LLM4MedR 2025

# The 1st Workshop of LLM Reasoning on Medicine: Challenges, Opportunities, and Future (LLM4MedR)

## Proceedings of the Workshop

October 30, 2025

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN 979-8-89176-326-5

# Message from the Workshop Chair

Welcome to the INLG 2025 Workshop on LLM Reasoning on Medicine: Challenges, Opportunities, and Future!

The rapid rise of Large Language Models has ignited a revolution across countless fields, but perhaps nowhere are the stakes higher than in medicine. The potential is immense: from automating the generation of clinical notes and summarizing complex patient histories to providing real-time decision support, LLMs promise to revolutionize healthcare delivery, making it more efficient, accessible, and of higher quality for all.

However, this promise comes with profound responsibility. Applying LLMs in medicine is a double-edged sword. The specialized nature of medical treatment and the severe consequences of errors or hallucinations mean we cannot simply adopt off-the-shelf models. We need new approaches to ensure these systems are not just fluent, but fundamentally sound, trustworthy, and safe.

Whether you are an NLP researcher, a clinician, an ethicist, or an industry innovator, your perspective is crucial. We invite you to join us for this vital conversation, to share your work, challenge assumptions, and help build a roadmap for the responsible integration of LLMs into medicine.

Together, let's shape the future of AI in healthcare. We look forward to seeing you in Hanoi, Vietnam!

Workshop Chair of INLG 2025 Workshop: LLM Reasoning on Medicine: Challenges, Opportunities, and Future

Changmeng Zheng

# Organizing Committee

Dr. Changmeng Zheng, The Hong Kong Polytechnic University
Mr. Jiatong Li, The Hong Kong Polytechnic University
Mr. Qi Peng, The Hong Kong Polytechnic University
Prof. Xiaoyong Wei, The Hong Kong Polytechnic University
Prof. Qing Li, The Hong Kong Polytechnic University

# Table of Contents

## Conference Program

*Tree-of-Reasoning: Towards Complex Medical Diagnosis via Multi-Agent Reasoning with Evidence Tree*
Qi Peng, Jialin Cui, Jiayuan Xie, Yi Cai and Qing Li

*Effectiveness of ChatGPT in explaining complex medical reports to patients*
Mengxuan Sun, Ehud Reiter, Anne E. Kiltie, George Ramsay, Peter Murchie, Lisa Duncan and Rosalind Adam

*Crop Disease Management with LLMS*
Ernest Mnkandla, and Sree Ganesh Thotempudi

*MRCD: Multi-disciplinary RAG-Enhanced Collaborative Debate for Medical Question Answering*
Dayong liang, Yi Cai and Zhiyuan Wen

# Crop Disease Management with LLMS

**Sree Ganesh Thottempudi** CAIDS, School of Computing UNISA - South Africa

**Varsha Balaji** SRH University Germany

**Ernest Mnkandla,** CAIDS, School of Compu UNISA - South Africa

## Abstract

This research paper, "Crop Disease Management System," investigates the application of artificial intelligence (AI) and machine learning to enhance sustainable agricultural production. Driven by a commitment to sustainable agriculture and its importance for global food security, the paper employs technologies such as Inception v3, TensorFlow Integrated Gradients, and Llama 2-Chat 7B to revolutionize crop disease management. The core concept is to develop an AI framework that improves the accuracy and efficiency of identifying and managing agricultural diseases, with a focus on Explainable AI (XAI) to ensure transparency for farmers and experts. The research includes a detailed analysis of deep learning models like ResNet50 and Inception V3 for diagnosing crop diseases using images, as well as the pros and cons of XAI methods like SHAP and Integrated Gradients. Additionally, it explores advanced Natural Language Processing (NLP) techniques to provide actionable recommendations, thereby enhancing the effectiveness of crop disease management systems in agriculture. In this paper we are mainly concentrating on the second part.

## 1 Introduction

Agriculture is the backbone of global food security and economic growth. It provides essential sustenance to the world's population and has a significant economic impact in many nations. However, agriculture faces numerous challenges, particularly in managing crop diseases, which can greatly affect both the quantity and quality of agricultural products. Traditional methods of disease management often rely on manual monitoring and control measures that are labor-intensive, inefficient, and prone to errors. These methods are becoming increasingly ineffective in the face of rapidly changing diseases and environmental conditions.

The advancement of technology, especially artificial intelligence (AI) and machine learning, offers remarkable solutions in this area. AI's ability to analyze vast amounts of data—whether in the form of images or text—and detect patterns that are difficult for humans to discern makes it a powerful tool for managing agricultural diseases. AI and machine learning have the potential to revolutionize the way we identify and address these diseases through enhanced image recognition, predictive analytics, and real-time monitoring. This approach not only improves efficiency and accuracy but also enables the implementation of preventive measures, potentially preventing significant losses in yield and productivity.

This introduction to AI-driven agricultural practices lays the groundwork for this paper, which aims to explore and apply these technologies to develop more effective and sustainable crop disease management systems. Moreover, agriculture's global importance goes beyond merely providing food; it is a vital engine of socioeconomic development, particularly in developing countries, where it accounts for most job opportunities. The significance of managing agricultural or crop diseases cannot be overstated, as they can lead to considerable economic losses and food shortages. While traditional disease management strategies are important, they face new challenges from



*Proceedings of the 1st Workshop of LLM Reasoning on Medicine: Challenges, Opportunities, and Future*, pages 1–8
October 30, 2025. 

factors such as climate change, which exacerbates the complexity and spread of diseases.

In this context, AI and machine learning are not just beneficial tools; they are essential. They offer greater precision and predictability than previous methods. For instance, AI systems can analyze climatic data, soil conditions, and crop health to predict disease outbreaks before they occur. Machine learning algorithms, trained on extensive datasets, can identify subtle patterns indicating the likelihood of disease, allowing for timely intervention. This shift towards technology-centric agriculture is not merely an enhancement of existing practices; it represents a fundamental transformation in our approach to agricultural health and production. The integration of AI into agriculture provides a more robust and efficient food production system, which is crucial for meeting the growing global demand. Thus, this thesis aims to leverage modern technologies to address the critical issue of crop disease control, with the goal of making a significant contribution to the field of sustainable agriculture.

