ClinicalRAG: Enhancing Clinical Decision Support through Heterogeneous Knowledge Retrieval

Yuxing Lu, Jinzhuo Wang*
Department of Big Data and Biomedical AI
College of Future Technology
Peking University
Beijing, China
yxlu0613@gmail.com wangjinzhuo@pku.edu.cn

Xukai Zhao
Department of Landscape Architecture
School of Architecture
South China University of Technology
Guangzhou, China
zhaoxukai0208@163.com

Abstract

Large Language Models (LLMs) have revolutionized text generation across diverse domains, showcasing an ability to mimic human-like text with remarkable accuracy. Yet, these models frequently encounter a significant hurdle: producing hallucinations, a flaw particularly detrimental in the healthcare domain where precision is crucial. In this paper, we introduce ClinicalRAG, a novel multi-agent pipeline to rectify this issue by incorporating heterogeneous medical knowledge—both structured and unstructured—into LLMs to bolster diagnosis accuracy. ClinicalRAG can extract related medical entities from user inputs and dynamically integrate relevant medical knowledge during the text generation process. Comparative analyses reveal that ClinicalRAG significantly outperforms knowledge-deficient methods, offering enhanced reliability in clinical decision support. This advancement marks a pivotal proof-of-concept step towards mitigating misinformation risks in healthcare applications of LLMs.

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities in generating human-like text across various domains (Achiam et al., 2023; Touvron et al., 2023; Singhal et al., 2023). However, these models often produce hallucinations—generating inaccurate or entirely fictitious information. This issue is particularly critical in sensitive domains like healthcare, where misinformation can have dire repercussions (Zawiah et al., 2023). The underlying cause of such hallucinations largely stems from the model's insufficient domain-specific knowledge.

Medical domain is characterized by its vast array of knowledge, which includes both structured information (such as knowledge graphs, medical databases) and unstructured information (like online resources) (Kreimeyer et al., 2017). These knowledge are inherently heterogeneous, spanning various subfields and formats, which poses significant challenges for traditional models that rely on a one-size-fits-all approach to knowledge integration and application. As a result, the discrepancies among knowledge sources impede the models' ability to utilize knowledge prompts from all available sources effectively. In light of this, we aim to propose a method that can seamlessly integrate and accommodate all source of medical knowledge.

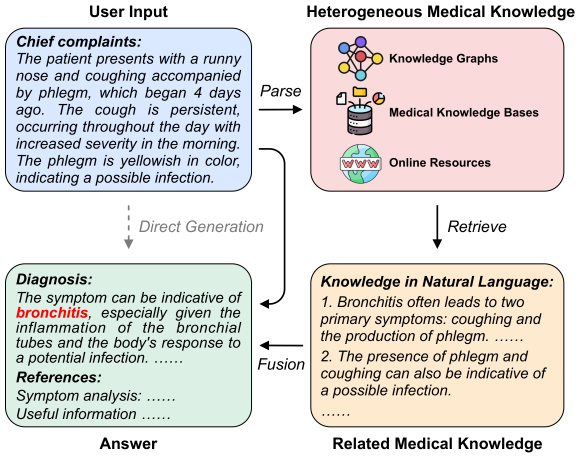

Figure 1: **Overview of ClinicalRAG.** Different from direct generation without any knowledge enhancement, ClinicalRAG utilizes heterogeneous medical knowledge to parse and cross-reference user inputs. It then integrates this to formulate diagnoses and provide relevant references, thereby supporting clinical decision-making.

Retrieval-Augmented Generation (RAG) offers a powerful approach for harnessing the implicit knowledge embedded within LLMs alongside diverse explicit knowledge sources (Hu and Lu, 2024). Through real-time retrieval of pertinent information during the generation phase, RAG models are adept at delivering outputs that are both precise and contextually relevant (Wu et al., 2024). Consequently, it empowers the models to efficiently access domain-specific information, enhancing the quality of the generated content.

