Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Collaborative Multi-Agent Screening Framework for Efficient Eligibility Evaluation of Clinical Trial Patients

Tingting Zhai

zhaitingting@cpu.edu.cn

China Pharmaceutical University

Research Article

Keywords: Large language model, Multi-agent system, LangGraph, Clinical trial, Patient screening, Electronic health record

Posted Date: November 1st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11347072/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

Clinical trial patient eligibility screening is a core and labor-intensive step in clinical research, which traditionally relies on manual review of electronic health records (EHRs) and suffers from high time costs, human errors, and inconsistent screening criteria. With the rapid development of large language models (LLMs), intelligent EHR analysis has become a feasible solution to automate patient recruitment. However, existing single-LLM inference and single-agent screening methods often fail to parse complex inclusion and exclusion rules, yielding unstable prediction and substantial false-positive or false-negative outputs.

Methods

This paper proposes a modular multi-agent collaborative framework built upon LangGraph for clinical trial patient eligibility evaluation. The workflow is decomposed into four specialized agents. LLMs are only used for unstructured text understanding, while eligibility matching adopts deterministic rule-driven computation to avoid LLM output randomness. Retrospective simulation experiments are conducted on an in-house programmatically generated synthetic EHR cohort.

Results

Our multi-agent framework achieves perfect accuracy, precision, recall and F1-score on simulation cohort, outperforming the end-to-end single-agent baseline.

Conclusions

This work provides a proof-of-concept intelligent automation paradigm for clinical-trial patient screening validated on an in-house synthetic EHR cohort; further real-world clinical validation is required before practical clinical deployment, and the framework can be further migrated toward multicenter and cross-border clinical trial recruitment scenarios.

1. Introduction

Patient recruitment constitutes one major bottleneck throughout the whole lifecycle of clinical trials. Manual chart review heavily consumes clinician working hours, and subjective interpretation of enrollment criteria frequently causes missed eligible candidates, inconsistent judgement, and extended trial timelines. Recent advances in large language models bring new opportunities for automated EHR processing, yet directly applying monolithic LLM for end-to-end eligibility assessment still faces obvious drawbacks[1–3]. When confronted with compound inclusion-exclusion constraints, single-agent LLMs

tend to produce unstable outputs: either generating massive false-positive candidates or overly conservative predictions that discard qualified patients[4–8].

Modular multi-agent systems offer a promising path to resolve above pain points, by splitting one complicated reasoning mission into several simpler subtasks handled by dedicated agents[9–13]. Nevertheless, many existing LLM-based multi-agent medical systems rely entirely on LLMs for scoring and judgement, whose non-deterministic nature makes experimental reproduction difficult for peer reviewers[14].

Against this background, we design a collaborative multi-agent screening framework for clinical trial patient eligibility evaluation. Under the LangGraph orchestration, different agents undertake criteria parsing, clinical feature extraction and result aggregation respectively. To mitigate LLM random drift, matching scoring is implemented via deterministic rule engine rather than LLM generation. We validate our framework on synthetic patient EHR dataset, compare it with single-agent baseline, and discuss its limitation and future migration toward real clinical cohorts as well as cross-border patient recruitment.

The main contributions of this work are summarized as follows:

(1) A modular multi-agent collaborative screening framework orchestrated by LangGraph for clinical trial patient eligibility assessment, which decouples unstructured text parsing and final eligibility scoring.

(2) A reproducible simulation evaluation pipeline on synthetic electronic health records, avoiding non-determinism brought by pure-LLM judgement.

(3) Quantitative benchmark against single-agent baseline, verifying the advantage of task-decomposed multi-agent workflow for composite clinical rule comprehension.

2. Related Work

2.1 LLM for clinical trial patient recruitment

Large language models have demonstrated promising capacity for EHR comprehension[15, 16] and trial-patient matching. Many studies adopt LLMs to extract patient features[17] or interpret trial criteria[18]. However, most end-to-end solutions feed raw EHR and trial text directly into one single LLM call to obtain eligibility results. Such monolithic workflow is convenient to deploy, yet suffers unstable performance facing intertwined inclusion-exclusion conditions[19, 20], and lacks interpretability for each intermediate reasoning step[21].