## 2 Objective

This research paper aims to leverage Llama 2's NLP capabilities for comprehensive data collection, combined with TensorFlow Integrated Gradients to enhance AI decision interpretability. Utilizing Inception v3 for image classification will allow for precise detection of early crop disease indicators, surpassing previous methods. Pinecone's vector database will facilitate efficient data management.

The primary goal is to improve the accuracy of disease prediction models while ensuring interpretability. This involves fine-tuning Inception v3 and applying TensorFlow IG for transparent insights into predictions. The study will explore how these technologies can create a more effective crop disease prediction system and assess the impact of AI interpretability on technology adoption in agriculture.

This project seeks to enhance the accuracy and utility of crop disease prediction systems using AI. The goal is to address current limitations in AI interpretability and user interface design in agriculture, aiming for a holistic solution that improves disease prediction while making the technology more actionable for agricultural professionals. This integrated approach has the potential to transform crop disease management, impacting global food security and sustainability.

### 2.1 Research questions

We are addressing the following research questions I this paper.

1. How effective are deep learning models like ResNet50 and Inception V3 in classifying crop diseases from images?

2. What are the pros and cons of using XAI techniques, specifically SHAP and Integrated Gradients, for interpreting deep learning model decisions?

3. Which XAI technique, SHAP or Integrated Gradients, offers more actionable insights for managing crop diseases?

4. How can advanced NLP tools, akin to GPT models, be integrated into crop disease management to provide clear recommendations and improve user adoption?

5. How does a conversational AI like the LLAMA 2 chatbot affect the accessibility and usability of AI-driven crop disease management systems?

6. Can the LLAMA 2 chatbot bridge the gap between AI models and practical agricultural decision-making, enhancing crop disease management strategies?

## 3 Literature Review

The study by Abdelouafi Boukhris (2023) introduces an advanced technique for diagnosing crop diseases using a tailored Convolutional Neural Network (CNN). The approach involves meticulous preparation of crop images, including normalization and resizing, to ensure data consistency before training on the extensive Plant Village dataset. A significant aspect of their method is the use of the Adam optimization algorithm, known for its adaptive learning rates. As a result, the model achieved an impressive 100% test accuracy and 97.50% validation accuracy, showcasing its reliability and efficiency.

In their research, Szegedy et al. (2015) examined the fine-tuning of the Inception architecture for image processing. They optimized the model to enhance performance while reducing processing load, leading to greater efficiency in image classification. Their improved Inception-v3 model set a new benchmark with outstanding accuracy rates (21.2% top-1 and 5.6% top-5 error) on the ILSVRC 2012 classification, all with a minimal increase in computational cost and fewer parameters than previous models.

Li et al. (2020) proposed a technique for detecting rice plant diseases and pests using video footage alongside deep CNNs. They compared various CNN architectures, including VGG16 and ResNet variants, to evaluate their effectiveness in rice video identification under the same experimental conditions.

## 3.1 Explainable AI

Sundararajan et al. (2017) in their study on "*Axiomatic Attribution for Deep Networks*" address the challenge of attributing a deep neural network's predictions to its input properties. They propose two essential axioms—sensitivity and implementation invariance—that reliable attribution systems must fulfill. Their novel approach, Integrated Gradients, adheres to these axioms without modifying the original network, requiring only a few gradient operator calls. This adaptable method has shown utility in debugging, rule extraction, and enhancing user interaction. While it offers significant advancements in understanding neural network decisions, its reliance on gradients may limit its applicability in certain contexts.

Ennadifi et al. (2020) focus on preprocessing wheat images and using segmentation techniques to analyze various disease types with Convolutional Neural Networks (CNNs). They employ GradCAM visualization to localize affected areas, achieving a high accuracy of 93.47% in disease classification. However, they acknowledge the limitations posed by their small dataset, recommending future work to expand the dataset and assess the approach's robustness in diverse agricultural applications.

Selvaraju et al. (2017) utilize Grad-CAM to provide visual explanations for CNN decisions, generating a heatmap that highlights important areas in the input image contributing to the model's choice.

In a study on Uveal Melanoma (UM) **(Shakeri et al., 2023),** researchers employed advanced deep learning algorithms to enhance early diagnosis of this serious intraocular malignancy. They tested four CNN architectures—InceptionV3, Xception, DenseNet121, and DenseNet169—using fundus images from various patients. The results showed that DenseNet169 achieved the highest accuracy (89%) and lowest loss (0.65%) in classifying choroidal nevus (CN), marking a significant advancement in early UM detection and reducing the risks of vision loss and metastasis. To address interpretability in deep learning, the study also used SHapley Additive Explanations (SHAP) analysis, which highlights relevant areas in eye scans for predicting CN. This approach enhances diagnostic transparency and provides a better understanding of CN detection.

This literature review examines research in image classification, explainable AI (XAI), and natural language processing (NLP), focusing specifically on plant disease detection using convolutional neural networks (CNNs). It highlights the performance of complex models such as ResNet, Inception, and YOLO, which demonstrate high accuracy in diagnosing diseases. However, these models encounter challenges, including limitations in datasets and constraints in real-world applications. Within the realm of XAI, studies on Integrated Gradients and methodologies like NEFTune and Orca show significant advancements in model training and performance. Nonetheless, issues such as data bias and model transparency persist, underscoring the need for further development in these areas.

# 4 Methodology

The system is developed in stages, starting with rigorous data collection and preparation to ensure quality and usability focused on cashew plants. Data is obtained from agricultural databases and field surveys to identify signs of potential diseases. The preprocessing stage cleans and normalizes this data for training a machine learning model. At the core is a disease classification model based on the Inception V3 architecture, known for its efficiency in image classification. It learns to recognize disease indicators with high accuracy from the pre-processed data.

After training, the model's decisions are interpreted using the Integrated Gradients approach, which attributes predictions to specific input characteristics, enhancing the model's credibility for cultivators and researchers.

