MedLA: A Logic-Driven Multi-Agent Framework for Complex Medical Reasoning with Large Language Models

Siqi Ma[1*], **Jiajie Huang**[1*], **Fan Zhang**[3], **Yue Shen**[2*], **Jinlin Wu**[3], **Guohui Fan**[4], **Zhu Zhang**[4], **Zelin Zang**[1,2,3†]
[1] Westlake University
[2] Ant Group
[3] CAIR, Hong Kong Institute of Science and Innovation (HKISI), CAS
[4] China-Japan Friendship Hospital

Abstract

Answering complex medical questions requires not only domain expertise and patient-specific information, but also structured and multi-perspective reasoning. Existing multi-agent approaches often rely on fixed roles or shallow interaction prompts, limiting their ability to detect and resolve fine-grained logical inconsistencies. To address this, we propose MEDLA, a logic-driven multi-agent framework built on large language models. Each agent organizes its reasoning process into an explicit logical tree based on syllogistic triads (major premise, minor premise, and conclusion), enabling transparent inference and premise-level alignment. Agents engage in a multi-round, graph-guided discussion to compare and iteratively refine their logic trees, achieving consensus through error correction and contradiction resolution. We demonstrate that MEDLA consistently outperforms both static role-based systems and single-agent baselines on challenging benchmarks such as MedDDx and standard medical QA tasks. Furthermore, MEDLA scales effectively across both open-source and commercial LLM backbones, achieving state-of-the-art performance and offering a generalizable paradigm for trustworthy medical reasoning.

Code — https://github.com/alexander2618/MedLA
Extended version — https://arxiv.org/abs/2509.23725

Introduction

LLM has demonstrated significant advantages in the field of medical reasoning, and its deep learning architecture can efficiently extract knowledge from massive amounts of literature and clinical cases to provide intelligent assistance in diagnostic decision-making, and is expected to significantly improve the accessibility and popularization of medical knowledge(Chang et al. 2024; Kasneci et al. 2023). Answering complex medical questions with large language models (LLMs) remains difficult. A practical system must integrate domain knowledge (Liu et al. 2024a), patient information (Yang et al. 2022), and explicit logical reasoning (Goh et al. 2024). Recent general-purpose and domain-adapted models achieve strong open-benchmark scores (Kim, Wang et al. 2024; Tang et al. 2024). In clinical use, however, they may hallucinate drug dosages, misapply guidelines, or draw invalid causal links, reducing diagnostic reliability (Liu, Li et al. 2023).

LLM-based medical reasoning follows two main paradigms. *(a) Knowledge fine-tuning (Singhal et al. 2023; Wang et al. 2025)* retrains models on large medical corpora, improving accuracy at the cost of data, compute, and deployment agility. *(b) Reasoning stimulation (Liévin et al. 2024)* has been tried using multi-agent role-playing to accomplish tasks through discussion and cooperation, which is considered a flexible and low-cost solution (Moor, Huang et al. 2023; Kim et al. 2024). However, we found that most current multi-agent systems only engage in positional discussions based on their rulings and cannot argue about logical details in depth. Current frameworks are not able to effectively localize logic/rule conflicts, thus be difficult to improve the performance.

Inspired by the 'major premise-minor premise-conclusion' paradigm of the classical syllogism, we use the syllogism as a minimal reasoning unit (Khemlani and Johnson-Laird 2012; Jiang and Yang 2023). Each triad consists of a generalized medical law (major premise), a patient-specific fact (minor premise), and the corresponding conclusion. By concatenating or parallelizing multiple triads, we construct an inference tree whose leaf nodes store empirical observations or domain rules, internal nodes hold intermediate inferences, and the root node gives the final clinical decision (Howson 2005). Such inference trees offer two major advantages: (i) Traceability - each conclusion can be traced back to its supporting premises; (ii) Comparability - each intelligence can align the reasoning tree (logical tree) to pinpoint conflicts or omissions at the premise level. Embedding this explicit structure into a multi-agent framework provides an easily auditable logical template for systematic cross-intelligence error correction and consistency verification.

To this end, we propose a medical logical tree-based multi-agent framework (termed MEDLA) that enables multi-agent collaboration and discussion through a logic tree structure (in Fig. 1(a)). *Although LLM cannot explicitly handle tree logic, we can organize the answers to medical questions in the form of a logic tree by guiding them with prompt words.* We design premise agents to extract the

[*] equal contribution
[†] corresponding author.

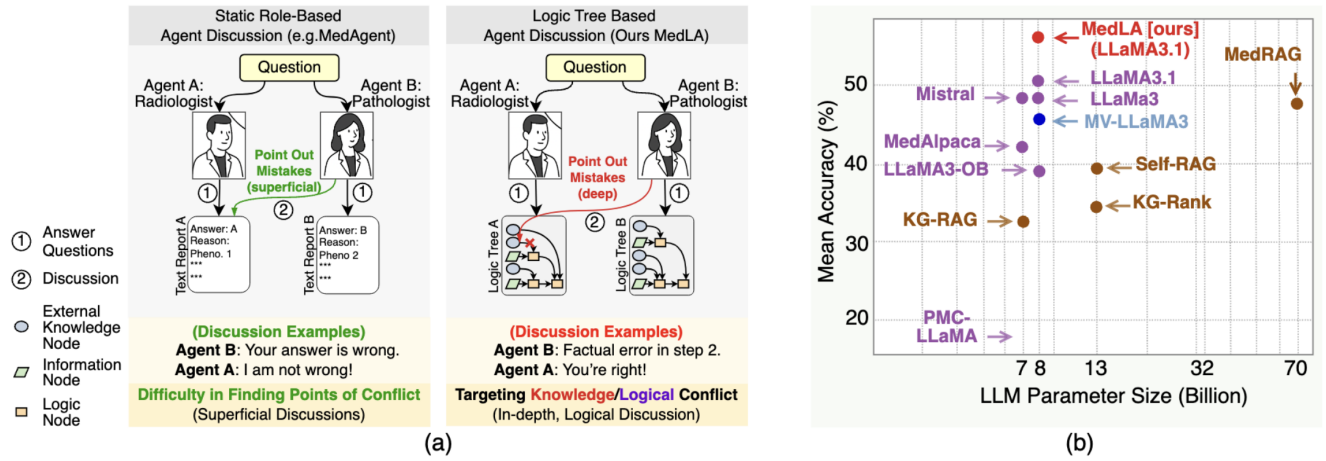

Figure 1: (a) Comparison between traditional role-based agent discussions and our proposed logic-based framework. (b) Performance and parameter comparison of MEDLA with existing systems. (a-Left) Traditional systems (e.g., MedAgent) assign agents fixed roles and aggregate their conclusions, leading to superficial discussions and difficulty identifying the root of disagreement. (a-Right) Our approach models each agent's reasoning as a logic tree, enabling inter-agent analysis of logical and knowledge-based inconsistencies. (b) MEDLA outperforms existing systems in the average accuracy of two benchmarks, demonstrating its effectiveness in handling complex medical reasoning tasks.