In this paper, we introduce a Clinical Retrieval Augmented Generation (ClinicalRAG) pipeline (Fig. 1), a novel framework designed to enhance

clinical decision-making by leveraging medical knowledge from a variety of sources. Our contributions are threefold and can be summarized as follows: 1) We develope a multi-agent integration approach, where each agent is responsible for a specific part of the ClinicalRAG process. This ensures the efficiency and robustness of the pipeline. 2) We design an effective solution for the extraction and integration of heterogeneous medical knowledge, which, compared to long text inputs, allows for the low-cost acquisition of high-quality information. 3) Experimental results demonstrate that our ClinicalRAG pipeline outperforms traditional methods such as simple prompt learning and direct generation. It also provides relevant references, facilitating a more effective clinical decision support.

2 Related work

Recent literature on knowledge-enhanced Clinical Decision Support (CDS) systems showcases a plethora of innovative approaches aimed at leveraging technology to improve healthcare outcomes. Anadani et al. (2023) implements ant colony optimization methods within CDS systems to customize treatment plans for patients, thus enriching the knowledge base for making optimal clinical decisions. Zhang et al. (2023) leverage a knowledge graph and an attribute graph to generate better medicine recommendations. Recently, Lu et al. (2023a,b) have introduced prompt learning methods for integrating heterogeneous medical knowledge. Moreover, the development of LLMs enables a more precise and effective way to utilize current medical knowledge. One useful method is Chain-of-Thought (Wei et al., 2022) which mimics human problem-solving processes by breaking down complex questions into simpler, manageable parts. Based on this, Tree-of-Thought (Yao et al., 2024) and Graph-of-Thought (Besta et al., 2023) methods are proposed to deal with more complex question-solving flow. Additionally, by integrating external knowledge, RAG significantly improves the quality of the generated content, making it more informative and accurate across various tasks (Ye et al., 2024), demonstrating its effectiveness in enriching language model outputs with detailed and precise information.

3 Methods

The detailed pipeline of ClinicalRAG is shown in Fig. 2. We employ a multi-agent strategy in ClinicalRAG, each agent is designed to carry out different task.

3.1 Medical entity extraction

The Medical Entity Extraction (MEE) agent's task is to parse and discern pertinent medical entities from the input. This preliminary step is critical, as it establishes the foundational context required for subsequent knowledge retrieval processes.

Given a user input I, the MEE agent seeks to identify a set of entities $E = \{e_1, e_2, \ldots, e_n\}$, where each entity e_i is associated with a specific medical concept. This can be formalized using a function f_{MEE}^{LLM} powered by an LLM. This can be denoted as:

$$f_{MEE}^{LLM}(I) = \{(e_i, c_i) | e_i \in I, c_i \in C\} \quad (1)$$

where e_i denote the i^{th} entity within the input, and c_i represents the category of the entity, drawn from a predefined set of categories C (e.g., symptoms, diseases, treatments). All the extracted entites are sent into the Heterogeneous Knowledge Index (HKI) engine for knowledge retrieval.

3.2 Heterogeneous knowledge index

The HKI engine is engineered to index and retrieve medical knowledge from diverse sources using entities identified from the MEE agent. This is crucial for dynamically augmenting LLMs' responses with accurate, context-specific medical information.

For each source S (e.g., knowledge graph G, knowledge base B and online resources O), we construct an entity-based index. Entities E extracted from the user input serve as the retrieval keys. Each entity $e \in E$ is associated with a vector representation $\vec{v_e}$ obtained via embedding techniques such as BERT (Devlin et al., 2018). Given a query entity e, the HKI retrieves relevant information by computing similarity scores across all indexed entities in G, B, and O. The retrieval is conducted separately for each source, leveraging their respective indexing systems.

$$Score(e, e') = \frac{\vec{v}_e \cdot \vec{v}_{e'}}{\|\vec{v}_e\| \|\vec{v}_{e'}\|}, \quad \forall e' \in S \quad (2)$$

where e' is an entity in the source S, and $Score(e, e')$ denotes the cosine similarity between the query entity and entities in the source.

The HKI employs a dynamic integration mechanism to compile and synthesize information from G, B and O based on relevance scores. This process ensures that the most pertinent and comprehensive knowledge is selected for supporting the LLM's generation process.

