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.

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

*Corresponding author

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

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

Our approach draws inspiration from multidisciplinary 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.

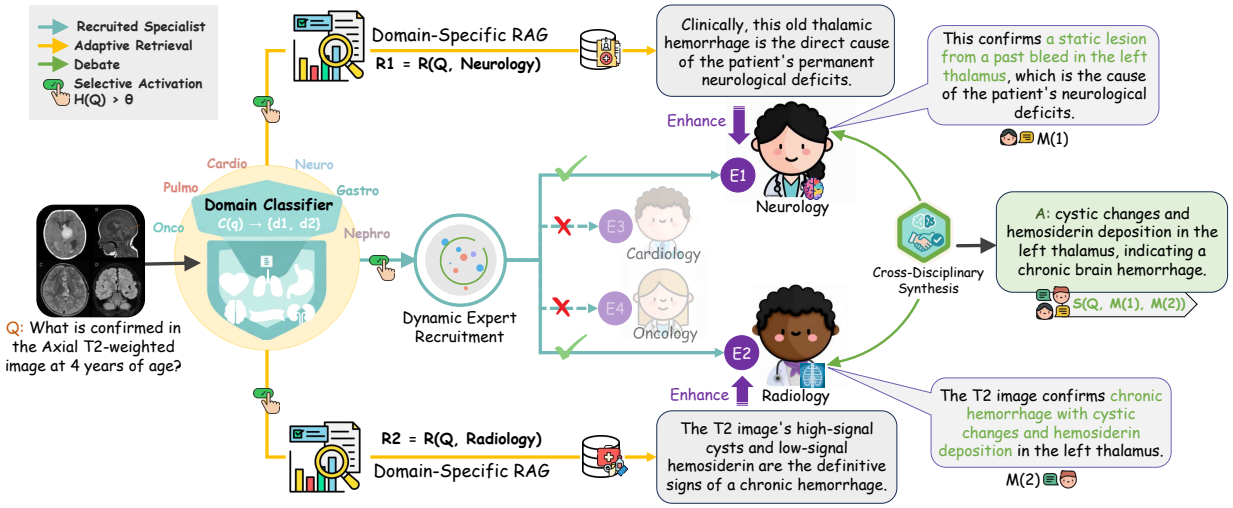

Figure 2: Overview of Multi-disciplinary RAG-enhanced Collaborative Debate(MRCD). A domain classifier recruits multi-specialty doctor agents (Neurology highlighted as the first), each using domain-specific RAG; agents debate with adaptive retrieval and then form a consensus answer.

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 τ is a synthesis template. The template τ 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.