major premise and minor premise from the questions. Subquestions are generated by the premise agents and sent to the logical agents. Then we use medical agents to reason about the logical nodes and generate the logical tree with the final answer. Based on the logical tree, a multi-round discussion mechanism is designed to allow agents to discuss the logical tree and correct each other's errors. The final answer is generated based on the logical tree and the discussion results.

The experimental results show that our MedLA outperforms existing MedDDx benchmarks (Su et al. 2025a), medical QA benchmarks (Xiong et al. 2024), and medical reasoning benchmarks (Zuo et al. 2025) by a large margin on both open-source and commercial LLM (Fig. 1(b)).

Contributions. (a) We introduce MEDLA, the first multi-agent framework for medical reasoning that represents each agent's thought process as an explicit logical tree. This design enables fine-grained traceability of inferences and systematic detection of premise-level conflicts. (b) We develop a multi-round, graph-guided discussion mechanism in which agents iteratively compare and revise their logical trees, leading to robust cross-agent error correction and convergence to high-confidence, self-consistent reasoning structures. (c) We perform comprehensive evaluations on both differential diagnosis (MEDDDX) and standard medical QA benchmarks, demonstrating that MEDLA outperforms static role-based multi-agent systems and single LLM baselines.

Methods

Problem Definition, Syllogism, and Logical Tree

In this work, each clinical QA task dataset $\mathcal{D}$ is defined as a set of tuples $\{(Q_i, \mathcal{O}_i, C_i)\}_{i=1}^{N}$, where Q_i = 'What is the correct diagnosis/management for the patient?' is the question text, $\mathcal{O}_i = \{O_{i_1}, O_{i_2}, \ldots, O_{i_{N_K}}\}$ is the set of candidate answers, N_K is the number of candidate answers, and $C_i \in \mathcal{O}_i$ is the ground-truth correct answer. The task for our system is to predict the correct answer from $\mathcal{O}_i$ given the input $(Q_i, \mathcal{O}_i)$. We primarily consider the Multiple-choice Question (MCQ) format (Pal, Umapathi, and Sankarasubbu 2022).

Syllogism and Syllogism-based logical tree. A minimal reasoning unit in medical diagnosis can be abstracted as the classical syllogism (Smiley 1973), v : (p^{maj} - major premise) $\wedge$ (p^{min} - minor premise) $\rightarrow$ (C - conclusion), where v is the syllogism node, which is a tuple of the major premise p^{maj}, the minor premise p^{min}, and the conclusion C. By chaining or paralleling multiple syllogisms, we obtain a logical tree (Khemlani and Johnson-Laird 2012; Revlis 2015),

$$\begin{aligned} &\mathcal{T} = (V, E), E \subseteq V \times V, \\ &V = \{v_1, v_2, \cdots, v_i, \cdots, v_{N_K}\}, \end{aligned} \tag{1}$$

where $\mathcal{T}$ denotes the syllogism-based logical tree, V is the set of syllogism nodes,v_i is defined in Eq. (1), and E is the set of directed edges. Each $(v_{i_1}, v_{i_2}) \in E$ means that v_{i_1} is a necessary antecedent of v_{i_2}. By constructing a logical tree, we can represent the reasoning process in a structured manner, allowing for better understanding and analysis of the reasoning steps involved in concluding.We further establish the theoretical properties of this reasoning framework, such as its convergence and stability, with proofs provided in the Appendix.

Agent Designs in MedLA

Based on the above definitions of logical tree, we propose a multi-agent framework for complex medical reasoning, called MEDLA, which is designed to handle complex medical questions by dynamically invoking specialized agents

