LLM-Match: An Open-Sourced Patient Matching Model Based on Large Language Models and Retrieval-Augmented Generation

Xiaodi Li, Ph.D.[1], Shaika Chowdhury, Ph.D.[1], Chung Il Wi, M.D.[2], Maria Vassilaki, M.D., Ph.D.[3], Xiaoke Liu, M.D., Ph.D.[4], Terence T Sio, M.D., M.S.[5], Owen Garrick, M.D.[6], Young J Juhn, M.D., M.P.H.[2], James R Cerhan, M.D. Ph.D.[7], Cui Tao, Ph.D.[1], Nansu Zong, Ph.D.[1,*]

[1]Department of Artificial Intelligence and Informatics Research, Mayo Clinic, Rochester, MN, USA; [2]Department of Pediatric and Adolescent Medicine, Mayo Clinic, Rochester, MN, USA; [3]Department of Quantitative Health Sciences, Mayo Clinic, Rochester, MN, USA; [4]Cardiology Department, Mayo Clinic Health System, La Crosse, WI, USA; [5]Department of Radiation Oncology, Mayo Clinic, Rochester, MN, USA; [6]Clinical Trials, Mayo Clinic, Rochester, MN, USA; [7]Department of Quantitative Health Sciences, Mayo Clinic, Rochester, MN, USA

ABSTRACT

Patient matching is the process of linking patients to appropriate clinical trials by accurately identifying and matching their medical records with trial eligibility criteria. We propose LLM-Match, a novel framework for patient matching leveraging fine-tuned open-source large language models. Our approach consists of four key components. First, a retrieval-augmented generation (RAG) module extracts relevant patient context from a vast pool of electronic health records (EHRs). Second, a prompt generation module constructs input prompts by integrating trial eligibility criteria (both inclusion and exclusion criteria), patient context, and system instructions. Third, a fine-tuning module with a classification head optimizes the model parameters using structured prompts and ground-truth labels. Fourth, an evaluation module assesses the fine-tuned model's performance on the testing datasets. We evaluated LLM-Match on four open datasets, n2c2, SIGIR, TREC 2021, and TREC 2022, using open-source models, comparing it against TrialGPT, Zero-Shot, and GPT-4-based closed models. LLM-Match outperformed all baselines.

INTRODUCTION

LLM-Match is a novel framework that automates patient-trial matching by integrating retrieval-augmented generation (RAG) with fine-tuned large language models (LLMs). It begins by outlining the inefficiencies of manual and rule-based recruitment systems and the limitations of earlier NLP and deep-learning methods that struggled with adaptability, long clinical text, and privacy concerns. Recent LLM-based solutions such as TrialGPT[1] improved flexibility but suffered from data-sharing risks, computational cost, and poor interpretability due to generative outputs. LLM-Match addresses these gaps through four components: (1) a RAG module that retrieves relevant EHR information, (2) structured prompt generation combining patient data and eligibility criteria, (3) fine-tuning with a classification head for precise eligibility prediction, and (4) comprehensive evaluation on four benchmark datasets (n2c2 2018[2], SIGIR 2016[3], TREC 2021, 2022[4, 5]). Using small open-source models (e.g., LLaMA 3 8B[6]), LLM-Match achieves state-of-the-art accuracy while ensuring transparency, scalability, and data privacy.

METHODS

Figure 1 provides an overview of our proposed LLM-Match framework. The framework consists of four key modules that work together to process unstructured EHR data and eligibility criteria, ensuring robust performance in real-world clinical settings: Retrieval-Augmented Generation (RAG) Module: This module retrieves the most relevant patient records from large-scale unstructured EHR data, ensuring that only pertinent information is fed into the LLM. It enhances model efficiency by filtering irrelevant data and providing contextualized input. Prompt Generation Module: To facilitate effective patient-trial matching, this module dynamically constructs structured prompts based on the retrieved clinical data and eligibility criteria. By leveraging domain-specific templates and contextual embeddings, it ensures optimal LLM comprehension and response generation. Fine-Tuning Module: This module adapts the LLM to patient-trial matching tasks using domain-specific datasets. It includes supervised fine-tuning with labeled clinical data and incorporates a classification head for eligibility prediction, enhancing interpretability and decision-making. Evaluation Module: To validate model performance, this module employs benchmark datasets and real-world clinical trial data. It assesses effectiveness using key metrics such as precision, recall, Macro-F1, and AUROC, ensuring reliable and scalable patient-trial matching.

