A Layered Debating Multi-Agent System for Similar Disease Diagnosis

Yutian Zhao[1,*], **Huimin Wang**[1*], **Yefeng Zheng**[3], **Xian Wu**[1†]
[1] Jarvis Research Center, Tencent YouTu Lab Shenzhen, China
[3] Medical Artificial Intelligence Lab, Westlake University, Hangzhou, China
{yutianzhao,hmmmwang,kevinxwu}@tencent.com

Abstract

Distinguishing between extremely similar diseases is a critical and challenging aspect of clinical decision-making. Traditional classification, contrastive learning, and Large Language Models (LLMs) based methods fail to detect the subtle clues necessary for differentiation. This task demands complex reasoning and a variety of tools to identify minor differences and make informed decisions. This paper probes a novel framework that leverages LLMs and a multi-agent system to achieve accurate disease diagnosis through a process of repeated debate and reassessment. The approach aims to identify subtle differences between similar disease candidates. We structure patient information and integrate extensive medical knowledge to guide the analysis towards discerning these differences for precise diagnosis. Comprehensive experiments were conducted on two public datasets and two newly introduced datasets, JarvisD2-Chinese and JarvisD2-English, to validate the effectiveness of our method. The results confirm the efficacy of our approach, demonstrating its potential to enhance diagnostic precision in healthcare.

1 Introduction

In recent years, AI-assisted clinical diagnosis has significantly enhanced the efficiency and accuracy of medical assessments. Swift and precise disease prediction is crucial for timely and effective treatment, ultimately saving lives. Diagnosing diseases that present with prominent symptoms is relatively straightforward. However, diagnosing conditions that exhibit very similar symptoms is more challenging and carries a higher risk of misdiagnosis. In clinical practice, when faced with the potential for misdiagnosis (also known as similar diseases), medical experts employ a method known as "differential diagnosis". This involves compiling a comprehensive list of all possible diseases that could cause the observed symptoms and systematically narrowing down this list through further medical examinations until the most likely disease is identified. For instance, Cardiovascular diseases like Myocarditis, Heart Failure, and Myocardial Infarction share symptoms such as chest pain, shortness of breath, fatigue, and palpitations, but have distinct causes and treatments. Accurate diagnosis is crucial to prevent serious complications. A key differentiator is the duration of symptoms: Heart Failure is long-term, while Myocardial Infarction and Myocarditis have different temporal patterns. Diagnosing these conditions requires extensive medical knowledge and expert reasoning to identify subtle differences.

Traditional methods for disease diagnosis include classification based methods that predict diseases using trained classification networks (Prince, 1996; Green et al., 2006; Atkov et al., 2012; Yang et al., 2022b,b); contrastive learning based methods that separate diseases using contrastive learning strategies (Chen et al., 2022; Wu et al., 2022; Zhao et al., 2024b); Large Language Models (LLMs) based methods that conduct disease diagnosis through pre-training or prompt learning based on LLMs (Liu et al., 2021; Li et al., 2020; Rasmy et al., 2021; Wang et al., 2023a, 2024a; Jin et al., 2024; Zhao et al., 2024a). However, these methods may fail to capture the subtle clues necessary for differential diagnosis, as these clues are often too subtle to detect and many require consequential decision-making.

In this paper, we propose a novel framework that leverages **M**ultiple **L**LM-based **A**gents working collaboratively to achieve accurate disease **D**iagnosis (denoted as **MLAD**). The key insight of MLAD lies in identifying subtle distinctions between similar disease candidates through a cycle of iterative debating and reflecting, all guided by comprehensive medical knowledge to facilitate ef-

[*] Equal Contribution
[†] Corresponding author

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers), pages 539–549

fective differential diagnosis. The process involves engaging agents specialized in different disease domains to present their perspectives, participate in debate, and reflect on the diagnosis. The process continues until the agents' diagnoses converge. Furthermore, we employ a highly effective structured mechanism, *imap* (Wang et al., 2024b), to restructure patient information, emphasizing crucial information like symptoms and lab results. Throughout the procedure, the agents have access to various resources, such as medical knowledge graph searches, to assist in pinpointing the correct diagnosis.

To evaluate MLAD, we first compare its performance on two publicly available medical exam datasets in both English and Chinese. To address the lack of challenged similar disease options and potential data leakage in public datasets, we enhanced two public datasets by revising the options to create a more robust similar disease diagnosis dataset. To generate options that include more differential diagnoses, we consider candidates derived from various sources such as medical knowledge graph, LLMs and ICD-10 [1].

