Multi-Agent Medical Reasoning with Large Language Models Guided by Cognitive Uncertainty

Anonymous ACL submission

Abstract

Multi-agent systems (MAS) based on Large Language Models (LLMs) show potential for complex medical reasoning through collaboration. However, current methods face two challenges: (1) they neglect agent-specific cognitive proficiency, resulting in suboptimal expert selection; (2) they ignore to model uncertainty during collaboration, allowing unreliable information to mislead subsequent decisions. To address these issues, we propose *MedCUR*, which leverages cognitive uncertainty to guide the overall reasoning process. We first quantify each agent's cognitive proficiency by jointly modeling task accuracy and uncertainty across medical question categories, providing a principled basis for selecting appropriate doctor agents for each question. The framework consists of a meta-agent and multiple doctor agents. Given a medical question, the meta-agent identifies required expert types and selects appropriate agents based on proficiency. During the discussion, agents explicitly express uncertainty in their statements. The meta-agent then distinguishes reliable information from uncertain viewpoints and dynamically introduces additional third-party experts when needed. Experimental results demonstrate that *MedCUR* significantly improves medical reasoning performance.

1 Introduction

Multi-agent systems (MAS) based on Large Language Models (LLMs) have demonstrated promising performance in solving complex medical reasoning tasks by leveraging LLMs' strong reasoning capabilities and extensive medical knowledge (Wang et al., 2025a; Peng et al., 2025).

Existing LLM-based medical reasoning studies often adopt multi-agent systems to simulate multidisciplinary collaboration in clinical practice. Typical approaches assign specific disciplinary or functional roles to different agents while sharing the same backbone model (Li et al., 2025; Chen et al., 2025b), and some works further adjust the scale of agent collaboration dynamically according to task difficulty (Kim et al., 2024; Chen et al., 2025a). In addition, some studies enable agents to explicitly express their own uncertainty during communication, thereby guiding subsequent agents to consider the received information more cautiously (Wang, 2025; Lu et al., 2024).

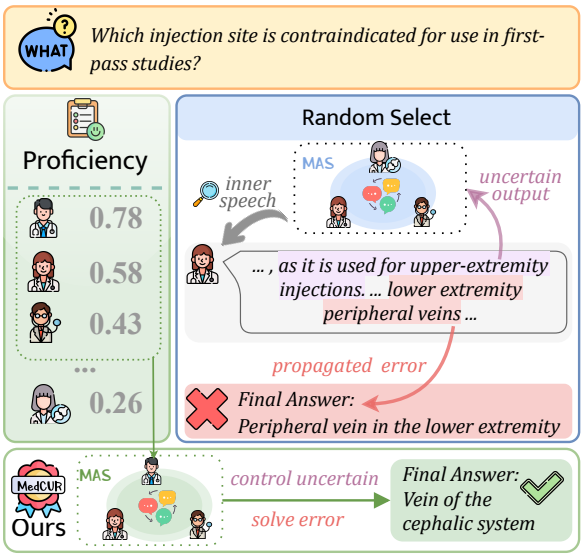


Figure 1: Random agent selection in multi-agent systems under unknown agent proficiency and the propagation of uncertain information can lead to incorrect answers.

However, existing methods still face two challenges: (1) ***Ignoring cognitive proficiency***: Although some methods assign predefined doctor roles to agents, these role assignments are independent of the agents' fine-grained cognitive proficiency. Consequently, agent selection and dynamic orchestration fail to leverage individual strengths, resulting in suboptimal collaboration in complex medical reasoning tasks. (2) ***Uncertainty propagation***:Current approaches lack explicit uncertainty management during conversational control, allow-

ing unreliable or doubtful information to propagate throughout the discussion and potentially compromise final medical decisions. As shown in Figure 1, if agents are selected randomly without considering their cognitive proficiency, the framework may fail to choose agents that are capable of providing reliable analysis and correct answers. Moreover, when agents express uncertain information without explicit uncertainty indicators, subsequent agents may mistakenly treat it as factual knowledge and build upon it in subsequent discussions. This can cause errors to accumulate and ultimately lead to incorrect final answers.

Therefore, we propose *MedCUR*, a framework consisting of a meta-agent and a set of doctor agents to enhance agent assignment and discussion control based on cognitive uncertainty. To address the first challenge, we design a risk-aware utility score, inspired by risk-sensitive objectives (Fei et al., 2020), to estimate cognitive proficiency. Specifically, we conduct multiple inference runs on medical datasets using widely used LLMs across different medical categories. We then statistically measure accuracy and self-consistency to jointly compute cognitive proficiency. For a given question, the meta-agent first analyzes the required medical categories and determines the number of doctor agents needed to solve the problem.

To address the second challenge, during the discussion, doctor agents are required to report the cognitive uncertainty of their responses and explicitly identify unresolved doubts that are critical for solving the question. The meta-agent identifies uncertain information and unresolved doubts from the discussion and decides whether to introduce an external expert to reduce uncertainty and resolve doubts. Doctor agents are subsequently guided to focus on the remaining uncertainties in subsequent discussion.Finally, inspired by consensual decision-making (Hartmann et al., 2009), the final decision is made by the leader by aggregating the opinions of other members with proficiency-based weights. The meta-agent integrates the viewpoints of all doctors based on their cognitive proficiency, rather than treating all agents equally.

We conduct extensive experiments to demonstrate the effectiveness of our framework, which improves performance across multiple benchmarks by approximately 4% on average compared to the single-agent baseline.

Our contributions can be summarized as follows:

- We propose *MedCUR*, a framework that dynamically orchestrates the multi-agent system based on cognitive proficiency for each medical question.
- We leverage cognitive uncertainty to regulate the discussion process and resolve uncertain information.
- Experimental results demonstrate that our method achieves superior performance on medical reasoning benchmarks, especially on challenging reasoning tasks.

2 Related Works

LLMs have been widely adopted for medical diagnosis through single-agent frameworks that integrate clinical knowledge, retrieval mechanisms, and structured reasoning. Systems such as Med-PaLM (Singhal et al., 2023), GPT-4-based medical question answering models (Nori et al., 2023), and retrieval-augmented diagnostic assistants (Yu et al., 2025) achieve strong performance on benchmarks such as MedQA (Jin et al., 2021) and PubMedQA (Jin et al., 2019). These systems effectively support a wide range of clinical tasks (Hu et al., 2025; Wang et al., 2025b), including medical question answering, report generation, and clinical decision making.

