A Proof-of-Concept Large Language Model Application to Support Clinical Trial Screening in Surgical Oncology

Samantha M. Lai, BA

Alysala M. Malik, BA

Tejas S. Sathe, MD

Caitlin J. Silvestri, MD

Gulam A. Manji, MD, PhD

Michael D. Kluger, MD, MPH

Corresponding Author and Reprint Requests: sl5033@cumc.columbia.edu

Abstract

Introduction

Clinical trials are the gold standard for advancing the forefront of medical knowledge and rely on consistent patient accrual for success. However, patient screening for clinical trials is complex and resource intensive. There is a need to increase the scalability of trial recruitment while maintaining or improving upon the sensitivity of the current process. We hypothesized that we could use a state-of-the-art large language model, prompt engineering, and publicly available clinical trial data to predict patient eligibility for clinical trials from clinic notes. Here, we present pilot data demonstrating the accuracy of our tool in a cohort of patients being evaluated for pancreas cancer.

Methods

We used a cohort of patients who were screened for clinical trials at a single institution for this study. We developed an LLM application using LangChain and the GPT-4o model to assist in clinical trial screening. Deidentified patient data from clinical notes and trial eligibility criteria from ClinicalTrials.gov were used as inputs. For each patient and trial, the model determined inclusion or exclusion with selected eligibility criteria as well as the trial overall. Model responses were graded programmatically against a human rater standard. We recorded time elapsed and cost of running each analysis.

Results

Of the 24 patients in the test set, 19 were eligible for at least one trial. There were 43 eligible patient-trial matches in our data set. Our model correctly predicted 39 out of 42 (90.7%) of these matches. There were 105 individual eligibility criteria evaluated per patient for a total of 2520

binary criteria. GPT-4o agreed with the raters for 2438 out of 2520 (96.7%) binary eligibility criteria.

Sensitivity to overall trial eligibility by trial ranged from 87.5% to 100% for 8 out of 9 trials. The application incorrectly screened the single patient in the CA-4948 trial test set, leading to a sensitivity of 0% for this trial. Specificity ranged from 73.3% to 100% over all nine trials. The median cost for screening a patient was 0.67 USD (0.63-0.74). Median time elapsed was 137.66 seconds (130.04-146.04). Median total token usage across three assistants was 112,266.5 tokens (102982.0-122174.2).

Conclusion

Overall, our model showed high efficiency and accuracy in selecting patients for appropriate clinical trials. Our results showed promise with a small cohort and future studies are needed to assess its accuracy with a larger sample of patients and trials. This model can be applied systematically to widen the surface area covered by the screening process as well as adapted to any set of clinical trials, opening the door for expansion to other surgical and medical disciplines.

Keywords

Large language models; LLMs; clinical trials; pancreatic cancer; screening; artificial intelligence; prompting; eligibility

Introduction

Clinical trials are the gold standard for advancing the forefront of medical knowledge and rely on consistent patient accrual for success. Patient screening for clinical trials is a complex, resource intensive task with consequences ranging from a patient's access to the newest advances in medical care to the significant economic burden felt by academic medical centers suffering from low accrual.[1–3] Thus, there is room to grow both in terms of improving patient access to trials as well as reducing waste in the system at large.[1,4] Successful screening and recruitment require cross-referencing vast amounts of clinical information in patient charts against specific trial eligibility criteria. An ASCO study summarizing effort required for screening estimated that finding, screening, and enrolling a patient took up to eight hours and could cost hundreds of dollars depending on the phase of the trial.[3] At our institution, screening often occurs in an ad hoc manner, ranging from physician input at interdisciplinary tumor boards to a dedicated clinical screener splitting time between recruiting patients over the phone and searching for potentially eligible patients in the electronic medical record. Fragmented and resource intensive efforts such as these may explain in part the low rate of enrollment in cancer trials in the United States.[5] There is a need to increase the scalability of trial recruitment while maintaining or improving upon the sensitivity of the current process.