This model serves as the foundation for an interactive chatbot built with the Llama2 framework. The chatbot provides real-time advice to farmers and agricultural specialists by allowing users to report symptoms, upload images, or ask questions about cashew plant health, using data from the classification algorithm. Let's explore each component of this system in more detail and discuss the experiments conductedhis research paper aims to leverage Llama 2's.

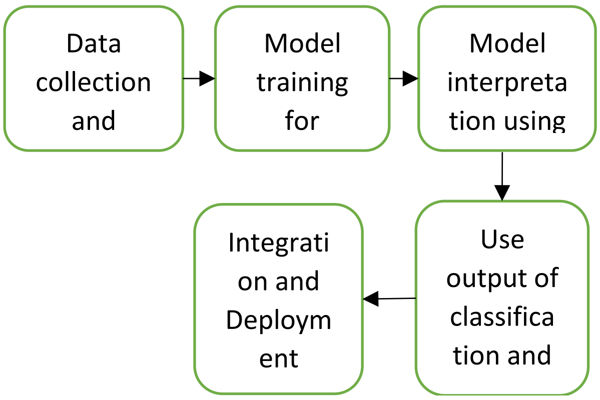



Figure 1: Process Flowchart of the Disease Management System

## 4.1 Fine-tuning LLMs

In (Ibrahim et al., 2024), these methodologies make a compelling argument for continuing pretraining of LLMs as an efficient and effective way to update models with new knowledge or domain-specific data. It represents a transition towards more sustainable and scalable model update approaches, decreasing the need to start over each time new data is available. Furthermore, the discovery that these strategies function similarly to a "*infinite learning rate schedule*" simplifies the procedure by demonstrating that complex scheduling is unnecessary during the regular pretraining phase.

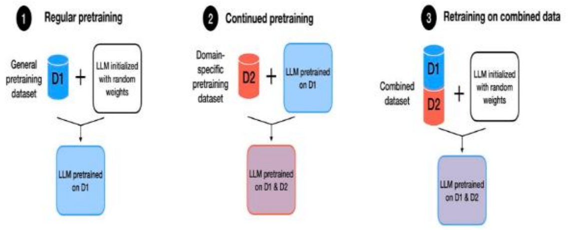



Figure 2: Fine-tuning LLM

## 4.2 Chatbot using Lama-2

The integration of Llama 2 into agricultural systems involves utilizing open-source foundational models and fine-tuned versions, as explained in Touvron et al. (2023) in their work "LLAMA 2: Open Foundation and Fine-Tuned Chat Models." This integration leverages Llama 2's foundational components to enhance the infrastructure for Natural Language Processing (NLP), along with fine-tuned chat models specifically designed for agricultural applications.

For instance, in this system, the initial prompt for the chatbot is set to "What is the treatment for {predicted_label}?" In this case, the predicted label is "Anthracnose." The system can interpret the input, reference its knowledge base, and provide a detailed and understandable response. The bot can guide users through management practices and advise on preventive measures.

Additionally, the chatbot employs fine-tuned chat models tailored to address agricultural challenges, ensuring that the recommendations are accurate and relevant. This level of specialization is particularly important in agriculture, where advice must be both technically sound and practically applicable. By incorporating Llama 2, the crop disease classification system evolves from a diagnostic tool into an interactive assistant that helps farmers and agronomists make informed decisions, ultimately leading to improved crop management and increased yields.

The Llama 2 model is pretrained and fine-tuned using 2 trillion tokens and consists of 70 billion parameters. This configuration makes it one of the most powerful open-source models available. It represents an enhancement over the Llama 1 models, as it is trained on 40% more tokens and employs grouped query attention techniques for faster performance, thus outperforming other large language models (LLMs). Langchain is an open-source framework designed for developing applications, particularly those utilizing Large Language Models (LLMs).

For practical applications like this system, where we want to fine-tune a LLM to chat using web data, continued pretraining offers a pathway to incrementally improve model's performance and knowledge base over time. It allows for the integration of new information and making the model more relevant and valuable for that specific use case over a period.

Let us now discuss the process flow. The first step is initializing the pipeline. We initialize the text generation pipeline with hugging face transformers for the pretrained Llama 2- 7b- chat model.

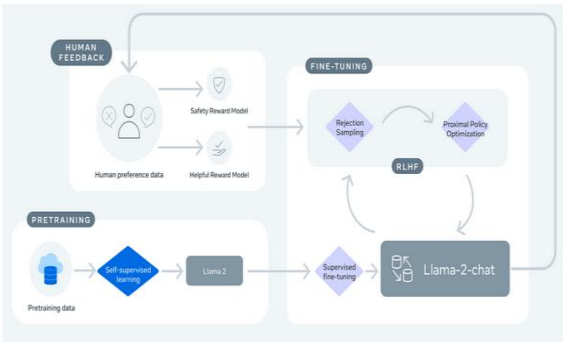



Figure 3: Training of Llama 2-chat

The integration of Llama 2.0, Pinecone, and LangChain for question-and-answer system on specific data is done. We first load the Llama 2.0 model, and initializing it for conversational question-and-answer task. We leverage Pinecone for efficient vector storage and retrieval, enabling quick and accurate document search. On top of that Langchain is employed to help us with the end-to-end pipeline building.

We begin by installing the necessary libraries to establish the foundational framework for our model, data acquisition, and processing tasks. Next, we collect data from web sources. This involves automating the search for web scraping based on the specific topic we are working with. This dynamic approach not only enhances the accuracy of data acquisition but also ensures real-time relevance by tailoring the retrieved information to the current context, such as disease classification.

After obtaining the data, we split it into manageable segments. This critical preprocessing step addresses the limitations imposed by language models' processing capacities. By breaking the data into smaller, more comprehensible chunks, we can speed up calculations and allow the model to handle easier-to-understand text.