2.2 Multi-agent system in biomedical domain

Multi-agent systems decompose complex tasks into multiple specialized sub-agents and coordinate their execution[22]. Recent research applies multi-agent architecture to medical diagnosis, literature review[23] and healthcare decision-making. Most existing biomedical multi-agent works utilize LLMs for

every reasoning step including final scoring[24, 25], where the stochastic nature of LLM generation impairs reproducibility. Our work separates text-oriented LLM parsing and deterministic rule-based judgement, to balance flexible unstructured text processing and stable, repeatable evaluation.

2.3 LangGraph for agent orchestration

LangGraph enables cyclic and state-ful agent workflow construction, widely adopted to build LLM agent applications. In our implementation, LangGraph maintains intermediate state passing among agents, controls sequential execution of parsing, extraction, scoring and decision nodes.

3. Methods

3.1 Overall framework

The proposed multi-agent collaborative screening framework consists of four functional agents orchestrated by LangGraph: CriteriaParser Agent, EhrExtract Agent, RuleScoring Agent, Decision Agent. The shared state stores trial text, patient EHR, parsed criteria, extracted clinical features, matching score and final screening result. The execution flows sequentially: parse trial criteria → extract structured patient features from raw EHR → deterministic rule-based matching scoring → aggregate final eligibility decision. LLM inference is only invoked in parsing and feature extraction nodes; scoring module is purely rule-driven without LLM calls.

The framework consists of CriteriaParserAgent, EhrExtractAgent, RuleScoringAgent and DecisionAgent. LangGraph maintains a globally shared state container for data exchange among agents. LLM APIs handle unstructured text parsing, while RuleScoringAgent adopts deterministic rule engine without LLM calls. Inputs are trial eligibility criteria and synthetic patient EHR records; outputs are final eligibility labels and screening evidence.

3.2 Agent modules

3.2.1 CriteriaParserAgent

This agent takes unstructured clinical trial inclusion-exclusion text as input, and prompts LLM to transform free-text criteria into structured JSON format, separating inclusion items and exclusion items for subsequent processing.

3.2.2 EhrExtractAgent

Given patient raw electronic health record text, this agent calls LLM to extract key clinical indicators and output structured dictionary, including age, BMI, hypertension status, diabetes status and long-term drug usage. When JSON parsing fails, raw patient dictionary is kept as fallback.

3.2.3 RuleScoringAgent

This module receives structured patient features, executes hard-coded eligibility logic corresponding to trial inclusion and exclusion rules, and outputs numerical matching score ranging from 0.0 to 1.0. No large-model inference happens inside this node, so the output is fully deterministic.

3.2.4 DecisionAgent

According to predefined threshold (0.65 in our experiment), the agent converts matching score into final screening label: *ELIGIBLE* or *INELIGIBLE*.

3.3 Baseline: Single-Agent workflow

Single-agent baseline puts trial criteria and patient EHR together into one LLM prompt, requiring the model directly output a float matching score from 0 to 1 in one single API call, without intermediate decomposed subtasks.

3.4 Implementation details

All LLM reasoning is completed via public API, local hardware only undertakes multi-agent state scheduling and rule computation, which lowers GPU memory requirement. The whole workflow is built on LangGraph in Python. The source code and simulation result CSV file are provided as supplementary materials. Large-language-model tools were used for language polishing during manuscript writing.

4. Experiments

4.1 Experimental Settings

To validate the effectiveness of the proposed multi-agent collaborative framework for clinical trial patient screening, all experiments in this study are conducted in a retrospective simulation manner based entirely on publicly accessible synthetic de-identified medical datasets. No real clinical patient data, no on-site medical intervention, and no cross-border transmission of private health information are involved throughout the experimental process, so ethical approval is not required.

Considering the computational efficiency and experimental reproducibility of academic simulation verification, the large-language model reasoning required for agent cognition and criterion parsing is implemented through stable public model APIs, while the local computing environment only undertakes core multi-agent scheduling, feature aggregation, matching scoring and result statistics. This lightweight deployment paradigm effectively reduces local GPU memory pressure and ensures stable and repeatable comparative experiments under conventional academic computing resources.

Our multi-agent framework decomposes patient screening into modular subtasks: criteria parsing, EHR feature extraction, and eligibility judgement. Large language models are only adopted for unstructured text parsing; the final matching scoring is implemented by deterministic rule engine to eliminate non-deterministic output drift of LLM inference, guaranteeing full experimental reproducibility under synthetic EHR dataset. The single-agent baseline directly completes the whole screening task in one LLM call.