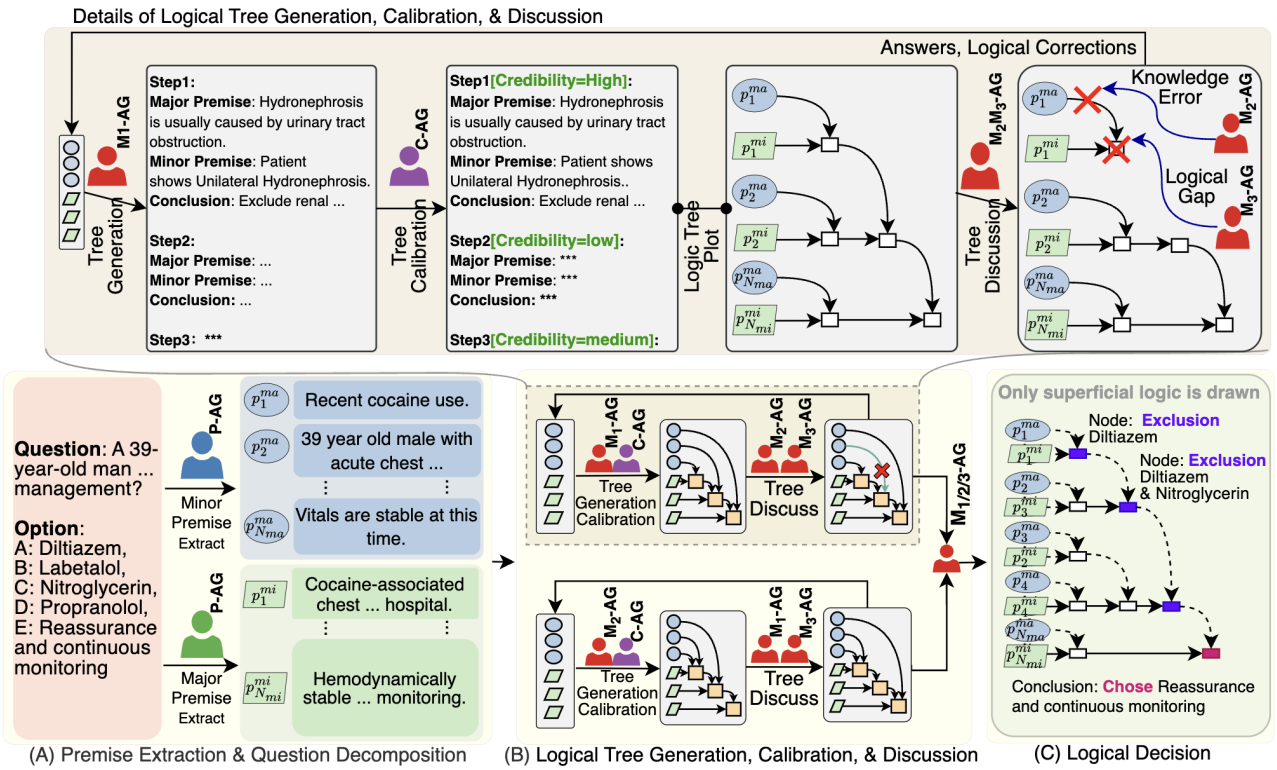

Figure 2: Overview of the proposed MedLA for complex medical reasoning. The system decomposes a medical query into logical sub-tasks, dynamically invokes specialized agents, and engages in collaborative reasoning to generate answers.

to collaboratively reason and generate comprehensive answers (in Fig. 2). The proposed framework consists of four main components, including a Premise Agent (P-Agent) for major/minor premise extraction, a Decompose Agent (D-Agent) for question splitting, multiple Medical Agents (M-Agents) for recursive tree generation, and a Credibility Agent (C-Agent) for node calibration. The system is designed to be modular and flexible, allowing for dynamic task decomposition and multi-agent collaboration.

Premise Agent (P-Agent) for major/minor premise extraction. The first step maps free text Q and external knowledge into major/minor premises that seed later syllogisms. From the question Q we extract entity-relation facts to obtain a patient fact set $\mathcal{P}^{\text{min}}$, and retrieve relevant rules $\mathcal{P}^{\text{maj}}$ from medical knowledge bases,

$$\begin{aligned}\mathcal{P}^{\text{maj}} &= \{p_1^{\text{maj}}, \ldots, p_r^{\text{maj}}, \ldots, p_{|N_{\text{maj}}|}^{\text{maj}}\} = \mathcal{O}_{\text{P-Agent}}^{\langle \text{pro}_{P_{\text{maj}}} \rangle}(Q), \\ \mathcal{P}^{\text{min}} &= \{p_1^{\text{min}}, \ldots, p_r^{\text{min}}, \ldots, p_{|N_{\text{min}}|}^{\text{min}}\} = \mathcal{O}_{\text{P-Agent}}^{\langle \text{pro}_{P_{\text{min}}} \rangle}(Q),\end{aligned} \tag{2}$$

where $\mathcal{P}^{Q,k}$ is the knowledge-major-premise set, $\mathcal{P}^{Q,\text{min}}$ is the patient-minor-premise set; each $p^{\text{maj},(n)}$ denotes the n-th retrieved medical rule, and each $p^{\text{min},(m)}$ denotes the m-th patient fact extracted from Q, $\mathcal{O}_{\text{P-Agent}}(\cdot)$ is the P-Agent function, and $\langle \text{pro}_{P_{\text{maj}}} \rangle$ and $\langle \text{pro}_{P_{\text{min}}} \rangle$ are the prompt templates. $|N_{\text{maj}}^{Q}|$ and $|N_{\text{min}}^{Q}|$ are the number of major and minor premises, respectively. The fixed premise extract prompt is listed in the Appendix.

Decompose Agent (D-Agent) for question splitting. Complex diagnostic problems often span multiple causal chains; decomposing these problems into atomic subproblems allows for more comprehensive thinking and discussion. D-Agent recursively splits Q into item questions $\{q_1, q_2, \ldots\}$,

$$\mathcal{S} = \{s_1, s_2, \ldots\} = \mathcal{O}_{\text{D-Agent}}^{\langle \text{pro}_D \rangle}(Q), \tag{3}$$

where $\mathcal{O}_{\text{D-Agent}}(\cdot)$ is the D-Agent function and $\langle \text{pro}_D \rangle$ is the prompt template (in the Appendix). The D-Agent prompt is designed to elicit a tree-like structure of questions, where each question q_s is a subproblem that can be answered independently. In this work, we adopt an elimination-based reasoning strategy to construct the question tree: if candidate answers are provided, the agent considers the plausibility of each option in turn using a process of elimination; for open-ended questions, the agent is first prompted to generate multiple possible answers and then considers each as an independent hypothesis.

Medical Agents (M-Agents) for logical tree generation. To obtain diverse and complementary reasoning perspectives, we run multiple M-Agents $\mathcal{M} = \{M^{(1)}, \ldots, M^{(j)}, \ldots, M^{(N_M)}, \}$ in parallel, where N_M is the number of M-Agents. Agent $M^{(j)}$ independently generates the next batch of derivable conclusions,

$$\mathcal{T}_{M^{(j)}} = \{V, E\} = \mathcal{O}_{\text{M(j)-Agent}}^{\langle \text{pro}_M \rangle}\left(\mathcal{P}^{\text{maj}}, \mathcal{P}^{\text{min}}, \mathcal{S}, \mathcal{T}_{\text{other}}\right), \tag{4}$$

where $\mathcal{O}_{\mathrm{M^{(j)}\text{-}Agent}}(\cdot)$ is j-th M-Agent, and $\langle \mathrm{pro}_M \rangle$ is the prompt template (in the Appendix). The prompt will guide the model to further split the subproblems and thus form a more complex logical tree $\mathcal{T}_{M^{(j)}}$. $\mathcal{T}_{M^{(j)}}$ contains a set of syllogism nodes $V = \{v_1, v_2, \ldots, v_i, \ldots, v_{|N_V|}\}$ and a set of directed edges E. $\mathcal{T}_{\text{other}}$ is an optional input, which is used to provide the tree structure of other M-Agents in the previous round for agent-agent discussion. It is worth noting that in this paper, since LLMs cannot directly produce structured tree outputs, the v_i we obtain by guiding LLM is a text block containing a syllogism(check Appendix).

