. Consequently, after retrieval and re-ranking, the parent disease tends to rank higher.

Misunderstanding of Examination Indicators Specific examination indicators, such as numerical values or symbols, are challenging to effectively map to knowledge graph nodes (e.g., "EF 27%" may be mapped to multiple nodes like "EF value elevated," "EF value reduced," or "EF value normal"). This ambiguity makes accurate mapping difficult. Additionally, LLMs struggle with recognizing and using such indicators, often generating hallucinations when interpreting numerical ranges.

Gap Between LLM Reasoning and Physician Diagnostic Labeling This error stems from dataset structural issues. For example, some physicians include past diseases in current diagnosis records, even when irrelevant to the present complaint, but our framework does not account for this. Moreover, abbreviations used by doctors for convenience may not be accurately interpreted by LLMs, leading to the exclusion of correct candidate diseases.

5 Conclusion

In this paper, we proposed medIKAL, a framework that seamlessly integrates LLMs with knowledge graphs to enhance clinical diagnosis on EMRs, with its key innovation being the weighted importance assignment to medical entities and a resnet-like integration approach. Experimental results showed that medIKAL significantly outperforms baselines, demonstrating its potential to improve diagnostic accuracy and efficiency in real-world clinical settings. Our medIKAL has offered a promising direction for AI-assisted clinical diagnosis, paving the way for more advanced healthcare applications.

Limitations

The limitations of collected CMEMR dataset. Although we have meticulously examined, desensitized, and verified the CMEMR dataset with medical experts, occasionally, the quality of the medical records may still fall short in actual experiments. Additionally, due to the limited sources of data, our medical record dataset exhibits an uneven distribution across departments.

The limitations of proposed medIKAL framework. Although medIKAL has demonstrated its effectiveness and great potential in the healthcare field, it still has some limitations. Firstly, while it is not strictly limited to EMR format inputs, it requires a high amount of information from the input data samples. When the input data information is sparse, the improvement in model reasoning performance by medIKAL decreases, and there is also an increased risk of hallucinations. Furthermore, medIKAL is unable to fully utilize special types of medical examination indicators (e.g., numerical or symbolic types). Addressing this issue is a key problem that needs to be solved in our future work.

Ethical Consideration

In our study on the application of LLMs in clinical disease diagnosis, ethical considerations are of paramount importance. We acknowledge the potential impacts of our work and have taken measures to address these concerns. To mitigate risks such as privacy breaches and the exposure of personal information, we have thoroughly reviewed and de-identified the data. Regarding copyright concerns, we plan to split the dataset by medical departments and store each EMR sample using its specific ID. Researchers will be able to access the full EMR data conveniently via the corresponding ID and the URL of the original source website.

Acknowledgments

This work was supported in part by the National Key Research and Development Program of China (No.2021YFF1201200), the Science and Technology Major Project of Changsha (No.kh2402004). This work was carried out in part using computing resources at the High-Performance Computing Center of Central South University.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

John W Ayers, Adam Poliak, Mark Dredze, Eric C Leas, Zechariah Zhu, Jessica B Kelley, Dennis J Faix, Aaron M Goodman, Christopher A Longhurst, Michael Hogarth, et al. 2023. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA internal medicine*, 183(6):589–596.

Jinheon Baek, Soyeong Jeong, Minki Kang, Jong C Park, and Sung Hwang. 2023a. Knowledge-augmented language model verification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1720–1736.

Jinheon Baek, Soyeong Jeong, Minki Kang, Jong C Park, and Sung Ju Hwang. 2023b. Knowledge-augmented language model verification. *arXiv preprint arXiv:2310.12836*.

Isaac A Bernstein, Youchen Victor Zhang, Devendra Govil, Iyad Majid, Robert T Chang, Yang Sun, Ann Shue, Jonathan C Chou, Emily Schehlein, Karen L Christopher, et al. 2023. Comparison of ophthalmologist and large language model chatbot responses to online patient eye care questions. *JAMA Network Open*, 6(8):e2330320–e2330320.

Zhihao Fan, Jialong Tang, Wei Chen, Siyuan Wang, Zhongyu Wei, Jun Xi, Fei Huang, and Jingren Zhou. 2024. Ai hospital: Interactive evaluation and collaboration of llms as intern doctors for clinical diagnosis. *arXiv preprint arXiv:2402.09742*.

Junyi Gao, Cao Xiao, Yasha Wang, Wen Tang, Lucas M Glass, and Jimeng Sun. 2020. Stagenet: Stage-aware neural networks for health risk prediction. In *Proceedings of The Web Conference 2020*, pages 530–540.

Yanjun Gao, Ruizhe Li, John Caskey, Dmitriy Dligach, Timothy Miller, Matthew M Churpek, and Majid Afshar. 2023. Leveraging a medical knowledge graph into large language models for diagnosis prediction. *arXiv preprint arXiv:2308.14321*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Rikui Huang, Wei Wei, Xiaoye Qu, Wenfeng Xie, Xianling Mao, and Dangyang Chen. 2024. Joint multi-facts reasoning network for complex temporal question answering over knowledge graph. *arXiv preprint arXiv:2401.02212*.

Pengcheng Jiang, Cao Xiao, Adam Cross, and Jimeng Sun. 2023a. Graphcare: Enhancing healthcare predictions with open-world personalized knowledge graphs. *arXiv preprint arXiv:2305.12788*.

