A Smart Multimodal Healthcare Copilot with Powerful LLM Reasoning

Xuejiao Zhao[1,2] , **Siyan Liu**[1,2] , **Su-Yin Yang**[3,4] , **Chunyan Miao**[1,2*]
[1]Joint NTU-UBC Research Centre of Excellence in Active Living for the Elderly (LILY), NTU
[2]College of Computing and Data Science, Nanyang Technological University (NTU), Singapore
[3]Tan Tock Seng Hospital, Singapore
[4]Woodlands Health, Singapore
{xjzhao,siyan.liu,ascymiao}@ntu.edu.sg, su_yin_yang@wh.com.sg

Abstract

Misdiagnosis causes significant harm to healthcare systems worldwide, leading to increased costs and patient risks. MedRAG is a smart multimodal healthcare copilot equipped with powerful large language model (LLM) reasoning, designed to enhance medical decision-making. It supports multiple input modalities, including non-intrusive voice monitoring, general medical queries, and electronic health records. MedRAG provides diagnostic, treatment, medication, and follow-up questioning recommendations. Leveraging retrieval-augmented generation enhanced by knowledge graph-elicited reasoning, it retrieves and integrates critical diagnostic insights, reducing the risk of misdiagnosis. MedRAG is evaluated on public and private datasets, outperforming existing models and offering more specific and accurate healthcare assistance. The MedRAG demonstration video is available at *https://www.youtube.com/watch?v=PNIBDMYRfDM*. The code is available at *https://github.com/SNOWTEAM2023/MedRAG*

1 Introduction

Misdiagnosis remains a critical challenge in healthcare, leading to significant patient harm and increased healthcare costs [Newman-Toker *et al.*, 2024; Dixit *et al.*, 2023]. In clinical practice, decision-making relies on integrating diverse information sources, yet existing AI-assisted diagnostic systems struggle to effectively process and reason across multiple modalities [Lee *et al.*, 2021; Microsoft, 2024; Rao *et al.*, 2024]. To address this, we present MedRAG [Zhao *et al.*, 2025], a smart multimodal healthcare copilot equipped with powerful large language model (LLM) reasoning, designed to enhance medical decision-making through multimodal integration and knowledge graph (KG)-elicited reasoning.

Through interviews with healthcare professionals, we identified key requirements for an effective AI-driven diagnostic assistant. Doctors emphasized that an ideal system should incorporate three primary input modalities to comprehensively support clinical workflows [OpenAI, 2023; Amballa, 2023; Zakka *et al.*, 2024; Wei *et al.*, 2018; Ren *et al.*, 2024]:

- Non-intrusive voice monitoring – Seamlessly captures real-time doctor-patient conversations during consultations without disruption. This enables instant follow-up questioning and context-aware diagnostic recommendations to support decision-making [Belle, 2023].
- General medical queries – Allows doctors to interactively refine differential diagnoses, seek clarifications, and receive personalized treatment suggestions in real time. This supports flexible information retrieval for both clinical and patient-facing decision support.
- Electronic Health Records (EHRs) – Analyzes similar cases to provide reasoning-enhanced diagnostics and personalized treatment recommendations, ensuring data-driven support for complex decisions.

While retrieval-augmented generation (RAG) has been proposed for medical AI applications, existing heuristic-based RAG models often fail to differentiate between diseases with similar manifestations [Wu *et al.*, 2024b; Guu *et al.*, 2020; Edge *et al.*, 2024]. Doctors noted that these models tend to generate vague or incorrect recommendations, lacking structured reasoning capabilities [Zelin *et al.*, 2024; Li *et al.*, 2023; Wu *et al.*, 2024a]. To overcome this limitation, we introduce KG-elicited reasoning, a key technology in MedRAG that enhances diagnostic accuracy by integrating structured medical knowledge with patient data.

MedRAG systematically constructs a hierarchical diagnostic KG, capturing subtle yet critical diagnostic differences. This KG is dynamically queried based on patient-specific manifestations and integrated with retrieved EHRs, allowing the system to reason through uncertainties and generate precise, context-aware diagnostic suggestions. Additionally, MedRAG proactively proposes follow-up questions to refine ambiguous cases, further supporting clinical workflows.

We evaluate MedRAG on both public (DDXPlus) and private (CPDD) datasets collected from Tan Tock Seng Hospital of Singapore. The results demonstrate its superiority over existing RAG approaches in diagnostic accuracy, specificity, and reasoning-based decision support. Our demo highlights how KG-elicited reasoning transforms MedRAG into a powerful, intelligent, and adaptable healthcare copilot, capable of assisting doctors across diverse clinical scenarios.