Credibility Agent (C-Agent) for node calibration. To ensure the credibility of the generated logical tree, we introduce a C-Agent to evaluate the confidence of each syllogism node. The C-Agent is designed to assess the logical consistency and relevance of each node in the tree, providing a credibility score for each syllogism.

$$\mathcal{C}_{M^{(j)}} = \{c_1, c_2, \ldots, c_i, \ldots, c_{|N_V|}\} = \mathcal{O}_{\text{C-Agent}}^{\langle \mathrm{pro}_C \rangle}\big(T_{M^{(j)}}\big), \tag{5}$$

where $\mathcal{O}_{\text{C-Agent}}(\cdot)$ is the function of C-Agent, and $\langle \mathrm{pro}_C \rangle$ is the prompt template (in the Appendix). In this paper, we define $c_i \in \{\text{High credibility, Medium credibility, Low credibility}\}$, where i is the index of the syllogism node v_i in the logical tree $\mathcal{T}_{M^{(j)}}$.

Logical Reasoning Workflow of MedLA

Based on the above agent designs, we propose a three-stage pipeline for logical reasoning in complex medical QA tasks.

Phase A: Premise Extraction & Question Decomposition. The system begins with the P-Agent: given the raw question Q, it uses fixed prompt templates to extract a set of knowledge-major premises $\mathcal{P}^{\text{major}}$ and patient-minor premises $\mathcal{P}^{\text{minor}}$. Next, the D-Agent recursively splits Q into atomic sub-questions $\{q_1, q_2, \ldots\}$, creating placeholder nodes for each. If a sub-question coincides with a candidate answer, it is marked as a terminal goal; otherwise, it remains pending for further inference.

Phase B: Logical Tree Generation, Calibration, & Discussion. A cohort of M-Agents runs in parallel. Each M-Agent independently takes $\mathcal{P}^{\text{major}}$, $\mathcal{P}^{\text{minor}}$, and the set of placeholders, then emits a batch of syllogistic nodes and directed edges in one LLM call, forming a provisional local logical tree. Subsequently, the C-Agent reevaluates each node's confidence. Nodes labeled as low confidence are flagged and retained as discussion material for the next phase; all medium-high confidence nodes are locked into the local tree to preserve core reasoning structure. Then the system enters a discussion phase, where each M-Agent exchanges its local tree with others, allowing them to compare and contrast reasoning paths. These discussions repeat until all agents have shared their trees. The system then identifies any discrepancies between the trees, focusing on the flagged low-confidence nodes. Each agent is prompted to review and revise its tree based on the feedback from its peers, using a revision prompt ($\langle \mathrm{pro}_{\mathrm{Rev}} \rangle$) to validate, add or remove premises, and re-score affected nodes.

Phase C: Logical Decision. Once the discussion phase is complete, the system synthesizes the final logical tree by merging all local trees. The final tree is then used to generate the final answer. The answer is generated by traversing the logical tree and aggregating the conclusions from each syllogism node. The system can also provide a detailed explanation of the reasoning process, including the major and minor premises, the syllogisms used, and the final conclusion.

Experiments

Datasets & Benchmarks. We evaluate on three complementary benchmarks. **(i) MedDDx benchmarks.** To stress differential-diagnosis reasoning, we use MedDDx benchmark(Su et al. 2025a) : Tests differential diagnosis across **Basic**, **Intermediate**, and **Expert** tiers, with difficulty defined by the semantic similarity of distractors from STaRK-Prime (Wu et al. 2024). **(ii) Multi-choice medical QA benchmarks** (Xiong et al. 2024). A suite combining **MMLU-Med**, **MedQA-US**, and **BioASQ-Y/N** to test general medical knowledge. It covers factual recall, guideline interpretation, and clinical decision-making. **(iii) Expert-Level Medical Reasoning and Understanding benchmark**(Zuo et al. 2025).To assess expert-level reasoning, we use **MedXpertQA**, a challenging and comprehensive benchmark for advanced medical knowledge (details in Appendix). The benchmark serves as our general-purpose test bed. The details of the benchmarks are shown in Appendix.

Baselines. We benchmark MedLA against four representative paradigms. (i) **Graph-based reasoning** methods ground answers on biomedical knowledge graphs—QAGNN (Yasunaga et al. 2021), JointLK (Sun et al. 2022), and DRAGON (Yasunaga et al. 2022). (ii) **Multi-agent voting** systems aggregate LLM outputs without explicit logic trees—Majority Voting, DyLAN (Liu et al. 2024c), MedAgents (Tang et al. 2024), and MDAgents (Liu et al. 2024b). (iii) **Stand-alone LLMs** rely purely on parametric knowledge, including LLaMA-2-7B/13B, Mistral-7B, MedAlpaca-7B, PMC-LLaMA-7B, LLaMA 3-8B, LLaMA 3-OB-8B, and the stronger LLaMA 3.1-8B; we also report their chain-of-thought (CoT) variants. (iv) **Retrieval-augmented generation (RAG)** couples an LLM with external retrievers—Self-RAG (7B/13B) (Asai et al. 2023), KG-Rank (Yang et al. 2024), KG-RAG (Soman et al. 2023), and MedRAG (70B) (Xiong et al. 2024). Together, these baselines span parametric, retrieval-augmented, graph-grounded, and naïve multi-agent strategies, furnishing a comprehensive backdrop against which to gauge MedLA's contributions. We did not include methods that require additional data and require fine-tuning of the larger model (e.g., KGAREVION (Su et al. 2025b)).

