2025 Annual Computer Security Applications Conference Workshops (ACSAC Workshops) | 979-8-3315-4536-9/25/$31.00 ©2025 IEEE | DOI: 10.1109/ACSACW69556.2025.00033

AI on Trial: LLM-as-a-Judge for Private and Reliable Clinical Decision-Making

Tanveer Khan
Department of Computing Sciences
Tampere University
Tampere, Finland
tanveer.khan@tuni.fi

Maaz Rafiq
Data Science
Appedology Pvt. Ltd.
Karachi, Pakistan
maaz.rafique.75@gmail.com

Antonis Michalas
Department of Computing Sciences
Tampere University
Tampere, Finland
antonios.michalas@tuni.fi

***Abstract*—Artificial Intelligence (AI) is changing the way healthcare works by helping doctors make better decisions. Large Language Models (LLMs) play a key role in this shift, not only by answering medical questions, but also by actively collecting critical information through follow-up questions. However, effectively applying LLMs to complex Clinical Decision-Making (CDM) tasks remains a challenge. In this paper, we propose `AI on Trial`– an enhanced framework built on top of MDAgents. It assigns LLMs to work alone or in teams based on the complexity of medical queries. To improve the reliability of responses, we extend the original MDAgent system by incorporating LLM-as-a-Judge, a clinically informed evaluator that scores responses based on quality, completeness, safety and clarity, while also providing the reasoning behind its assessment. This judge also explains its reasoning using Chain-of-Thoughts (CoT's). We also add a second judge that acts as a team of expert reviewers. It uses four LLM agents to carefully check the response and give a final score and confidence level. This two-step review process makes the system more accurate and trustworthy. Our results show that integrating the multi-agent evaluation with MDAgents improves performance, making it a strong tool for AI-based clinical support.**

***Index Terms*—Clinical-Decision Making, Large Language Models, LLM-as-a-Judge**

I. Introduction

Clinical Decision-Making (CDM) [1] is about using the best information to make the right decision for patient care. It combines medical knowledge, doctor experience, and patient needs. This approach helps doctors pick treatments that work well and fit each patient.

However, practicing CDM is becoming more difficult. Medical research is growing rapidly. There are many studies, reports, and articles to read [2]. Searching through all this information, checking its quality, and applying it correctly takes alot of time and skills. Tools such as the PICO (Patient, Intervention, Comparison, Outcome) framework help structure questions and filter studies [3], but the process is still slow and complex. Given the potential consequences of clinical error, this process must be both thorough and accurate.

Large Language Models (LLMs) offer a promising way to support CDM [4]. These models can understand and generate human-like text, helping with tasks like summarizing studies, answering clinical questions, and finding important information quickly [5]. Models like GPT-4 [6] have shown strong performance on medical exams and other healthcare tasks, showing that they can assist doctors make better decisions.

Despite their promise, general-purpose LLMs face key challenges in clinical use. They often lack the deep medical knowledge required to handle complex or highly specialized clinical tasks [7]. In addition, these models can "hallucinate" answers that sound confident but are factually wrong or unsafe [8]. This is especially risky in healthcare settings. To address this, researchers have turned to multi-agent LLM frameworks [9]. These systems mimic how real physicians work together when making clinical decisions. Some examples are RareAgents [10], MDAgents [11], AMSC [12], AMIE [13], Dr-House [14], MEDAIDE [15], MEDIQ [16], Exploring the Inquiry [17]. For example, *Exploring the Inquiry* [17] simulates and AI hospital workflow that covers the full process from patient consultation to diagnosis, while MEDIQ [16] supports multi-round doctor-patient interactions to iteratively collect missing information and refine diagnosis.

Inspired by how real clinical teams collaborate, MDAgents [11] introduces an adaptive decision-making framework that uses LLMs to emulate hierarchical diagnostic workflows, from the individual doctor to the collaborative clinical teams. MDAgents work in four steps *(i)* assessing the complexity of a case *(ii)* Recruiting appropriate agents based on that medical complexity; *(iii)* Performing collaborative analysis and synthesis, and *(iv)* producing a final response.

However, a key limitation of MDAgents lies in the lack of a second-level validation step. Once an answer is made, either by a solo doctor or by a clinical team, it is treated as the final decision. There is no mechanism to review, critique, or improve the response. This lack of feedback is problematic, especially in medicine, where LLM-generated responses must be carefully evaluated for accuracy, completeness, and safety before they can be trusted in clinical settings.

Another big challenge in using LLMs for CDM is keeping patient information private. Medical reports, case reports, and trial data contain sensitive personal information. Sharing these data with large models can expose private information. This raises serious concerns about data privacy. There are privacy-preserving techniques to protect privacy, such as using cryptographic methods [18]–[25] and data anonymization [26]. This helps keep patient information safe. Although this is an

979-8-3315-4536-9/25/$31.00 ©2025 IEEE
DOI 10.1109/ACSACW69556.2025.00033

Authorized licensed use limited to: Tampere University. Downloaded on May 14,2026 at 11:42:55 UTC from IEEE Xplore. Restrictions apply.

important research direction, our current work does not focus on privacy issues.