RESULTS

We evaluate our model on four publicly available datasets: the n2c2 2018 cohort selection benchmark[2,] the Special Interest Group on Information Retrieval (SIGIR) in 2016[3], and the 2021 and 2022 Clinical Trials (CT) tracks[4, 5] of the Text REtrieval Conference (TREC).

As shown in Table 1, LLM-Match with the classification head achieves the best Macro-F1 score (0.86), outperforming baselines and prior SOTA. Its combination of RAG and fine-tuning improves balance across classes and generalizability, while using open-source LLaMA-3-8B proves that lightweight, transparent models can rival closed-source systems like GPT-4 in clinical trial matching.

As shown in Table 2, LLM-Match with the classification head achieves the highest AUROC (0.8010), surpassing all baselines, including GPT-4-based TrialGPT-Ranking. This shows strong discrimination between eligible and ineligible patients. By combining RAG with fine-tuned classification, LLM-Match offers a scalable, transparent, and cost-effective open-source alternative to proprietary models for clinical trial recruitment.

As shown in Table 3, across SIGIR, TREC 2021, and TREC 2022, LLM-Match with the classification head consistently outperforms the version without it. The ablation results show that the head improves feature extraction and decision boundaries, enhancing robustness and adaptability across diverse datasets.

TABLE 1: Experimental Results on the n2c2 2018 Cohort Selection Challenge.

Model	Precision	Recall	Overall Macro-F1
Basic Baseline	0.69	0.78	0.43
Prior SOTA	0.88	0.91	0.75
zero-shot (Llama-2-70b)	0.82	0.41	0.46
zero-shot (Mixtral-8x7B)	0.72	0.83	0.64
zero-shot (GPT-3.5)	0.74	0.80	0.59
zero-shot (GPT-4)	0.91	0.92	0.81
LLM-Match w/o classification head (Llama-3-8B-Instruct)	0.84	0.80	0.85
LLM-Match w/ classification head (Llama-3-8B-Instruct)	0.84	0.83	**0.86**

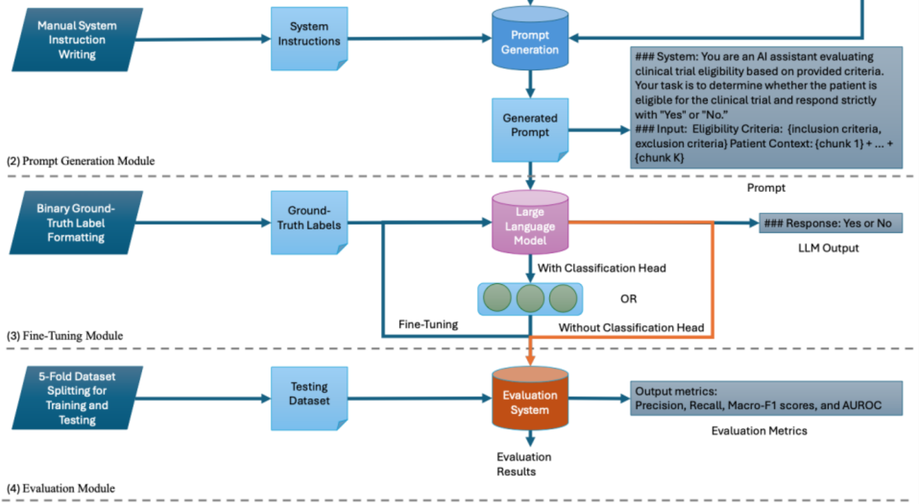
TABLE 2: Experimental Results on the SIGIR, TREC 2021, and TREC 2022 Datasets (Average).