Xinke Jiang, Ruizhe Zhang, Yongxin Xu, Rihong Qiu, Yue Fang, Zhiyuan Wang, Jinyi Tang, Hongxin Ding, Xu Chu, Junfeng Zhao, et al. 2023b. Think and retrieval: A hypothesis knowledge graph enhanced medical large language models. *arXiv preprint arXiv:2312.15883*.

Bowen Jin, Chulin Xie, Jiawei Zhang, Kashob Kumar Roy, Yu Zhang, Suhang Wang, Yu Meng, and Jiawei Han. 2024. Graph chain-of-thought: Augmenting large language models by reasoning on graphs. *arXiv preprint arXiv:2404.07103*.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Hyunsu Lee. 2023. The rise of chatgpt: Exploring its potential in medical education. *Anatomical sciences education*.

Peter Lee, Sebastien Bubeck, and Joseph Petro. 2023. Benefits, limits, and risks of gpt-4 as an ai chatbot for medicine. *New England Journal of Medicine*, 388(13):1233–1239.

Wendi Li, Wei Wei, Xiaoye Qu, Xian-Ling Mao, Ye Yuan, Wenfeng Xie, and Dangyang Chen. 2023. Trea: Tree-structure reasoning schema for conversational recommendation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2970–2982.

Wenge Liu, Yi Cheng, Hao Wang, Jianheng Tang, Yafei Liu, Ruihui Zhao, Wenjie Li, Yefeng Zheng, and Xiaodan Liang. 2022. " my nose is running."" are you also coughing?": Building a medical diagnosis agent with interpretable inquiry logics. *arXiv preprint arXiv:2204.13953*.

Dingkun Long, Qiong Gao, Kuan Zou, Guangwei Xu, Pengjun Xie, Rui Guo, Jianfeng Xu, Guanjun Jiang, Luxi Xing, and P. Yang. 2022. Multi-cpr: A multi domain chinese dataset for passage retrieval.

Ashwin Nayak, Matthew S Alkaitis, Kristen Nayak, Margaret Nikolov, Kevin P Weinfurt, and Kevin Schulman. 2023. Comparison of history of present illness summaries generated by a chatbot and senior internal medicine residents. *JAMA Internal Medicine*, 183(9):1026–1027.

Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2024a. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*.

Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2024b. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*.

Constance Percy, Valerie van Holten, Calum S Muir, World Health Organization, et al. 1990. *International classification of diseases for oncology*. World Health Organization.

Jiashuo Sun, Chengjin Xu, Lumingyuan Tang, Saizhuo Wang, Chen Lin, Yeyun Gong, Heung-Yeung Shum, and Jian Guo. 2023. Think-on-graph: Deep and responsible reasoning of large language model with knowledge graph. *arXiv preprint arXiv:2307.07697*.

Yu Sun, Shuohuan Wang, Shikun Feng, Siyu Ding, Chao Pang, Junyuan Shang, Jiaxiang Liu, Xuyi Chen, Yanbin Zhao, Yuxiang Lu, et al. 2021. Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. *arXiv preprint arXiv:2107.02137*.

Chaojie Wang, Yishi Xu, Zhong Peng, Chenxi Zhang, Bo Chen, Xinrun Wang, Lei Feng, and Bo An. 2023a. keqing: knowledge-based question answering is a nature chain-of-thought mentor of llm. *arXiv preprint arXiv:2401.00426*.

Xiaochen Wang, Junyu Luo, Jiaqi Wang, Ziyi Yin, Suhan Cui, Yuan Zhong, Yaqing Wang, and Fenglong Ma. 2023b. Hierarchical pretraining on multimodal electronic health records. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2839–2852.

Xidong Wang, Guiming Hardy Chen, Dingjie Song, Zhiyi Zhang, Zhihong Chen, Qingying Xiao, Feng Jiang, Jianquan Li, Xiang Wan, Benyou Wang, et al. 2023c. Cmb: A comprehensive medical benchmark in chinese. *arXiv preprint arXiv:2308.08833*.

Xinyu Wang, Yong Jiang, Nguyen Bach, Tao Wang, Zhongqiang Huang, Fei Huang, and Kewei Tu. 2021. Improving named entity recognition by external context retrieving and cooperative learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1800–1812.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

He sicheng Wang Yuxin, Sun Qingxuan. 2023. M3e: Moka massive mixed embedding model.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Yilin Wen, Zifeng Wang, and Jimeng Sun. 2023. Mindmap: Knowledge graph prompting sparks graph of thoughts in large language models. *arXiv preprint arXiv:2308.09729*.

Jiageng Wu, Xian Wu, and Jie Yang. 2024. Guiding clinical reasoning with large language models via knowledge seeds. *arXiv preprint arXiv:2403.06609*.

Ran Xu, Wenqi Shi, Yue Yu, Yuchen Zhuang, Bowen Jin, May D Wang, Joyce C Ho, and Carl Yang. 2024. Ram-ehr: Retrieval augmentation meets clinical predictions on electronic health records. *arXiv preprint arXiv:2403.00815*.

Ran Xu, Yue Yu, Chao Zhang, Mohammed K Ali, Joyce C Ho, and Carl Yang. 2022. Counterfactual and factual reasoning over hypergraphs for interpretable clinical predictions on ehr. In *Machine Learning for Health*, pages 259–278. PMLR.

Lian Yan, Yi Guan, Haotian Wang, Yi Lin, and Jingchi Jiang. 2023. Efficient evidence-based dialogue system for medical diagnosis. In *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 3406–3413. IEEE.