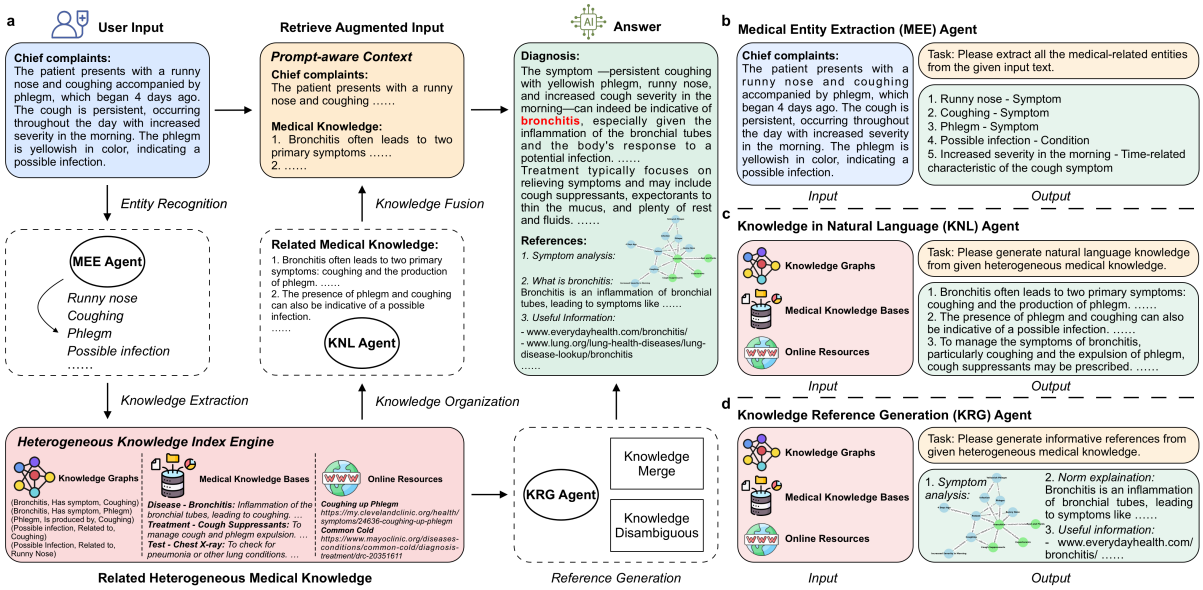

Figure 2: **ClinicalRAG framework.** a) The pipeline of ClinicalRAG. Patients' chief complaints are first sent to MEE agent to extract related medical entities. Heterogeneous medical knowledge are retrieved from different sources and converted into natural language by KNL agent. User input and medical knowledge are fused and sent to generate high-quality answers, with KRG agent provide proper references. b) MEE agent helps extract important medical entities from patient's input. c) KNL agent convert heterogeneous medical knowledge into unified natural language form. d) KRG agent provides useful and disambiguous references from heterogeneous medical knowledge.

$$K = \bigcup_{S \in \{G,B,R\}} Top_k(Score(e,S)), \quad \forall e \in E \tag{3}$$

where Top_k selects the top k items from each source S based on the retrieval score, and K represents the integrated knowledge set ready for utilization in the following generation process.

3.3 Knowledge to natural language

Once heterogeneous medical knowledge is retrieved and compiled, the Knowledge to Natural Language (KNL) agent converts this information into natural language. This conversion process can be represented as a function f_{KNL} that maps a set of knowledge pieces $K = \{k_1, k_2, \ldots, k_m\}$ to a natural language representation N with a template-based transformation T and a natural language generation model G:

$$n_i = T(k_i) \oplus G(k_i), \forall k_i \in K \tag{4}$$

where $\oplus$ denotes concatenating template-based text with generated text to form a comprehensive natural language description n_i for each piece of knowledge k_i. The set of all n_i forms the natural language representation $N = \{n_1, n_2, \ldots, n_m\}$, which serves as enriched context for the LLM, enabling it to generate more accurate and contextually relevant responses in CDS systems.

3.4 Knowledge reference generation

KRG agent aims to aggregate the relevant medical knowledge K retrieved by HKI into a standardized reference format that can be seamlessly integrated into the output of the LLM. This process ensures that the information provided is not only accurate and relevant but also properly cited, enhancing the credibility of the generated content.