Evaluation Metric & Testing Protocol & Implementation Details. Following (Su et al. 2025a), the performance is measured by accuracy (Acc), averaged over three independent runs ($\pm$ std). For every run, we shuffle the benchmark order and reset the LLM's sampling state. All experiments used the officially provided base model weights and configurations, and vLLM (v0.7.2). 8-card A100-80GB GPU

	Method	Reference	MedDDx Benchmarks (Su et al. 2025a) Basic Acc.(±std)	Intermediate Acc.(±std)	Expert Acc.(±std)	AVE
Graph Based Methods	QAGNN	NAACL2021	29.5(±0.3)	26.5(±0.2)	25.3(±0.3)	27.1
	JointLK	NAACL2022	24.7(±0.4)	25.3(±0.4)	24.4(±0.4)	24.8
	Dragon	NeurIPS2022	28.6(±0.3)	24.7(±0.2)	24.0(±0.4)	25.8
Multi Agents Methods	MV-LLaMA3.1(8B)	-	39.6(±1.0)	32.8(±0.6)	30.2(±0.8)	34.2
	DyLAN	COLM2024	39.3(±1.5)	33.5(±0.9)	31.1(±0.7)	34.6
	MedAgents	ACL2024	41.0(±0.7)	35.7(±1.1)	32.9(±1.5)	36.5
	MDAgents	NeurIPS2024	42.1(±1.3)	37.5(±0.9)	33.4(±0.6)	37.7
General & Medical LLMs	Mistral(7B)	Mistral2023	41.2(±0.3)	35.6(±0.3)	37.5(±0.7)	38.1
	MedAlpaca(7B)	BHT2023	39.9(±1.2)	32.5(±0.4)	31.1(±0.9)	34.5
	PMC-LLaMA(7B)	SJTU2024	8.7(±1.5)	8.6(±0.2)	7.9(±0.6)	8.4
	LLaMA3(8B)	Meta2024	42.8(±0.5)	31.9(±0.2)	30.6(±0.9)	35.1
	LLaMA3.1(8B)[baseline]	Meta2024	43.4(±1.8)	36.8(±0.2)	30.6(±2.1)	36.9
General & Medical LLMs with COT	CoT-Mistral(7B)	Mistral2023	40.4(±1.0)	36.8(±2.3)	37.9(±2.7)	38.4
	CoT-MedAlpaca(7B)	BHT2023	39.5(±0.7)	32.1(±1.1)	31.2(±1.0)	34.3
	CoT-PMC-LLaMA(7B)	SJTU2024	8.8(±0.2)	7.7(±0.4)	6.3(±0.5)	7.6
	CoT-LLaMA3(8B)	Meta2024	43.4(±0.9)	36.8(±0.4)	31.3(±0.3)	37.2
	CoT-LLaMA3.1(8B)	Meta2024	43.9(±1.7)	39.3(±0.5)	32.2(±1.4)	38.5
RAG & Based Methods	Self-RAG(7B)	ICLR2024	23.8(±0.7)	19.9(±3.7)	22.4(±4.5)	22.0
	Self-RAG(13B)	ICLR2024	24.9(±1.0)	29.0(±1.8)	26.6(±3.1)	26.8
	KG-Rank(13B)	ACL-w2024	25.3(±2.1)	25.6(±1.3)	23.4(±1.0)	24.8
	MedRAG(70B)	Oxon2024	36.5(±0.8)	34.8(±1.1)	32.7(±0.3)	34.7
Logic Based	MedLA+LLaMA3.1(8B)	Ours	**48.2(±1.2)**	**43.0(±2.1)**	**41.7(±0.8)**	**44.3 (↑ 7.4)**

Table 1: The performance of our MedLA on MedDDx Benchmarks. The table includes the accuracy along with the standard deviation(±std) for each metric. The results demonstrate the effectiveness of MedLA in addressing complex medical reasoning tasks across different datasets. OB means OpenBioLLM, MV means Majority Voting, and AVE means averaged accuracy. The best results are highlighted in bold, while the best results of each method's block are marked with an underline. Reference indicates the paper where the method was first introduced. The results of the baselines are taken from (Su et al. 2025a).

servers are used for testing. The raw data of the dataset involved in the experiments is adopted from MIRAGE[1]. More details can be found in the Appendix.

[Overall Performance Analysis] MedLA significantly outperforms baselines under open-source LLM settings. To evaluate the effectiveness of MedLA, we conducted comprehensive experiments on a series of medical reasoning benchmarks, all based on open-source large language models (e.g., LLaMA). To ensure the objectivity of the comparison, we directly refer to the baseline results reported in (Su et al. 2025a). We also provide a detailed summary in the Appendix.

Analysis: (a) MedLA outperforms all baselines on both standard and challenging benchmarks, achieving state-of-the-art (SOTA) results across QA datasets and excelling in expert-level diagnosis on MedDDx. These results demonstrate that MedLA not only enhances factual reasoning but also improves differential diagnostic performance in real-world medical settings. (b) MedLA Beyond CoT and Baselines: MedLA surpasses both base and CoT-enhanced models under the same open-source foundation, showing that our logic extraction strategy effectively distills more reliable knowledge and supports dynamic reasoning. (c) MedLA Stronger than Multi-Agent Systems: MedLA outperforms multi-agent baselines, indicating that its logic tree enables deeper interaction and better coordination among agents even without explicit role-based decomposition. (d) MedLA Outperforms RAG without External Knowledge: MedLA exceeds RAG-based models despite not using external retrieval, proving its strong internal reasoning ability through structured logic alone. (e) MedLA Remains Effective Under New Challenges:On a newly introduced test set,MedXpertQA, MedLA's performance remains outstanding, proving that its logic-enhanced reasoning ability is generalizable, rather than merely an optimization for known tasks, and is capable of effectively tackling new medical challenges.