To further enhance performance on multidisciplinary collaborative reasoning tasks, recent studies have proposed multi-agent medical diagnosis systems to simulate the clinical reasoning process of multidisciplinary teams (MDTs). Frameworks such as MDAgents (Kim et al., 2024), MedAgents (Tang et al., 2024), and AI Hospital (Fan et al., 2025) assign multiple LLM-based agents to different clinical roles or diagnostic stages. Building upon this, MDTeamGPT (Chen et al., 2025a) further incorporates historical consultation experience into multi-round collaborative reasoning to provide reference for subsequent decision making, while MAC (Chen et al., 2025b) introduces supervisory agents to coordinate decision making and control the progress of discussions. However, existing multi-agent medical reasoning frameworks remain insufficient in modeling cognitive capability differences among agents and in employing uncertainty-driven dialogue control during collaboration.

3 Method

In this paper, we propose *MedCUR* to better guide multi-agent collaboration for solving complex medical questions. Figure 2 illustrates the overall frame-

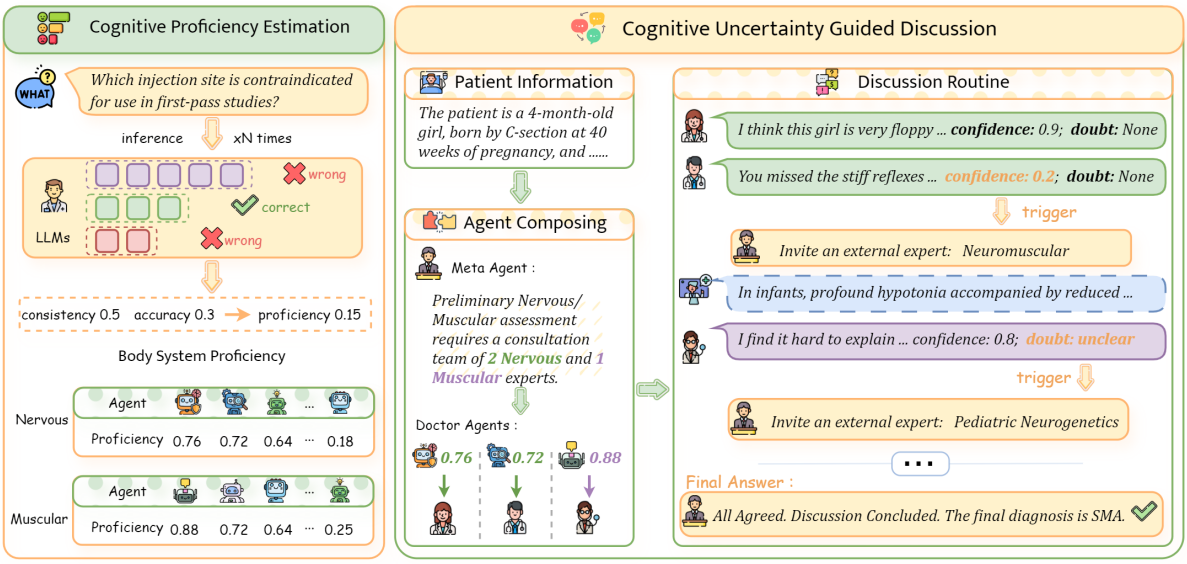


Figure 2: The framework of *MedCUR*. We first estimate the cognitive proficiency of existing popular LLMs. The meta-agent then dynamically selects agents based on cognitive proficiency and regulates the discussion using confidence-based uncertainty signals, where lower confidence indicates higher cognitive uncertainty.

work of our method. We first estimate the cognitive proficiency of multiple LLMs across medical categories. Based on the input question, the meta-agent then selects a customized group of doctor agents and leverages the uncertainty scores of their responses to regulate the discussion process. The detailed prompt templates used for each module are provided in Appendix A.

3.1 Cognitive proficiency

Since accuracy alone cannot distinguish consistently correct reasoning from accident success, while uncertainty does not inherently reflect correctness, integrating both is necessary to more reliably assess cognitive reliability in multi-agent medical reasoning frameworks.

To ground agent selection in the LLM's medical cognitive level, we reflect the LLM's proficiency q across different body-system categories that jointly considers accuracy and cognitive uncertainty. The body-system taxonomy follows the 11 categories utilized in the MedXpertQA dataset (Zuo et al., 2025). Specifically, we evaluate existing popular LLMs on a subset of the dataset consisting of 244 instances, where each question is inferred 10 times. The accuracy over the 10 responses is then computed and denoted as a. We then design the risk-aware utility score to comparatively estimate cognitive proficiency by incorporating a dis-utility contribution matrix (Howard and Matheson, 1972) that accounts for risk-based rewards:

$$p = t_p * e^{-\beta r}, \tag{1}$$

where t_p denotes the transition probability, β is the risk-sensitive parameter, r represents the uncertainty-based reward. We use accuracy to approximate t_p. The uncertainty reward r is reflected using self-consistency s (Xiong et al., 2024), defined as $r = 1 - s = 1 - \frac{\sum_i f_i - f_m}{\sum_i f_i}$, where f_i denotes the frequency of each candidate answer generated by the model across multiple samples, and f_m is the frequency of the majority answer. A smaller r indicates higher consistency, which corresponds to a higher proficiency p.

When $\beta > 0$, the objective is risk-averse, penalizing uncertainty more strongly. When $\beta = 0$, the formulation reduces to the risk-neutral expected performance. To balance risk and performance, we adopt the risk-neutral setting. We approximate the risk-sensitive objective in the limit ($\beta \to 0$) using a first-order Taylor expansion, reducing the utility to a linear form:

$$\begin{aligned} p &= a * (1 - \beta r + O(\beta^2)) \\ &\approx a * (1 - \beta + \beta s). \end{aligned} \tag{2}$$

Since only the relative ordering of models matters, $1 - \beta + \beta s$ is linear and monotonically increasing in s, so the constant term can be ignored. Therefore, the final utility score, which preserves

relative utility and measures cognitive proficiency, can be formulated as:

$$p = a * s. \tag{3}$$

This formulation provides a simple and theoretically grounded measure of cognitive proficiency that accounts for both accuracy and uncertainty in LLM responses. A higher score indicates that the agent is more capable of providing accurate and stable information.