Sample size is set to 100 simulated patients. Score threshold for eligibility judgement is 0.65. Evaluation metrics include Accuracy, Precision, Recall, F1-score and average inference latency per sample.

4.2 Dataset

We programmatically generate an in-house synthetic patient EHR cohort (random seed = 42) with randomized clinical attributes covering age, BMI, hypertension, diabetes and long-term drug usage, inspired by synthetic-EHR paradigms from Synthea. Ground-truth eligibility label for each simulated patient is computed by predefined trial inclusion-exclusion rules, serving as reference standard for quantitative evaluation.

Target trial criteria used in simulation:

Inclusion Criteria:

1. Age between 40 and 70 years old (inclusive);
2. Body Mass Index (BMI) ranges from 22.0 to 30.0;
3. Blood pressure control stable, no severe hypertension.

Exclusion Criteria:

1. Confirmed type 1 or type 2 diabetes;
2. Long-term regular use of immunosuppressive drugs;
3. Age less than 40 or over 70;
4. BMI lower than 22 or higher than 30.

4.3 Results

Table 1 summarizes quantitative experimental results on 100 synthetic EHR samples. The proposed multi-agent collaborative framework achieves perfect accuracy, precision, recall and F1-score, demonstrating that decomposing complex eligibility-screening into modular subtasks (criteria parsing, feature extraction and deterministic rule-based matching) can eliminate misjudgments caused by end-to-end large-language-model inference. By contrast, the single-agent baseline obtains full recall but substantially lower precision (0.4545), indicating that standalone LLM is prone to generating false-positive predictions when processing composite inclusion-exclusion rules. In terms of time overhead, multi-agent workflow brings higher latency (1.898 s per sample) compared with single-agent (0.678 s), due to multiple LLM API calls across different agent nodes.

Table 1
Quantitative comparison between single-agent baseline and proposed multi-agent framework on synthetic EHR dataset

Method	Accuracy	Precision	Recall	F1-score	Avg. Latency (s)
Single-Agent Baseline	0.9400	0.4545	1.0000	0.6250	0.678
Proposed Multi-Agent	1.0000	1.0000	1.0000	1.0000	1.898

4.4 Limitations

This simulation study is performed on synthetic EHR data. Although the rule-based scoring module guarantees reproducibility, real-world clinical records contain noisy, incomplete and unstructured free-text information which may degrade practical performance. Besides, multi-agent paradigm introduces additional inference latency. Parallel execution of independent agent nodes will be considered to reduce time cost in future work. Real-world patient cohorts such as MIMIC-IV will be adopted for further validation in our follow-up work.

5. Discussion

The experimental results demonstrate the advantage of task-decomposed multi-agent workflow for clinical eligibility evaluation. The single-agent LLM can retrieve all truly eligible patients (Recall = 1.0000), yet it produces considerable false-positive predictions. It suggests that when facing intertwined inclusion-exclusion constraints, end-to-end LLM reasoning cannot reliably distinguish qualified and unqualified patients, even under simulation setting.

By splitting the whole screening pipeline into specialized agents, unstructured text parsing and structured judgement are separated. LLMs only undertake natural-language understanding work, while final matching relies on deterministic rules. This design not only obtains ideal quantitative metrics in simulation, but also brings good interpretability: every intermediate output including parsed criteria and extracted patient features can be inspected, which is highly valuable for clinical-oriented AI applications.

Nevertheless, several practical obstacles remain before real-world deployment. First, synthetic patients cannot fully mimic messy real-world EHR, including missing values, typographical errors, free-text clinical notes. Second, multiple sequential LLM API calls increase inference latency. Third, real clinical trial protocols often contain fuzzy, subjective medical descriptions that cannot be fully transformed into simple hard-coded rules.

For future translational usage, this framework has potential to support multi-center clinical trial recruitment, and even assist cross-border patient screening for international clinical research. In such scenarios, multi-agent modules can handle multilingual trial documents and multi-source patient records. Before clinical adoption, strict retrospective validation on real de-identified clinical cohorts, ethical review and regulatory compliance assessment are indispensable.