Generative artificial intelligence is an emerging technology that shows promise in meeting this need. There is evidence that large language models (LLMs) can synthesize clinical information as well as capture information from unstructured clinical materials such as notes and imaging reports.[6] Many studies have shown that LLMs can glean both the individual criteria necessary to determine eligibility and reason effectively about an overall clinical picture as it relates to eligibility.[7–11] Despite the growing mountain of evidence supporting large language models' abilities to glean information from the electronic medical record, limitations in the size of the

context window of many general use LLMs as well as the ability to send protected health information through the OpenAI application programmer interface (API) have resulted in a gap in the literature regarding the performance of large language models for clinical trials screening on raw, minimally processed clinical notes.[9] A large language model has not yet been shown to effectively screen a patient for both individual eligibility criteria and overall trial eligibility using full-length, free-text clinical notes and unstructured eligibility criteria.

We hypothesized that we could use a state-of-the-art large language model, prompt engineering, and publicly available clinical trial data to predict patient eligibility for clinical trials from clinic notes. Here, we present pilot data demonstrating the accuracy of our tool in a cohort of patients being evaluated for pancreas cancer.

Methods

Chart review

We used patients who were consented to be screened for enrollment in nine selected clinical trials at a single institution for this study. Thirty-two patients were identified for further study based on treatment at the Pancreas Center and availability of medical and surgical oncology notes. Eight charts representing eight of nine trials were set aside as training data for developing the prompts, leaving 24 patients in the test set. Characteristics of the patients in the training and test sets are available in Supplementary Table 1.[12] A medical oncology note capturing the necessary information to determine clinical trial eligibility was identified in the medical record. If available, a surgical oncology note was also identified. All identifying information was removed manually by study personnel in accordance with the Safe Harbor method for deidentification.[13] Additionally, any references to specific clinical trials by an identifiable name or by NCT number were removed.

Two-rater Standard for Eligibility Criteria Review

Eligibility criteria were retrieved from ClinicalTrials.gov for each trial using the trial NCT number and deconstructed into one or multiple binary criteria that, taken together, captured the information necessary to determine whether the patient met the ClinicalTrials.gov criterion. For example, the inclusion criterion “Histologically or cytologically confirmed metastatic pancreatic ductal adenocarcinoma” was separated into “The patient’s diagnosis is pancreatic ductal adenocarcinoma” and “The patient’s disease is metastatic.” The deidentified notes were then used by two independent raters to determine eligibility for nine clinical trials as well as 105 binary criteria derived using the method described above (Supplementary Table 1). Eligibility for inclusion in a trial was defined as meeting all binary inclusion criteria while not meeting any

exclusion criteria. Raters were trained in clinical trials screening but were not board-certified medical or surgical oncologists. All inter-rater discrepancies in answers were resolved.

Model Development

We built a LLM application using the Python LangChain library and the OpenAI Generative Pre-trained Transformer 4 Omni (GPT-4o) model with all default configuration settings (Figure 1).[14,15] The application consisted of three specialized chains, or sequences of action, each of which invoked the large language model with a different prompt. The first chain, called the eligibility chain, received two items as input: the deidentified notes from one patient as input and the eligibility criteria for all the clinical trials. The eligibility criteria were retrieved by NCT number using the ClinicalTrials.gov API. The chain was then prompted to evaluate each of the 105 binary eligibility criteria including eligibility for each of the clinical trials by providing the patient's eligibility status with respect to each criterion or trial and a brief sentence to explain each of its answers (Supplementary Table 2).

