MedCoScientist: A Multi-Agent LLM Framework for Clinical Decision Support

Demonstration Track

Gleb Solovev
ITMO University
Saint Petersburg, Russia
glebsolo46@gmail.com

Ivan Gurev
ITMO University
Saint Petersburg, Russia
ivanzo17898@gmail.com

Tatyana Polevaya
ITMO University
Saint Petersburg, Russia
tpolevaya@itmo.ru

Zubanenko Aleksei
ITMO University
Saint Petersburg, Russia
aazubanenko@itmo.ru

Nikolay Nikitin
ITMO University
Saint Petersburg, Russia
nnikitin@itmo.ru

ABSTRACT

Direct application of large language models (LLMs) in clinical practice is constrained by response instability and the risk of hallucinations; therefore, a more verifiable form of human-centered AI support is required. We present a demonstrative case study of the MedCoScientist system—a multi-agent framework designed to support clinical decision-making[1]. Using a rare endocrine emergency, pituitary apoplexy, as an example, the system ingests pituitary MRI and a brief clinical history, extracts and interprets key findings, formulates and validates a differential diagnosis, and then automatically retrieves relevant PubMed publications with study-type labels and PICO extraction. This case illustrates the practical utility of multi-agent systems in a clinician-in-the-loop setting: rather than delivering a "final answer," the system provides traceable, evidence-based support while leaving the ultimate decision to the physician.

KEYWORDS

Multi-Agent Systems; Large Language Models; Clinical Medicine

ACM Reference Format:
Gleb Solovev, Ivan Gurev, Tatyana Polevaya, Zubanenko Aleksei, and Nikolay Nikitin. 2026. MedCoScientist: A Multi-Agent LLM Framework for Clinical Decision Support: Demonstration Track. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 3 pages. https://doi.org/10.65109/GDIW9780

1 INTRODUCTION

Large language models (LLMs) are increasingly explored for use in high-stakes domains such as clinical medicine, yet their direct deployment remains problematic[11]. State-of-the-art models are not ready for autonomous clinical decision-making: their outputs can vary with phrasing, information volume, they integrate poorly into established clinical workflows [3]. Importantly, access to advanced models (e.g., GPT-4) has not been shown to reliably improve physicians' diagnostic reasoning[7]. Moreover, LLM "hallucinations"—plausible but incorrect statements—pose a particularly serious risk in healthcare[2].

Recent works indicates that multi-agent systems built on LLMs can outperform single-model approaches by decomposing tasks, enabling internal cross-checks, and leveraging external tools[5, 6]. Building on this direction, we present MedCoScientist[2], an extension of the CoScientist[3] framework previously validated in drug-molecule discovery. MedCoScientist combines an LLM-based multi-agent architecture with specialized medical computational and predictive tools, allowing the system to verify and refine generated outputs rather than relying solely on free-form text generation. Using pituitary apoplexy as a case study, we demonstrate how this multi-agent framework can operate as a clinically oriented decision-support system for physicians managing a rare endocrine emergency.

2 DEMONSTRATION CASE STUDY

Clinicians may experience difficulties when identifying pathologies that are rarely encountered in routine practice [1, 8]. In recent years, artificial intelligence–based approaches have been proposed to address this challenge, achieving moderate success [10]. In medicine, the reliability of AI recommendations is critical; however, it is well known that, without validation and the use of specialized tools, large language models (LLMs) may hallucinate and produce outputs that mislead clinicians.

To enhance the clinical usability of large language models (LLMs), we propose a multi-agent pipeline and evaluate it on pituitary apoplexy (PA). The system ingests a T1-weighted pituitary MRI and a brief clinical history, classifies DICOM modality, and extracts imaging features that are interpreted in clinical context. The resulting assessment is validated, and a diagnosis plus salient keywords are produced to support downstream literature retrieval. Relevant PubMed articles are then retrieved to ground recommendations in existing evidence and are presented with practitioner-oriented

[1] https://youtu.be/ezTswtNTbNA

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/GDIW9780

[2] https://github.com/ITMO-NSS-team/CoScientist/tree/main/MedCoScientist
[3] https://github.com/ITMO-NSS-team/CoScientist

tags. Each paper is automatically classified by study type: observational, experimental, or review. Observational studies are further labeled by temporal design (prospective/retrospective) and cohort structure (cross-sectional/cohort/case–control). Experimental studies are characterized by setting (clinical/in vivo/in vitro/in silico) and allocation strategy (randomized/non-randomized/propensity score–matched). Reviews are categorized as narrative, systematic, or meta-analytic. Finally, the system extracts PICO elements[4] within the evidence-based medicine (EBM)[9] framework (Population/Problem, Intervention, Comparison, Outcome) from each publication.