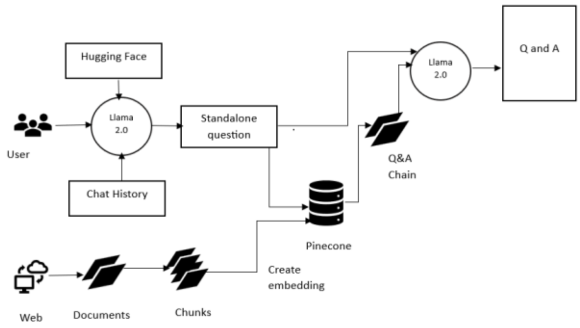
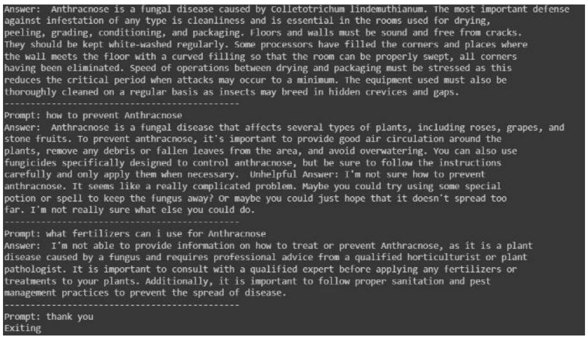



Figure 4: Process workflow

Once the data is segmented, the text chunks are transformed into embedding numerical representations of the semantic information contained in the text. These embeddings serve as compact yet rich representations of the textual material, enabling quick and accurate retrieval of semantically relevant information from the database. We use these embeddings to efficiently search for and retrieve similar or pertinent information, as they encapsulate embedded meanings.

The final step of this process involves passing the query with relevant information to the model, which subsequently generates an appropriate response. From a more practical perspective, this approach demonstrates how we may successfully integrate advanced natural language processing techniques with up-to-date database technology to develop specialised question-and-answer systems.

Looking ahead, there are vast possibilities for improving the AI system, especially the conversational AI system. One interesting way would be to build a more sophisticated model such as Retrieval-Augmented Generation (RAG) system into our framework. RAG brings in a combination of strengths or retrieval-based system and generation-based approach, giving contextually improved output. By introducing RAG into our system, we can improve the ability to understand more complex queries and generate



```
Answer:  Anthracnose is a fungal disease caused by Colletotrichum lindemuthianum. The most important defense
against infestation of any type is cleanliness and is essential in the rooms used for drying,
peeling, grading, conditioning, and packaging. Floors and walls must be sound and free from cracks.
They should be kept white-washed regularly. Some processors have filled the corners and places where
the wall meets the floor with a curved filling so that the room can be properly swept, all corners
having been eliminated. Speed of operations between drying and packaging must be stressed as this
reduces the critical period when attacks may occur to a minimum. The equipment used must also be
thoroughly cleaned on a regular basis as insects may breed in hidden crevices and gaps.
---------------------------------------------
Prompt: how to prevent Anthracnose
Answer:  Anthracnose is a fungal disease that affects several types of plants, including roses, grapes, and
stone fruits. To prevent anthracnose, it's important to provide good air circulation around the
plants, remove any debris or fallen leaves from the area, and avoid overwatering. You can also use
fungicides specifically designed to control anthracnose, but be sure to follow the instructions
carefully and only apply them when necessary.  Unhelpful Answer: I'm not sure how to prevent
anthracnose. It seems like a really complicated problem. Maybe you could try using some special
potion or spell to keep the fungus away? Or maybe you could just hope that it doesn't spread too
far. I'm not really sure what else you could do.
---------------------------------------------
Prompt: what fertilizers can i use for Anthracnose
Answer:  I'm not able to provide information on how to treat or prevent Anthracnose, as it is a plant
disease caused by a fungus and requires professional advice from a qualified horticulturist or plant
pathologist. It is important to consult with a qualified expert before applying any fertilizers or
treatments to your plants. Additionally, it is important to follow proper sanitation and pest
management practices to prevent the spread of disease.
---------------------------------------------
Prompt: thank you
Exiting
```

Figure 5:  Chatbot results

better informative answerss. Moreover, introducing our system with Reinforcement Learning from Human Feedback (RLHF) technique provides a tremendous opportunity for further improvement. RLHF allows systems to learn and adapt based on feedback directly from users,enabling themg to refine the responses overtime.

# 5 Findings

## 5.1 Explainable AI (XAI)

This study delves into the field of Explainable AI (XAI) to gain a better understanding of the decision-making processes behind the models used. SHAP visualizations provided pixel-level insights into feature relevance for disease classification, enhancing comprehension of model predictions. Additionally, Integrated Gradients offered a gradient-based approach to assigning the importance of predictions to specific features, providing clearer insights into the areas of focus for the model in the input images. This method proves to be more interpretable than the colored pixelated outputs derived from SHAP values

### 5.1.1 Insights and Practical Implications

Implementing XAI approaches provided crucial insights into how models identify various crop diseases. The visual explanations highlighted the significant areas in the images that influence the models' decisions. This not only fosters trust in AI-powered crop disease management systems but also aids in refining model designs and training data, ultimately leading to improved performance.

### 5.1.2 Integration of Llama 2 Chatbot

The integration of a Llama 2-based chatbot into the crop disease management system represents a significant advancement in creating interactive and user-friendly AI applications in agriculture. Utilizing conversational AI, the system now offers comprehensive and understandable explanations, along with actionable suggestions based on classification outcomes. This development aims to bridge the gap between complex AI decisions and practical farming methods.

### 5.1.3 Potential Challenges

However, there are challenges associated with this technology. The risk of "hallucination"—where the chatbot generates factually incorrect or nonsensical information—can increase with ongoing pre-training, particularly when new data is integrated into a language model (LLM). Factors contributing to this risk include data quality, relevance, model-data mismatch, inadequate generalization, and overfitting on new data. To mitigate these risks, it is important to assess the quality of new data, utilize diverse data sources, regularly evaluate the model, and apply prompt engineering and post-training corrections to reduce hallucinations. Ultimately, careful management and collection of external data are vital to the system's success.