Quan Yan, Junwen Duan, and Jianxin Wang. 2024a. Multi-modal concept alignment pre-training for generative medical visual question answering. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 5378–5389.

Weixiang Yan, Haitian Liu, Tengxiao Wu, Qian Chen, Wen Wang, Haoyuan Chai, Jiayi Wang, Weishan Zhao, Yixin Zhang, Renjun Zhang, et al. 2024b. Clinicallab: Aligning agents for multi-departmental clinical diagnostics in the real world. *arXiv preprint arXiv:2406.13890*.

Rui Yang, Haoran Liu, Qingcheng Zeng, Yu He 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. *arXiv preprint arXiv:2403.05881*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2024. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36.

Ningyu Zhang, Mosha Chen, Zhen Bi, Xiaozhuan Liang, Lei Li, Xin Shang, Kangping Yin, Chuanqi Tan, Jian Xu, Fei Huang, et al. 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.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. Ernie: Enhanced language representation with informative entities. *arXiv preprint arXiv:1905.07129*.

Yinghao Zhu, Changyu Ren, Shiyun Xie, Shukai Liu, Hangyuan Ji, Zixiang Wang, Tao Sun, Long He, Zhoujun Li, Xi Zhu, et al. 2024. Realm: Rag-driven enhancement of multimodal electronic health records analysis via large language models. *arXiv preprint arXiv:2402.07016*.

A Detailed Information of the CMEMR dataset

Specific information on the CMEMR dataset is shown in Table 5.

Department	Num	Avg Len
Gynaecology	411	627.46
Otolaryngology	212	967.99
Obstetrics&Gynecology	1316	489.15
Nursing	52	584.88
Emergency	87	552.96
Psychiatry	127	867.66
Rehabilitation	284	631.13
Dentistry	130	342.56
Anesthesiology	232	634.25
Internal Medicine	3590	528.72
Dermatology	286	518.08
Neurosurgery	3152	531.82
Ophthalmologic	100	453.24
Oncology	471	855.66
Total	10450	558.60

Table 5: Departments distribution of the collected EMRs. "Num" denotes the total number of EMRs of the department. "Avg Len" denotes the average number of words per record.

Following the way of (Yan et al., 2024b), we select a representative sample from CMEMR as an illustrative example, and present both the original Chinese version (Figure 5) and the corresponding English translation (Figure 6).

B Algorithms for medIKAL

We summarize the comprehensive algorithmic procedure of Entity Type-driven Candidate Disease Localization and Filtering and Path-based Reranking, as shown in Algorithm 1 and 2.

Algorithm 1 Entity Type-driven Candidate Disease Localization and Filtering

Require: Entity Set $\mathcal{E}_\mathcal{Q}$, Knowledge graph $\mathcal{G}$, Number of candidate diseases $topm$
Ensure: Candidate disease set $\mathcal{D}_{can}$

Initialize the set of diseases $\mathcal{D} \leftarrow \emptyset$
for each entity $e_i \in \mathcal{E}_\mathcal{Q}$ **do**
  Assign a contribution weight w_{t_i} according to its entity type t_i
  Obtain 1-hop neighbor triplets in $\mathcal{G}$ to locate relevant diseases $\mathcal{D}_i = \{d_{i1}, d_{i2}, \ldots, d_{in}\}$
  for each disease $d_{ij} \in \mathcal{D}_i$ **do**
    if $d_{ij} \in \mathcal{D}$ **then**
      Add w_{t_i} to the score of d_{ij}
    else
      Add d_{ij} to $\mathcal{D}$ with an initial score w_{t_i}
    end if
  end for
end for
Sort the diseases in $\mathcal{D}$ in descending order based on their scores
Select the $topm$ diseases to form $\mathcal{D}_\mathcal{G}$
Merge $\mathcal{D}_\mathcal{G}$ with $\mathcal{D}_{\text{LLM}}$ to form $\mathcal{D}_{can} \leftarrow \mathcal{D}_{\text{LLM}} \cup \mathcal{D}_\mathcal{G}$
return $\mathcal{D}_{can}$

Algorithm 2 Candidate Disease Reranking Based on Paths

Require: Subgraph $\mathcal{G}_s = (V, E)$, Set of candidate diseases $\mathcal{D}_{can}$, Set of entities $\mathcal{E}_\mathcal{Q}$, Number of reranked candidate diseases $topn$
Ensure: Reranked candidate diseases $\mathcal{D}_{rerank}$

Initialize an empty list scores
for each disease $\mathcal{D}_i \in \mathcal{D}_{can}$ **do**
  Initialize score $\leftarrow 0$
  for each entity $e_j \in \mathcal{E}_\mathcal{G}$ **do**
    Compute the shortest path $\text{dist}(\mathcal{D}_i, e_j)$
    if $\text{dist}(\mathcal{D}_i, e_j) = \infty$ **then**
      score $\leftarrow$ score $+ 0$
    else
      score $\leftarrow$ score $+ \frac{1}{\text{dist}(\mathcal{D}_i, e_j)}$
    end if
  end for
  Append $(\mathcal{D}_i, \text{score})$ to scores
end for
Sort scores by the second element (score) in descending order
$\mathcal{D}_{rerank} \leftarrow$ Select the first $topn$ elements from scores
return $\mathcal{D}_{rerank}$

Retriever	R	P	F1
bm25	40.37	29.86	34.32
tf-idf	40.25	29.68	34.16
m3e	41.95	32.63	36.70
all-mpnet	42.01	32.75	36.80
bge	**42.20**	32.81	36.91
corom	42.16	**32.86**	**36.93**
bge + bm25	41.62	30.57	35.24
corom + bm25	41.75	30.46	35.22

Table 6: Performances of medIKAL using different retrieval methods during entity-node matching on CMEMR dataset.