Using this demonstration task, we show that a multi-agent system combining LLMs and vision–language models (VLMs) can serve as a clinical assistant by supporting diagnostic reasoning and providing targeted literature, while keeping responsibility for the final diagnosis with the physician. In contrast to producing an unverified "final answer," our approach emphasizes traceability, evidence grounding, and decision support.

3 SYSTEM DESCRIPTION

Overall Design. The proposed system implements a modular, agent-oriented architecture for complex biomedical and scientific reasoning tasks. The architecture combines LLMs, tool-augmented agents, and a graph-based memory subsystem within a structured planning–execution– reflection loop. The system is designed to decompose high-level user queries into executable steps, dynamically dispatch specialized agents, and iteratively refine intermediate results until a final response is produced.

At a high level, the system consists of four tightly coupled subsystems: (1) data and state representations, (2) a set of control agents responsible for planning and orchestration, and (3) domain-specific execution agents implemented as callable nodes.

Control and Orchestration Agents. The system's execution flow is governed by a hierarchy of control agents. These agents are powered by the qwen3-235b-a22b model.

Chat Agent acts as the primary entry point. It determines whether a query can be answered directly or requires multi-step reasoning. If necessary, it delegates planning to the planner agent. *Planner Agent* transforms the user query into an explicit, ordered plan consisting of atomic execution steps. *Supervisor Agent* executes the plan by sequentially dispatching domain-specific agents. It manages tool invocation, collects intermediate task results, and handles execution failures. *Re-planner Agent* evaluates intermediate results and decides whether plan refinement is required. This enables adaptive execution in cases of incomplete or unsatisfactory intermediate outputs. *Summary Agent* aggregates completed task results into a concise, coherent summary, which is then transformed into the final user-facing response.

Domain-Specific Execution Agents. The system employs several scenario agents, each encapsulating a well-defined biomedical reasoning capability: *Hypothesis PICO Agent*, powered by qwen3-235b-a22b, decomposes a scientific or clinical hypothesis into structured PICO elements. It uses an LLM for argument extraction and dedicated tool nodes for formal PICO parsing. The output is stored as a structured intermediate result and reused in downstream tasks. *Image Analyzer Agent* processes medical imaging studies (e.g., MRI) together with brief symptom descriptions and patient history. It uses a fine-tuned VLM Gemma 27B and a DICOM-focused prompt tailored to endocrinology and cardiology. The agent recognizes modality/sequence, detects abnormalities, generates differential diagnoses, and validates results in a structured, role-based pipeline.

The chain-of-thought pipeline has four stages: Classification - identifies modality and acquisition/sequence features using clinical context. Finding extraction—summarizes key anatomical/pathological features with severity, confidence, artifacts, and attention to rare conditions. Interpretation - maps findings to ICD codes, produces a differential, and outputs a brief clinical report and prognosis. Validation—checks conclusions against guidelines (e.g., ACR), refines the differential (especially rare cases), and generates PubMed search keywords for follow-up review.

PubMed Literature Agent automates literature discovery via PubMed and uses the GPT-4o model. Based on extracted keywords or hypotheses, it retrieves relevant publications, performs taxonomy classification, and extracts PICO elements for each paper. Retrieved articles and metadata are stored in the system state for subsequent reasoning or summarization.

Execution Flow. Upon receiving a user query, the system evaluates whether multi-step reasoning is required. If so, a plan is generated and executed step-by-step by the supervisor agent, which dynamically invokes the appropriate domain-specific agents. Intermediate results may trigger replanning. Once execution converges, results are summarized and returned to the user.

4 EXPERIMENTAL STUDIES

We ran a small pilot comparing GPT-5.2, DeepSeek-V3.2, and Gemini-3-Flash. For MRI interpretation with patient symptoms, GPT-5.2 performed best, giving the correct primary diagnosis but a limited differential. Gemini also found the correct diagnosis with little justification, while DeepSeek missed it. All models extracted PICO elements, but PubMed retrieval varied: GPT-5.2 returned the most relevant articles, Gemini returned more but less relevant results, and DeepSeek returned few with limited relevance. In contrast, MedCoScientist provides an end-to-end pipeline and, with its RePlanner agent, supports stepwise verification and iterative refinement for more reliable decisions than a single LLM.