In summary, our contributions can be outlined as follows:

- To improve differential diagnosis, we proposed a new framework, MLAD, where multiple LLM-based agents engage in iterative debating and reflecting, guided by comprehensive medical knowledge, to identify subtle distinctions between similar diseases.
- To assess the differential diagnosis abilities, we created two challenged disease diagnosis datasets by revising options using specialized strategies derived from two public datasets.
- To validate the superiority of MLAD, we conducted extensive experiments and made in-depth analyses, demonstrating the effectiveness of our methods.

2 Methods

The key insight of MLAD lies in its ability to uncover subtle differences between similar diseases through iterative debate and reflection, guided by essential medical knowledge and tools. As illustrated in Figure 1, MLAD begins with an initialization phase that highlights the input text with patient information using *imap*—a data structure for key information extraction introduced by (Wang et al., 2024b). It also equips the LLM-based agents with different disease backgrounds. The process then moves into the debating phase, which includes an inner-group discussion among agents with the same diagnosis to consolidate their reasoning, followed by an inter-group debate to compare differing diagnostic views. Subsequently, the tool utilization phase allows agents to use resources like search engines to acquire additional medical knowledge and evaluate the perspectives of other agents. After this, all agents are given the opportunity to reflect on their points and re-evaluate their diagnoses. This cycle continues until a consensus on the diagnosis is reached. The detailed process is as follows.

2.1 Initialization

The initialization process reshapes the patient information for denoising and key information extraction, aligning agents from diverse backgrounds to simulate an expert panel. We use *imap*, a data structure that distills medical text into term-value pairs, enhancing the diagnosis process by capturing essential data from the records. This guides agents to focus on symptom comparison and distinct diagnoses. However, LLM-based agents may lack specialist expertise. To mitigate this, we equip LLMs with specialized disease knowledge profiles from a Medical Knowledge Graph [2], transforming them into distinct specialist agents as shown in Figure 1. Each agent specializes in a single disease domain, enhancing initial answer variety and facilitating critical discussion.

2.2 Tools Augmented Layered Debating

In this phase, agents participate in several rounds of intra- and inter-group discussions, drawing on the summarized perspectives of other agents to inform their individual decisions. Differing from the conventional debate-based diagnosis methods (Lu et al.), MLAD critically examines the diagnostic results and reasoning, integrating evidence provided by peers and the use of diagnostic tools.

Each agent A_i begins with a freely chosen initial disease D_i and adheres to the following procedure: A_i participates in an inner-group discussion with other agents who have also selected D_i. A_i presents its reasoning r_i, which is amalgamated with the reasoning of other inner-group agents to produce a combined reasoning report R_i. Subse-

[1] https://icd.who.int/browse10/2019/en

[2] https://jarvislab.tencent.com/kg-intro.html

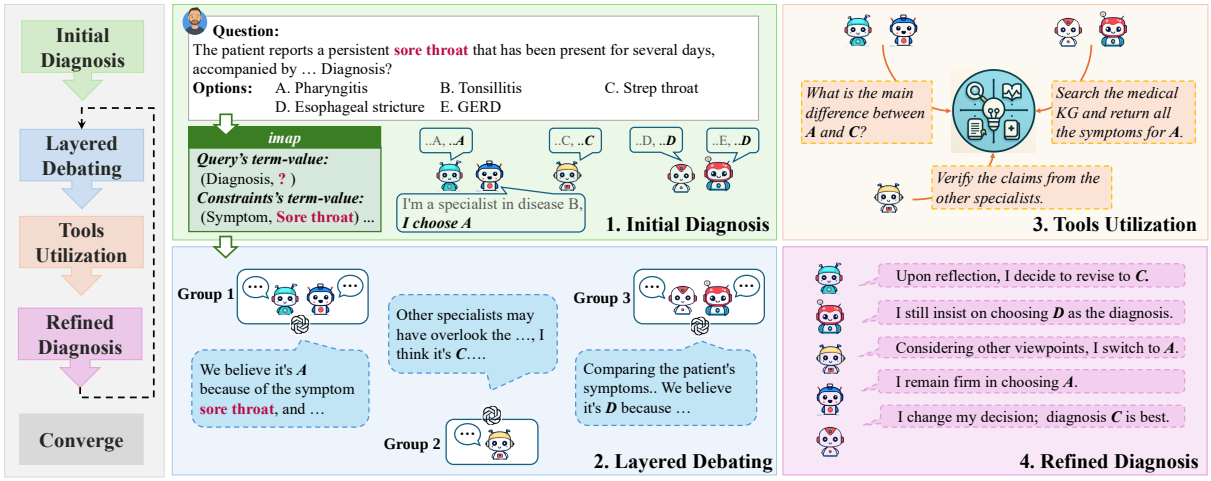

Figure 1: Overview of MLAD. Initially, agents with diverse disease backgrounds diagnose based on structured patient information extracted by *imap*. The process then involves several rounds of inner-group discussions, inter-group debates, tool utilization, and self-reflection, all guided by *imap*, to complete the diagnosis task.