In our proposed approach, we build upon the MDAgents framework by adopting a multi-agent LLM architecture for CDM. The process begins with the system evaluating the complexity of the medical query. Based on clinical criteria, the case is categorized as basic, intermediate, and advanced. Depending on this classification, the framework activates a single Primary Care Physician (PCP) for basic queries, or Multi-disciplinary Team (MDT) or Integrated Care Team (ICT) for intermediate and advanced medical queries.

After collaborative reasoning and synthesis, the final response generated by the agents is submitted to a specialized evaluation component called LLM-as-a-Judge. This judging model takes as input both the original medical query and the generated answer. It then assigns an evaluation score based on several key attributes: the inclusion of essential medical information, accuracy of the answer, completeness, safety, and clarity. This evaluation step ensures that the final response is not only clinically sound but also aligned with best practices in CDM.

a) Contribution: The main contributions of this paper are:

- We extend the MDAgent framework by adding an LLM-as-a-Judge, that evaluates the quality, completeness, safety, and clarity of the answers generated by the medical LLM agents. The judge assigns a score and explains its evaluation using clinical reasoning.
- Also, we introduce a multi-agent LLM evaluator that simulates expert peer review by critically analyzing and validating the generated responses. This ensures a more transparent and robust decision-making process.
- Our approach addresses a key limitation of MDAgent by adding a second-level validation step. If the judge finds the answer weak, it is flagged as unreliable.
- By combining automatic evaluation with human feedback when needed, our proposed approach makes multi-agent medical LLMs more reliable and easier to understand.

II. Related Work

Digital healthcare is changing the way patients receive care through technologies such as wearable devices and AI tools. These allow physicians to monitor patients remotely, create personalized treatments, and predict health issues [27]. ML tracks how diseases progress, while language models understand medical notes to support better decisions. Together, these tools make healthcare more efficient, accurate, and accessible [28].

LLMs are very good in generating human-like text and understanding complex language. In healthcare, LLMs are used for tasks such as answering medical exam questions [29], conducting biomedical research [30], and assisting diagnosis [31]. They have also been tested with the generation of medical report summaries, clinical decision support, and even virtual health assistants. By saving physicians time, LLMs contribute to more efficient and patient-friendly care [32].

Interactive healthcare conversational systems are important to provide patients with accurate and relevant information. Datasets and benchmarks like MedQA [33] and MedMCQA [34] have helped improve question-answering systems in medicine. These systems use large datasets to fine-tune LLMs, making their responses more relevant and patient-centered.

Recent research shows that using multi-agent LLMs, can outperform single models [35]–[37]. In these systems, each agent has a specific role, and they break down complex tasks into smaller parts [35], [37]. Another approach involves agents debating each other's answer to arrive at better solutions, often improving accuracy and reasoning [38]–[40]. Other forms of collaboration include group voting [41], expert discussions [42], or even negotiation [43]. Although these approaches are promising, they usually work with fixed settings that may not work well across all tasks [44]. In some setups, this can reduce performance or use up extra resources on easy tasks.

To address this, MDAgents [11] propose a dynamic approach that adjusts the collaboration strategy according to the complexity of the task. It classifies clinical tasks as low, moderate, or high complexity using CDM techniques. This is useful in CDM, where teamwork, flexibility, and reasoning are essential. However, while MDAgents use different collaboration strategies, they lack an explicit mechanism for evaluating and assessing the quality of the generated response, which is critical in healthcare.

To fill this gap, we introduce an LLM-as-a-Judge component – a clinically informed evaluation system that reviews agent outputs for accuracy and quality. Different types of LLM-as-a-judge exist [45], some compare two answers and choose the better one; others score a single answer directly; and some use reference answers to guide the grading. This judging approach has several benefits, such as allowing a faster and larger-scale evaluation and providing an explanation along with scores, making its assessment transparent. RAGAS [46] – the LLM-as-a-Judge framework, has gained widespread adoption for the evaluation of Question Answering (QA) systems. It receives the generated answer, the correct answer, and sometimes the question. Then, using clear scoring rules and carefully designed prompts, the judge scores the response.

III. AI on Trial

A. Agent Roles

In our approach, we use different types of agents. Each agent has a clear role in the system:

- **Moderator:** This agent acts as a general physician. It looks at the medical query and decides how complex it is. Based on this, it chooses whether the case should be handled by a doctor (PCP), a specialist team (MDT), or a more coordinated care team (ICT).
- **Recruiter:** Once the complexity is known, the recruiter puts together the right team. For simple cases, it assigns a single PCP, For more difficult cases, it brings in either an MDT or ICT, depending on how complex the query is.