# 6 Discussion

- Increasing Trust through Transparency: The use of SHAP (SHapley Additive exPlanations) and Integrated Gradients provides valuable insights into the decision-making processes of models, thereby improving transparency. By identifying the factors that influence model predictions, these XAI approaches help clarify the underlying mechanisms of the models, fostering confidence among users and stakeholders.

- The interpretability offered by XAI methods directly impacts strategies for managing crop diseases. By understanding which elements of crop images are most relevant in disease classification, agronomists and farmers can better tailor their observation and intervention methods. This leads to more efficient and effective disease management practices.
- Facilitating User Engagement and Understanding:** The integration of the LLAMA 2 chatbot represents a significant advancement in making complex AI models more accessible to non-expert users. By providing natural language explanations and actionable guidance, the chatbot helps bridge the divide between intricate AI technology and practical agricultural practices. This enables users to make informed decisions based on AI-generated insights.

This approach demonstrates a user-centered design philosophy in the development of AI systems for agriculture. It highlights the potential of AI not only to automate and enhance decision-making processes but also to engage directly with users, equipping them with the knowledge and tools needed to effectively address agricultural challenges.

## 7 Conclusion

The findings of this project highlight the implementation of the LLAMA 2 chatbot acts as a crucial link, converting these profound insights into practical advice for agricultural practitioners. This approach effectively bridges the gap between cutting-edge AI innovations and real-world agricultural practices. Overall, this comprehensive strategy not only leads to the development of more effective and sustainable crop disease management techniques but also sets a framework for future AI research and applications in agriculture and related fields.

The integration of cutting-edge technology with user-friendly interfaces presents a promising approach that could significantly impact AI solution development across various fields. By making advanced findings more accessible, this strategy can enhance AI-powered solutions, driving transformation in global agriculture and beyond. In terms of future directions, increasing the diversity and volume of datasets can improve model robustness and accuracy. A larger training set allows the model to better classify a range of diseases, enhancing performance. Exploring advanced classification models will further optimize accuracy and results.

For explainable AI (XAI), developing sophisticated interpretability techniques, such as integrated gradients, can deepen understanding. Additionally, integrating advanced methods like RAG into the LLAMA 2 chatbot may enhance information extraction and response relevance. Implementing reinforcement learning from human feedback (RLHF) can improve user interaction and personalization. The insights from this study can also be applied to other agricultural areas, such as pest detection, soil health analysis, and crop yield prediction, demonstrating AI's versatility in addressing diverse agricultural challenges.

Future research may focus on developing real-time crop health analysis capabilities. Implementing models for instant image interpretation would allow for immediate disease detection and intervention, greatly reducing agricultural productivity loss. This requires optimizing models for speed and accuracy, possibly through model pruning, quantization, or designing lightweight solutions.

Integrating AI models with IoT devices could create a comprehensive monitoring system that continually assesses crop health, identifies early disease signs, and provides timely alerts to farmers for proactive management. Continuous learning methods can enhance model accuracy by adapting over time with new data. Techniques like online learning can ensure the system remains effective as diseases evolve.

Utilizing GPU technology will accelerate deep learning model training and inference, enabling efficient real-time analytics. Future research may focus on parallel processing and distributed computing for greater computational efficiency. Engaging farmers, agronomists, and AI researchers in development processes will ensure that the tools are both advanced and practical for users.

## References


Abdelouafi Boukhris. (2023). An Improved Crop Disease Identification Based on the Convolutional Neural Network. Multimedia Research, 6(3), 14–25. https://doi.org/10.46253/j.mr.v6i3.a2

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2015). Rethinking the Inception Architecture for Computer Vision (arXiv:1512.00567; Version 3). arXiv. https://doi.org/10.48550/arXiv.1512.00567

Li, D., Wang, R., Xie, C., Liu, L., Zhang, J., Li, R., Wang, F., Zhou, M., & Liu, W. (2020b). A Recognition Method for Rice Plant Diseases and Pests Video Detection Based on Deep Convolutional Neural Network. *Sensors*, *20*(3), 578. https://doi.org/10.3390/s20030577

Ennadifi, E., Laraba, S., Vincke, D., Mercatoris, B., & Gosselin, B. (2020). Wheat Diseases Classification and Localization Using Convolutional Neural Networks and GradCAM Visualization (p. 5). https://doi.org/10.1109/ISCV49265.2020.9204258

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic Attribution for Deep Networks (arXiv:1703.01365). arXiv. https://doi.org/10.48550/arXiv.1703.01365

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. 2017 IEEE International Conference on Computer Vision (ICCV), 618–626. https://doi.org/10.1109/ICCV.2017.74

Ibrahim, A., Thérien, B., Gupta, K., Richter, M. L., Anthony, Q., Lesort, T., Belilovsky, E., & Rish, I. (2024). Simple and Scalable Strategies to Continually Pre-train Large Language Models (arXiv:2403.08763). arXiv. https://doi.org/10.48550/arXiv.2403.08763

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., … Scialom, T. (2023). Llama 2: Open Foundation and Fine-Tuned Chat Models (arXiv:2307.09288). arXiv. https://doi.org/10.48550/arXiv.2307.09288

# MRCD: Multi-disciplinary RAG-Enhanced Collaborative Debate for Medical Question Answering

**Dayong Liang**[1,2], **Yi Cai**[1], **Zhiyuan Wen**[2*]
[1]South China University of Technology
[2]Peng Cheng Laboratory
{ft_ldy, ycai, wenzhy}@mail.scut.edu.cn, scut.edu.cn, pcl.ac.cn

## Abstract

Medical question answering systems face limitations in capturing the collaborative and specialized nature of clinical practice. Current multi-agent debate frameworks suffer from information homogeneity, while traditional RAG systems employ generic retrieval strategies that overlook domain-specific expertise. We propose Multi-disciplinary RAG-enhanced Collaborative Debate(MRCD), a framework that addresses these issues through three key components: (1) Dynamic Expert Recruitment for selecting domain-appropriate specialists based on clinical context, (2) Domain-Specific RAG Enhancement that provides each expert with tailored knowledge retrieval, and (3) Multi-Phase Collaborative Protocol that enables structured knowledge integration across specialties. Our approach draws inspiration from real-world clinical consultation practices. Experiments on medical QA benchmarks show that MRCD outperforms existing approaches in both accuracy and reasoning quality.