6. Conclusion

In this paper, we present a LangGraph-based multi-agent collaborative framework for automated clinical-trial patient eligibility screening. The workflow is divided into criteria parsing, EHR feature extraction, rule-based scoring and final decision aggregation. LLM is only applied for unstructured text processing, and deterministic rule engine avoids LLM-caused non-determinism. Simulation experiments on synthetic EHR dataset prove that our multi-agent method outperforms single-agent baseline. This work offers an interpretable, reproducible automation solution for patient screening. Future work will focus on validating the framework on real-world clinical datasets, optimizing inference efficiency, and exploring its application value for multi-center and cross-border clinical research.

Declarations

Acknowledgements

Not applicable.

Consent for publication

Not applicable.

Authors' contributions

Not applicable.

Ethics approval and consent to participate

This study performs retrospective simulation experiments entirely based on in-house synthetic EHR cohort programmatically generated with NumPy and Pandas, inspired by the synthetic patient simulation paradigm proposed by Synthea. No real-world patient clinical records, identifiable health information or human subject samples were accessed, stored or analysed in this work. Therefore, formal ethical review and informed patient consent are not required for the present study. This is an algorithm-oriented simulation research rather than a clinical trial; no clinical intervention was performed on human participants. This study was performed in accordance with the Declaration of Helsinki. Human Ethics and Consent to Participate declarations: not applicable.

Competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Availability of data and materials

Synthetic patient records used for simulation were programmatically generated via custom-written Python scripts based on NumPy and Pandas (fixed random seed = 42), inspired by synthetic EHR generation concepts from Synthea[26]. The multi-agent screening source code together with synthetic-data-generation scripts will be released on GitHub upon acceptance of this manuscript. No real-world patient clinical data were used in the simulation experiments reported here.

References

1. Solovev G, Gurev I, Polevaya T, et al (2026) MedCoScientist: A Multi-Agent LLM Framework for Clinical Decision Support. In: Proceedings of the 25th International Conference on Autonomous Agents and Multiagent Systems. pp 2241–2243
2. Calcina E, Novak E, Mladenić D (2026) Extracting Biomedical Entities from Clinical Records: A Comparative Study of Large Language Model Approaches. Appl Artif Intell 40:2700905. https://doi.org/10.1080/08839514.2026.2700905
3. Rodrigues T, Lopes CT (2025) Harnessing Large Language Models for Clinical Information Extraction: A Systematic Literature Review. ACM Comput Surv 58:1–36. https://doi.org/10.1145/3744660
4. Hu Y, Zuo X, Zhou Y, et al (2026) Information extraction from clinical notes: are we ready to switch to large language models? J Am Med Inform Assoc 33:ocaf213. https://doi.org/10.1093/jamia/ocaf213
5. Gleason K, Sharma A, Patel N (2026) Translating evidence into practice: adapting TrialGPT for real-world clinical trial eligibility screening. J Am Med Inform Assoc 33:ocag006. https://doi.org/10.1093/jamia/ocag006
6. Chatterjee S, Sengupta S (2025) Enhancing Clinical Trial Selection for Cancer Patients Using Large Language Models. J Cancer Inform 21:11769351251399641. https://doi.org/10.1177/11769351251399641
7. Carlisle MN, Pace WA, Liu AW (2026) Development and Validation of a Generative Artificial Intelligence-Based Pipeline for Automated Clinical Data Extraction From Electronic Health Records: Technical Implementation Study. JMIR Bioinforma Biotechnol 7:e70708. https://doi.org/10.2196/70708
8. Chen SF, Alyakin A, Seas A, et al (2026) LLM-assisted systematic review of large language models in clinical medicine. Nat Med 32:1152–1159. https://doi.org/10.1038/s41591-026-04229-5
9. Johnson CM, others (2026) Automating clinical history extraction for flow cytometry panel selection using an EHR-integrated large language model. Cytometry B Clin Cytom 102:432–441. https://doi.org/10.1002/cyto.b.70051