3.2 Cognitive Uncertainty Guided Discussion

To better guide the discussion progress, we leverage cognitive uncertainty throughout the whole discussion by estimating the uncertainty of agents' statements and adaptively controlling the MAS structure.

3.2.1 Agent Composing

The discussion process consists of a meta-agent and a set of doctor agents. The meta-agent is responsible for coordinating the overall discussion, including analyzing the medical question, assigning appropriate experts, and integrating the final decisions. While each doctor agent focuses on medical reasoning from its specialized perspective.

To adaptively construct the discussion team for a given question, the meta-agent first analyzes the medical query through the instruction in Appendix A.2 and identifies the fine-grained expertise required. It then dynamically determines both the expert categories and the corresponding number of doctor agents n for each category according to the complexity of the question. For each selected category, the framework assigns the top-n doctor agents with the highest proficiency in the corresponding expertise area to participate in the discussion. This adaptive assignment strategy enables the framework to allocate specialized reasoning resources more effectively, improving both the diversity and reliability of medical reasoning.

3.2.2 Discussion Routine

Agent Speech. The discussion begins with the doctor agent selected by the meta-agent as most relevant to the given question. This doctor agent is responsible for the following requirements:

- Actively engage in discussions with other doctors, considering reliable viewpoints to solve the question.
- Offer valuable insights into uncertain viewpoints to advance the discussion and continuously refine your answer.
- Explicitly identify professional doubts that you cannot resolve independently but that are critical for solving the question. If no such doubts exist, indicate "None".

During the discussion, each doctor agent is required to report a confidence score c, where lower confidence corresponds to higher cognitive uncertainty of its response, and to identify unresolved professional doubts that cannot be addressed independently but are critical for solving the question. Above procedure is guided by the prompt provided in Appendix A.4.

After receiving the doctor agent's response, the meta-agent first separates reliable viewpoints from uncertain viewpoints based on the discussion content and propagates them to all agents.

Uncertainty Mitigation. When an agent's confidence score falls below the alert threshold ($c < \tau$) or unresolved doubts are reported, the meta-agent is triggered to handle the uncertain information. Specifically, the meta-agent identifies the fine-grained subfield expertise required to resolve the uncertain information under the guidance of the prompt in Appendix A.5. Rather than relying solely on coarse medical categories, the meta-agent generates a detailed description of the divergent concepts that require clarification to specify the appropriate expert. This description is mapped to the corresponding medical category, and the LLM with the highest cognitive proficiency in that category is assigned the role of a third-party expert.

The third-party expert is instructed to provide expert-level conceptual explanations without directly answering the original question. This design ensures that the expert serves in a knowledge support role rather than as a decision-maker.

After receiving expert feedback, the meta-agent broadcasts it to all doctor agents and continues the discussion.

3.2.3 Meta-decision Integration

The discussion terminates when all doctor agents reach a consensus or when the maximum number of discussion rounds is reached. Based on consensual decision-making (Hartmann et al., 2009), the final decision is aggregated by the leader as a weighted combination of the opinions of other members. Therefore, the meta-agent aggregates the

final responses from all doctor agents and weights them according to their cognitive proficiency to produce the final answer based on the prompt in Appendix A.6.

Through the above design, *MedCUR* adaptively determines the MAS structure both before and during the discussion, making multi-agent collaboration more effective and flexible.

4 Experimental Setting

4.1 Models

To estimate LLM cognitive proficiency, we evaluate nine models spanning diverse parameter scales and architectures, including standard transformers and Mixture-of-Experts (MoE): DeepSeek-R1 (DeepSeek-AI, 2025), Claude Sonnet 4.5 (Anthropic, 2025), Gemini 2.5 pro (Team, 2025a), GPT-5.1 (OpenAI, 2025), GPT-oss-120B, GPT-oss-20B, Qwen3-Max (Qwen Team, 2025), Qwen3-30B-A3B (Team, 2025b), and Qwen3-8B.

4.2 Benchmarks

We employ the following benchmarks to assess medical reasoning performance across varying difficulty levels:

• **RareDisease**(Chen et al., 2025b)**:** A challenging non-multiple-choice benchmark for rare disease diagnosis with 302 questions, emphasizing long-chain reasoning.

• **MedXpertQA**(Zuo et al., 2025)**:** An expert-level multiple-choice benchmark covering diverse body systems. We use its sampled version with 244 questions.

• **MedMCQA**(Pal et al., 2022)**:** A large-scale multiple-choice dataset from Indian medical entrance exams.

• **MedQA**(Jin et al., 2021)**:** A USMLE-style multiple-choice benchmark.

4.3 Baselines

For baselines, we select several representative multi-agent methods.

• **MedAgents** (Tang et al., 2024) simulates multidisciplinary consultations through multi-round discussions among doctor agents, emphasizing role-based collaborative reasoning.

• **MAC** (Chen et al., 2025b) introduces a supervisory agent to coordinate interactions among different doctor agents and is primarily designed for medical diagnosis scenarios.

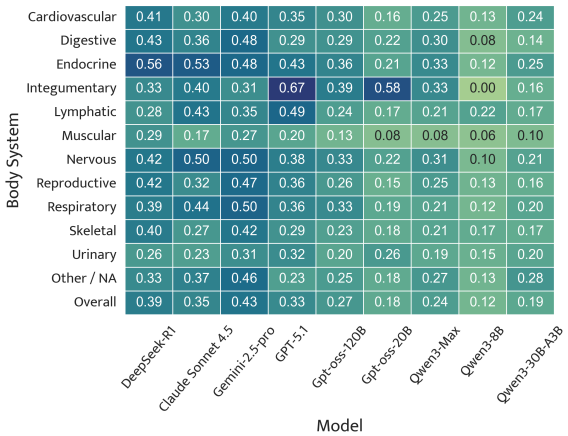


Figure 3: The cognitive proficiency of existing popular LLMs.