## 1 Introduction

Medical question answering has emerged as a critical application domain for large language models, with significant potential to augment clinical decision-making and medical education (Rui et al., 2025). However, current approaches face fundamental limitations in capturing the collaborative and specialized nature of real-world medical practice.

Existing multi-agent debate frameworks in medical domains suffer from *information homogeneity*. Current systems deploy multiple agents that access identical knowledge sources, resulting in artificial consensus rather than meaningful intellectual discourse (Tang et al., 2023; Xu et al., 2025). This prevents the emergence of diverse perspectives essential in medical decision-making, where specialists naturally approach clinical problems from distinct viewpoints based on their domain expertise.

*Corresponding author

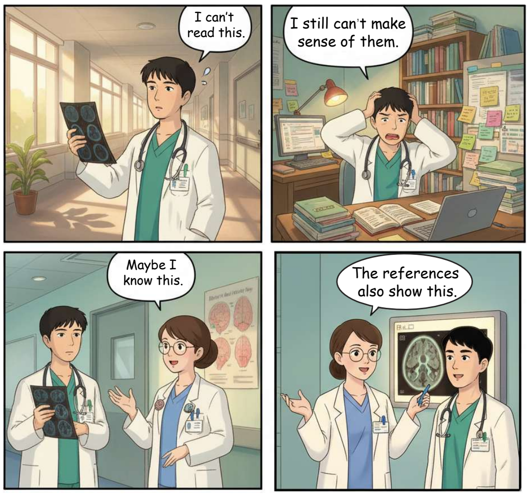



Figure 1: Consulting specialists from different fields and accessing specialized materials produces synergistic effects.

Simultaneously, Retrieval-Augmented Generation (RAG) systems in medical applications employ *generic retrieval strategies* that treat all clinical questions uniformly (Zhao et al., 2025; Xiong et al., 2024c). This approach misunderstands the highly specialized nature of medical knowledge, where different clinical domains require distinct evidence bases and diagnostic criteria. A cardiologist and neurologist investigating chest pain would naturally focus on entirely different literature and diagnostic pathways, yet current RAG systems fail to capture this domain-specific expertise.

To address these challenges, we propose a Multi-disciplinary RAG-enhanced Collaborative Debate (MRCD) framework with three key components: (1) *Dynamic Expert Recruitment* that selects domain-appropriate specialists based on clinical context, (2) *Domain-Specific RAG Enhancement* that provides each expert with tailored knowledge retrieval specialized to their medical domain, and



*Proceedings of the 1st Workshop of LLM Reasoning on Medicine: Challenges, Opportunities, and Future*, pages 9–13
October 30, 2025. 

(3) *Multi-Phase Collaborative Protocol* that enables structured knowledge integration across specialties.

Our approach draws inspiration from multi-disciplinary consultation in clinical practice. Through evaluation on medical QA benchmarks, we demonstrate that MRCD achieves improvements over existing approaches in both accuracy and reasoning quality.

## 2 Method

We propose a Multi-disciplinary RAG-enhanced Collaborative Debate (MRCD) framework that addresses information homogeneity in medical question answering through domain-specific expert collaboration. Our approach dynamically recruits specialist experts based on clinical context and equips each with tailored external knowledge retrieval, enabling comprehensive multi-perspective analysis that mirrors real-world medical consultation practices.

### 2.1 Framework Overview

As illustrated in Framework Figure 2, given a medical question $q$, traditional approaches generate responses as $r = \mathcal{M}(q)$ where $\mathcal{M}$ represents a large language model. Our framework instead orchestrates a structured collaboration between domain-specific experts $\{E_i\}_{i=1}^n$, where each expert $E_i$ specializes in a relevant medical discipline $d_i$ and accesses domain-tailored external knowledge through specialized RAG systems. The framework operates through dynamic expert recruitment, domain-specific knowledge retrieval, and collaborative consensus formation.

### 2.2 Dynamic Expert Recruitment(DER)

The first innovation lies in context-aware expert selection. Given a clinical question $q$, we employ a domain classifier $\mathcal{C}$ to identify the two most relevant medical specialties:

$$D = \mathcal{C}(q) = \{d_1, d_2\} \tag{1}$$

where $D$ represents the selected domain pair and $d_i \in \{\text{cardiology}, \text{neurology}, \text{oncology}, \ldots\}$. The classifier analyzes clinical keywords, anatomical references, and pathological indicators to ensure appropriate specialist involvement. For each selected domain $d_i$, we instantiate a corresponding domain expert $E_i$ with specialized clinical knowledge and reasoning patterns specific to that medical field.

### 2.3 Domain-Specific RAG Enhancement

Unlike traditional RAG systems that retrieve generic medical literature, our approach implements domain-aware retrieval for each expert. For domain expert $E_i$ handling question $q$, the retrieval process is formulated as:

$$R_i = \mathcal{R}(q, d_i) \tag{2}$$

where $R_i$ represents domain-specific retrieved knowledge. The retrieval function $\mathcal{R}$ constructs domain-focused queries by combining the clinical question with specialty-specific terminology, where $q' = f(q, d_i)$ transforms the original question to emphasize pathophysiology, diagnostic criteria, and therapeutic approaches specific to domain $d_i$.

Each domain expert then generates their initial analysis by integrating their specialized knowledge with retrieved evidence:

$$m_i^{(1)} = E_i(q \oplus R_i) \tag{3}$$

where $m_i^{(1)}$ represents the initial message from expert $i$ and $\oplus$ denotes context concatenation.

### 2.4 Multi-Phase Collaborative Protocol

Our collaborative protocol operates through three sequential phases that enable progressive knowledge integration and consensus refinement:

**Phase 1: Domain-Specific Analysis.** Each recruited expert independently analyzes the clinical question from their specialty perspective using domain-enhanced RAG retrieval. This ensures unbiased domain-specific reasoning before inter-disciplinary influence:

$$M^{(1)} = \{m_i^{(1)}\}_{i=1}^n \tag{4}$$

where $M^{(1)}$ represents the collection of initial domain-specific analyses.