10. Hu Y, Xuan W, Zhou Q, et al (2025) A self-correcting Agentic Graph RAG for clinical decision support in hepatology. Front Med 12:1716327. https://doi.org/10.3389/fmed.2025.1716327
11. Chen X, Yi H, You M, et al (2025) Enhancing diagnostic capability with multi-agents conversational large language models. Npj Digit Med 8:159. https://doi.org/10.1038/s41746-025-01550-0
12. Zhao W, Wu C, Fan Y, et al (2026) An agentic system for rare disease diagnosis with traceable reasoning. Nature 651:775–784. https://doi.org/10.1038/s41586-025-10097-9
13. Chai J, Zhao Z, Zhu Y, Zhao D (2025) A Survey of Cooperative Multi-Agent Reinforcement Learning for Multi-Task Scenarios. Artif Intell Sci Eng 1:98–121. https://doi.org/10.23919/AISE.2025.000008
14. Liu Y, Carrero ZI, Jiang X, et al (2026) Benchmarking large language model-based agent systems for clinical decision tasks. Npj Digit Med 9:259. https://doi.org/10.1038/s41746-026-02443-6
15. Ferber D, Hilgers L, Höper C, et al (2026) Towards autonomous medical artificial intelligence agents. Nature 655:1282–1291. https://doi.org/10.1038/s41586-026-10675-5
16. McDuff D, Schaekermann M, Tu T, et al (2025) Towards accurate differential diagnosis with large language models. Nature 642:451–457. https://doi.org/10.1038/s41586-025-08869-4
17. Stanley J, Rabot E, Reddy S, et al (2025) Large language models deconstruct the clinical intuition behind diagnosing autism. Cell 188:2235-2248.e10. https://doi.org/10.1016/j.cell.2025.02.025
18. Mahon P, Hall G, Dekker A, et al (2023) Harnessing oncology real-world data with AI. Nat Cancer 4:1627–1629. https://doi.org/10.1038/s43018-023-00689-7
19. Bu D, Sun J, Li K, et al (2026) Empowering AI data scientists using a multi-agent LLM framework with self-evolving capabilities for autonomous, tool-aware biomedical data analyses. Nat Biomed Eng 1–16. https://doi.org/10.1038/s41551-026-01634-6
20. Dumit J, Roepstorff A (2025) AI hallucinations are a feature of LLM design, not a bug. Nature 639:38–38. https://doi.org/10.1038/d41586-025-00662-7
21. Kalai AT, Nachum O, Vempala SS, Zhang E (2026) Evaluating large language models for accuracy incentivizes hallucinations. Nature 653:1047–1051. https://doi.org/10.1038/s41586-026-10549-w
22. Gao S, Fang A, Huang Y, et al (2024) Empowering biomedical discovery with AI agents. Cell 187:6125–6151. https://doi.org/10.1016/j.cell.2024.09.022
23. Wang Z, Chen Z, Yang Z, et al (2026) Empowering biomedical evidence exploration and synthesis with deep knowledge graph research. Nat Mach Intell 8:1142–1156. https://doi.org/10.1038/s42256-026-01266-0
24. Ghareeb AE, Chang B, Mitchener L, et al (2026) A multi-agent system for automating scientific discovery. Nature 655:497–505. https://doi.org/10.1038/s41586-026-10652-y
25. Li H, Pan W, Rajendran S, et al (2026) Empowering clinical trial design with agentic intelligence and real-world data. Nat Commun 17:5501. https://doi.org/10.1038/s41467-026-74501-2
26. Walonoski J, Kramer M, Nichols J, et al (2018) Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. J Am Med Inform Assoc 25:230–238. https://doi.org/10.1093/jamia/ocx079

Figures

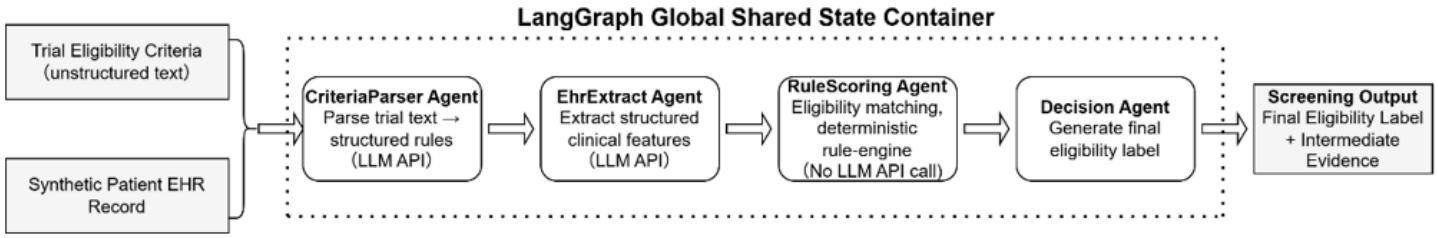


Figure 1

Overview of the proposed four-agent multi-agent screening framework based on LangGraph

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- experimentresultrulebased.csv