quently, the inter-group debate commences. Each group begins by examining the reports submitted by their counterparts. They are allowed to utilize tools such as a medical knowledge graph to collect supplementary information like symptoms associated with a particular disease. They can also compare two diseases using online searches or ask a Language Learning Model (LLM) to provide a summary. Armed with this newly acquired evidence and the initial viewpoints from other groups, the agents are then able to refine and rearticulate their diagnosis. This iterative process persists until the agents arrive at a preliminary consensus or an early stopping mechanism is activated.

2.3 Consensus Diagnosis and Early-Stopping

In an ideal scenario, agents will achieve a formal consensus by integrating the refined answers and reasoning derived from the inter-group debate stage. This consensus signifies that all agents agree on a single disease diagnosis, leveraging their combined domain expertise to validate the final determination. The debate and reflection process ensures a robust, well-analyzed final decision.

Once all agents reach a consensus, a definitive and reliable diagnosis is delivered. To enhance the efficiency of inter-group debating, we implement an early-stopping mechanism, which operates under two conditions: **1)** If one disease receives all votes, early stopping is triggered; **2)** If all diseases receive an equal number of votes for more than 3 consecutive rounds, a new agent is brought in to cast a deciding vote, thereby ending the debate. This mechanism terminates communication when agents consistently confirm their reasoning with high confidence, thereby reducing unnecessary computations.

3 Experiment Result

3.1 Datasets and Baselines

The JarvisD2-Chinese and JarvisD2-English datasets, containing 10,953 and 248 question-answer pairs respectively, are created from various medical references. To test differential diagnosis, the datasets are expanded with more challenging misdiagnosed options, followed by expert manual verification and voting. Details on the original and enhanced datasets are provided in Appendix A.1.

We compared MLAD with various models including Embedding-based methods, General LLMs and Specialized LLMs. Details for each baseline and example prompts are in Appendix A.2.

3.2 Main Results

Table 1 illustrates the diagnostic prediction performance of various models, highlighting a decrease in accuracy when shifting from standard to enhanced datasets. LLMs show an average accuracy drop of 18.3% on JarvisD2-Chinese and 17.3% on JarvisD2-English, emphasizing the challenge of diagnosing easily confused diseases and the need for enhanced datasets. The use of MLAD significantly improves LLMs' accuracy on both dataset versions, increasing performance by 6.4% on standard and 8.5% on enhanced versions. This indicates MLAD's effectiveness in distinguishing similar diseases, thus enhancing accuracy in complex

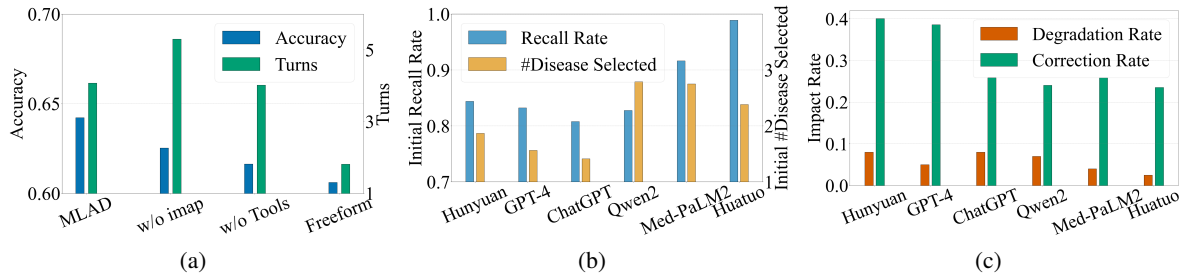

Figure 2: (a) Impact of *imap*, tools, and debating mechanisms on the average accuracy and debating turns across all models on enhanced JarvisD2. (b) Average recall rate of the correct answer and the number of diseases selected at the initial diagnosis stage on enhanced JarvisD2. (c) Performance alteration proportion for each model using MLAD on enhanced JarvisD2.

clinical situations. MLAD improves general LLMs by an average of 6.8%, while specialized LLMs see a larger increase of 8.8%, suggesting that they can leverage MLAD more effectively. Hunyuan and Qwen2 notably outperform other LLMs on the JarvisD2-Chinese dataset. Given the open-source nature of JarvisD2's data, these models may have been trained on this dataset. However, MLAD still significantly enhances their accuracy.