• **ReConcile** (Chen et al., 2024) is a general multi-agent deliberation framework that improves reasoning quality through iterative discussions and consensus building.

• **Healthcare Agent** (Ren et al., 2025) adopts a predefined multi-agent collaboration workflow in healthcare settings to support complex medical tasks.

4.4 Framework Settings

Following prior methods, we set the maximum number of discussion rounds to 10. To prevent unfair comparisons due to an excessive number of agents, we limit the meta-agent to assigning a maximum of three doctor agents. Considering both performance and computational cost, we use DeepSeek-R1 as the meta-agent. Since medical reasoning tasks require high-precision reasoning, we set the confidence alert threshold τ to 0.8. For fair comparison, all multi-agent baselines are implemented using the same backbone LLM (DeepSeek-R1), unless otherwise specified.

5 Experimental Results

5.1 LLM's Cognitive Level

We first analyze the cognitive proficiency of existing LLMs as the basis for agent selection. Figure 3 shows that Gemini 2.5 pro achieves the best overall performance, indicating its strong advantage in general medical reasoning ability. Meanwhile, different LLMs exhibit distinct strengths across body systems, suggesting that cognitive proficiency is not uniformly distributed but is highly domain-dependent. In other words, even if a model performs well overall, it does not necessarily mean that

it is consistently optimal across all body-system categories; instead, some models may be more stable and reliable in specific categories.

This observation supports the necessity of assigning different LLMs to different questions based on their domain-specific expertise. In this way, MedCUR can avoid the mismatch problem caused by random agent selection and increase the likelihood that each medical question is handled by models with stronger proficiency in the relevant domains. We further observe that models with smaller parameter sizes generally struggle with complex medical reasoning tasks. Considering the trade-off between performance and computational cost, we adopt DeepSeek-R1 as the meta-agent in our framework, as it provides competitive cognitive proficiency while maintaining lower inference costs.

5.2 MAS Results

Table 1 shows the results of different methods on four benchmarks. We can observe that *MedCUR* significantly improves performance on all benchmarks, demonstrating the effectiveness of our framework in complex medical reasoning scenarios. This improvement is not limited to a single dataset, but consistently appears across medical reasoning tasks with different formats and difficulty levels, indicating that MedCUR has good stability and generalizability.

In contrast, existing multi-agent methods tend to perform well on easier benchmarks but often exhibit degraded performance on more challenging ones, sometimes performing even worse than single-agent baselines. We hypothesize that this degradation derives from the uncontrolled propagation of uncertain information during multi-agent discussions, where agents may uncritically adopt earlier incorrect reasoning, leading to error amplification. In particular, for more challenging datasets such as RareDisease and MedXpertQA, the questions often require long-chain reasoning and cross-domain medical knowledge. Once an early agent provides an uncertain or incorrect judgment, subsequent agents may continue reasoning around this misleading conclusion, ultimately reducing the quality of the final answer. For easy tasks, a single agent is often sufficiently confident to solve the problem independently, making redundant or inaccurate information pose little risk. By explicitly modeling cognitive uncertainty and introducing expert-guided clarification, *MedCUR* effectively mitigates this issue, prevents error propagation, and maintains strong performance even on difficult cases.

6 Analysis

In this section, we analyze the impact of each component in our framework, intermediate confidence, and LLM invocation behavior. In addition, we provide the case study in Appendix C.

6.1 Ablation Study

6.1.1 Ablation Settings

We perform an ablation study to investigate the effectiveness of each design in our framework. First, we randomly select agents instead of using cognitive proficiency to guide agent selection. We consider the following two settings:

- *Random agents I*: All LLMs mentioned in the experimental setting are used as candidates.
- *Random agents II*: Only the top six best-performing LLMs are used as candidates to mitigate the negative impact of weaker or smaller LLMs on random selection.

Additionally, we remove *uncertainty reporting* (i.e., confidence-based uncertainty signals) during doctor agents' discussions and disable the involvement of *external experts* for resolving uncertain information, respectively.

6.1.2 Ablation Analysis

Table 2 presents the ablation results. The results show that each component of our framework contributes to the overall performance. In particular, randomly selecting agents leads to a performance drop of over 10%, demonstrating that selecting appropriate experts based on cognitive proficiency is crucial for solving complex medical reasoning problems. Even when weaker LLMs are included as candidates, our framework still maintains effective collaboration. Notably, this setting still outperforms baselines, demonstrating the effectiveness of our framework.

Removing uncertainty reporting also results in performance decrease, highlighting the importance of explicitly generating and utilizing cognitive uncertainty to guide the discussion. Furthermore, disabling external expert consultation leads to further performance degradation. This suggests that while uncertainty reporting and meta-agent control help regulate discussion direction, they are insufficient to alleviate the negative impact of uncertain information. Introducing external experts effectively

Methods	RareDisease	MedXpertQA	MedMCQA	MedQA	Average
Single Agent					
DeepSeek-R1	36.42%	44.26%	79.90%	88.67%	62.31%
Multi-Agent					
MedAgents	32.78% ↓3.64	36.07% ↓8.19	79.67% ↓0.23	92.33% ↑3.66	60.21% ↓2.10
ReConcile	30.79% ↓5.63	35.25% ↓9.01	80.86% ↑0.96	89.33% ↑0.66	59.06% ↓3.25
MAC	33.11% ↓3.31	34.43% ↓9.83	80.14% ↑0.24	91.33% ↑2.66	59.75% ↓2.56
Healthcare Agent	30.46% ↓5.96	31.97% ↓12.29	77.22% ↓2.68	89.67% ↑1.00	57.33% ↓4.98
MedCUR (Ours)	**37.74%** ↑1.32	**52.05%** ↑7.79	**82.05%** ↑2.15	**93.33%** ↑4.66	**66.29%** ↑3.98

Table 1: Results of different methods on medical reasoning benchmarks.

Method	MedXpertQA	Decrease
Random agents I	37.29%	↓ 14.76
Random agents II	38.52%	↓ 13.53
- Uncertainty reporting	44.26%	↓ 7.79
- External expert	47.95%	↓ 4.10
MedCUR	52.05%	-

Table 2: The ablation study results.