5 CONCLUSION AND FUTURE WORKS

We introduced MedCoScientist, a multi-agent framework that mitigates key limitations of autonomous LLMs in clinical settings by emphasizing traceability, evidence grounding, and iterative verification. Using pituitary apoplexy as a representative scenario of a rare medical emergency, we demonstrated that the system can function as a clinical decision support system, outperforming single LLMs in both the accuracy and justification of its responses. This work is a step toward practical, clinician-in-the-loop assistants for rare conditions.

In future work, we plan to integrate agents' reasoning directly into the chat interface to improve the interpretability of results; conduct larger-scale experiments with comparisons against a broader set of comparable systems; and incorporate evaluation by expert clinicians to measure trust and the practical applicability of MedCoScientist's recommendations.

ACKNOWLEDGMENTS

This work supported by the Ministry of Economic Development of the Russian Federation (IGK 000000C313925P4C0002), agreement No139-15-2025-010.

REFERENCES

[1] Karolina Budych, Thomas M Helms, and Carsten Schultz. 2012. How do patients with rare diseases experience the medical encounter? Exploring role behavior and its impact on patient–physician interaction. *Health policy* 105, 2-3 (2012), 154–164.

[2] Shan Chen, Mingye Gao, Kuleen Sasse, Thomas Hartvigsen, Brian Anthony, Lizhou Fan, Hugo Aerts, Jack Gallifant, and Danielle S Bitterman. 2025. When helpfulness backfires: LLMs and the risk of false medical information due to sycophantic behavior. *npj Digital Medicine* 8, 1 (2025), 605.

[3] Paul Hager, Friederike Jungmann, Robbie Holland, Kunal Bhagat, Inga Hubrecht, Manuel Knauer, Jakob Vielhauer, Marcus Makowski, Rickmer Braren, Georgios Kaissis, et al. 2024. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature medicine* 30, 9 (2024), 2613–2622.

[4] Xiaoli Huang, Jimmy Lin, and Dina Demner-Fushman. 2006. Evaluation of PICO as a knowledge representation for clinical questions. In *AMIA annual symposium proceedings*, Vol. 2006. 359.

[5] 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 (2024), 79410–79452.

[6] Dongliang Li and Syaheerah Lebai Lutfi. 2026. Large Language Model–Based Virtual Patient Systems for History-Taking in Medical Education: Comprehensive Systematic Review. *JMIR Med Inform* 14 (2 Jan 2026), e79039. https://doi.org/10.2196/79039

[7] Ziwei Niu, Shuyi Ouyang, Shiao Xie, Yen-wei Chen, and Lanfen Lin. 2024. A survey on domain generalization for medical image analysis. *arXiv preprint arXiv:2402.05035* (2024).

[8] Christine Phillips, Anne Parkinson, Tergel Namsrai, Anita Chalmers, Carolyn Dews, Dianne Gregory, Elaine Kelly, Christine Lowe, and Jane Desborough. 2024. Time to diagnosis for a rare disease: managing medical uncertainty. A qualitative study. *Orphanet Journal of Rare Diseases* 19, 1 (2024), 297.

[9] David L Sackett, William MC Rosenberg, JA Muir Gray, R Brian Haynes, and W Scott Richardson. 1996. Evidence based medicine: what it is and what it isn't. , 71–72 pages.

[10] Hongbin Yu, Tian Chen, Xin Zhang, Yunfan Yang, Qinyu Liu, Chenlu Yang, Kai Shen, He Li, Wenjiao Tang, Xushu Zhong, et al. 2025. Performance of large language models in diagnosing rare hematologic diseases and the impact of their diagnostic outputs on physicians: combined retrospective and prospective study. *Journal of Medical Internet Research* 27 (2025), e77334.

[11] Juexiao Zhou, Haoyang Li, Siyuan Chen, Zhangtianyi Chen, Zhongyi Han, and Xin Gao. 2025. Large language models in biomedicine and healthcare. *npj Artificial Intelligence* 1, 1 (2025), 44.