Table 1: Diagnosis accuracy (%) comparison with baselines on JarvisD2-Chinese and JarvisD2-English Datasets: Standard and Enhanced Versions. Our method backed by different LLMs is indicated by blue, and the best result for each dataset is highlighted in underline.

Methods	JarvisD2-Chinese				JarvisD2-English			
	Standard		Enhanced		Standard		Enhanced	
	Baseline	MLAD	Baseline	MLAD	Baseline	MLAD	Baseline	MLAD
Embedding-Based								
MedBERT	22.2	-	20.1	-	21.8	-	21.4	-
KEPT	24.0	-	23.0	-	26.2	-	23.4	-
GP	23.2	-	20.0	-	22.2	-	20.0	-
MKeCL	27.6	-	24.6	-	31.3	-	28.6	-
General LLMs								
Hunyuan	94.4	95.6	83.7	86.5	74.6	81.1	50.8	57.5
Qwen2	97.8	98.5	78.7	85.0	78.6	83.9	56.5	65.2
ChatGPT	64.2	69.2	39.2	52.7	56.0	72.2	29.8	42.4
GPT-4	80.7	85.5	60.0	66.3	84.7	89.5	53.2	60.3
Specialized LLMs								
MedPaLM-2	71.8	76.1	59.0	64.9	56.8	79.6	40.3	58.3
Huatuo2	88.9	91.9	67.2	78.2	66.5	68.8	50.0	53.2

3.3 Analysis and Discussion

Ablative Study We perform an ablative study on MLAD to investigate the impact of *imap*, tools, and debating mechanisms. Remarkably, about 72% of debates achieved full consensus within the pre-established maximum of 10 turns. As illustrated in Figure 1(a), *imap* significantly enhanced both efficiency and accuracy by directing agents' attention to crucial patient data. Furthermore, adding tools enhances accuracy while maintaining a similar average turn with the MLAD. Freeform debating, lacking inner- and inter-group settings, led to a 3.6% accuracy drop due to conformity issues in LLMs (Zhang et al., 2023b). Agents, aware of the support each disease candidate had, often converged on the initially popular but incorrect diagnoses. Layered debating, involving intra- and inter-group discussions, mitigated this issue. Agents knew the disease candidates but not the support each had, reducing conformity pressure and increasing diagnosis accuracy.

Agent Behavior in Initial Diagnosis In the initial diagnosis phase, all LLMs achieve at least an 80% recall rate for including the correct disease, with Huatuo2 leading at 98.9%. If the correct disease is not initially selected, it is excluded from further discussions, leading to incorrect conclusions. Even if the correct disease is included in later debates, LLMs often fail to recognize it, indicating an internal knowledge conflict that prevents reevaluation. This may necessitate new training data for accuracy improvement. Additionally, Hunyuan, GPT-4, and ChatGPT typically select fewer than two disease candidates initially, while Qwen2 starts with around three.

MLAD's Impact on Correcting Diagnosis Errors Figure 2(c) showcases the MLAD method's impact on various models, with all models improving their accuracy by at least 20%. Hunyuan and GPT-4 notably corrected nearly 40% of initial errors. Despite introducing some confusion, causing a few correct answers to be marked incorrect, the error rate stayed below 10% for all models. Thus, MLAD significantly enhanced overall accuracy.

Case Study A case study on how MLAD enhances LLMs' ability to distinguish between similar diseases is provided in Figure 3 of Appendix A.3.

4 Conclusion

This paper proposes a collaborative framework named MLAD, which utilizes multiple LLM-based agents for accurate differential diagnosis. The method involves iterative debating and reflecting, guided by extensive medical knowledge, to identify subtle distinctions between similar diseases. Empirical results on two public datasets and two newly introduced challenging dataset demonstrate the effectiveness of MLAD. Especially, MLAD outperforms other methods on the challenging dataset and demonstrates strong generalizability in differentiating similar diseases.

Limitations

We acknowledge two limitations of our study.

First, our study relies solely on publicly available datasets, which differ significantly from real clinical medical records. Due to privacy policies, we are unable to access actual health records from hospitals. Future research could extend our experiments to real clinical datasets to further validate the superiority of the proposed framework.