Authorized licensed use limited to: Tampere University. Downloaded on May 14,2026 at 11:42:55 UTC from IEEE Xplore. Restrictions apply.

- **PCP:** This agent works alone to answer simple and straightforward medical questions. It is a fast and efficient way to handle low-complexity cases.
- **MDT:** This is a group of agents where each one acts as a medical specialist. They work together to solve intermediate cases. First, the team talks about the question and shares their opinions. If they do not agree on an answer, a moderator steps in, reviews their discussion, and gives helpful feedback. Each agent then updates their response based on that feedback. This cycle of discussion and feedback continues until the team agrees or finishes the discussion. In the end, the moderator decides the final answer.
- **ICT:** This team is used for more advanced cases. It acts like a group of doctors and specialists working closely together, which is helpful when a patient has many health problems or needs complex care. Each ICT team writes a report and the final answer is made using all these reports together.
- **LLM-as-a-Judge:** In our proposed approach, we deployed LLM-as-a-Judge as a critical evaluation component to ensure the reliability, correctness, and safety of the responses generated by the MDAgent system. It is implemented in two forms
  - **Single LLM with Chain-of-Though Evaluation:** A single LLM acts as a judge and evaluates the final response using a CoT's reasoning approach. It consider multiple aspects, such as *(i)* **Key Information Identifier:** Does the response cover all important elements from the query?; **Answer Assessment:** Is the response relevant?; **Completeness:** Does it fully address the question?; **Safety:** Does it avoid unsafe recommendations? and **Clarity:** Is the answer clearly communicated?.
  - **Multiagent LLM Evaluation:** In this approach, we simulate a team of four specialist agents who collaboratively review the response generated by the MDAgents. Each agent has a specific role: *(i)* **Discriminator:** identifies flaws, unsafe recommendations, and logical gaps *(ii)* **Fact Checker:** Verifies the response against medical knowledge and guidelines; *(iii)* **Justifier:** provides reasoning to support and justify the given response, ensuring that it is well-grounded; and *(iv)* **Moderator:** aggregates the reviews, highlights points of agreement and disagreement, and assigns an evaluation score (from 1 - 5) along with a confidence level (high, medium, and low).

B. Medical complexity classification

Like in MDAgents, our proposed approach first checks how difficult the medical question is. It places each query into one of three levels:

- **Low:** Simple and clear medical questions that can be answered by a single PCP agent.
- **Medium:** These questions have multiple parts or need different types of expertise, so they are handled by an MDT.
- **High:** These are difficult queries that require a fully coordinated ICT to manage.

C. Workflow

CDM requires not only accurate responses, but also reliable validation and reasoning. Although MDAgents have demonstrated strong performance in medical QA tasks, their approach lacks a mechanism to verify the quality of generated responses. Once a medical agent, be it PCP, MDT, or ICT, produces a response, there is no further review. This can be problematic in domains like healthcare, where errors in judgment can lead to incorrect diagnosis.

MDAgent

Medical Task → Agent Assignment → Group/Solo Response

To address this limitation, we propose `AI on Trial`– an enhanced framework by integrating LLM-as-a-Judge into the MDAgent framework. This component acts as a second layer of evaluation that systematically reviews the responses generated by the MDAgents. The judge is designed with medical reasoning in mind and uses a set of clearly defined criteria to evaluate each response. In our approach, we consider two types of judges: one that uses step-by-step thinking (CoT's) and another that uses a team of multiple agents working together (see Figure 1).

1) Step-by-step Evaluation with CoT's Reasoning:: For each query, the judging process begins with identifying key medical facts. The judge then assesses whether the response generated by MDAgent accurately addresses the core clinical query. It evaluates the response based on the following criteria: key information identifier, answer assessment, completeness, safety, and clarity. All these terms are discussed above.

After applying these checks, the judge assigns a score between 1 and 5: where 1 indicates an unacceptable or unsafe response, while 5 indicates a high quality, medically sound and professional response. If the score falls below a set threshold (3 or low), the system flags the response as potentially unreliable. In addition, the judge provides a step-by-step explanation of how it reached its decision, followed by a concise summary that outlines the strengths and weaknesses of the response.

2) Multi-agent LLM-as-a-Judge:: To further strengthen the evaluation process, we introduce a second version of LLM-as-a-Judge that works like a multi-agent review panel. In this setup, four distinct LLM agents work together to evaluate each response produced by the MDAgent framework. Each agent has a specific role, making the review process more organized and thorough. The Discriminator identifies flaws and logical gaps in the response. The Fact Checker verifies the information based on established medical knowledge and guidelines. The Justifier explains the reasoning behind the response. Finally, the Moderator collects all feedback, identifies areas of agreement and disagreement between agents, and assigns a final

Authorized licensed use limited to: Tampere University. Downloaded on May 14,2026 at 11:42:55 UTC from IEEE Xplore. Restrictions apply.