C Detailed Setting-ups for Different Modules in medIKAL Workflow

C.1 Details of the NER Model

The RaNER (Wang et al., 2021) model we use in this paper is released by Tongyi-Laboratory, which is trained on the CMeEE dataset (Zhang et al., 2022). RaNER adopts the Transformer-CRF model, using StructBERT as the pre-trained model base, integrating the relevant sentences recalled by external tools as additional context, and employing Multi-view Training for training. It can recognize a total of 9 types of entities, including body (bod), department (dep), disease (dis), drugs (dru), medical equipment (equ), medical examination items (ite), microorganisms (mic), medical procedures (pro), and clinical symptoms (sym).

C.2 Details of Retrieval Method

In the entity-node matching process mentioned in section 3.2.1, we used a dense retrieval method to link EMR's entities to KG's nodes. In order to better explore the appropriate retrieval method, we implemented three types of retrieval methods based on the retriv library[2]: sparse retrieval, dense retrieval, and hybrid retrieval.

- Sparse Retrieval: We evaluated two representative methods, namely bm25 and tf-idf.
- Dense Retrieval: We evaluated several representative embedding models, namely m3e-large (Wang Yuxin, 2023), all-mpnet-base-v2, bge-large-zh-v1.5, and CoROM.
- Hybrid Retrieval: We evaluated two combinations: "bge + bm25" and "corom + bm25".

[2] https://github.com/AmenRa/retriv

The results are shown in Table 6. As we expected, the effect of dense retrieval is better than that of sparse retrieval and hybrid retrieval, because when the entity to be retrieved contains a large number of Chinese characters, sparse retrieval methods are prone to mismatching due to the lack of consideration of word order and semantics. According to the results, we choose the CoROM model as embedding model of the dense retrieval process.

The CoROM Chinese-medical text representation model we use in this paper is also released by Tongyi-Laboratory. It employs the classic dual-encoder text representation model and is trained on medical domain data with Multi-CPR (Long et al., 2022). The training process is divided into two stages – in the first stage, negative sample data is randomly sampled from the official document set, and in the second stage, difficult negative samples are mined via Dense Retrieval to augment the training data for retraining.

C.3 Details of Other HyperParameters

For the threshold θ mentioned in Section 3.3.2, we set it to 60% of the total score. This parameter generally has a minor impact on performance, as for most samples, the knowledge retrieved from KGs and the patient information extracted from EMRs are sufficient for LLMs to make a confident judgment (i.e., the evaluation score is either close to full marks or close to zero), and variations in the θ value do not significantly affect the final result. But for a small number of samples with higher uncertainty, LLMs tend to provide an evaluation score close to 50% of the total score. So after experimental comparison, we finally set θ to 60% of the total score.

To explore the influence of the number of candidate diseases *Top-k* on medIKAL's performance, we conduct experiments under settings with *Top-k* ranging in [1, 2, 3, 5]. The results are shown in Table 7. According to the results, the Recall gradually decreases with the increase of *Top-k*, while the Precision increases. When the *Top-k* is set very large or very small, although it can get a higher recall or precision rate accordingly, but from the practical clinical application scenario, too large or too small *Top-k* is not conducive to assisting doctors in clinical diagnosis and decision-making. Therefore, in this paper we set *Top-k* to 3 on CMEMR dataset, and 2 on CMB-Clin, GMD and CMD datasets.

Top-k	R	P	F1
1	27.27	56.74	36.83
2	34.15	41.21	37.34
3	42.16	32.86	36.93
5	49.42	24.27	32.55
10	60.85	13.92	22.74

Table 7: Performances of medIKAL with different numbers of candidate diseases (denoted as *Top-k*) on CMEMR dataset.

C.4 Detailed Settings about Knowledge Graph

The knowledge graph we use in this paper is CPubMedKG-v1(Large-scale Chinese Open Medical Knowledge Graph)[3] developed by Harbin Institute of Technology (Shenzhen). It is currently the largest fully open Chinese medical knowledge graph in China. The knowledge is derived from over 2 million high-quality Chinese core medical journals under the umbrella of the Chinese Medical Association. It is regularly updated and conforms to mainstream Chinese medical standards in terms of entity and relationship specifications. The sources of entities and relationships are clearly defined, traceable, and easily distinguishable. The graph contains a total of 4,383,910 disease-centered triples. It includes 523,052 disease entities, 188,667 drug entities, 145,908 symptom entities, and a total of 1,728,670 entities. There are more than 40 types of relationships covering drug treatment, complications, laboratory tests, indications, risk factors, affected populations, mortality rates, and more. The total number of structured knowledge triples reaches 3.9 million.

For the entity-type weights, we obtain the entity-type weight allocation scores through the following two methods:

- We extract paragraphs related to diagnosis from the medical textbooks provided by (Jin et al., 2021). A specific example can be found in Table 8-(1).
- We selected 500 medical records with detailed diagnostic evidence from our collection and collected all diagnostic evidence. These samples will be excluded in subsequent evaluation phases. A specific example can be found in Table 8-(2).