Model	Single ↑	*MedCUR* ↑	Cost ↓
Claude Sonnet 4.5	39.34%	50.81% ↑11.47	$0.21
DeepSeek-R1	44.26%	52.05% ↑7.79	$0.05
Gemini 2.5 Pro	49.10%	54.92% ↑5.82	$0.12

Table 3: Impact of using different LLMs for the Meta-Agent.

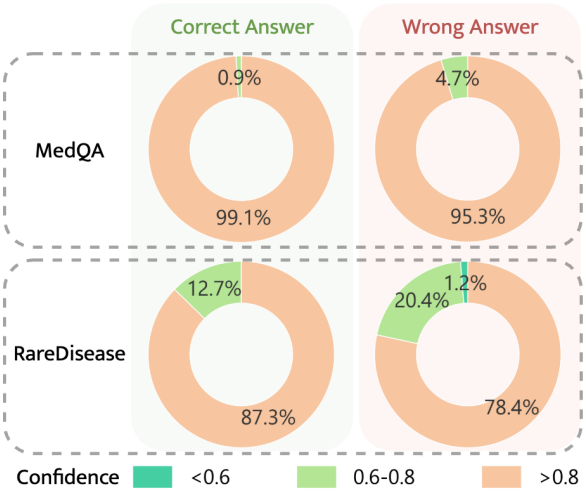


Figure 4: Confidence analysis across datasets of different difficulty levels.

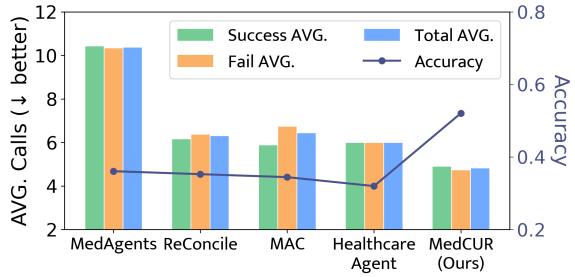
mitigates this issue by providing targeted clarification, thereby improving overall reasoning quality.

6.2 Meta Agent Impaction

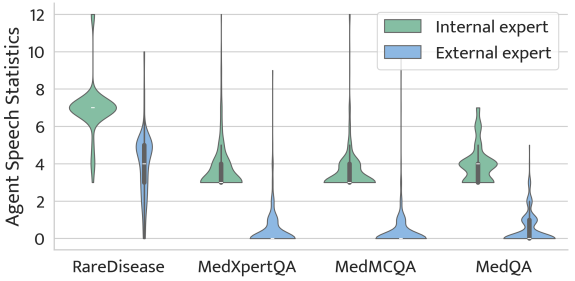
To further analyze how the meta-agent influences medical reasoning performance, we evaluate three different LLMs as meta-agent: Claude Sonnet 4.5, DeepSeek-R1, and Gemini 2.5 pro. Table 3 reports the results of *MedCUR* on MedXpertQA as well as the average cost for each question. The results show that higher inherent cognitive capability of the meta-agent generally leads to better overall reasoning performance. Although Gemini 2.5 pro achieves the best results, its cost is approximately 2.4 times higher than that of DeepSeek-R1. While Claude Sonnet 4.5 exhibits relatively weaker medical task performance, it still shows significant performance improvements, indicating that our framework is effective across different meta-agent choices. However, the performance gain brought by a stronger meta-agent should be considered together with its inference cost. In this sense, DeepSeek-R1 provides a favorable trade-off between effectiveness and cost, supporting its use as the meta-agent in our framework.

6.3 Confidence Analysis

To assess whether agent-reported confidence reliably reflects cognitive uncertainty, we analyze confidence distributions across datasets with different difficulty levels in Figure 4. Results show that on the more challenging RareDisease dataset, agents produce a substantially higher proportion of low-confidence predictions than on MedQA, indicating that confidence decreases as problem difficulty increases. This trend aligns with human experts' decision-making behavior in complex medical cases. Moreover, within the same dataset, incorrect predictions consistently exhibit lower average confidence than correct ones, suggesting that although self-reported confidence cannot precisely



(a) Average number of LLM calls across different methods.



(b) Average number of LLM calls of our method across different benchmarks.

Figure 5: The statistic of LLMs call.

measure correctness, it can serve as an effective signal for reasoning reliability and help identify high-risk reasoning that requires further verification, especially in medical scenarios that are highly sensitive to error propagation.

We also observe a strong bias toward high-confidence outputs, with most predictions exceeding 0.8. Motivated by this observation, MedCUR adopts a relatively high confidence threshold to trigger uncertainty-driven control, treating low-confidence responses as high-uncertainty signals to enable more robust handling of high-risk medical reasoning tasks.

6.4 LLMs Calls Analysis

Cost Efficiency Comparison. We illustrate the distribution of LLM calls on the MedXpertQA benchmark in Figure 5. Figure 5a shows the average number of LLM calls for different methods. We observe that our method requires the fewest LLM calls per question, indicating higher efficiency and lower cost than other methods. In contrast, MedAgents requires the largest number of LLM calls, and Healthcare Agent uses a constant number of LLM calls for all questions. These methods lack the ability to adapt their structures to question complexity. Other multi-agent methods can terminate discussions once consensus is reached, resulting in fewer LLM calls than MedAgents. However, this operation only reduces computational cost without improving accuracy. In contrast, our method adaptively determines the agent structure based on cognitive proficiency and confidence-based uncertainty signals, achieving both lower computational cost and better reasoning performance.

Adaptive Resource Allocation. We further analyze the LLM call distribution of our method across different benchmarks in Figure 5b. The results show that *MedCUR* adaptively adjusts the number of LLM calls according to question difficulty. For easier datasets, the framework typically requires fewer internal discussion rounds and external expert consultation, since the assigned doctor agents can reach high-confidence conclusions quickly. For more challenging benchmarks, the number of LLM calls increase noticeably. This adaptive behavior allows the framework to allocate computational resources where they are most needed.

Moreover, the average number of doctor agents allocated is 2.58, 2.07, 1.39, and 1.76 for RareDisease, MedXpertQA, MedMCQA, and MedQA, respectively. This indicates that *MedCUR* can adjust the structure size based on question difficulty without sacrificing performance. This further confirms that *MedCUR* achieves cost efficiency through difficulty-aware resource allocation rather than uniform computation.