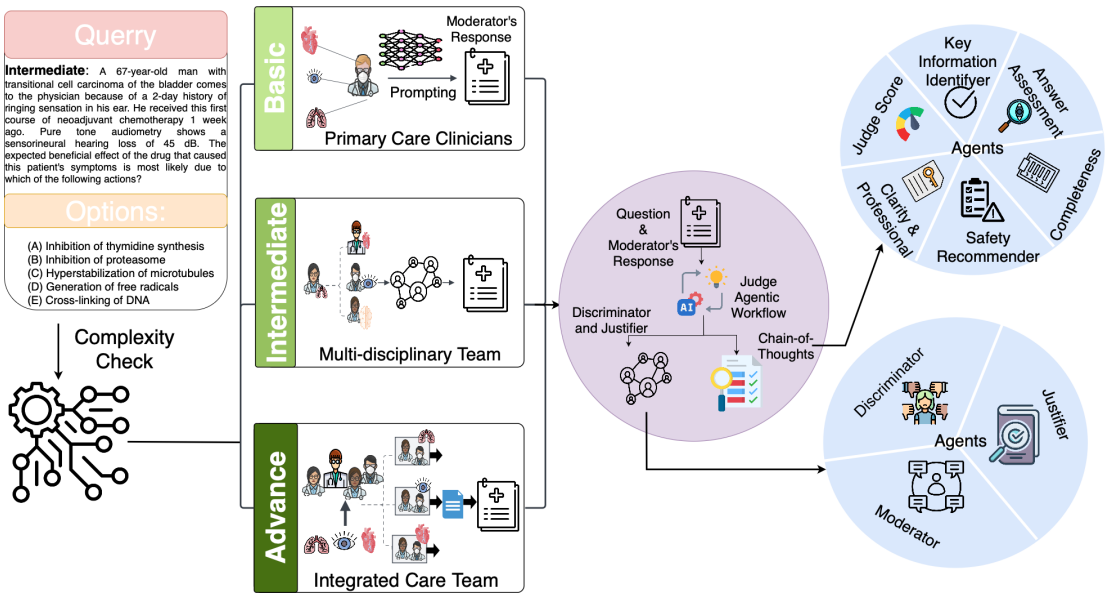

Fig. 1. AI on Trial: The client submits a query which the moderator classifies by complexity (basic, intermediate, or advanced). Based on this, the query is routed to the appropriate team – PCP, MDT, or ICT. The generated response is then evaluated by an LLM-as-a-Judge. If the score falls below a set of threshold a medical doctor is brought in for review

evaluation score between 1 and 5, along with confidence level marked as low, medium, or high.

This collaborative evaluation process improves the reliability of the system. It mimics a real-world peer review model, where diverse medical perspectives are used to refine CDM. The added confidence level helps to show how certain the system is about its judgment, which is especially helpful in ambiguous cases.

This extended evaluation framework makes the MDAgent system more trustworthy and suitable for CDM. By adding both individual and team-based judges, we ensure that the generated answers are not only accurate, but also complete and safe.

AI-on-Trial

Medical Task → Agent Assignment → Group/Solo Response → LLM-as-a-Judge → Final Answer

IV. Performance Analysis

A. Experimental Results

a) System Specification: For the experiments, we used a machine running Ubuntu 22.04 LTS, equipped with an AMD Ryzen 7 6800H processor, 16 GB of RAM, and an NVIDIA GeForce RTX 3060 GPU. The code is implemented in python.

b) Datasets: We tested our proposed framework using the MedQA dataset [33]. It consists of real medical exam questions from the US, Mainland China, and Taiwan. Our study focuses on the English test set (taken from the MDAgent work [11]), which has 1,273 multiple choice questions from the US Medical Licensing Exam (USMLE). Each question has five answer options.

B. Analysis

Clinical Query 1:

A 67-year-old man with transitional cell carcinoma of the bladder visits the physician with a 2-day history of ringing in his ears. He recently completed his first course of neoadjuvant chemotherapy. Audiometry shows a sensorineural hearing loss of 45 dB. **Question:** Which drug action is most likely responsible for his symptoms? **Complexity:** Intermediate **Options:**

A. Inhibition of thymidine synthetase
B. Inhibition of proteasome
C. Hyperstabilization of microtubules
D. Generation of free radicals
E. Cross-linking of DNA

MDAgent Analysis: Option (C) is most likely. Taxanes (e.g., paclitaxel) hyperstabilize microtubules and are known to cause ototoxicity (hearing loss). Other agents like alkylators or free radical generators are less commonly associated with this effect.
Final Answer: C) Hyperstabilization of microtubules
LLM-as-Judge Score (CoT's): 4
LLM-as-Judge Score (multi-agent): 3
LLM-as-Judge answer (multi-agent): (E) Cross-linking of DNA
Ground Truth: E) Cross-linking of DNA

Authorized licensed use limited to: Tampere University. Downloaded on May 14,2026 at 11:42:55 UTC from IEEE Xplore. Restrictions apply.