[Overall Performance Analysis] MedLA demonstrates strong performance advantages on commercial LLMs, validating its generality and robustness. Our previous experiments were primarily conducted on open-source large language models, where MedLA had already shown promising results across various medical reasoning tasks. To further assess the adaptability and competitiveness of MedLA under more powerful backbones, we conducted additional evaluations using state-of-the-art commercial LLMs such as DeepSeek. All baseline methods were also implemented on the same commercial model to ensure fair comparison. For

[1]https://github.com/Teddy-XiongGZ/MIRAGE

	Method	Reference	Multi-choice medical QA benchmarks (Xiong et al. 2024) MMLU-Med Acc.(±std)	MedQA-US Acc.(±std)	BioASQ-Y/N Acc.(±std)	AVE
Graph Based Methods	QAGNN	NAACL2021	31.7(±0.6)	47.0(±0.3)	70.7(±0.6)	49.8
	JointLK	NAACL2022	28.8(±0.6)	42.5(±0.2)	70.6(±0.5)	47.3
	Dragon	NeurIPS2022	31.9(±0.3)	47.5(±0.2)	70.6(±0.3)	50.0
Multi Agents Methods	MV-LLaMA3.1(8B)	-	60.2(±0.5)	46.8(±0.4)	65.2(±0.4)	57.4
	DyLAN	COLM2024	62.5(±0.3)	51.6(±0.6)	63.8(±0.5)	59.3
	MedAgents	ACL2024	64.3(±0.4)	53.2(±0.3)	64.1(±0.6)	60.5
	MDAgents	NeurIPS2024	65.0(±0.2)	53.4(±0.2)	64.0(±0.3)	60.8
General & Medical LLMs	Mistral(7B)	Mistral2023	63.4(±0.4)	47.7(±0.7)	64.4(±0.1)	58.5
	MedAlpaca(7B)	BHT2023	60.0(±0.4)	40.1(±0.1)	49.3(±3.4)	49.8
	PMC-LLaMA(7B)	SJTU2024	20.7(±1.1)	24.7(±0.4)	34.6(±1.7)	26.7
	LLaMA3(8B)	Meta2024	63.4(±0.5)	56.6(±0.4)	65.4(±0.6)	61.8
	LLaMA3.1(8B)[baseline]	Meta2024	67.7(±0.7)	56.3(±0.6)	68.7(±0.6)	64.2
General & Medical LLMs with COT	COT-Mistral(7B)	Mistral2023	63.4(±0.3)	47.4(±0.2)	65.1(±0.2)	58.6
	COT-MedAlpaca(7B)	BHT2023	60.3(±0.4)	39.9(±0.3)	48.5(±2.5)	49.6
	COT-PMC-LLaMA(7B)	SJTU2024	20.4(±0.8)	20.8(±0.2)	20.8(±0.6)	20.7
	COT-LLaMA3(8B)	Meta2024	65.1(±0.5)	55.2(±0.3)	64.2(±0.5)	61.5
	COT-LLaMA3.1(8B)	Meta2024	68.1(±0.5)	54.9(±0.3)	70.6(±0.5)	64.5
RAG Based Methods	Self-RAG (7B)	ICLR2024	32.2 (±1.9)	38.0 (±2.8)	59.4 (±1.2)	43.2
	Self-RAG (13B)	ICLR2024	50.2 (±0.4)	40.8 (±2.0)	64.6 (±5.0)	51.9
	KG-Rank (13B)	ACL-w2024	45.2 (±0.5)	36.2 (±1.1)	50.3 (±1.5)	43.9
	MedRAG (70B)	Oxon2024	57.9 (±1.5)	48.7 (±1.4)	71.9 (±1.8)	59.5
Logic Based	MedLA + LLaMA3.1(8B)	Ours	**70.7(±0.1)**	**62.6(±0.1)**	**76.5(±0.1)**	**69.9(↑5.7)**

Table 2: The performance of our proposed MedLA model on Multi-choice medical QA benchmarks (Xiong et al. 2024). The table includes the accuracy (Acc) along with the standard deviation (±std) for each metric. The results demonstrate the effectiveness of MedLA in addressing complex medical reasoning tasks across different datasets.

Model	Method	Number	Score
deepseek-r1	MedLA	60	36.0(±4.3)
	baseline	60	21.3(±4.9)
deepseek-v3	MedLA	60	25.6(±3.1)
	baseline	60	15.0(±0.0)

Table 3: Performance comparison of **MedLA** and **baseline** on DeepSeek-based reasoning evaluated on the MedXpertQA benchmark(Zuo et al. 2025).

Variant	MEDQA-US	MEDDDX	
	Acc±std	Basic	Expert
MEDLA (full)	**62.6**±0.1	**48.2**±2.1	**41.7**±0.8
-Revision loop	58.4±0.3	44.2±1.9	38.6±1.0
-Credibility	57.3±0.4	41.8±1.7	37.2±1.3
-LogicTree (CoTOnly)	56.1±0.4	38.7±1.5	34.9±1.2
MV	54.8±0.4	37.5±0.9	30.2±0.8

Table 4: Step-wise ablation of MEDLA. The last row Majority Voting (MV) is a naïve majority-vote ensemble that keeps only the common backbone (LLAMA 3.1-8B) and the same agent prompts. All numbers are three-seed means ± std.

objectivity, we referenced the baseline results reported in (Su et al. 2025a). The outcomes are presented in Table 3.

[Detailed Difficulty vs. Performance Analysis] MedLA gives greater lift to more difficult tasks. We examine how the proposed logic-tree framework scales with task difficulty by comparing MedLA to its LLaMA-3.1-8B backbone on the three graded subsets of MedDDx (basic, intermediate, expert). Three random seeds are evaluated per tier; mean accuracy and standard deviation are reported in Fig. 3-left. All experimental factors—retriever, decoding temperature, and candidate pool—are controlled, so any performance delta reflects the contribution of MedLA's multi-agent reasoning.

Analysis: The relative improvement grows monotonically with difficulty: +4.6 pp (percentage point) on basic, +6.4 pp on intermediate, and +11.1 pp on the expert tier. Confidence intervals also narrow, indicating increased prediction stability. These results suggest that explicit logic-tree exchange and cross-agent revision become progressively more beneficial as diagnostic options converge semantically, reinforcing the value of structure-level reasoning for the most challenging clinical cases.