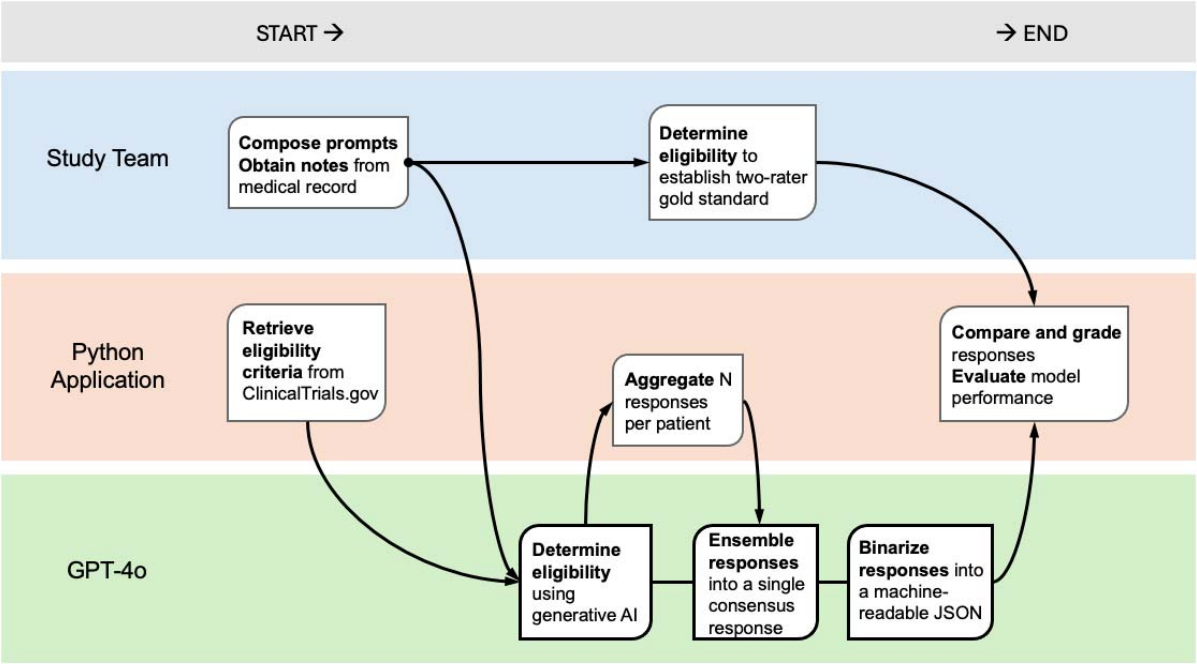
We then took steps to improve the reliability of generalized LLM output.[16] The eligibility screening process was run five times for each patient, resulting in five separate answers per patient. A second chain, called the ensembling chain, received the five answers for each binary criterion and output a consensus answer by consolidating the five answers based on the majority answer to reduce the effect of any single hallucination.

Finally, a third chain, called the binarization chain, was prompted to receive the output of the ensembling chain and output a JSON in a standardized format, specified in the prompt (Supplementary Table 2). Prompts did not contain input and output examples, relying instead on

the zero-shot learning capabilities of GPT-4o. We iteratively adjusted prompts based on mistakes made on the test set until trial eligibility performance on the test set reached a plateau.

Figure 1. Depiction of data flow, inputs, and outputs

The workflow depicts the flow of the data from preprocessing of clinical notes by study personnel and automated retrieval of eligibility criteria from the ClinicalTrials.gov API. GPT-4o denotes Generative Pre-trained Transformer 4 Omni; AI, artificial intelligence; JSON, JavaScript object notation.



Grading

The JSON files outputted by the LLM application were graded programmatically by a simple Python grading function that compared the model output to the two-rater gold standard. Sensitivity and specificity were calculated for each criterion and overall trial eligibility. A true positive was defined as identifying an individual to be eligible for a trial in alignment with a two-rater gold standard. A true negative was defined as identifying an individual to be not eligible in

alignment with a two-rater gold standard. Discrepancies between the model and the inter-rater gold standard (termed “errors”) were categorized by the raters. We identified several types of errors that resulted in a discordant outcome between the LLM tool and the human raters. A human error was defined as a discrepancy where the information needed to arrive at the correct answer was in the clinical note and extracted by GPT-4o, but incorrectly annotated by the raters. “Prompt Unclear” and “Note Unclear” were assigned to discrepancies where the information required to answer the criterion was deemed to be ambiguous by the raters. A “Clinical Judgment” error was invoked in cases where GPT-4o differed from the raters in appraisal of a piece of information in the note, such as the severity of a condition. A “Misinterpretation” error was defined as an instance where GPT-4o did not seem capable of extracting the relevant information from the note. A “Hallucination” was defined as an error where the answer provided by GPT-4o actively contradicted information stated in the note.

Results

Performance on Trial Eligibility and Individual Criteria

Of the 24 patients in the test set, 19 were eligible for at least one trial. There were 43 eligible patient-trial matches in our data set. Our model correctly predicted 39 out of 42 (90.7%) of these matches. There were 105 individual eligibility criteria evaluated per patient for a total of 2520 binary criteria. GPT-4o agreed with the raters for 2438 out of 2520 (96.7%) binary eligibility criteria.