(1) Example: **[Diagnosis]**: History of vitamin D overdose. Early elevation of blood calcium > 3 mmol/L (12 mg/dl), strong positive urinary calcium (Sulkowitch reaction), routine urinalysis shows positive urinary proteins, and in severe cases, red blood cells, leukocytes, and tubular patterns are seen.
(2) Example: **[Diagnostic Evidence]**: 1.history of prior radiotherapy for esophageal cancer, long history of hypertension, history of smoking. 2.left limb weakness for 1 day. 3.Examination revealed hypertension, decreased muscle strength of the left limb, and decreased tenderness. 4.Ancillary tests showed immediate elevated blood glucose, ECG T-wave abnormality, cervical vascular ultrasound and cranial CT and MRI suggestive of cerebral infarction.

Table 8: (1).A specific example of paragraphs related to diagnosis from the medical textbooks provided by (Jin et al., 2021). (2).A specific example of diagnostic evidences in our collected EMRs.

We calculate the entity-type proportions of all the segments above, obtaining initial entity-type weights. The setting of our experiments can be found in Table 9. It is important to note that entity-type weights are not fixed and can be adjusted according to different tasks, which is also the advantage of the method we propose.

For the shortest path algorithm in path-based reranking, we use the GraphDataScience [4] library to implement it.

Type	Weight
dis	.1638
pro	.0043
sym	.6297
dru	.1391
bod	.0212
ite	.0372
equ	.0029
mic	.0009
dep	.0004

Table 9: Entity-type weight settings in our experiments.

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

[4] https://neo4j.com/product/graph-data-science/

D Details of Representative Baseline Methods

In this paper, in addition to directly using LLMs, we compare our framework with two important paradigms, namely LLM⊕KG and LLM⊗KG (Sun et al., 2023). Below, we provide detailed explanations of the representative baseline methods corresponding to these two paradigms.

D.1 LLM⊕KG Baselines

MindMap (Wen et al., 2023): The process begins by identifying key entities in the question and retrieving the knowledge graph to form evidence subgraphs. The LLM then aggregates these into a reasoning graph and generates an answer, presenting its reasoning as a mind map.

HyKGE (Jiang et al., 2023b): The method follows a multi-stage pipeline: it uses LLMs' zero-shot capabilities to expand queries and identify anchor entities, retrieves reasoning chains (path, common ancestor, and co-occurrence), and re-ranks them for alignment with the query. The filtered knowledge is then combined with the query, and LLMs generate the final answer.

D.2 LLM⊗KG Baselines

ToG (Sun et al., 2023): In this framework, the LLM acts as an agent, using beam search to explore reasoning paths on the KG until enough information is gathered or the search depth limit is reached. ToG involves three stages: initialization, exploration, and reasoning. The LLM identifies initial entities, expands paths through search and pruning, and evaluates if the path suffices to generate an answer, repeating exploration if necessary.

Graph-CoT (Jin et al., 2024): The method simulates human thought by breaking complex graph reasoning into iterative steps: LLM reasoning to identify needed information, LLM-graph interaction to generate operations like node lookup, and graph execution to perform these operations and return results. This cycle repeats until the LLM reaches the final answer.

E Evaluation Metrics Calculation

Firstly, for the disease entities in the diagnosis results $\hat{\mathcal{D}}$ and the reference diagnosis results $\mathcal{R}$ in the medical records, we used a fuzzy matching process (with a predefined threshold of 0.5) to associate these disease entities with ICD-10 terms, thus mapping $\hat{\mathcal{D}}$ and $\mathcal{R}$ to two standardized disease sets $S_{\hat{\mathcal{D}}}$ and $S_{\mathcal{R}}$ respectively. We then define: **True Positives (TP):** The number of disease entities in the predicted result $S_{\hat{\mathcal{D}}}$ that correspond correctly with the reference diagnosis $S_{\mathcal{R}}$.

False Positives (FP): The number of disease entities that appear in the predicted result $S_{\hat{\mathcal{D}}}$ but do not match correctly with the reference diagnosis $S_{\mathcal{R}}$.

False Negatives (FN): The number of disease entities in the reference diagnosis $S_{\mathcal{R}}$ that do not appear in the predicted result $S_{\hat{\mathcal{D}}}$. Based on the above statistical values, we calculate the following evaluation metrics:

$$\text{Recall (R)} : R = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5}$$

$$\text{Precision (P)} : P = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{6}$$

$$\text{F1 Score (F1)} : F = \frac{2 \times P \times R}{P + R} \tag{7}$$

F The prompt templates used in this paper

Data Example (Chinese)