[Detailed Base Model vs. Performance Analysis] MedLA Continuous Enhancement on a Stronger Base Model. To assess MedLA's benefits beyond a single back-

Method	Component time (sec.)				Total
	FT	RT	GBT	IFT(sec.)	
Majority Voting	–	–	–	1 853	1 853
MedAgents	–	–	–	2 793	2 793
KG-RAG	–	603	–	1 845	2 448
KGAREVION	10k+	–	–	1 821	10k+
MedLA (ours)	–	–	–	3 657	3 657

Table 5: Time Consumption Analysis. Wall-clock (sec.) latency on BIOASQ-Y/N. FT = extra fine-tuning, RT = retrieval, GBT = logic-graph build (incl. revision loop), IFT = pure LLM inference. '–' indicates no time cost.

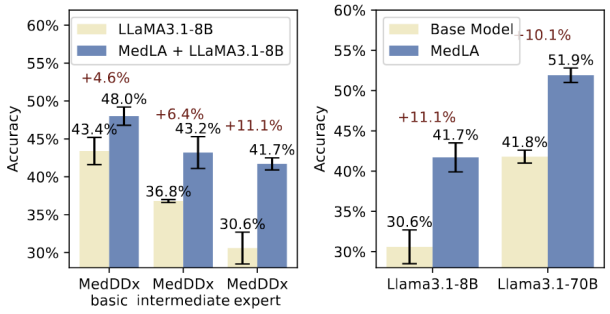

Figure 3: Performance comparison of MedLA with LLaMA3.1-8B at different levels of difficulty on the MedDDx benchmark. Error bars represent SD.

bone, we compare its gains over both the 8-bit and 70-bit variants of LLaMA-3.1 on the MedDDx-Expert tier (in Fig. 3-right).

Analysis: MedLA yields consistent, substantial improvements on both backbones: from 30.6% to 41.7% (+11.1 pp) for 8B, and from 41.8% to 51.9% (+10.1 pp) for 70 B. These results demonstrate that our structured, multi-agent reasoning adds value even as the underlying LLM scales up, underscoring the generality and robustness of the logic-tree approach across model sizes.

[Ablation Study] Each MedLA module contributes additively to final accuracy. To evaluate the independent contributions of the three modules of logic tree, confidence calibration, and multi-round correction, we eliminated each component one by one under the same LLaMA-3.1-8B backbone, the same cueing and decoding hyperparameters (see Table 4), and averaged them with 3 random seeds.

Analysis: Dropping the Revision loop lowers accuracy by 2.2 pp on MedQA-US and roughly 1.8 pp on both MedDDx tiers, confirming that structured peer feedback yields tangible gains. Suppressing the Credibility Agent causes an additional 1.1-1.4 pp decline, showing that calibrated confidence scores steer agents towards more reliable updates. Removing the entire Logic-tree scaffold—thereby falling back to plain chain-of-thought outputs—produces the steepest drop (-4.5 pp on MedQA-US, -4.3 pp on MedDDx-Expert).

[Time Consumption Analysis] The complexity and time consumption of MedLA is manageable. To quantify the latency of the different methods in a real inference process, we recorded the wall-clock time on the BIOASQ-Y/N dataset. All models are deployed on the same A100-80GB, and the decoding temperature is kept consistent with the number of concurrent threads to avoid interference from hardware differences. Table 5 splits the total elapsed time into four components: additional fine-tuning (FT), external retrieval (RT), logic graph construction and revision (GBT), and pure language model inference (IFT). For KGAREVION, which is only parameter fine-tuning, we add the officially reported 10 k-second training time to the total cost.

Analysis: Most baselines perform forward inference only once, with latency linearly proportional to model size; KGAREVION performs rapid inference but requires several additional hours of fine-tuning and has the highest overall cost. medLA performs stepwise inference through 17 subagents, which is about 2 times higher than the simple majority-voting scheme but much lower than KGAREVION, which requires offline training and is within the acceptable range. within the acceptable interval. More importantly, MedLA does not introduce additional retrieval or offline fine-tuning sessions.

Conclusion

We present MedLA, a logic-driven multi-agent system that decomposes clinical questions into structured syllogistic trees and enables iterative, cross-agent discussions to resolve logical and knowledge conflicts. Notably, these improvements require no fine-tuning or external retrieval, highlighting the value of structured logic and collaborative reasoning.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 62306313), the Chinese Academy of Medical Sciences Innovation Fund for Medical Sciences (2024-I2M-TS-035), and the Zhejiang Province Selected Funding for Postdoctoral Research Projects (ZJ2025113). Additional support was provided by the InnoHK program, and Ant Group through the CAAI-Ant Research Fund. We also thank the Funding from the ROOTCLOUD TECHNOLOGY CO.,LTD.

We thank the China-Japan Friendship Hospital for providing medical expertise and resources, and Zhipu AI for their API support. We are grateful to Prof. Zhen Lei from the CAIR, Hong Kong Institute of Science and Innovation (HKISI) and Prof. Stan Z. Li from Westlake University for their valuable suggestions and comments.

References

Asai, A.; Wu, Z.; Wang, Y.; Sil, A.; and Hajishirzi, H. 2023. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. *arXiv preprint arXiv:2310.11511*.

Chang, Y.; Wang, X.; Wang, J.; Wu, Y.; Yang, L.; Zhu, K.; Chen, H.; Yi, X.; Wang, C.; Wang, Y.; et al. 2024. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3): 1–45.

Goh, E.; Gallo, R.; Hom, J.; Strong, E.; Weng, Y.; Kerman, H.; Cool, J. A.; Kanjee, Z.; Parsons, A. S.; Ahuja, N.; et al.

2024. Large language model influence on diagnostic reasoning: a randomized clinical trial. *JAMA Network Open*, 7(10): e2440969–e2440969.

Howson, C. 2005. *Logic with trees: an introduction to symbolic logic*. Routledge.

Jiang, C.; and Yang, X. 2023. Legal syllogism prompting: Teaching large language models for legal judgment prediction. In *Proceedings of the nineteenth international conference on artificial intelligence and law*, 417–421.

Kasneci, E.; Seßler, K.; Küchemann, S.; Bannert, M.; Dementieva, D.; Fischer, F.; Gasser, U.; Groh, G.; Günnemann, S.; Hüllermeier, E.; et al. 2023. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and individual differences*, 103: 102274.

Khemlani, S.; and Johnson-Laird, P. N. 2012. Theories of the syllogism: A meta-analysis. *Psychological bulletin*, 138(3): 427.