In this section, we discuss how well MDAgents and our proposed approaches perform in CDM. We also highlight their limitations and explain why they sometimes make mistakes. Understanding these errors is important to improve future systems. Mistakes may occur because the model lacks medical knowledge or because its reasoning is flawed.

We begin by analyzing the performance of the original MDAgents framework. One common issue we observe is misclassifications. For example, in the case shown in **Clinical Query 1**, the correct answer to the query is option **E**. However, MDAgents incorrectly selected option **C** as the final answer.

Clinical Query 2:

A 32-year-old G1P1 patient presents to her obstetrician after having a positive pregnancy test at home. She reports that she and her husband had been trying to have a child for the past three months. She has no history of sexually transmitted disease, intravenous drug use, or blood transfusions, and she has never traveled outside of the United States. She was up-to-date on all immunizations before her pregnancy. Ultrasound is consistent with an 8-week gestational sac. The patient requests as few tests as possible, although she does not want to compromise the health of her fetus.

Question: Which of the following screening tests should be performed on all pregnant women?

Complexity: Intermediate

Options:

A. Reassure and recommend vaccination against rubella postpartum

B. Recommend pregnancy termination

C. Arrange a chorionic villus sampling

D. Recommend additional serologic testing for parvovirus B19

E. Test for rubella virus PCR

MDAgent Answer: E

LLM Judge Answer (multi-agent): A

LLM Judge Score: 2

Ground Truth: A

In this example, the incorrect classification was performed by a single PCP. However, in another case, Clinical Query 2, both the majority opinion and the final decision pointed to the wrong answer.

There are few possible reasons for these misclassifications. First, the individual agent may not have had enough medical knowledge to properly understand the query. Second, the agent selected to respond might not have been the most appropriate for that specific query complexity. Third, collaborative discussion among agents could have been influenced by incorrect interpretations, leading the conversation in an unintended direction.

To improve this, we extended MDAgents by adding LLM-as-a-Judge that uses CoT's reasoning to evaluate responses. However, in the same example, even our proposed approach failed. The judge still gave a high score to the incorrect response generated by the MDAgents.

This shows a major issue. The judge was evaluating how well the answer was written, not whether it was actually correct. Since it used the flawed MDAgent answer as its starting point, its judgment was also flawed. This suggests that CoT-based judges can be biased if the input they are reviewing is already wrong.

To solve this, we tested the same query using our second version of LLM-as-a-Judge, which uses multi-agent LLMs for evaluation. In this setup, four different agents work together: a **discriminator** to spot flaw, a **Fact Checker** to verify accuracy, a **Justifier** to explain the reasoning, and a Moderator to make the final judgment. As shown in the Clinical Query 1, this version correctly identified option **E** as the right answer. Similarly, for Clinical Queries 2 (intermediate) and 3 (Advanced) , it also selects the correct option.

The reason this multi-agent Judge performed better is because it did not rely solely on the final response. Instead, it carefully reviewed both the query and the generated response. The Discriminator highlighted flaws, the Fact Checker ensured medical accuracy, and the Moderator considered all viewpoints before assigning a score and confidence level.

Note: We are not claiming that the multi-agent judge fixed all the misclassifications made by MDAgents and the CoT-based LLM-as-a-Judge. But it was able to fix most of them.

In summary, this discussion highlights an important finding: even well-structured evaluation systems can fail if they depend too much on flawed inputs. Our multi-agent LLM-based Judge is more robust because it brings multiple perspectives into the evaluation process. This makes the final judgment more accurate and trustworthy, especially in critical domains like healthcare.

We evaluated LLM-as-a-Judge on the MedQA dataset, which contains 1,273 samples. The results show a consistent improvement in diagnostic accuracy when integrating the **LLM-as-a-Judge** with the baseline **MDAgents** framework. LLM Judge successfully corrected 49 misclassified samples predicted by baseline MDAgents, covering all three complexity levels (easy, medium, hard). This demonstrates the ability to improve reasoning consistency and factual correctness in diverse clinical contexts. A smaller subset of 20 samples remained challenging for both systems, often due to inherent ambiguity or overlapping medical reasoning. These findings confirm that incorporating a clinically informed multi-agent evaluation layer not only reduces uncorrected diagnostic errors but also improves the robustness, interpretability, and trustworthiness of AI-assisted clinical decision-making.

Authorized licensed use limited to: Tampere University. Downloaded on May 14,2026 at 11:42:55 UTC from IEEE Xplore. Restrictions apply.