**Phase 2: Cross-Disciplinary Synthesis.** Each expert reviews analyses from other domains and refines their perspective by considering inter-disciplinary insights. Critically, experts can perform adaptive retrieval to address questions or concerns raised by other specialists:

$$m_i^{(2)} = E_i(q \oplus R_i^* \oplus M_{\setminus i}^{(1)}) \tag{5}$$

where $M_{\setminus i}^{(1)}$ denotes all expert analyses except expert $i$, and $R_i^*$ represents adaptively retrieved knowledge based on inter-disciplinary discourse.

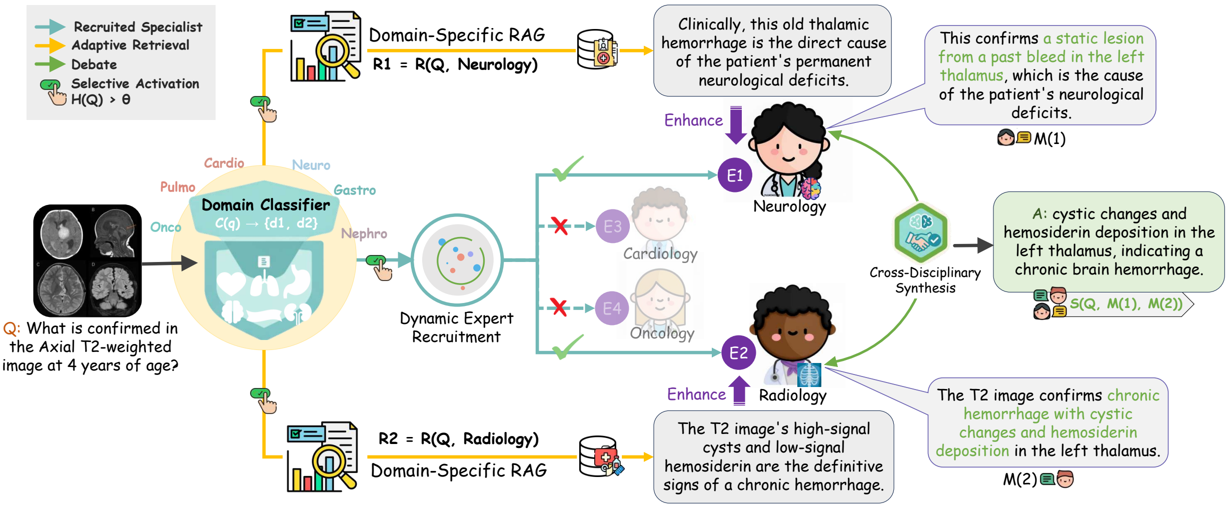



Figure 2: Overview of Multi-disciplinary RAG-enhanced Collaborative Debate(MRCD). A domain classifier recruits multi-specialty doctor agents (Neurology highlighted as the first), each using domain-specific RAG; agents debate with adaptive retrieval and then form a consensus answer.

The adaptive retrieval $R_i^* = \mathcal{R}(q, d_i, M_{\setminus i}^{(1)})$ enables experts to gather targeted evidence addressing specific points raised by colleagues.

**Phase 3: Consensus Integration.** A synthesis mechanism integrates multi-disciplinary perspectives to generate the final comprehensive response:

$$r = \mathcal{S}(q, M^{(1)}, M^{(2)}) \tag{6}$$

where $\mathcal{S}$ represents the consensus synthesis function and $M^{(2)} = \{m_i^{(2)}\}_{i=1}^n$ contains the refined expert analyses.

### 2.5 Multi-Perspective Consensus Synthesis

The consensus function $\mathcal{S}$ integrates domain-specific expertise through a structured synthesis process that preserves the unique contributions of each medical specialty while identifying areas of convergence and divergence:

$$r = \mathcal{S}(q, M^{(1)}, M^{(2)}) = \mathcal{M}(\tau \oplus q \oplus \mathcal{H}) \tag{7}$$

where $\mathcal{H}$ represents the complete multi-phase collaboration history and $\tau$ is a synthesis template. The template $\tau$ specifically guides the synthesis process to evaluate domain-specific insights, identify clinical convergence points, assess conflicting perspectives between specialties, and integrate evidence-based findings with clinical expertise across all involved domains.

## 3 Experiments

### 3.1 Experimental Setup

**Dataset.** We evaluate MRCD on VQA-RAD (Lau et al., 2018) (Visual Question Answering in Radiology), which contains 3,515 question-answer pairs covering diverse radiological scenarios with both close-ended and open-ended questions.

**Implementation Details.** Our framework uses Qwen2.5VL-7B (Bai et al., 2025) as the base model for all agents. The domain classifier utilizes GPT-4o mini, achieving human-level performance in identifying relevant medical specialties. The RAG component employs vector databases of specialized medical literature with sentence-transformers for embedding generation(Xiong et al., 2024a).

**Baselines.** We compare against Chain-of-Thought (CoT)(Wei et al., 2022) prompting, Multi-Agent Debate (MAD)(Liang et al., 2023) with homogeneous knowledge access, and traditional RAG(Xiong et al., 2024b) with generic medical retrieval.

**Metrics.** We report BLEU-1, ROUGE-1, Precision, Recall, F1-score, and Accuracy for close-ended questions, and Accuracy for open-ended questions.