Kim, D.; Wang, X.; et al. 2024. MDAgents: An Adaptive Collaboration of LLMs for Medical Decision-Making. *arXiv preprint arXiv:2406.06782*.

Kim, Y.; Park, C.; Jeong, H.; Chan, Y. S.; Xu, X.; McDuff, D.; Lee, H.; Ghassemi, M.; Breazeal, C.; and Park, H. W. 2024. MDAgents: An Adaptive Collaboration of LLMs for Medical Decision-Making. ArXiv:2404.15155 TLDR: The novel framework, Medical Decision-making Agents, aims to address the gap in strategic deployment of LLMs by automatically assigning the effective collaboration structure for LLMs, and explores the dynamics of group consensus, offering insights into how collaborative agents could behave in complex clinical team dynamics.

Liévin, V.; Hother, C. E.; Motzfeldt, A. G.; and Winther, O. 2024. Can large language models reason about medical questions? *Patterns*, 5(3).

Liu, C.; Li, C.; et al. 2023. LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day. *arXiv preprint arXiv:2306.00890*.

Liu, J.; Zhang, C.; Guo, J.; Zhang, Y.; Que, H.; Deng, K.; Liu, J.; Zhang, G.; Wu, Y.; Liu, C.; et al. 2024a. Ddk: Distilling domain knowledge for efficient large language models. *Advances in Neural Information Processing Systems*, 37: 98297–98319.

Liu, Z.; Zhang, Y.; Li, P.; Liu, Y.; and Yang, D. 2024b. Dynamic LLM-Agent Network: An LLM-agent Collaboration Framework with Agent Team Optimization.

Liu, Z.; Zhang, Y.; Li, P.; Liu, Y.; and Yang, D. 2024c. A Dynamic LLM-Powered Agent Network for Task-Oriented Agent Collaboration. arXiv:2310.02170.

Moor, M.; Huang, Q.; et al. 2023. Med-Flamingo: A Multimodal Medical Few-shot Learner. *arXiv preprint arXiv:2307.15189*.

Pal, A.; Umapathi, L. K.; and Sankarasubbu, M. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, 248–260. PMLR.

Revlis, R. 2015. Syllogistic reasoning: Logical decisions from a complex data base. In *Reasoning: Representation and process*, 93–133. Psychology Press.

Singhal, K.; Azizi, S.; Tu, T.; Mahdavi, S. S.; Wei, J.; Chung, H. W.; Scales, N.; Tanwani, A.; Cole-Lewis, H.; Pfohl, S.; et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972): 172–180.

Smiley, T. J. 1973. What is a syllogism? *Journal of philosophical logic*, 136–154.

Soman, K.; Rose, P. W.; Morris, J. H.; and et al. 2023. Biomedical Knowledge-Graph-Optimized Prompt Generation for Large Language Models. *arXiv preprint arXiv:2311.17330*.

Su, X.; Wang, Y.; Gao, S.; Liu, X.; Giunchiglia, V.; Clevert, D.-A.; and Zitnik, M. 2025a. KGARevion: An AI Agent for Knowledge-Intensive Biomedical QA. In *The Thirteenth International Conference on Learning Representations*.

Su, X.; Wang, Y.; Gao, S.; Liu, X.; Giunchiglia, V.; Clevert, D.-A.; and Zitnik, M. 2025b. KGARevion: An AI Agent for Knowledge-Intensive Biomedical QA. ArXiv:2410.04660 [cs].

Sun, Y.; Shi, Q.; Qi, L.; and Zhang, Y. 2022. JointLK: Joint Reasoning with Language Models and Knowledge Graphs for Commonsense Question Answering. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 5049–5060. Seattle, USA: Association for Computational Linguistics.

Tang, X.; Zou, A.; Zhang, Z.; et al. 2024. MedAgents: Large Language Models as Collaborators for Zero-shot Medical Reasoning. *arXiv preprint arXiv:2311.10537*.

Wang, H.; Zhao, S.; Qiang, Z.; Li, Z.; Liu, C.; Xi, N.; Du, Y.; Qin, B.; and Liu, T. 2025. Knowledge-tuning large language models with structured medical knowledge bases for trustworthy response generation in Chinese. *ACM Transactions on Knowledge Discovery from Data*, 19(2): 1–17.

Wu, S.; Zhao, S.; Yasunaga, M.; Huang, K.; Cao, K.; Huang, Q.; Ioannidis, V.; Subbian, K.; Zou, J. Y.; and Leskovec, J. 2024. Stark: Benchmarking llm retrieval on textual and relational knowledge bases. *Advances in Neural Information Processing Systems*, 37: 127129–127153.

Xiong, G.; Jin, Q.; Lu, Z.; and Zhang, A. 2024. Benchmarking Retrieval-Augmented Generation for Medicine. *arXiv preprint arXiv:2402.13178*.

Yang, R.; Liu, H.; Marrese-Taylor, E.; and et al. 2024. KG-Rank: Enhancing Large Language Models for Medical QA with Knowledge Graphs and Ranking Techniques. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, 155–166. Association for Computational Linguistics.

Yang, X.; Chen, A.; PourNejatian, N.; Shin, H. C.; Smith, K. E.; Parisien, C.; Compas, C.; Martin, C.; Costa, A. B.; Flores, M. G.; et al. 2022. A large language model for electronic health records. *NPJ digital medicine*, 5(1): 194.

Yasunaga, M.; Bosselut, A.; Ren, H.; Zhang, X.; Manning, C. D.; Liang, P.; and Leskovec, J. 2022. DRAGON: Deep Bidirectional Language–Knowledge Graph Pretraining. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*. ArXiv:2210.09338.

Yasunaga, M.; Ren, H.; Bosselut, A.; Liang, P.; and Leskovec, J. 2021. QA-GNN: Reasoning with Language Models and Knowledge Graphs for Question Answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 535–546. Online: Association for Computational Linguistics.

Zuo, Y.; Qu, S.; Li, Y.; Chen, Z.; Zhu, X.; Hua, E.; Zhang, K.; Ding, N.; and Zhou, B. 2025. Medxpertqa: Benchmarking expert-level medical reasoning and understanding. *ICML*.