Clinical Query 3:

Two days after being admitted to the hospital following a fall, a 77-year-old woman complains of fatigue and headaches. During the fall she sustained a right-sided subdural hematoma. She has hypertension and hyperlipidemia. Her medications prior to admission were hydrochlorothiazide and atorvastatin. Vital signs are within normal limits. Physical and neurologic examinations show no abnormalities.
Laboratory studies show:
Serum

Na^+: 130 mEq/L
K^+: 4.0 mEq/L
Cl^-: 103 mEq/L
HCO_3^-: 24 mEq/L
Urea nitrogen: 14 mg/dL
Creatinine: 1.1 mg/dL
Osmolality: 270 mOsmol/kg H_2O

Urine

Sodium: 45 mEq/L
Osmolality: 326 mOsmol/kg H_2O

A CT scan of the head shows an unchanged right-sided subdural hematoma.
Question: Which of the following is the most appropriate next step in management?
Complexity: Advanced
Options:

A. Head elevation
B. Tolvaptan
C. Desmopressin
D. Fluid restriction
E. Mannitol

MDAgent Answer: B
LLM Judge Answer (multi-agent): D
LLM Judge Score (multi-agent): 2
Ground Truth: D

C. Open Scince and Reproducible Research

To support **open science and reproducible research**, our work, source code used for the evaluations is publicly available AI on Trial.

V. Conclusion and Future Direction

In this paper, we introduce `AI on Trial`, a framework built on top of MDAgents to improve CDM using LLMs. Like MDAgents, our method assigns agents to work alone or in teams depending on how difficult the task is. To extend the reliability of the system, we introduced an LLM-as-a-Judge component that evaluates the quality of responses and helps correct misclassified queries. This addition is especially important in healthcare, where incorrect classification can have serious consequences. We tested our approach in the MedQA dataset and found that `AI on Trial`was able to correctly classify several queries that MDAgents handled incorrectly. These results highlight the importance of integrating evaluation and feedback mechanisms into multi-agent LLM frameworks for safer and more accurate CDM.

In the future, we plan to enhance the proposed model by analyzing CoT's judge limitations, integrating privacy-preserving methods, and extending evaluation to real-world clinical data. Additionally, we will incorporate a human-in-loop component. When the automated Judge identifies a weak or uncertain response, a medical expert (such as a doctor) can step in to review and improve it. This expert feedback will not only help refine the final output, but will also be used to fine-tune both the agents and the judge in future iterations. This approach strikes a balance between automation and safety, which is especially important in sensitive domains like medicine.

Acknowledgment

This work was funded by 31000755 ERC-hankkeen vastinrahoitus (Brumley). The authors also sincerely thank the peer reviewers for their valuable comments and suggestions that helped improve the quality of this paper.

References

[1] M. Banning, "A review of clinical decision making: models and current research," *Journal of clinical nursing*, vol. 17, no. 2, pp. 187–195, 2008.
[2] J. Li, Z. Zhou, H. Lyu, and Z. Wang, "Large language models-powered clinical decision support: enhancing or replacing human expertise?" pp. 1–4, 2025.
[3] W. S. Richardson, M. C. Wilson, J. Nishikawa, and R. S. Hayward, "The well-built clinical question: a key to evidence-based decisions." *ACP journal club*, vol. 123, no. 3, pp. A12–3, 1995.
[4] E. Brügge, S. Ricchizzi, M. Arenbeck, M. N. Keller, L. Schur, W. Stummer, M. Holling, M. H. Lu, and D. Darici, "Large language models improve clinical decision making of medical students through patient simulation and structured feedback: a randomized controlled trial," *BMC medical education*, vol. 24, no. 1, p. 1391, 2024.
[5] L. Tang, Z. Sun, B. Idnay, J. G. Nestor, A. Soroush, P. A. Elias, Z. Xu, Y. Ding, G. Durrett, J. F. Rousseau *et al.*, "Evaluating large language models on medical evidence summarization," *NPJ digital medicine*, vol. 6, no. 1, p. 158, 2023.
[6] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.
[7] K.-H. Jung, "Large language models in medicine: Clinical applications, technical challenges, and ethical considerations," *Healthcare Informatics Research*, vol. 31, no. 2, pp. 114–124, 2025.
[8] Y. Zhang, Y. Li, L. Cui, D. Cai, L. Liu, T. Fu, X. Huang, E. Zhao, Y. Zhang, Y. Chen *et al.*, "Siren's song in the ai ocean: A survey on hallucination in large language models," *Computational Linguistics*, pp. 1–45, 2025.
[9] Z. Yao and H. Yu, "A survey on llm-based multi-agent ai hospital," 2025.
[10] X. Chen, Y. Jin, X. Mao, L. Wang, S. Zhang, and T. Chen, "Rareagents: Autonomous multi-disciplinary team for rare disease diagnosis and treatment," *arXiv e-prints*, pp. arXiv–2412, 2024.
[11] Y. Kim, C. Park, H. Jeong, Y. S. Chan, X. Xu, D. McDuff, H. Lee, M. Ghassemi, C. Breazeal, and H. W. Park, "Mdagents: An adaptive collaboration of llms for medical decision-making," *Advances in Neural Information Processing Systems*, vol. 37, pp. 79 410–79 452, 2024.
[12] H. Wang, S. Zhao, Z. Qiang, N. Xi, B. Qin, and T. Liu, "Beyond direct diagnosis: Llm-based multi-specialist agent consultation for automatic diagnosis," *arXiv preprint arXiv:2401.16107*, 2024.
[13] T. Tu, A. Palepu, M. Schaekermann, K. Saab, J. Freyberg, R. Tanno, A. Wang, B. Li, M. Amin, N. Tomasev *et al.*, "Towards conversational diagnostic ai," *arXiv preprint arXiv:2401.05654*, 2024.