【病例ID】: 63682
【科室】: 耳鼻咽喉科
【病历摘要】:
基本信息: 女，55岁。
主诉: 咽痛伴呼吸费力、吞咽困难1天。
现病史: 患者1天前无明显诱因出现咽部疼痛，无咳嗽、咳痰，无头痛、发热，无胸闷、心慌等其他不适。患者未予治疗，未行特殊处理。昨日夜间患者无明显诱因出现呼吸费力，吞咽感困难，咽部疼痛加重，自行口服消炎药物未见明显缓解，无意识障碍，无咳嗽，咳痰，无头痛、发热，无恶心、呕吐等其他不适。患者今为求进一步治疗，特来我科就诊，门诊拟急性会厌炎收治入院。入院患者自起病来精神软，饮食睡眠差，大小便未见明显异常。
既往史: 自诉糖尿病病史三年，甲状腺肿大，子宫肌瘤病史，日常规律服用阿卡波糖片三次每日，一粒每次。自诉餐前空腹血糖7，餐后血糖小于10，具体不详。早前曾行卵巢囊肿摘除术，自诉现一般情况可。自诉规律服用左旋甲状腺素片半片每天，既往常有早饭过晚感胃痛，日常有反酸病史，未正规就诊及规律用药，无肝炎结核等传染病史，无高血压，冠心病等病史，无外伤无输血史，无食物药物过敏史，按计划预防接种。
辅助检查:
血常规示: 中性细胞比率0.751↑、淋巴细胞比率0.185↓、嗜酸性粒细胞$0.04 \times 10^9/L$、血小板压积0.31; 血生化示: LDL胆固醇$3.81mol/L$↑、C反应蛋白$28.6mg/L$、总胆固醇$5.84mol/L$、葡萄糖$6.32mmol/L$、糖化血清白蛋白0.1681; 尿常规示: 白细胞$25 \times 10^9/L$↑、尿潜血$25/\mu L$; 甲功5项示: 促甲状腺素$0.184\mu IU/ml$↓; 咽拭子脓液培养示: 生长正常菌群; 胸片示: 两肺未见明显活动性病变; 副鼻窦CT示: 副鼻窦CT平扫未见明显异常，口咽部右侧粘膜稍增厚，请结合临床; 腹部B超示: 胆囊底部絮状回声，考虑泥沙样结石，肝脏、胰腺、脾脏未见明显占位性病变; 甲状腺彩超示: 甲状腺多发结节。其他未见明显异常。
【临床诊断】
初步诊断: 急性会厌炎，咽部水肿，2型糖尿病
诊断依据: 患者症状为咽痛伴呼吸费力、吞咽困难1天，查体咽部悬雍垂及软腭水肿，双侧扁桃体1度，咽后壁淋巴滤泡充血，间接喉镜下观：会厌下榻，舌面黏膜水肿呈球状，充血明显，双侧声带窥不清。患者糖尿病病史3年。
鉴别诊断: 根据患者症状.体征，检查等可初步诊断为急性会厌炎，咽部水肿，但亦不排除以下可能：1.会厌囊肿：常可见会厌舌面囊肿样物质隆起，表面光滑成球状，患者常异物感明显，无咽痛，无呼吸，吞咽等困难。2.会厌脓肿：查体会厌表面可见粘膜明显隆起，肿物内可见脓性分泌物流出，患者常感咽喉部疼痛，咽喉部异物感明显。
诊断结果: 急性会厌炎，咽部水肿，右侧扁桃体周围脓肿，2型糖尿病

Figure 5: A data example from CMEMR.

Data Example (English)

[Case ID]: 63682
[Department]: Otorhinolaryngology
[Case Summary]:
Basic Information: Female, 55 years old.
Chief Complaint: Sore throat accompanied by labored breathing and difficulty swallowing for 1 day.
History of Present Illness: One day ago, the patient developed a sore throat without obvious triggers, accompanied by no cough, sputum, headache, fever, chest tightness, or palpitations. She did not seek treatment or take special measures. Last night, labored breathing and difficulty swallowing emerged, with worsening throat pain. Self-administered anti-inflammatory medication was ineffective. She denied loss of consciousness, nausea, vomiting, or other symptoms. Due to worsening symptoms, she visited our department and was admitted with a preliminary diagnosis of acute epiglottitis. since the onset of illness, the patient has experienced lethargy, poor appetite, and sleep disturbances, with no significant abnormalities in urination or defecation.
Past Medical History: The patient reports a 3-year history of diabetes, thyroid enlargement, and uterine fibroids. She takes acarbose regularly (1 tablet, three times daily) and levothyroxine sodium (half a tablet daily). Fasting blood glucose is self-reported at 7 mmol/L, and postprandial glucose under 10 mmol/L, specifics unclear. She underwent ovarian cystectomy and reports stable health. She occasionally experiences gastric pain after delayed breakfast and has a history of acid reflux but has not sought formal treatment. She denies infectious diseases, hypertension, coronary artery disease, trauma, or blood transfusions. No known allergies; vaccinations are up-to-date.
Auxiliary Examinations:
Complete Blood Count: Neutrophil ratio 0.751↑, lymphocyte ratio 0.185↓, eosinophils $0.04 \times 10^9/L$, plateletcrit 0.31; Biochemical Panel: LDL cholesterol $3.81mol/L$↑, C-reactive protein $28.6mg/L$, total cholesterol $5.84mol/L$, glucose $6.32mmol/L$, glycated albumin 0.1681; Urinalysis: Leukocytes $25 \times 10^9/L$↑, urine occult blood $25/\mu L$; Thyroid Function Test (5 items): Thyroid-stimulating hormone $0.184\mu IU/ml$↓; Throat Swab Pus Culture: Normal flora growth; Chest X-ray: No significant active pulmonary lesions; Paranasal Sinus CT: No significant abnormalities in the sinuses; slight mucosal thickening on the right side of the oropharynx; clinical correlation suggested; Abdominal Ultrasound: Flocculent echo at the gallbladder fundus, suggestive of sludge-like stones; liver, pancreas, and spleen unremarkable; Thyroid Ultrasound: Multiple nodules in the thyroid gland. No other significant abnormalities.
[Clinical Diagnosis]
Preliminary Diagnosis: Acute epiglottitis, pharyngeal edema, Type 2 Diabetes Mellitus (T2DM)
Diagnostic Basis: The patient presents with a sore throat, labored breathing, and difficulty swallowing for 1 day. Examination reveals uvular and soft palate edema, Grade 1 bilateral tonsils, pharyngeal lymphoid follicular hyperemia, and indirect laryngoscopy showing epiglottic collapse with globular mucosal swelling and marked hyperemia. The patient has a 3-year history of diabetes.
Differential Diagnosis: Based on symptoms, signs, and examinations, the preliminary diagnosis is acute epiglottitis with pharyngeal edema. Differential considerations include: 1. **Epiglottic Cyst**: Typically characterized by cystic elevation on the lingual surface of the epiglottis, smooth and globular appearance, often with prominent foreign body sensation but without throat pain, respiratory or swallowing difficulties. 2. **Epiglottic Abscess**: Physical examination reveals significant mucosal elevation with purulent discharge from the lesion, accompanied by pronounced throat pain and foreign body sensation.
Final Diagnosis: Acute epiglottitis, pharyngeal edema, right peritonsillar abscess, Type 2 Diabetes Mellitus.