Second, the scope of our study is somewhat narrow, as it only investigates similar disease diagnosis in two languages. A logical progression of this research would involve expanding the range of diseases studied, exploring additional language systems, and testing models beyond the selected baselines.

Ethics Statement

Our work adheres to the ACL Ethics Policy. Meanwhile, this paper aims to underscore the differential diagnosis that may arise from the improper application of the proposed models within the medical domain. The primary objective of our research is to explore a multi-agent system for accurate disease diagnosis with LLMs. However, it is crucial to note that the proposed methods are not yet ready for deployment in real-world medical settings. The potential for these models to mislead users about the underlying reasons for their predictions is a significant concern. Misinterpretations could lead to incorrect decisions, with potentially serious implications for patient care and outcomes. Moreover, the ethical considerations of our work extend beyond the accuracy and reliability of the models. The privacy and security of sensitive medical data hold utmost importance. Throughout the data collection and utilization process, even when using publicly available datasets, we have enforced rigorous measures to safeguard this sensitive information. In conclusion, while our work holds promise for improving disease diagnosis, it is essential to approach its application with caution. We must continue to prioritize the ethical considerations of accuracy, transparency, data privacy, and security as we further develop and refine these models.

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*.

Oleg Yu Atkov, Svetlana G Gorokhova, Alexandr G Sboev, Eduard V Generozov, Elena V Muraseyeva, Svetlana Y Moroshkina, and Nadezhda N Cherniy. 2012. Coronary heart disease diagnosis by artificial neural networks including genetic polymorphisms and clinical parameters. *Journal of Cardiology*, 59(2):190–194.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Yan Cai, Linlin Wang, Ye Wang, Gerard de Melo, Ya Zhang, Yanfeng Wang, and Liang He. 2024. Medbench: A large-scale chinese benchmark for evaluating medical large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17709–17717.

Yuhao Chen, Yanshi Hu, Xiaotian Hu, Cong Feng, and Ming Chen. 2022. CoGO: a contrastive learning framework to predict disease similarity based on gene network and ontology structure. *Bioinformatics*, 38(18):4380–4386.

Michael Green, Jonas Björk, Jakob Forberg, Ulf Ekelund, Lars Edenbrandt, and Mattias Ohlsson. 2006. Comparison between neural networks and multiple logistic regression to predict acute coronary syndrome in the emergency room. *Artificial Intelligence in Medicine*, 38(3):305–318.

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

Mingyu Jin, Qinkai Yu, Chong Zhang, Dong Shu, Suiyuan Zhu, Mengnan Du, Yongfeng Zhang, and Yanda Meng. 2024. Health-llm: Personalized retrieval-augmented disease prediction model. *arXiv preprint arXiv:2402.00746*.

Yikuan Li, Shishir Rao, José Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. 2020. BEHRT: Transformer for electronic health records. *Scientific Reports*, 10(1):1–12.

Junling Liu, Peilin Zhou, Yining Hua, Dading Chong, Zhongyu Tian, Andrew Liu, Helin Wang, Chenyu You, Zhenhua Guo, Lei Zhu, et al. 2024. Benchmarking large language models on cmexam-a comprehensive chinese medical exam dataset. *Advances in Neural Information Processing Systems*, 36.

Ning Liu, Qian Hu, Huayun Xu, Xing Xu, and Mengxin Chen. 2021. Med-BERT: A pretraining framework for medical records named entity recognition. *IEEE Transactions on Industrial Informatics*, 18(8):5600–5608.

Meng Lu, Ho Brandon, Ren Dennis, and Xuan Wang. Triageagent: Towards better multi-agents collaborations for large language model-based clinical triage. In *ICML 2024 AI for Science Workshop*.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR.

Martin J Prince. 1996. Predicting the onset of Alzheimer's disease using Bayes' theorem. *American Journal of Epidemiology*, 143(3):301–308.

Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. 2021. Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ Digital Medicine*, 4(1):86.

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.

Haochun Wang, Sendong Zhao, Zewen Qiang, Nuwa Xi, Bing Qin, and Ting Liu. 2024a. Beyond direct diagnosis: Llm-based multi-specialist agent consultation for automatic diagnosis. *arXiv preprint arXiv:2401.16107*.

Huimin Wang, Wai-Chung Kwan, Kam-Fai Wong, and Yefeng Zheng. 2023a. Coad: Automatic diagnosis through symptom and disease collaborative generation. *arXiv preprint arXiv:2307.08290*.