Authorized licensed use limited to: Tampere University. Downloaded on May 14,2026 at 11:42:55 UTC from IEEE Xplore. Restrictions apply.

[14] B. Yang, S. Jiang, L. Xu, K. Liu, H. Li, G. Xing, H. Chen, X. Jiang, and Z. Yan, "Drhouse: An llm-empowered diagnostic reasoning system through harnessing outcomes from sensor data and expert knowledge," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 8, no. 4, pp. 1–29, 2024.

[15] J. Wei, D. Yang, Y. Li, Q. Xu, Z. Chen, M. Li, Y. Jiang, X. Hou, and L. Zhang, "Medaide: Towards an omni medical aide via specialized llm-based multi-agent collaboration," *arXiv preprint arXiv:2410.12532*, 2024.

[16] S. Li, V. Balachandran, S. Feng, J. Ilgen, E. Pierson, P. W. W. Koh, and Y. Tsvetkov, "Mediq: Question-asking llms and a benchmark for reliable interactive clinical reasoning," *Advances in Neural Information Processing Systems*, vol. 37, pp. 28 858–28 888, 2024.

[17] Z. Liu, Q. Tu, W. Ye, Y. Xiao, Z. Zhang, H. Cui, Y. Zhu, Q. Ju, S. Li, and J. Xie, "Exploring the inquiry-diagnosis relationship with advanced patient simulators," *arXiv preprint arXiv:2501.09484*, 2025.

[18] T. Khan and A. Michalas, "Learning in the dark: Privacy-preserving machine learning using function approximation," in *2023 IEEE 22nd International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*. IEEE, 2023, pp. 62–71.

[19] T. Khan, K. Nguyen, A. Michalas, and A. Bakas, "Love or hate? share or split? privacy-preserving training using split learning and homomorphic encryption," in *2023 20th Annual International Conference on Privacy, Security and Trust (PST)*. IEEE, 2023, pp. 1–7.

[20] T. Khan, K. Nguyen, and A. Michalas, "A more secure split: Enhancing the security of privacy-preserving split learning," in *Nordic Conference on Secure IT Systems*. Springer, 2023, pp. 307–329.

[21] T. Khan, M. Budzys, and A. Michalas, "Make split, not hijack: Preventing feature-space hijacking attacks in split learning," in *Proceedings of the 29th ACM on Symposium on Access Control Models and Technologies*, ser. SACMAT '24. New York, NY, USA: Association for Computing Machinery, 2024.

[22] T. Khan, M. Budzys, K. Nguyen, and A. Michalas, "Sok: Wildest dreams: Reproducible research in privacy-preserving neural network training," in *Proceedings of the 24th Privacy Enhancing Technologies Symposium (PETS'24)*. Berlin, Heidelberg: Springer-Verlag, 2024.

[23] T. Khan, M. Budzys, and A. Michalas, "Split happens: Combating advanced threats with split learning and function secret sharing," in *2025 IEEE Conference on Communications and Network Security (CNS)*. IEEE, 2025, pp. 1–9.

[24] T. Khan and A. Michalas, "Oops!... they stole it again: Attacks on split learning," *arXiv preprint arXiv:2508.10598*, 2025.

[25] K. Nguyen, T. Khan, H. Abdinasibfar, and A. Michalas, "A privacy-centric approach: Scalable and secure federated learning enabled by hybrid homomorphic encryption," *arXiv preprint arXiv:2507.14853*, 2025.

[26] H. Wimmer, V. Y. Yoon, and V. Sugumaran, "A multi-agent system to support evidence based medicine and clinical decision making via data sharing and data privacy," *Decision Support Systems*, vol. 88, pp. 51–66, 2016.

[27] Z. Jeddi and A. Bohr, "Remote patient monitoring using artificial intelligence," in *Artificial intelligence in healthcare*. Elsevier, 2020, pp. 203–234.