*Corresponding Authors

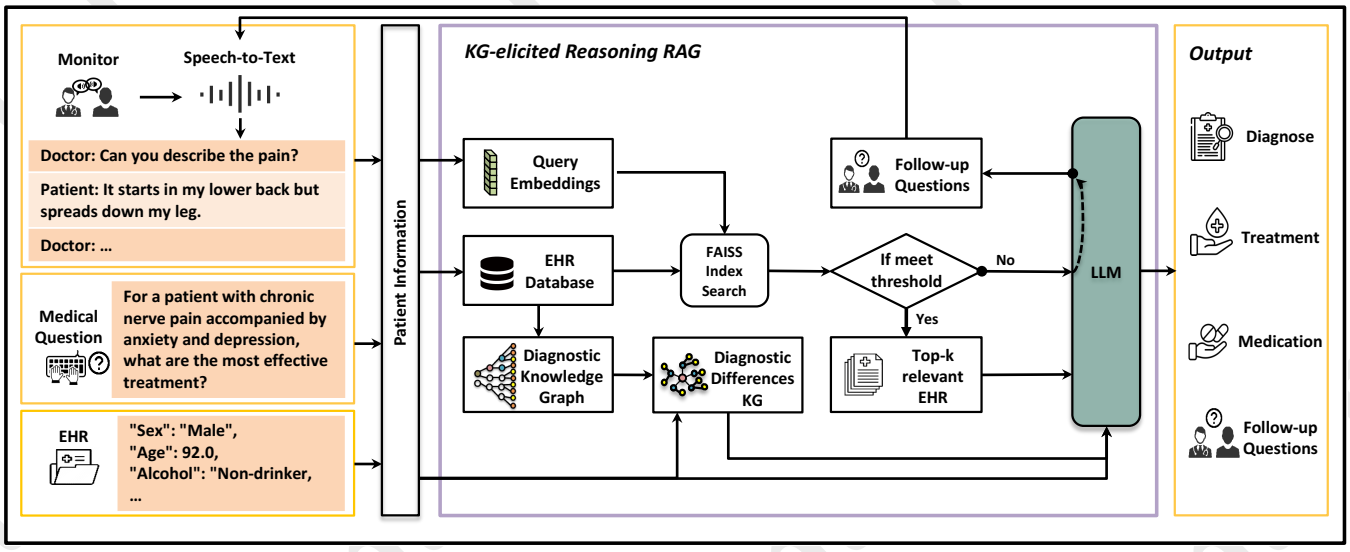

Figure 1: Framework of Multimodal Healthcare Copilot-MedRAG

2 System Design

As shown in Figure 1, MedRAG incorporates three different modes of input, a KG-elicited reasoning RAG module and four outputs. MedRAG can seamlessly support various open-source and close-source LLMs, ensuring high adaptability and easy deployment in medical settings.

2.1 Multimodal Input

MedRAG provides three input modes to accommodate different clinical scenarios, as illustrated in Figure 1. MedRAG monitors doctor-patient conversations during consultations without interruption using Google's Speech-to-Text API in real time [Google Cloud, 2025]. With one-click activation at the beginning of the consultation, MedRAG automatically handles information collection, analysis, follow-up questions and diagnostic recommendations, reducing doctors' cognitive and operational workload while allowing them to focus on patient interaction. Doctors can also upload existing files like undiagnosed EHRs or simply type questions by keyboard. All collected information is processed by the KG-elicited Reasoning RAG for further diagnostic analysis.

2.2 Knowledge Graph-elicited Reasoning RAG

Knowledge graph-elicited reasoning RAG serves as the core analytical module of MedRAG, it constructs a diagnostic KG from the existing EHR database and identifies the most relevant subgraph based on the patient's manifestations [Yang *et al.*, 2025; Lind *et al.*, 1985]. It elicits the reasoning ability of the RAG model by extracting relevant triplets as context, which are then fed to the backbone LLM along with retrieved relevant documents, enabling more accurate and structured diagnostic insights.

Diagnostic Knowledge Graph

Given the EHR database, MedRAG constructs a four-tiered diagnostic KG by clustering diseases with similar manifestations into hierarchical categories while manifestations of each disease are decomposed into unique features [Zhao *et al.*, 2017; Li *et al.*, 2018]. Features, diseases, subcategories, and categories are structured as nodes to form an undirected KG. Further, we apply GPT-4o to augment the differentiation of similar diseases by expanding more unique features of each disease within each subcategory. Given an undiagnosed patient's manifestations, MedRAG identifies the most relevant subcategory, and triplets ⟨disease, relation, feature⟩ associated with the identified subcategory are gathered as contextual information to elicit the reasoning capability of the backbone LLM.