Figure 6: A data example from CMEMR(translated).

Error Class 1: Overlapping Disease Characteristics
[Case 1]: Overlapping characteristics of different diseases **[Key Info]:** Fine scales visible on back rash; prominent petechiae on both lower limbs; fluocinolone acetonide cream **[Label]:** Chronic Lichenified Pityriasis **[Candidate Diseases]:** Psoriasis, Eczema, Pityriasis Rosea, Henoch-Schönlein Purpura **[LLM Final Decision]:** Psoriasis (✘), Pityriasis Rosea (✘)
[Case 2]: Overlapping characteristics within the same category **[Key Info]:** Fatigue; melena; fecal occult blood test positive (colloidal gold method) **[Label]:** Gastric cancer with bleeding, Iron-deficiency anemia, ... **[Candidate Diseases]:** Gastric cancer, Gastrointestinal bleeding, Anemia, ..., ~~Iron-deficiency anemia~~ (reprioritized, ranking >$topn$)... **[LLM Final Decision]:** Gastric cancer (✓), Gastrointestinal bleeding (✓), Anemia (✓)
Error Class 2: Misunderstanding of Examination Indicators
[Case 1]: Insufficient understanding of the meaning and utility of examination indicators **[Key Info]:** EF 27%; FS 12%; LAM light chain M-protein positive **[Label]:** Multiple Myeloma, Cardiac Amyloidosis **[Candidate Diseases]:** Multiple Myeloma, Chronic Heart Failure, ... **[LLM Final Decision]:** Multiple Myeloma (✓), Chronic Heart Failure (✘)
Error Class 3: Gap Between LLM Reasoning and Physician Diagnostic Labeling
[Case 1]: Diagnostic results include previous medical history **[Key Info]:** Deep ulcer in the middle of the gastric body; white ulcer scars in the duodenal bulb; pulmonary tuberculosis for over a year, already cured **[Label]:** Gastric ulcer, Duodenal bulb ulcer, ..., Inactive Pulmonary Tuberculosis **[Candidate Diseases]:** Gastric ulcer, Duodenal bulb ulcer, ..., ~~Inactive Pulmonary Tuberculosis~~ (not considered as a candidate disease) **[LLM Final Decision]:** Gastric ulcer (✓), Duodenal bulb ulcer (✓)
[Case 2]: Diagnostic results include abbreviations **[Key Info]:** Paroxysmal abdominal pain; urine glucose 3+, ketone 2+; normal serum amylase **[Label]:** FCPD, DKA **[Candidate Diseases]:** Diabetes, Chronic Pancreatitis, DKA **[LLM Final Decision]:** Diabetes (✓), Chronic Pancreatitis (✘)

Table 10: Examples of Classification Errors. For clarity, only part of the key information of the selected samples is presented. "**[Candidate Diseases]**" denotes the set of candidate diseases obtained after the Path-based Reranking stage. Diseases with underlines indicate that diagnoses annotated by the doctor being either misdiagnosed or missed by LLMs. "✓" indicates that the final decision of LLMs is consistent or partially consistent with the doctor's annotation, while "✘" indicates the opposite.