The KRG agent first identifies and removes duplicate knowledge entries from the set. This is achieved by comparing the content and source metadata of each knowledge piece. If two pieces k_i and k_j are found to be identical in content or exceedingly similar in the information provided, only one is retained for further processing. The non-duplicate knowledge pieces are then sorted in descending order of their relevance scores and formatted into a standardized reference style. This ordering ensures that the most pertinent references are prioritized in the final reference list.

4 Experiments

4.1 Dataset

We utilized a subset of the CBLUE EHR dataset (Zhang et al., 2021) for our proof-of-concept experiments. We filtered out all records containing multiple diagnoses and selected 2,000 records comprising patients' chief complaints along with their corresponding diagnoses to serve as the dataset for this study. In our research, we employ the DiseaseKG, an open-source Chinese medical knowledge graph available through OpenKG, as our primary knowledge graph. To supplement this, we

Table 1: Diagnosis performance comparison (Avg(SD)). The highest accuracy is highlighted in bold.

Model	Direct classification	ClinicalRAG pipeline
LSTM+Attn	69.17(0.88)	-
BERT	74.07(2.15)	-
MedKPL	77.78(2.51)	-
GPT-3.5-Turbo	80.04(0.41)	81.75(0.65)
GPT-4.0	**82.78(1.25)**	**84.94(1.48)**
Llama-2-7b	77.90(1.69)	79.47(2.50)
Llama-2-13b	78.93(1.02)	80.55(1.25)

construct a knowledge database from a selection of medical textbooks. Additionally, we utilize online medical information, predominantly sourced from Wikipedia, to enrich our data.

4.2 Experiment settings

The patient's chief complaint input, when combined with the medical-knowledge-aware context, was used as input to the LLM for text generation. In our experiments, we choose four mainstream available LLMs: GPT-3.5-Turbo (Ouyang et al., 2022), GPT-4.0 (Achiam et al., 2023), Llama-2-7b, Llama-2-13b (Touvron et al., 2023) in our experiments, where GPT-3.5 and GPT-4 are accessed through OpenAI API, and Llama-2-13b, Llama-2-13b with a token size of 4096 are deployed locally. In our experiments, the temperature parameter (Brown et al., 2020) was set to 0 for all LLMs. In our experiments, we calculate the diagnostic accuracy of LLM compared to the Clinical pipeline by checking whether the diagnostic results provided by the LLM are consistent with the labels in the dataset.

4.3 Diagnosis performance

We evaluated several different EHR diagnosis models, including direct classification approaches like LSTM model with attention mechanism (Chen et al., 2020), BERT model for text classification (Devlin et al., 2018), medical knowledge prompt learning (MedKPL) model (Lu et al., 2023a), and different generative LLMs (GPT-3.5-Turbo (Ouyang et al., 2022), GPT-4.0 (Achiam et al., 2023), Llama-2-7b, Llama-2-13b (Touvron et al., 2023)) under both direct diagnosis generation and the ClinicalRAG pipeline. We compared the diagnostic results generated by the model with the actual disease categories of the patients. The comparison results are shown in Table 1.

All LLMs outperform traditional methods in direct classification scenarios, with GPT-4.0 leading at an accuracy of 82.78(1.25)%. Furthermore, the implementation of the ClinicalRAG pipeline consistently enhances model performance, where nearly all LLMs achieved an accuracy improvement of over 2%, highlighting ClinicalRAG's significant role in augmenting medical diagnostic capabilities.

Table 2: Ablation study of different agents and input lengths (Avg(SD)).

	GPT-3.5-Turbo	Llama-2-7b
Full ClinicalRAG Pipeline	81.75(0.65)	79.47(2.50)
- w/o MEE agent	80.85(1.90)	79.03(1.53)
- w/o KNL agent	79.53(0.86)	78.64(1.66)
- w/o KRG agent	81.77(0.84)	79.44(1.62)
Input Length		
- 2048 tokens	80.81(1.28)	78.46(1.42)
- 1024 tokens	77.87(1.86)	77.53(1.57)

4.4 Ablation study

To quantitatively evaluate the contribution of different modules in ClinicalRAG, we conducted a series of ablation studies, the results are shown in Table 2.