Retrieval-Augmented Generation

To provide backbone LLM with case-specific information and mitigate hallucinations in generated outputs, we apply the RAG method, retrieving relevant documents before generation. In MedRAG, we use the EHR database as retrieval documents, as EHRs are systematically collected and structured within hospital databases. When patients' disease manifestations are fed into MedRAG, the system measures the semantic similarity between input information and EHRs using cosine similarity. The top 3 relevant EHRs are then selected to provide the contextual input for backbone LLM. OpenAI's text-embedding-3-large API is used as the text encoder to generate embeddings for both input information and EHRs.

Proactive Question Generation

When monitoring a medical consultation, MedRAG evaluates whether sufficient information is available for diagnostic reasoning by analyzing the semantic similarity of the input data and determining whether some EHRs meet a predefined threshold. If it is insufficient, MedRAG identifies the most critical unmentioned disease features in the diagnostic KG to differentiate between similar diseases and formulates follow-up questions. Otherwise, MedRAG proceeds to generate diagnostic recommendations.

3 User Interface and Evaluation

We present the user interface (UI) and conduct comprehensive evaluations, including a case study and expert evaluation by doctors, to assess the system's performance.

3.1 UI of MedRAG

We provide a user-friendly interface built with Streamlit and CSS as shown in Figure 2. The left panel displays the chat history, allowing quick access to past consultations. The main panel at right presents three input modes: Speaking, Uploading Files, and Typewriting. The bottom section provides a text input field for direct user queries.

3.2 Case Study

System	Diagnostic Suggestion
Query	Provide diagnosis suggestions for the following patient: Age: 47. Functional status: Difficulty walking [...] Description: Pain from right lower back radiates down to buttock and right posterior lower limb.
Llama3.1-8b **Mixtral-8x7b**	Lumbar Radiculopathy, Sciatica, [...]. It is possible that the patient is experiencing pain due to sciatica.
Qwen2.5-72b	Potential Diagnoses: Sciatica [...]; Lumbar Herniated Disc [...]; Spinal Stenosis: [...].
MedRAG (Ours)	Lumbar canal stenosis. You can further ask: Is the pain worse when standing or walking down hill?

Table 1: Comparison of Diagnostic Suggestions Across Systems

In Table 1, we compare MedRAG with other LLMs including Llama3.1-8b, Mixtral-8x7b and Qwen2.5-72b, which often provide incorrect or ambiguous diagnoses, such as sciatica or radiculopathy, and struggle to distinguish similar conditions. In contrast, MedRAG accurately identifies lumbar canal stenosis and proactively generates follow-up questions to help doctors further refine the diagnosis.

Backbone LLM	Modal	L1	L2	L3
GPT-4o	text	91.87	81.78	73.23
GPT-4o	voice	88.23	78.43	70.58
GPT-3.5-turbo	text	70.56	68.68	50.57
GPT-3.5-turbo	voice	64.70	60.78	45.09

Table 2: Evaluation of Different Modals on CPDD

Furthermore, a detailed demonstration scenario of a medical consultation with the corresponding suggested follow-up question, along with an end-to-end evaluation including voice modality, are provided in Appendix A-B[1] and Table 2.

3.3 Doctor Evaluation

To complement quantitative benchmarks with clinical insights, we incorporated a human evaluation involving four experienced doctors. These experts' feedback provides essential perspective on how MedRAG is perceived in clinical contexts, particularly in terms of trust and usability.

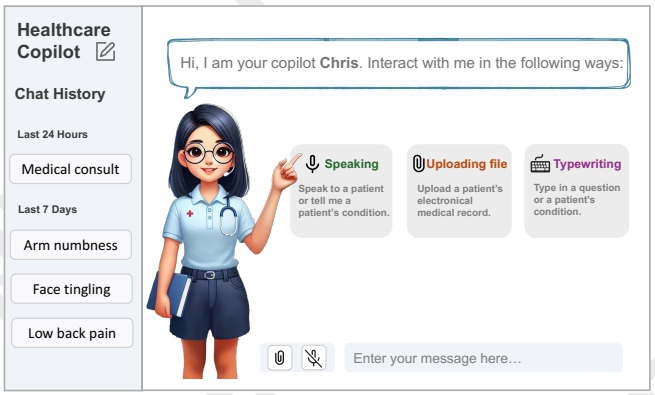

Figure 2: User Interface of MedRAG

For the evaluation, doctors assessed three representative test cases with responses from both MedRAG and GPT-4o, focusing on functional design, user interface, EHR analysis, and medical consultation analysis. Our evaluation uses five Human Factors criteria (e.g., Clinical Relevance and Trust) widely used to assess AI-assisted systems [Choudhury and Shamszare, 2023; Choudhury, 2022; Zhao *et al.*, 2021]. The details of the criteria definitions and specific questions are provided in Appendix C [1].

