

Real-world validation of a multimodal LLM-powered pipeline for High-Accuracy Clinical Trial Patient Matching

Corresponding Author: Mr Anatole Callies

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

This manuscript introduces a multimodal large language model (LLM)-based pipeline designed to streamline the pre-screening process for clinical trial eligibility using unprocessed electronic health record (EHR) documents. The authors leverage advanced AI technologies, including the Voyage AI Vision Language Model and OpenAI's o1 model, to address the complexities of clinical trial eligibility determination. The authors demonstrate strong performance on a multi-site, real-world dataset, achieving an impressive F1 score of 0.87. While prior efforts such as TrialGPT reported similar accuracy (87%) on specific cohorts, this work extends the evaluation to a more heterogeneous dataset, enhancing its generalizability. Notably, the pipeline enables users to review overall eligibility in under 9 minutes per patient on average, representing an 80% improvement over traditional manual chart reviews. The availability of the analysis code in Github is also a strength.

Areas for Improvement:

1. **Dataset Transparency:** The manuscript provides limited detail regarding the construction of the real-world dataset, noting only that "Following the release of our tool, we had it beta tested by 30 sites. We used their in-app feedbacks to evaluate our performance." While Table 3 offers some insight into the types of trials included, and Table 4 gives an overview of the types of criteria, the dataset description would benefit from additional categorization of the composition and complexity of the types of criteria found in the different trial types. This would significantly enhance the reproducibility and interpretability of the evaluation. Additionally, including more concrete examples of eligibility criteria as processed by the system would help clarify the types of inputs the pipeline can handle.
2. **Figure Readability:** Figures 3 and 4 are central to understanding the system's operation and the pre-screening validation tool. However, the current font size and layout make them difficult to read. I recommend increasing the font size and possibly reformatting the figures (e.g., splitting them or making them taller) to improve clarity.
3. **Clarification of Visual Data Processing:** The manuscript mentions the use of visual capabilities of LLMs to interpret medical records without lossy image-to-text conversions. However, it would be beneficial to elaborate on how the model processes visual data, as this is likely a new area for most readers.

Overall, this is a promising contribution that addresses a critical bottleneck in clinical trial workflows. With some improvements to clarity and presentation, the manuscript will be a strong addition to the literature.

Reviewer #2

(Remarks to the Author)

This paper introduces a patient-trial matching pipeline where chatgpt-4o is used to preprocess trial documents, vision LLM (instead of traditional text-based, or OCR-based approaches) is then used to embed and retrieve patient EHRs, then chatgpt-o1 is used to decide the match. The use of image for patient-trial matching seems innovation. There are several concerns.

1. As described, the pipeline uses images only because of two reasons: 1. 46% of EHRs contain visually challenging items like tables, graphs, and handwritten notes. 2. There are too many EHR vendors and it is therefore difficult to extract

structured information. Treating them all as images would make it easier to make universally compatible software. Given the development of HIT standards and the convergence of EHR vendors, the claim would require some additional justifications. For example, it is hard to think about there are handwritten notes for sites with active trials.

2. The experiment is not rigor. There seems no direct comparison between the proposed pipeline (or just the embedding/retrieval component) and existing work. So it is not easy to judge whether the adoption of VLM-based retrieval actually brings performance boost. The numbers seem pretty high (~0.9 F1 etc.), but there is no report of other methods. The reported metrics in Section 2.1 are not comparable because the datasets are different, therefore they are misleading.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

Authors have addressed all my concerns in this revision.

Reviewer #2

(Remarks to the Author)

My comments have been addressed.

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

Dear Reviewer,

Thank you for your thorough and constructive review of our manuscript, "Real-world validation of a multimodal LLM-powered pipeline for High-Accuracy Clinical Trial Patient Matching leveraging EHR data." We appreciate your feedback regarding the innovation of using image understanding for patient-trial matching. We have carefully considered all your comments. Below, we address each of your concerns and explain the modifications made.

Concern 1: Justification for leveraging images instead of traditional EHR integration.