[28] Y. Shen, J. Shao, X. Zhang, Z. Lin, H. Pan, D. Li, J. Zhang, and K. B. Letaief, "Large language models empowered autonomous edge ai for connected intelligence," *IEEE Communications Magazine*, vol. 62, no. 10, pp. 140–146, 2024.

[29] V. Liévin, C. E. Hother, A. G. Motzfeldt, and O. Winther, "Can large language models reason about medical questions?" *Patterns*, vol. 5, no. 3, 2024.

[30] S. Thapa and S. Adhikari, "Chatgpt, bard, and large language models for biomedical research: opportunities and pitfalls," *Annals of biomedical engineering*, vol. 51, no. 12, pp. 2647–2651, 2023.

[31] D. McDuff, M. Schaekermann, T. Tu, A. Palepu, A. Wang, J. Garrison, K. Singhal, Y. Sharma, S. Azizi, K. Kulkarni *et al.*, "Towards accurate differential diagnosis with large language models," *Nature*, pp. 1–7, 2025.

[32] S. Tripathi, R. Sukumaran, and T. S. Cook, "Efficient healthcare with large language models: optimizing clinical workflow and enhancing patient care," *Journal of the American Medical Informatics Association*, vol. 31, no. 6, pp. 1436–1440, 2024.

[33] D. Jin, E. Pan, N. Oufattole, W.-H. Weng, H. Fang, and P. Szolovits, "What disease does this patient have? a large-scale open domain question answering dataset from medical exams," *Applied Sciences*, vol. 11, no. 14, p. 6421, 2021.

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

[35] G. Li, H. Hammoud, H. Itani, D. Khizbullin, and B. Ghanem, "Camel: Communicative agents for" mind" exploration of large language model society," *Advances in Neural Information Processing Systems*, vol. 36, pp. 51 991–52 008, 2023.

[36] W. Chen, Y. Su, J. Zuo, C. Yang, C. Yuan, C.-M. Chan, H. Yu, Y. Lu, Y.-H. Hung, C. Qian *et al.*, "Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors," in *ICLR*, 2024.

[37] Q. Wu, G. Bansal, J. Zhang, Y. Wu, B. Li, E. Zhu, L. Jiang, X. Zhang, S. Zhang, J. Liu *et al.*, "Autogen: Enabling next-gen llm applications via multi-agent conversations," in *First Conference on Language Modeling*.

[38] Y. Du, S. Li, A. Torralba, J. B. Tenenbaum, and I. Mordatch, "Improving factuality and reasoning in language models through multiagent debate," in *Forty-first International Conference on Machine Learning*, 2023.

[39] H. Wang, X. Du, W. Yu, Q. Chen, K. Zhu, Z. Chu, L. Yan, and Y. Guan, "Learning to break: Knowledge-enhanced reasoning in multi-agent debate system," *Neurocomputing*, vol. 618, p. 129063, 2025.

[40] Y. Liu, Y. Liu, X. Zhang, X. Chen, and R. Yan, "The truth becomes clearer through debate! multi-agent systems with large language models unmask fake news," in *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2025, pp. 504–514.

[41] X. Wang, J. Wei, D. Schuurmans, Q. V. Le, E. H. Chi, S. Narang, A. Chowdhery, and D. Zhou, "Self-consistency improves chain of thought reasoning in language models," in *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. [Online]. Available: https://openreview.net/forum?id=1PL1NIMMrw

[42] J. C. Chen, S. Saha, and M. Bansal, "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), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, L. Ku, A. Martins, and V. Srikumar, Eds. Association for Computational Linguistics, 2024, pp. 7066–7085. [Online]. Available: https://doi.org/10.18653/v1/2024.acl-long.381

[43] F. Bianchi, P. Chia, M. Yuksekgonul, J. Tagliabue, D. Jurafsky, J. Zou *et al.*, "How well can llms negotiate? negotiationarena platform and analysis," *PROCEEDINGS OF MACHINE LEARNING RESEARCH*, vol. 235, pp. 3935–3951, 2024.

[44] Z. Liu, Y. Zhang, P. Li, Y. Liu, and D. Yang, "A dynamic llm-powered agent network for task-oriented agent collaboration," in *First Conference on Language Modeling*, 2024.

[45] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing *et al.*, "Judging llm-as-a-judge with mt-bench and chatbot arena," *Advances in neural information processing systems*, vol. 36, pp. 46 595–46 623, 2023.

[46] S. Es, J. James, L. E. Anke, and S. Schockaert, "Ragas: Automated evaluation of retrieval augmented generation," in *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, 2024, pp. 150–158.

Authorized licensed use limited to: Tampere University. Downloaded on May 14,2026 at 11:42:55 UTC from IEEE Xplore. Restrictions apply.