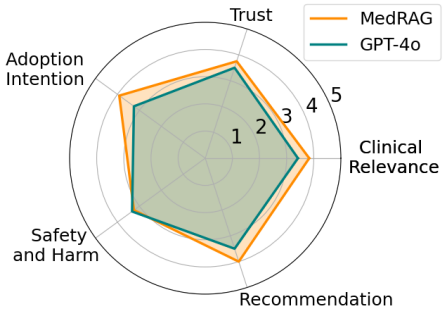

Figure 3: Result of Doctor Evaluation

The comparative results are presented in Figure 3. The results demonstrate that MedRAG outperforms GPT-4o across all criteria, with particularly outstanding performance in Adoption Intention. Some doctors emphasized that, since evidence-based practice is fundamental to medicine [Prasad and others, 2014], MedRAG stood out for its strong emphasis on evidence-based reasoning.

4 Conclusion

MedRAG is a smart multimodal healthcare copilot with powerful LLM reasoning, integrating multimodal inputs and KG-elicited reasoning to enhance diagnostic accuracy and decision support. The results of case studies and doctor evaluation have consistently demonstrated the effectiveness and reliability of MedRAG in medical decision-making contexts.

[1] *https://github.com/SNOWTEAM2023/MedRAG/tree/main/appendix*

Acknowledgments

This research is supported by the Joint NTU-UBC Research Centre of Excellence in Active Living for the Elderly (LILY) and the College of Computing and Data Science (CCDS) at NTU Singapore. It is also partially supported by the Singapore Ministry of Education Academic Research Fund Tier 1 (Grant No. 2017-T1-001-270). This research is also supported, in part, by the National Research Foundation, Prime Minister's Office, Singapore under its NRF Investigatorship Programme (NRFI Award No. NRF-NRFI05-2019-0002). Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not reflect the views of National Research Foundation, Singapore. This research is supported, in part, by the Singapore Ministry of Health under its National Innovation Challenge on Active and Confident Ageing (NIC Project No. MOH/NIC/HAIG03/2017). This research is supported, in part, by the RIE2025 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) (Award I2301E0026), administered by A*STAR, as well as partially supported by Alibaba Group and NTU Singapore through Alibaba-NTU Global e-Sustainability CorpLab (ANGEL). This work is partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

Contribution Statement

Xuejiao Zhao and Siyan Liu made equal contributions.

References

[Amballa, 2023] Durga Prasad Amballa. Ai-powered copilot for healthcare sales agents: Enhancing customer engagement and test recommendations. *Journal of Scientific and Engineering Research*, 10(10):164–167, 2023.

[Belle, 2023] Lin Belle. Openai expands healthcare push with color health's cancer copilot. *The Wall Street Journal*, 2023. Accessed: 2024-09-18.

[Choudhury and Shamszare, 2023] Avishek Choudhury and Hamid Shamszare. Investigating the impact of user trust on the adoption and use of chatgpt: survey analysis. *Journal of Medical Internet Research*, 25:e47184, 2023.

[Choudhury, 2022] Avishek Choudhury. Factors influencing clinicians' willingness to use an ai-based clinical decision support system. *Frontiers in digital health*, 4:920662, 2022.

[Dixit *et al.*, 2023] Ram A Dixit, Christian L Boxley, Sunil Samuel, Vishnu Mohan, Raj M Ratwani, and Jeffrey A Gold. Electronic health record use issues and diagnostic error: a scoping review and framework. *Journal of patient safety*, 19(1):e25–e30, 2023.

[Edge *et al.*, 2024] Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*, 2024.

[Google Cloud, 2025] Google Cloud. Speech-to-text api, 2025. Accessed: Feb 12, 2025.

[Guu *et al.*, 2020] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR, 2020.

[Lee *et al.*, 2021] Ching Hung Lee, Zehao Zhang, and Xuejiao Zhao. A survey of smart healthcare for the elderly based on user requirements and supply accessibility. In *5th International Conference on Crowd Science and Engineering*, pages 108–112, 2021.

[Li *et al.*, 2018] Hongwei Li, Sirui Li, Jiamou Sun, Zhenchang Xing, Xin Peng, Mingwei Liu, and Xuejiao Zhao. Improving api caveats accessibility by mining api caveats knowledge graph. In *2018 IEEE International Conference on Software Maintenance and Evolution (ICSME)*, pages 183–193, 2018.