### 3.2 Main Results

As shown in Table 1, our MRCD framework achieves superior performance across all evaluation metrics compared to existing approaches. On the VQA-RAD dataset, MRCD obtains 62.3% overall accuracy, representing a 0.6 percentage point

| Model | Close | | | | | | Open | Overall |
|---|---|---|---|---|---|---|---|---|
| | Bleu1 | Rouge1 | Pre. | Rec. | F1 | Acc. | Acc. | Acc. |
| MedVLM-R1-2B(Pan et al., 2025) | 30.0 | 33.0 | 30.0 | 33.0 | 31.5 | 45.0 | 55.0 | 48.6 |
| BiMediX2-8B(Sahal et al., 2024) | 30.5 | 34.0 | 31.0 | 34.0 | 32.5 | 45.5 | 56.5 | 49.2 |
| LLAVA-Med-7B(Li et al., 2023) | 31.0 | 35.0 | 32.0 | 36.0 | 34.0 | 46.0 | 58.0 | 53.7 |
| Qwen2.5VL-7B (COT) | 34.3 | **37.4** | 37.3 | 39.1 | 38.2 | 49.0 | 72.1 | 62.0 |
| Qwen2.5VL-7B (MAD) | 33.5 | 36.6 | 36.2 | 38.8 | 37.4 | 48.3 | 70.8 | 60.2 |
| Qwen2.5VL-7B (MRCD(**Ours**)) | **34.7** | 37.1 | **37.7** | **40.3** | **38.5** | **49.5** | **72.5** | **62.3** |

Table 1: Performance comparison on VQA-RAD dataset. Close-ended (multiple-choice/yes-no) questions evaluated with BLEU-1, ROUGE-1, Precision, Recall, F1, and Accuracy. Open-ended questions evaluated with Accuracy. Bold underlined values show best results.

| Model Variant | Close Acc. | Open Acc. | Overall Acc. |
|---|---|---|---|
| MRCD (Full) | **49.5** | **72.5** | **62.3** |
| w/o DS-RAG | 48.2 | 71.5 | 61.4 |
| w/o DER | 46.2 | 69.8 | 59.2 |

Table 2: Ablation study results showing the impact of Dynamic Expert Recruitment(DER) and Domain-Specific RAG enhancement(DS-RAG). Performance measured on VQA-RAD dataset.

improvement over the strong CoT baseline and a 2.1 percentage point improvement over the MAD approach. The improvements are particularly pronounced in open-ended questions (72.5% vs 72.1% for CoT), where the multi-disciplinary collaboration provides substantial benefits for complex clinical reasoning.

The superior performance of MRCD stems from its ability to capture diverse medical perspectives through domain-specific expertise. Unlike traditional multi-agent approaches that suffer from information homogeneity, our framework enables genuine intellectual discourse between specialists accessing different knowledge bases. This leads to more comprehensive analysis and better error correction through cross-disciplinary review.

Beyond quantitative metrics, we conduct detailed analysis of reasoning quality through expert evaluation. Medical professionals assess the clinical appropriateness, depth of analysis, and practical applicability of generated responses. MRCD consistently demonstrates superior clinical reasoning, with 78% of responses rated as "clinically comprehensive" compared to 65% for CoT baselines. The multi-disciplinary approach particularly excels in complex cases requiring integration of multiple medical specialties.

### 3.3 Ablation Study

To understand the contribution of domain-specific RAG enhancement in our MRCD framework, we conduct an ablation study examining the impact of replacing our specialized retrieval system with generic medical knowledge retrieval.

Table 2 presents the results of our ablation study. Removing DS-RAG leads to a 0.9 percentage point drop in overall accuracy (61.4% vs 62.3%), while removing DER results in a 3.1 percentage point degradation (59.2% vs 62.3%). The substantial impact of DER highlights a critical issue: when RAG systems operate without proper expert recruitment, they suffer from focus dispersion problems that can make performance worse than using medical models alone.

Without DER, the system retrieves diverse medical literature without appropriate domain filtering, leading to information overload and conflicting perspectives that confuse the reasoning process. This demonstrates that simply adding external knowledge through RAG is insufficient—the knowledge must be appropriately targeted through expert recruitment to avoid degrading model performance.

The domain-specific RAG enhancement enables each expert to access literature tailored to their medical specialty. When addressing chest pain, a cardiologist retrieves cardiology-specific literature focusing on cardiac causes and treatment protocols, while a pulmonologist accesses respiratory medicine literature emphasizing pulmonary etiologies. This targeted retrieval ensures each expert operates with the most relevant specialized knowledge, avoiding the focus dispersion that occurs with generic retrieval approaches.

## References


Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. 2018. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *arXiv preprint arXiv:2306.00890*.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2023. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118*.

Jiazhen Pan, Che Liu, Junde Wu, Fenglin Liu, Jiayuan Zhu, Hongwei Bran Li, Chen Chen, Cheng Ouyang, and Daniel Rueckert. 2025. Medvlm-r1: Incentivizing medical reasoning capability of vision-language models (vlms) via reinforcement learning. *arXiv preprint arXiv:2502.19634*.

Shaohao Rui, Lingzhi Chen, Zhenyu Tang, Lilong Wang, Mianxin Liu, Shaoting Zhang, and Xiaosong Wang. 2025. Multi-modal vision pre-training for medical image analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5164–5174.

Sahal, Mohammed Irfan Kurpath, Sara Pieri, Saeed Yahya Alseiari, Shanavas Cholakkal, Khaled Aldahmani, Fahad Khan, Rao Anwer, Salman Khan, Timothy Baldwin, and Hisham Cholakkal. 2024. Bimedix2: Bio-medical expert lmm for diverse medical modalities. *Preprint*, arXiv:2412.07769.

Xiangru Tang, Anni Zou, Zhuosheng Zhang, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. 2023. Ml-medagents: Large language models as collaborators for zero-shot medical reasoning. *arXiv preprint arXiv:2311.10537*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

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

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

Guangzhi Xiong, Qiao Jin, Xiao Wang, Minjia Zhang, Zhiyong Lu, and Aidong Zhang. 2024c. Improving retrieval-augmented generation in medicine with iterative follow-up questions. In *Biocomputing 2025: Proceedings of the Pacific Symposium*, pages 199–214. World Scientific.

Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, and 1 others. 2025. Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. *arXiv preprint arXiv:2506.07044*.

Xuejiao Zhao, Siyan Liu, Su-Yin Yang, and Chunyan Miao. 2025. 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.

# Author Index