7 Conclusion

In this paper, we propose *MedCUR*, a multi-agent framework that incorporates cognitive uncertainty to guide the design of agent structures for complex medical reasoning tasks. Specifically, we quantify the cognitive proficiency of individual agents by jointly modeling task accuracy and uncertainty across different medical categories. Given a specific question, the meta-agent dynamically orchestrates discussions by selecting appropriate doctor agents based on cognitive proficiency. During collaboration, all doctor agents explicitly provide cognitive uncertainty and doubts. The meta-agent leverages these as control signals to regulate the involvement of third-party experts and to manage the discussion process. Finally, the meta-agent integrates the agents' viewpoints based on their cognitive proficiency to produce the final answer. Extensive experiments on medical reasoning benchmarks demonstrate that *MedCUR* consistently improves both performance and efficiency compared with existing multi-agent baselines.

Limitations

Although MedCUR performs well on medical reasoning tasks, its extension to general-purpose tasks remains limited. The current method relies on predefined task categories to estimate agents' cognitive proficiency, making it more suitable for settings with clear categories and available evaluation data. In open-domain tasks, however, categories are often more complex and dynamic. When a question involves unseen categories, cross-category combinations, or cases that are difficult to map to existing categories, uncertainty estimation may become less reliable, which can further affect the meta-agent's selection of the optimal model or expert agent. Future work could explore dynamically extensible task representations and online calibration mechanisms to improve adaptability to general tasks and unseen categories.

References

Anthropic. 2025. Introducing Claude Sonnet 4.5. Accessed: 2026-04-23.

Justin Chen, Swarnadeep Saha, and Mohit Bansal. 2024. ReConcile: Round-table conference improves reasoning via consensus among diverse LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7066–7085, Bangkok, Thailand. Association for Computational Linguistics.

Kai Chen, Xinfeng Li, Tianpei Yang, Hewei Wang, Wei Dong, and Yang Gao. 2025a. Mdteamgpt: A self-evolving llm-based multi-agent framework for multi-disciplinary team medical consultation. *arXiv preprint arXiv:2503.13856*.

Xi Chen, Huahui Yi, Mingke You, WeiZhi Liu, Li Wang, Hairui Li, Xue Zhang, Yingman Guo, Lei Fan, Gang Chen, and 1 others. 2025b. Enhancing diagnostic capability with multi-agents conversational large language models. *NPJ digital medicine*, 8(1):159.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Zhihao Fan, Lai Wei, Jialong Tang, Wei Chen, Wang Siyuan, Zhongyu Wei, and Fei Huang. 2025. AI hospital: Benchmarking large language models in a multi-agent medical interaction simulator. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10183–10213, Abu Dhabi, UAE. Association for Computational Linguistics.

Yingjie Fei, Zhuoran Yang, Yudong Chen, Zhaoran Wang, and Qiaomin Xie. 2020. Risk-sensitive reinforcement learning: near-optimal risk-sample tradeoff in regret. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Stephan Hartmann, Carlo Martini, and Jan Sprenger. 2009. Consensual decision-making among epistemic peers. *Episteme*, 6(2):110–129.

Ronald A. Howard and James E. Matheson. 1972. Risk-sensitive markov decision processes. *Manage. Sci.*, 18(7):356–369.

Xiaobin Hu, Yunhang Qian, Jiaquan Yu, Jingjing Liu, Peng Tang, Xiaozhong Ji, Chengming Xu, Jiawei Liu, Xiaoxiao Yan, Xinlei Yu, and 1 others. 2025. The landscape of medical agents: A survey. *Authorea Preprints*.

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.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577.

Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik S Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae W Park. 2024. Mdagents: An adaptive collaboration of llms for medical decision-making. *Advances in Neural Information Processing Systems*, 37:79410–79452.

Rumeng Li, Xun Wang, Dan Berlowitz, Jesse Mez, Honghuang Lin, and Hong Yu. 2025. Care-ad: a multi-agent large language model framework for alzheimer's disease prediction using longitudinal clinical notes. *npj Digital Medicine*, 8(1):541.

Meng Lu, Brandon Ho, Dennis Ren, and Xuan Wang. 2024. TriageAgent: Towards better multi-agents collaborations for large language model-based clinical triage. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5747–5764, Miami, Florida, USA. Association for Computational Linguistics.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

OpenAI. 2025. Introducing GPT-5. Accessed: 2026-04-23.

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.

Qi Peng, Jialin Cui, Jiayuan Xie, Yi Cai, and Qing Li. 2025. Tree-of-reasoning: Towards complex medical diagnosis via multi-agent reasoning with evidence tree. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, page 1744–1753, New York, NY, USA. Association for Computing Machinery.

Qwen Team. 2025. Qwen3-Max: Just scale it. Accessed: 2026-04-23.

Zhiyao Ren, Yibing Zhan, Baosheng Yu, Liang Ding, Pingbo Xu, and Dacheng Tao. 2025. Healthcare agent: eliciting the power of large language models for medical consultation. *npj Artificial Intelligence*, 1.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, and 1 others. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. 2024. MedAgents: Large language models as collaborators for zero-shot medical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 599–621, Bangkok, Thailand. Association for Computational Linguistics.

Gemini Team. 2025a. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Qwen Team. 2025b. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Wenmiao Wang. 2025. Llm-based multi-agent system for neuro-ophthalmic diagnosis and personalized treatment planning. *Frontiers in Neuroscience*, 19:1688509.

Wenxuan Wang, Zizhan Ma, Zheng Wang, Chenghan Wu, Jiaming Ji, Wenting Chen, Xiang Li, and Yixuan Yuan. 2025a. A survey of LLM-based agents in medicine: How far are we from baymax? In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10345–10359, Vienna, Austria. Association for Computational Linguistics.

Wenxuan Wang, Zizhan Ma, Zheng Wang, Chenghan Wu, Jiaming Ji, Wenting Chen, Xiang Li, and Yixuan Yuan. 2025b. A survey of LLM-based agents in medicine: How far are we from baymax? In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10345–10359, Vienna, Austria. Association for Computational Linguistics.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *Preprint*, arXiv:2306.13063.