Huimin Wang, Yutian Zhao, Xian Wu, and Yefeng Zheng. 2024b. imapscore: Medical fact evaluation made easy. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 10242–10257.

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

Yawen Wu, Dewen Zeng, Zhepeng Wang, Yi Sheng, Lei Yang, Alaina J James, Yiyu Shi, and Jingtong Hu. 2022. Federated self-supervised contrastive learning and masked autoencoder for dermatological disease diagnosis. *arXiv preprint arXiv:2208.11278*.

Zhichao Yang, Sunjae Kwon, Zonghai Yao, and Hong Yu. 2022a. Multi-label Few-shot ICD Coding as Autoregressive Generation with Prompt. *arXiv preprint arXiv:2211.13813*.

Zhichao Yang, Shufan Wang, Bhanu Pratap Singh Rawat, Avijit Mitra, and Hong Yu. 2022b. Knowledge Injected Prompt Based Fine-tuning for Multi-label Few-shot ICD Coding. *arXiv preprint arXiv:2210.03304*.

Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Jianquan Li, Guiming Chen, Xiangbo Wu, Zhiyi Zhang, Qingying Xiao, et al. 2023a. Huatuogpt, towards taming language model to be a doctor. *arXiv preprint arXiv:2305.15075*.

Jintian Zhang, Xin Xu, and Shumin Deng. 2023b. Exploring collaboration mechanisms for llm agents: A social psychology view. *arXiv preprint arXiv:2310.02124*.

Yutian Zhao, Huimin Wang, Yuqi Liu, Wu Suhuang, Xian Wu, and Yefeng Zheng. 2024a. Can LLMs replace clinical doctors? exploring bias in disease diagnosis by large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13914–13935, Miami, Florida, USA. Association for Computational Linguistics.

Yutian Zhao, Huimin Wang, Xian Wu, and Yefeng Zheng. 2024b. Mkecl: Medical knowledge-enhanced contrastive learning for few-shot disease diagnosis. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11394–11404.

A Appendix

Table 2: Source distribution of enhanced JarvisD2-Chinese options and the proportion that misled an LLM. All values are multiplied by 100 for clarity.

Source	Medical KG	LLMs	ICD-10	Same Body Part	Varying Severity
Source %	8.74	89.32	2.91	17.48	33.98
Misled %	55.55	44.56	66.66	16.66	28.57

Table 3: Source distribution of enhanced JarvisD2-English options. All values are multiplied by 100 for clarity.

Source	Medical KG	LLMs	ICD-10	Same Body Part	Varying Severity
Source %	12.12	30.18	5.63	28.77	26.04
Misled %	52.66	54.14	60.12	41.57	45.31

A.1 Datasets

The JarvisD2-Chinese and JarvisD2-English datasets are created from various medical references, including CMExam (Liu et al., 2024), CMB (Wang et al., 2023b), MedQA (Jin et al., 2021), MedMCQA (Pal et al., 2022), and MedBench (Cai et al., 2024). Each question in both datasets includes five options. The number of distinct diseases covered in each dataset is 4,949 and 238 respectively.

A.1.1 Enhanced Dataset Construction

To test differential diagnosis, the datasets are expanded with more challenging misdiagnosed options through a five-step process: 1) Extracting similar diseases from a Medical Knowledge Graph; 2) Asking Large Language Models (LLMs) for probable diseases; 3) Randomly selecting diseases from the same ICD-10 section. 4) Identifying diseases affecting the same body part; 5) Selecting diseases of varying severity for the correct answers.

Three medical researchers from two universities are involved in the process of verifying the options to ensure they are both valid and challenging. These researchers are experts in their respective fields, bringing a wealth of knowledge and experience to the task. Before beginning the verification process, all participants underwent standardized training. This training was designed to ensure consistency and accuracy across all evaluations, minimizing the potential for subjective bias or individual discrepancies. The process of verification involves a consensus-based approach. For an option to be considered as an 'enhanced option', it must receive unanimous agreement from all three researchers. They must all agree that the option 1) represents a reasonable disease, 2) is similar to the correct answer, and 3) the answer still remains the most reasonable and accurate disease based on the content of the question.The first criterion ensures that the options are medically sound and plausible. The second criterion ensures that the options are not wildly different from the correct answer, thereby maintaining a level of challenge and complexity. The third criterion ensures that, despite the similarities with other diseases, the correct answer remains the most accurate and reasonable based on the information provided in the question.

Hunyuan, GPT-4, and Qwen2 vote on these options, with the top five, including the correct answer, becoming the final five options.