Model	AUROC
SciFive (encoder-decoder)	0.5895
BioBERT (dual-encoder)	0.5952
PubMedBERT (dual-encoder)	0.5976
SapBERT (dual-encoder)	0.5933
BioLinkBERT (cross-encoder)	0.6176
TrialGPT-Ranking (GPT-3.5)	0.6582
TrialGPT-Ranking (GPT-4)	0.7979
LLM-Match w/o classification head (Llama-3-8B-Instruct)	0.7736
LLM-Match w/ classification head (Llama-3-8B-Instruct)	**0.8010**

TABLE 3: Experimental Results on the SIGIR, TREC 2021, and TREC 2022 Datasets (Separate).

Model	Dataset	AUROC
LLM-Match w/o classification head (Llama-3-8B-Instruct)	SIGIR	0.7659
LLM-Match w/ classification head (Llama-3-8B-Instruct)	SIGIR	0.8100
LLM-Match w/o classification head (Llama-3-8B-Instruct)	TREC 2021	0.7645
LLM-Match w/ classification head (Llama-3-8B-Instruct)	TREC 2021	0.7829
LLM-Match w/o classification head (Llama-3-8B-Instruct)	TREC 2022	0.7903
LLM-Match w/ classification head (Llama-3-8B-Instruct)	TREC 2022	**0.8102**

FIGURE 1/ Overview of LLM-Match.



DISCUSSION

LLM-Match is the first framework to combine retrieval-augmented generation (RAG) with fine-tuning for patient-trial matching. Its main innovation is the addition of a classification head, which enhances predictive precision and interpretability by providing structured, task-specific outputs. Unlike generative heads that may produce inconsistent responses, the classification head supports clearer decision boundaries, improving reliability for clinical applications. As a result, LLM-Match achieves higher Macro-F1 and AUROC scores than baseline models, showing better balance between eligible and ineligible predictions.

Using open-source models such as LLaMA 3 8B offers additional advantages in transparency, cost-efficiency, and data privacy compared to proprietary systems like GPT-4. Fine-tuning smaller models allows customization and reproducibility while maintaining strong performance. This demonstrates that optimized open-source LLMs can rival or surpass closed models, supporting broader accessibility and collaboration in clinical AI research.

However, the framework still has limitations. Concatenating EHR notes during prompt generation may introduce noise and obscure key information. The model also lacks temporal modeling and operates with limited token size. Moreover, all experiments rely on synthetic datasets, which do not fully capture the complexity and variability of real-world EHRs. Future work should validate the framework on authentic clinical data to assess its generalizability and robustness. Future improvements may involve reinforcement learning (RL) to enhance adaptability.

CONCLUSIONS

LLM-Match is the first framework to integrate retrieval-augmented generation with fine-tuning and a classification head for clinical trial patient matching. The classification head significantly improves predictive performance by enabling more precise and structured decision-making compared to generative approaches, resulting in higher Macro-F1 and AUROC scores across benchmark datasets. Built entirely on open-source models like LLaMA 3 (8B), LLM-Match demonstrates that smaller, transparent models can outperform or rival proprietary alternatives such as GPT-4.

REFERENCES

1. Jin Q, Wang Z, Floudas CS, Chen F, Gong C, Bracken-Clarke D, et al. Matching patients to clinical trials with large language models. Nature communications. 2024;15(1):9074.

2. Stubbs A, Filannino M, Soysal E, Henry S, Uzuner Ö. Cohort selection for clinical trials: n2c2 2018 shared task track 1. Journal of the American Medical Informatics Association. 2019;26(11):1163-71.

3. Koopman B, Zuccon G, editors. A test collection for matching patients to clinical trials. Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval; 2016.

4. Roberts K, Demner-Fushman D, Voorhees EM, Bedrick S, Hersh WR, editors. Overview of the TREC 2021 Clinical Trials Track. TREC; 2021.

5. Roberts K, Demner-Fushman D, Voorhees EM, Bedrick S, Hersh WR, editors. Overview of the TREC 2022 Clinical Trials Track. TREC; 2022.

6. Dubey A, Jauhri A, Pandey A, Kadian A, Al-Dahle A, Letman A, et al. The Llama 3 Herd of Models. ArXiv. 2024;abs/2407.21783.