Yi Yu, Lingli Li, and Yaqin Li. 2025. Augmenting large language models and retrieval-augmented generation with an evidence-based medicine-enabled agent system. *medRxiv*, pages 2025–10.

Yuxin Zuo, Shang Qu, Yifei Li, Zhang-Ren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. 2025. MedxpertQA: Benchmarking expert-level medical reasoning and understanding. In *Forty-second International Conference on Machine Learning*.

A Prompt Templates

This appendix presents all prompt templates used in the MedCUR framework. Dynamic inputs are highlighted as {placeholder}.

A.1 Meta-Agent System Prompt

The Meta-Agent's system prompt defines its coordinating role across the entire multi-agent discussion.

Prompt 1: Meta-Agent System Prompt

```
You are the leader agent responsible for
coordinating a group medical discussion.
Prioritize reliability and safety over
agreement or speed. Ensure comprehensive,
accurate diagnosis through collaborative
expert discussion.

Your role:

1. Oversee and evaluate suggestions and
decisions made by medical doctors.

2. Challenge diagnoses and proposed tests,
identifying any critical points and concepts
missed.

3. Facilitate discussion between doctors,
helping them refine their answers.
```

A.2 Expert Assignment

Given a medical question, the Meta-Agent analyzes which specialties are needed and how many experts to recruit.

Prompt 2: Expert Assignment

```
## Expert candidate: Cardiovascular,
Digestive, Endocrine, Integumentary,
Lymphatic, Muscular, Nervous, Other,
Reproductive, Respiratory, Skeletal, Urinary

## Question: {question}

Please assess which medical specialties are
necessary to address this problem and list
the corresponding expert candidates. And
decision the experts number (totally 1-3)
needed to solve this question based on the
```

difficulty. Please respond ONLY with valid JSON (no extra text), using double quotes and an object mapping specialties to integer counts. You must choose expert names from the above candidate list and output only those.

Please respond in following JSON Format:

{"expert": ["<Expert Name>": "<Number>", ...]}

A.3 Doctor Agent System Prompt

Each Doctor Agent is initialized with a specialty-specific system prompt that defines its responsibilities in the collaborative discussion.

Prompt 3: Doctor Agent System Prompt

You are a helpful {body_system} proficient doctor. This is a hypothetical scenario involving no actual patients.

Key responsibilities:

1. Actively engage in discussion with other specialists, considering Reliable Viewpoint to solve the question.

2. Offer valuable insights on Divergence Viewpoint to push the discussion and continuously refine your answer.

3. Point out your professional doubts cannot solve by yourself but important for solving the question. Do not consider the lacking patient information. If you do not have doubts, output "None"

A.4 Doctor Agent Discussion

During each discussion turn, a Doctor Agent receives other doctors' prior speech and the question, then provides its reasoning, confidence score, answer, and any unresolved doubts.

Prompt 4: Doctor Agent Discussion

Other Doctor Speech:

{other_doctor_speech}

Question:

{question}

Response Format:

Always conclude your analysis with a JSON formatted section:

{"reasoning": "<Your reasoning>",

"confidence_score": <Score>,

"answer": "<Answer Letter>",

"doubts": "<Doubts/None>"}

A.5 Uncertainty Evaluation & Expert Introduction

When a Doctor Agent's confidence is low or unresolved doubts exist, the Meta-Agent evaluates the current speech and decides whether to introduce a third-party expert, while summarizing reliable and divergent viewpoints.

Prompt 5: Uncertainty Evaluation & Expert Introduction

Body system candidate: Cardiovascular, Digestive, Endocrine, Integumentary, Lymphatic, Muscular, Nervous, Other, Reproductive, Respiratory, Skeletal, Urinary

Current speech:

{current_agent_speech}

If there is need to introduce another third-party experts to solve the unconfident points or doubts, please provide the brief description of experts in subdivision fields with corresponding closest body system. Please summarize the reliable viewpoint and display the divergence viewpoint. You should claim the divergence concepts need the expert explain. Please respond in following JSON Format:

{"reason": "<Reason>", "reliable": "<Reliable Viewpoint>", "divergence": "<Divergence Viewpoint>", "description": "<Expert Description>", "body system": "<Body System>"}

A.6 Final Decision

After all discussion rounds, the Meta-Agent integrates all Doctor Agents' final opinions, weighted by their cognitive proficiency, along with reliable and divergent viewpoints, to produce the final answer.

Prompt 6: Final Decision Integration

Doctor final speech:

{doctor_final_speech}

Reliable viewpoint:

{reliable_info}

Divergence viewpoint:

{divergence_info}

Doctor cognitive proficiency (0-1):

{doctor_proficiency_json}

Question:

{question}

Please integrate all doctors' final opinions, weighting each doctor's contribution by their cognitive proficiency. Take into account the coherent viewpoints across doctors, carefully handle divergent opinions, and generate the final answer accordingly. Please respond in following JSON Format:

{"reason": "<Reason>", "answer": "<Answer Letter>"}

B Detailed Cognitive Proficiency Data

Figure 3 in the main text presents the overall cognitive proficiency scores ($p = a * s$). This section provides the two constituent metrics used to compute the proficiency score: Accuracy and Self-Consistency, enabling a more fine-grained examination of each model's performance across body system categories.

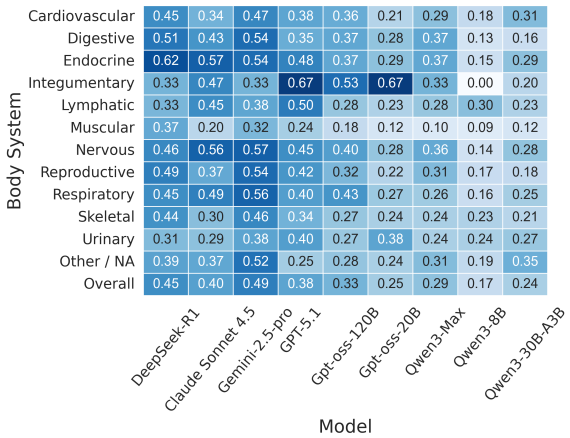


Figure 6: Accuracy of each model across 11 body system categories. Each value is computed as the proportion of correct answers over 10 repeated inference runs.