We appreciate your request for additional justification regarding our approach of leveraging images. We provide clarifications below:

1. **Prevalence of Visually Challenging Elements:** As established by Ekowo et al. (cited in section 2.2.3 of our paper), handwritten notes remain very prevalent in clinical records. Furthermore, as we witnessed in our own real-world dataset, visually challenging elements, including images, tables, graphs, and handwritten notes, are present in more than 40% of medical records (Figure 2). These elements often contain critical information that is not possible to extract using traditional text-based approaches. To make it more concrete, we have added such examples in Appendix B.
2. **EHR Market Fragmentation and Non-EHR Sites:** Despite some convergence in the healthcare IT landscape, the EHR market remains significantly fragmented, especially at a global scale. This fragmentation makes it challenging to develop universally compatible software that relies on system APIs for data extraction. Pursuing that integration approach would place an added burden on sites, especially as many of them rely on several systems to store patient data. Our approach resolves that problem. Besides, as we have now further explained in section 2.2.4, our solution is also designed to cater to research-only sites that typically do not leverage an EHR system and often rely on Clinical Trial Management Systems (CTMS) to store medical records, primarily in PDF format. Treating these documents as images allows for a more universally compatible and robust solution, regardless of the underlying EHR vendor or the absence of an EHR system.

Concern 2: Comparison with existing work.

We thank you for your valuable feedback regarding the need for direct comparisons. We acknowledge the importance of demonstrating the performance boost achieved by our VLM-based retrieval. In response, we wish to clarify that a comparison with existing work was indeed performed on a common dataset and is detailed in the manuscript.

As explained in Section 4., we specifically used the same publicly available Harvard dataset as the Wornow et al. paper (citation #7 of the manuscript). Our results are given in Section 5.1, and the comparison is made in section 6.1.1. We show that this previous approach, using text-only non-reasoning LLMs, achieved 89% and 74% recall on the unaltered version of the eligibility criteria, while we achieved 92% precision and 93% recall. This deliberate choice ensures that our experimental results are directly comparable to theirs, allowing for a clear assessment of the performance gains achieved by our proposed pipeline.

Thank you once again for your time and thoughtful review. We look forward to your further consideration of our revised manuscript.

Sincerely,

Anatole Callies, on behalf of all co-authors

Dear Reviewer,

We would like to express our sincere gratitude for your constructive feedback on our manuscript. We appreciate your positive assessment of our work, particularly your recognition of its promising contribution to addressing a critical bottleneck in clinical trial workflows.

We have carefully considered all your suggestions and have revised the manuscript accordingly. Below, we outline how we have addressed each of your points:

1. **Dataset Transparency:** We understand your concern regarding the limited detail on dataset construction. While we are unable to share the entire dataset of criteria due to Non-Disclosure Agreements covering the trial protocols, we have taken steps to improve transparency within the given constraints. We have added concrete examples of criteria for each modality in Tables 4, 5, and 6. Besides, we would like to emphasize on the fact that the dataset is composed of complete sets of eligibility criteria taken from protocols. There was no hand-picking, and as such they are representative of the complexity encountered in real-world trials.
2. **Figure Readability:** We agree that Figures 3 and 4 are central to understanding our system's operation and the pre-screening validation tool. As per your recommendation, we have increased the font size and reformatted both figures to improve clarity and readability significantly.
3. **Clarification of Visual Data Processing:** We appreciate your suggestion to elaborate on the model's processing of visual data, recognizing that this is a new area for many readers. In response, we have added a brief explanation of how visual processing works within our pipeline in Section 3.2. More importantly, we have referenced emblematic publications that readers can consult for a more in-depth understanding of the underlying visual language model technologies.

We believe that these revisions have strengthened the manuscript, improving its clarity, presentation, and overall impact. We are confident that the updated version addresses all your valuable feedback.

Thank you once again for your time and thoughtful review. We look forward to your further consideration of our revised manuscript.

Sincerely,

Anatole Callies, on behalf of all co-authors