A.1.2 Enhanced Dataset Analysis

88% and 98% of the questions from each dataset had their options modified for enhancement, with an average of 1.77 and 2.75 options altered per question, respectively. As shown in Table 2 and Table 3, these modifications resulted in a diverse source distribution of the final challenging options in both JarvisD2-Chinese and JarvisD2-English. It's important to note that a single question could contain options derived from multiple sources, adding to the complexity of the task.

In the JarvisD2-Chinese dataset, the majority of the challenging options (89.32%) were sourced from the direct answers provided by Large Language Models (LLMs), indicating their potential to generate complex and challenging diagnostic possibilities. On the other hand, the source distribution in the JarvisD2-English dataset was more evenly spread, suggesting a broader range of challenging options.

Interestingly, the options that most frequently led to mistakes by the LLMs were those sourced from diseases within the same ICD-10 section, across both datasets. This suggests that diseases with similar classifications tend to be more confusing for the models. Furthermore, options related to diseases affecting the same body part and those of varying severity had a higher rate of misleading the LLMs in the JarvisD2-English dataset compared to the JarvisD2-Chinese dataset.

A.2 Baselines and Implementation

We compared MLAD with various models: 1) Embedding-based methods like MedBERT (Rasmy et al., 2021), KEPT (Yang et al., 2022b), GP (Yang et al., 2022a), and MKeCL (Zhao et al., 2024b); 2) General LLMs such as Hunyuan-70B [3], Qwen2-72B (Bai et al., 2023), ChatGPT, and GPT-4 (Achiam et al., 2023); and 3) Specialized LLMs fine-tuned for the medical domain, including MedPaLM-2 (Singhal et al., 2023) and Huatuo2-34B (Zhang et al., 2023a).

All models are instructed using the same prompts, as shown in Table 4 - 7, with a maximum of 10 debating turns allowed. Three tools are included: medical knowledge, GPT-4, and a search engine.

A.3 Case Study

[3] https://hunyuan.tencent.com/

Initial Diagnosis Prompt:
<Role and Background>:
You are a doctor and a patient has come to you for a diagnosis. The patient's medical record is as follows:
Medical record: [Record]
Possible diseases: [Diseases]

Given your experience with disease [Disease$_i$], you have identified the following background knowledge for it:
[Disease$_i$ Info]

<Task>:
First, please combine your knowledge with the medical record information to choose the most likely diagnosis for this patient, and provide a reason. Please output:

Diagnosis:
Reason:

Table 4: Initial Diagnosis Prompt.