[Li *et al.*, 2023] Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Cureus*, 15(6), 2023.

[Lind *et al.*, 1985] E. Lind, O. Fausa, K. Elgjo, and E. Gjone and. Clinical manifestations. *Scandinavian Journal of Gastroenterology*, 20(6):665–670, 1985. PMID: 4035286.

[Microsoft, 2024] Microsoft. Microsoft copilot in healthcare. https://www.avanade.com/en/services/artificial-intelligence/ai-copilot-hub/health-ai-copilot, 2024. Accessed: 2024-10-11.

[Newman-Toker *et al.*, 2024] David E Newman-Toker, Najlla Nassery, Adam C Schaffer, Chihwen Winnie Yu-Moe, Gwendolyn D Clemens, Zheyu Wang, Yuxin Zhu, Ali S Saber Tehrani, Mehdi Fanai, Ahmed Hassoon, et al. Burden of serious harms from diagnostic error in the usa. *BMJ Quality & Safety*, 33(2):109–120, 2024.

[OpenAI, 2023] OpenAI. Color health's cancer copilot, 2023. Accessed: 2024-09-18.

[Prasad and others, 2014] Kameshwar Prasad et al. Fundamentals of evidence based medicine. Technical report, Springer, 2014.

[Rao *et al.*, 2024] Haocong Rao, Minlin Zeng, Xuejiao Zhao, and Chunyan Miao. A survey of artificial intelligence in gait-based neurodegenerative disease diagnosis. *arXiv preprint arXiv:2405.13082*, 2024.

[Ren *et al.*, 2024] Zhiyao Ren, Yibing Zhan, Baosheng Yu, Liang Ding, and Dacheng Tao. Healthcare copilot: Eliciting the power of general llms for medical consultation. *arXiv preprint arXiv:2402.13408*, 2024.

[Wei *et al.*, 2018] Sidong Wei, Xuejiao Zhao, and Chunyan Miao. A comprehensive exploration to the machine learning techniques for diabetes identification. In *2018 IEEE 4th World Forum on Internet of Things (WF-IoT)*, pages 291–295. IEEE, 2018.

[Wu *et al.*, 2024a] Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Weidi Xie, and Yanfeng Wang. Pmc-llama: toward building open-source language models for medicine. *Journal of the American Medical Informatics Association*, page ocae045, 2024.

[Wu *et al.*, 2024b] Junde Wu, Jiayuan Zhu, and Yunli Qi. Medical graph rag: Towards safe medical large language model via graph retrieval-augmented generation. *arXiv preprint arXiv:2408.04187*, 2024.

[Yang *et al.*, 2025] Xiao Yang, Xuejiao Zhao, and Zhiqi Shen. A generalizable anomaly detection method in dynamic graphs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(20):22001–22009, Apr. 2025.

[Zakka *et al.*, 2024] Cyril Zakka, Joseph Cho, Gracia Fahed, Rohan Shad, Michael Moor, Robyn Fong, Dhamanpreet Kaur, Vishnu Ravi, Oliver Aalami, Roxana Daneshjou, et al. Almanac copilot: Towards autonomous electronic health record navigation. *arXiv preprint arXiv:2405.07896*, 2024.

[Zelin *et al.*, 2024] Charlotte Zelin, Wendy K Chung, Mederic Jeanne, Gongbo Zhang, and Chunhua Weng. Rare disease diagnosis using knowledge guided retrieval augmentation for chatgpt. *Journal of Biomedical Informatics*, 157:104702, 2024.

[Zhao *et al.*, 2017] Xuejiao Zhao, Zhenchang Xing, Muhammad Ashad Kabir, Naoya Sawada, Jing Li, and Shang-Wei Lin. Hdskg: Harvesting domain specific knowledge graph from content of webpages. In *2017 ieee 24th international conference on software analysis, evolution and reengineering (saner)*, pages 56–67. IEEE, 2017.

[Zhao *et al.*, 2021] Xuejiao Zhao, Huanhuan Chen, Zhenchang Xing, and Chunyan Miao. Brain-inspired search engine assistant based on knowledge graph. *IEEE Transactions on Neural Networks and Learning Systems*, 34(8):4386–4400, 2021.

[Zhao *et al.*, 2025] Xuejiao Zhao, Siyan Liu, Su-Yin Yang, and Chunyan Miao. Medrag: Enhancing retrieval-augmented generation with knowledge graph-elicited reasoning for healthcare copilot. In *Proceedings of the ACM on Web Conference 2025*, pages 4442–4457, 2025.