First, we tested the impact of each agent on the ClinicalRAG generation effect by removing the corresponding agents. The results show that the KNL agent plays the most important role in the entire ClinicalRAG pipeline, with a relative decrease in model performance of 1.53% after removing KNL. The importance of the MEE agent comes next (0.67%), while KRG, as the agent providing medical references, has a smaller impact on the diagnostic effect of ClinicalRAG.

We then look into the impact of input length on the ClinicalRAG generation performance, where we limit the input length to 2048 and 1024 tokens respectively. We found that as the input length decreases, the performance of the model also shows a downward trend, especially in the process of reducing from 2048 (-0.98%) to 1024 (-2.91%).

5 Conclusion

In this paper, we presented ClinicalRAG, a novel multi-agent pipeline that significantly enhances the accuracy and reliability of clinical decision support provided by LLMs. By seamlessly integrating heterogeneous medical knowledge—ranging from structured knowledge graphs and to unstructured medical knowledge bases and online resources—ClinicalRAG addresses the critical challenge of hallucinations and inaccuracies in LLM-generated content within the healthcare domain. Our comprehensive experiments have demonstrated the superior diagnosis performance of the ClinicalRAG pipeline over traditional methods.

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*.

Ishwa Anadani, Pavi Sharma, and Anand Sharma. 2023. Aco based clinical decision support system for better medical care. *International Journal on Recent and Innovation Trends in Computing and Communication*.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Michal Podstawski, Hubert Niewiadomski, Piotr Nyczyk, et al. 2023. Graph of thoughts: Solving elaborate problems with large language models. *arXiv preprint arXiv:2308.09687*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Che-Wen Chen, Shih-Pang Tseng, Ta-Wen Kuan, and Jhing-Fa Wang. 2020. Outpatient text classification using attention-based bidirectional lstm for robot-assisted servicing in hospital. *Information*, 11(2):106.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Yucheng Hu and Yuxing Lu. 2024. Rag and rau: A survey on retrieval-augmented language model in natural language processing. *arXiv e-prints*, pages arXiv–2404.

Kory Kreimeyer, Matthew Foster, Abhishek Pandey, Nina Arya, Gwendolyn Halford, Sandra F Jones, Richard Forshee, Mark Walderhaug, and Taxiarchis Botsis. 2017. Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review. *Journal of biomedical informatics*, 73:14–29.

Yuxing Lu, Xiaohong Liu, Zongxin Du, Yuanxu Gao, and Guangyu Wang. 2023a. Medkpl: a heterogeneous knowledge enhanced prompt learning framework for transferable diagnosis. *Journal of Biomedical Informatics*, page 104417.

Yuxing Lu, Xukai Zhao, and Jinzhuo Wang. 2023b. Medical knowledge-enhanced prompt learning for diagnosis classification from clinical text. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 278–288.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

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.

Kevin Wu, Eric Wu, Ally Cassasola, Angela Zhang, Kevin Wei, Teresa Nguyen, Sith Riantawan, Patricia Shi Riantawan, Daniel E. Ho, and James Zou. 2024. How well do llms cite relevant medical references? an evaluation framework and analyses.

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.

Haoran Ye, Jiarui Wang, Zhiguang Cao, and Guojie Song. 2024. Reevo: Large language models as hyperheuristics with reflective evolution. *arXiv preprint arXiv:2402.01145*.

Mohammed Zawiah, Fahmi Y Al-Ashwal, Lobna Gharaibeh, Rana Abu Farha, Karem H Alzoubi, Khawla Abu Hammour, Qutaiba A Qasim, and Fahd Abrah. 2023. Chatgpt and clinical training: perception, concerns, and practice of pharm-d students. *Journal of multidisciplinary healthcare*, pages 4099–4110.

Ningyu Zhang, Mosha Chen, Zhen Bi, Xiaozhuan Liang, Lei Li, Xin Shang, Kangping Yin, Chuanqi Tan, Jian Xu, Fei Huang, et al. 2021. Cblue: A chinese biomedical language understanding evaluation benchmark. *arXiv preprint arXiv:2106.08087*.

Yingying Zhang, Xian Wu, Quan Fang, Shengsheng Qian, and Changsheng Xu. 2023. Knowledge-enhanced attributed multi-task learning for medicine recommendation. *ACM Transactions on Information Systems*, 41(1):1–24.