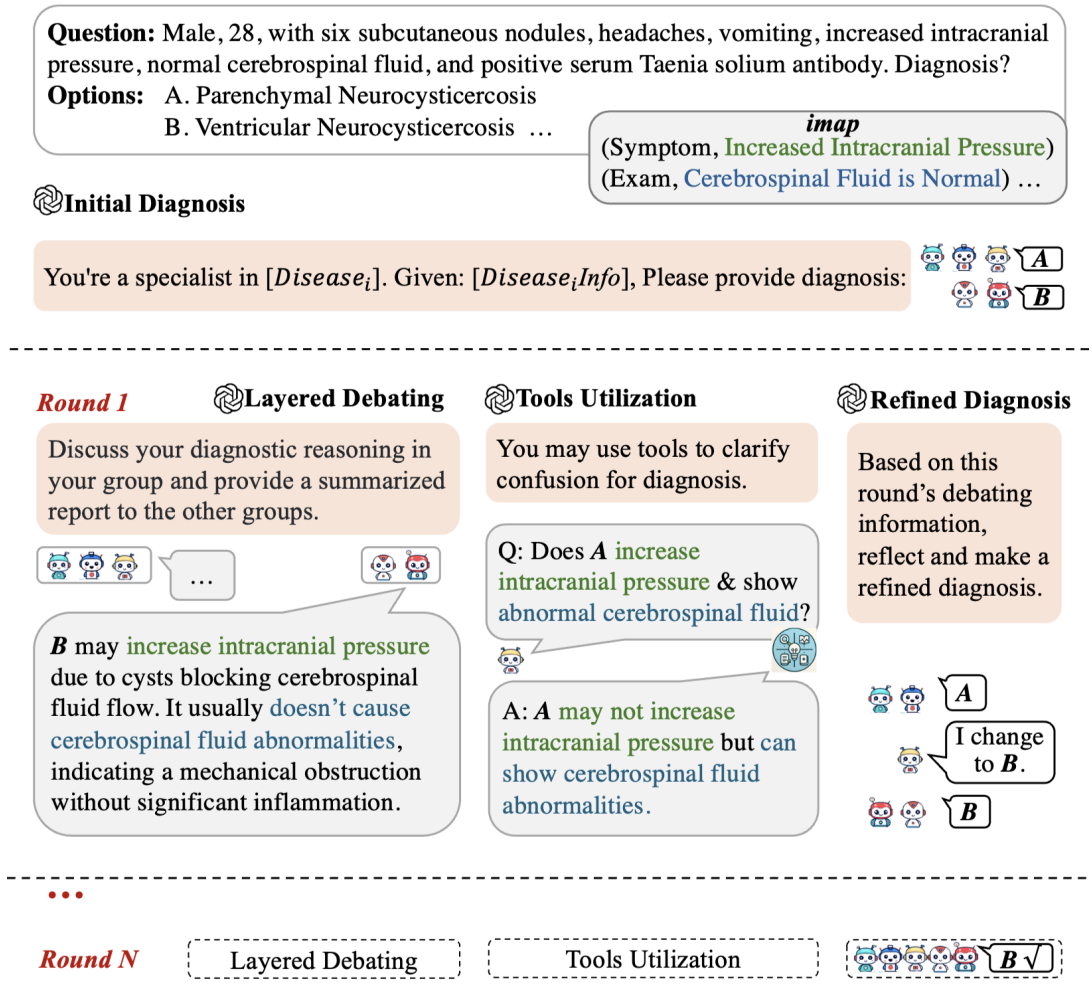

Figure 3: A case study on how MLAD enhances LLMs' ability to distinguish between similar diseases.

Layered Debating Prompt:
Next, you need to consult with other experts who have different diagnostic opinions. Please refer to the following example to output your argument.

<Example>:
Medical record: Male, 31 years old. Sudden severe headache for 1 hour, mainly in the occipital region, accompanied by projectile vomiting 3 times. Physical examination: painful expression, sweating all over, positive meningeal irritation signs.
Possible diseases: Rupture of basilar artery aneurysm with subarachnoid hemorrhage, subarachnoid hemorrhage

Expert_1:
Diagnosis: Rupture of basilar artery aneurysm with subarachnoid hemorrhage
Argument: According to the medical history, the patient is a 31-year-old male with a sudden severe headache, mainly in the occipital region, accompanied by projectile vomiting and positive meningeal irritation signs. These symptoms highly suggest subarachnoid hemorrhage (SAH), and aneurysm rupture is one of the common causes of SAH.

Expert_2:
Diagnosis: Subarachnoid hemorrhage
Argument: Although the patient's symptoms could be due to a rupture of a basilar artery aneurysm with subarachnoid hemorrhage, there is no specific imaging evidence or other diagnostic methods (such as CT, MRI, cerebral angiography) in the medical history to clearly indicate a basilar artery aneurysm rupture. Therefore, based solely on clinical symptoms and signs, the most reasonable preliminary diagnosis should be: subarachnoid hemorrhage

Below are the diagnosis and reason given by each disease expert:
Expert_1:
Diagnosis: [Disease_1]
Reason: [Reason_1]

Expert_2:
Diagnosis: [Disease_2]
Reason: [Reason_2]
...
Based on your previous individual analysis and the last round diagnosis and reasons of the other experts, provide your argument for why you believe the patient's diagnosis is [Disease_i], rather than the other possible diseases.

Table 5: Layered Debating Prompt.

Tools Utilization Prompt:
Below are the summarized arguments given by the other experts during the previous stage: [Summarized Arguments]

Please begin by integrating the information gathered from previous stages, which should include the valid points from other experts' arguments. Reflect on your own arguments to identify any potential gaps or omissions. Then, objectively reassess which disease has a higher diagnostic accuracy.

If you find that you still lack the necessary medical knowledge to make a definitive diagnosis, consider using tools to help clarify your concerns or questions. This could involve distinguishing between diseases that have similar symptoms or characteristics.

If you have any questions or uncertainties, you can choose to query the Medical Knowledge Graph or use a search engine to gain a deeper understanding.

Table 6: Tools Utilization Prompt.

Refined Diagnosis Prompt:

Please integrate the insights from other experts and the new information you've gathered using various tools to determine the most probable diagnosis for this patient. This process should involve a thorough review and consideration of all available data.

Please output:
Diagnosis:
Reason: (Your explanation for the diagnosis, including the key pieces of information that led you to this conclusion, any significant points from your discussions with other experts, and the new knowledge you've gained from your research.)

Table 7: Refined Diagnosis Prompt.