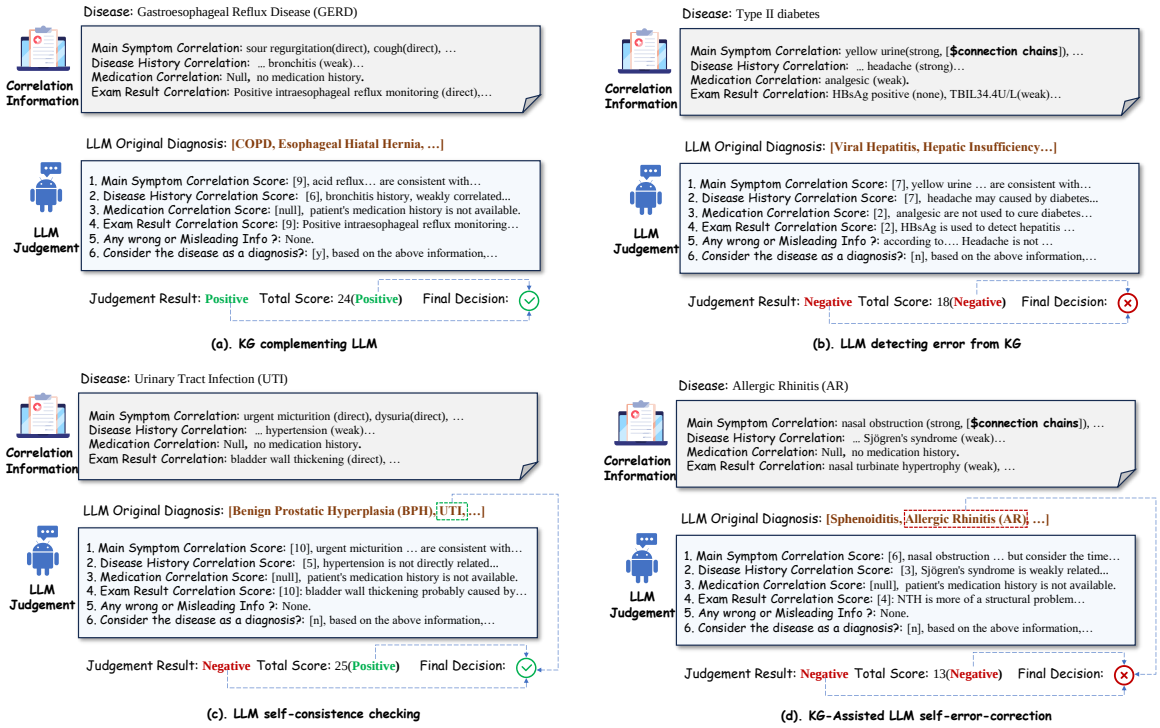

Figure 7: Case study.

[**Role**]<SYS>
You are an outstanding AI medical expert. You can summarize critical information for diagnosis based on the content of the patient's medical records.

[**Role**]<USR>
Below is a portion of the electronic medical record of a real patient. Please read the following content carefully to understand the patient's basic condition.
Patient Medical Record Content
"""
"History of Present Illness": ${HPI}
"Past Medical History": ${PMH}
"""
Task:
Based on the above content, please summarize the key information useful for diagnosis and treatment and generate a summary report.
Report Format Requirements:
Please fill in the "[]" sections according to the following format to complete the report. Use concise language whenever possible.
"""
1. Main symptoms: []
2. Recent medical visits: [] (if none, write "none")
3. Past medical history: [] (if none, write "none")
4. Past surgical history: [] (if none, write "none")
5. Medication usage: [] (if none, write "none")
"""
Output:
${}

Table 11: The default prompt for the LLM Summarization module (for the patients' basic condition) .

[**Role**]<SYS>
You are an excellent AI medical expert. You can summarize key information useful for diagnosis based on the patient's examination results.

[**Role**]<USR>
Task:
Please summarize and generalize the key information useful for diagnosis based on the patient's examination results.
Example
"""
[Patient's Examination Results]
"Physical Examination": Bilateral waistline symmetry, no tenderness in the bilateral ureteral regions, bladder area distended, no palpable mass, no redness or abnormal discharge at the urethral opening, no abnormalities in the scrotum, and no abnormalities in the bilateral testicles and epididymis. Digital rectal exam: Prostate approximately 4.0×5.0cm in size, soft, central area slightly shallow, small nodules palpable.
"Laboratory and Aided Examination": Ultrasound results show 1. Bilateral kidney cysts 2. Prostatic hyperplasia 3. No abnormalities in the ureters and bladder.
"""
Please refer to the above example to summarize the patient's examination results.
[Patient's Examination Results]
"Physical Examination": ${PE}
"Laboratory and Aided Examination": ${LAE}
##Output:
${}

Table 12: The default prompt for the LLM Summarization module (for the patients' exam results).

[**Role**]<SYS>
You are an outstanding AI medical expert. You can perform a preliminary disease diagnosis based on the patient's condition.

[**Role**]<USR>
##Patient Information
"""
[General Condition]: ${summary_1}
[Examination Findings]: ${summary_2}
"""
##Task
Based on the patient's symptoms, medical visit history, past medical history, and examination results, predict the possible diseases the patient may have (you can provide the top-${n} possible predictions). Please only output the prediction results, do not output any other content.
##Prediction Results
Predicted Disease 1: ${} Predicted Disease 2: ${} Predicted Disease 3: ${} ...

Table 13: The default prompt for the LLM Direct Diagnose Module.

[**Role**]<SYS>
You are an experienced medical expert. You can evaluate the reasonableness of existing diagnostic results by considering the patient's symptoms, medical history, medication usage, and examination results.

[**Role**]<USR>
##Patient Information
"""
[General Condition]: ${summary_1}
[Examination Findings]: ${summary_2}
"""
A doctor has made a preliminary diagnosis based on the above information, with the diagnosis being: ${disease}
You need to consider whether this diagnosis is correct. To do this, you queried a medical knowledge graph and obtained the following information:
##Correlation Information
"""
Correlation between diagnosis ${disease} and patient's main symptoms: ${correlation_1}
Correlation between diagnosis ${disease} and patient's medical history: ${correlation_2}
Correlation between diagnosis ${disease} and patient's medication usage: ${correlation_3}
Correlation between diagnosis ${disease} and patient's examination results: ${correlation_4}
"""
##Task
Based on the patient's condition and the above information, and in combination with your own knowledge, please quantitatively evaluate the reasonableness of the diagnosis ${disease}.
##Requirements
"""
1.Consistency with the patient's chief complaint score: [?] (out of 10)
2.Correlation with the patient's medical history score: [?] (out of 10)
3.Correlation with the patient's medication usage score: [?] (out of 10)
4.Correlation with the patient's examination results score: [?] (out of 10)
5.Are there any errors or misleading information in the "Correlation Information" section ?
6.Can this disease be used as a diagnostic result: [?] (y/n)
"""
##Output:
${}

Table 14: The default prompt for the LLM Diagnosis Evaluation Module.