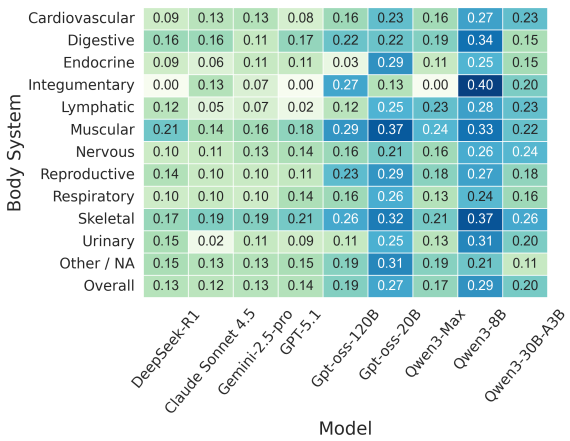


Figure 7: Self-Consistency of each model across 11 body system categories. Higher values indicate more stable outputs and lower cognitive uncertainty.

C Case Study

This appendix presents a representative case study, illustrating how MedCUR corrects an initially wrong prediction through uncertainty-guided expert intervention.

Case Description

A 68-year-old woman presents with a 6-month history of intermittent headaches and dizziness. She describes the dizziness as a "spinning room" sensation. Previous evaluations by her primary care provider, neurologist, and otolaryngologist were unremarkable. The Dix-Hallpike test is negative, and the neurological examination is normal. Structural examination shows that the right greater wing of the sphenoid is superior to the left, with physiologic motion. Additional cervical findings are also observed.
Answer choices: (A) Benign paroxysmal positional vertigo; (B) OA ESRRL cervical somatic dysfunction; (C) Sphenobasilar synchondrosis dysfunction; (D) Vestibular neuritis with cervical involvement; (E) Right torsion cranial strain somatic dysfunction; (F) Left lateral strain cranial somatic dysfunction; (G) C6–C7 NSRRR cervical somatic dysfunction; (H) Temporal bone dysfunction with superior semicircular canal dehiscence; (I) Cervicogenic vertigo with upper cervical dysfunction; (J) Atlas rotation somatic dysfunction with vascular compromise.
Gold answer: E. Right torsion cranial strain somatic dysfunction.

Step 1: Agent Composing

The meta-agent first analyzes the question and identifies that the case requires both nervous-system and skeletal-system expertise.
Assigned experts:
1. Nervous-system expert
2. Skeletal-system expert

Step 2: Discussion Routine – Initial Doctor Response

Doctor 0 first rules out benign positional vertigo because the Dix-Hallpike test is negative. The doctor then focuses on the cervical findings and makes an initial prediction.
Initial answer: **I. Cervicogenic vertigo with upper cervical dysfunction**
Confidence score: **0.65**
Reported doubt:
The osteopathic nomenclature for the C6–C7 finding is unclear. The doctor is uncertain whether the described motion restriction corresponds to a specific dysfunction pattern, and

whether this finding or the upper cervical finding should be considered primary.

Step 3: Uncertainty Evaluation & Expert Introduction

The meta-agent detects that the initial answer has low confidence and contains unresolved professional doubt. Instead of directly propagating this uncertain diagnosis, the meta-agent separates reliable information from uncertain points.
Reliable information:
1. Benign positional vertigo is less likely because the Dix-Hallpike test is negative.
2. Central causes are less likely because the neurological examination is normal.
Uncertain point:
The primary source of the symptoms remains unclear, especially the relation between the cervical findings and the key structural finding.
Action:
The meta-agent introduces an external osteopathic neuromusculoskeletal expert to clarify the uncertain point.

Step 4: Third-Party Expert Feedback

The third-party expert identifies the key evidence: the right greater wing of the sphenoid is superior to the left. This finding directly indicates a right torsion cranial strain.
Expert conclusion:
The most likely primary diagnosis is **E. Right torsion cranial strain somatic dysfunction**.
The expert further explains that this diagnosis better accounts for both dizziness and headache than the initial cervical-only diagnosis.

Step 5: Discussion Routine – Updated Doctor Response

After receiving the expert clarification, Doctor 1 agrees with the corrected diagnosis.
Updated answer: **E. Right torsion cranial strain somatic dysfunction**
Confidence score: **0.95**
Reasoning:
The finding that the right greater wing of the sphenoid is superior to the left is the defining clue for a right torsion cranial strain. This diagnosis is more specific than the initial cervical diagnosis and is directly supported by the examination.

Step 6: Fine-grained Uncertainty Handling

Doctor 1 raises a remaining uncertainty about whether the cranial finding or the cervical finding should be regarded as the primary source. The meta-agent identifies this as a fine-grained uncertainty and introduces another expert with expertise in cranial and cervical biomechanics. The second expert explains that the cranial finding is more likely to be the primary abnormality, while the cervical findings are better interpreted as secondary compensations.
The expert again supports answer **E**.

Step 7: Meta-decision Integration

After the expert intervention, Doctor 0 revises the initial wrong answer.
Revised answer: **E. Right torsion cranial strain somatic dysfunction**
Confidence score: **0.85**
Final decision:
The meta-agent aggregates the final opinions and selects **E. Right torsion cranial strain somatic dysfunction**.

Case Analysis. This case demonstrates the key advantage of MedCUR. Instead of simply aggregating agent responses, MedCUR detects low-confidence reasoning and unresolved uncertainty, then introduces a suitable external expert to clarify the critical point. As a result, the initially incorrect answer is corrected, and the agents converge on the correct diagnosis. This shows that MedCUR improves reasoning reliability through uncertainty detection, expert intervention, and answer correction.

D LLM Versions

This appendix lists all Large Language Models (LLMs) used in our experiments with their confirmed versions.

Table 4: Large Language Models (LLMs) and Their Versions

Model	Version
DeepSeek-R1	deepseek-r1-0528
Claude Sonnet 4.5	claude-sonnet-4-5-20250929
Gemini 2.5 Pro	gemini-2.5-pro
GPT-5.1	gpt-5.1
GPT-oss-120B	gpt-oss-120b
GPT-oss-20B	gpt-oss-20b
Qwen3-Max	qwen3-max
Qwen3-30B-A3B	qwen3-30b-a3b-thinking-2507
Qwen3-8B	qwen3-8b