Sensitivity to overall trial eligibility by trial ranged from 87.5% to 100% for 8 out of 9 trials. The application incorrectly screened the single patient in the CA-4948 trial test set, leading to a sensitivity of 0% for this trial. Specificity ranged from 73.3% to 100% over all nine trials (Figure 2). Performance of all individual criteria are included in Supplementary Table 3.

Table 1. Selected Individual Eligibility Criteria

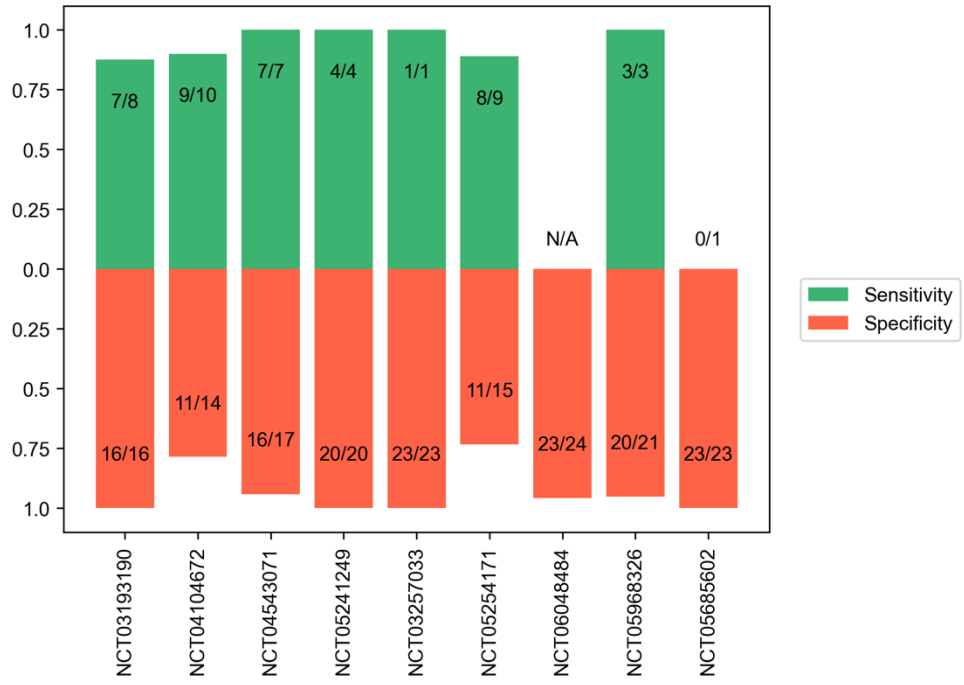
This data represents a single experiment run on September 9, 2024. Each row represents sensitivity and specificity for GPT-4o determination of individual criterion status. The number of patients included in each calculation is indicated.

Criterion	**Sensitivity (TP for exclusion)**	**Sensitivity Calculation**	**Specificity (TN for inclusion)**	**Specificity Calculation**
The patient's age is less than 18	0.0	N/A	1.000000	24/24
The patient's pancreatic cancer is resectable	1.0	5/5	1.000000	19/19
The patient's pancreatic cancer is metastatic	1.0	15/15	1.000000	9/9

The patient has had disease progression on 5-FU for metastatic or unresectable pancreatic ductal adenocarcinoma.	0.0	0/1	1.000000	23/23
ECOG performance status (PS) is greater than 2.	0.5	1/2	1.000000	22/22
The patient is taking metformin.	0.2	1/5	0.947368	18/19
The patient's total bilirubin is >1.5x upper limit of normal or >3x upper limit or normal in patients with Gilbert disease	0.5	1/2	0.954545	21/22
The patient has active or symptomatic Hepatitis B	0.0	0/1	1.000000	23/23

Figure 2. Sensitivity and Specificity for Trial Eligibility

This data represents a single experiment run on September 9, 2024. Each pair of green and orange bars represents sensitivity and specificity for GPT-4o determination of trial eligibility across nine trials for pancreatic cancer, with each trial denoted by the National Clinical Trial (NCT) number on the X axis. The number of patients included in each calculation and odds ratios are indicated as N(predicted)/N(total).



Cost and Time Elapsed

The median cost for screening a patient was 0.67 USD (0.63-0.74). Median time elapsed was 137.66 seconds (130.04-146.04). Median total token usage across three assistants was 112,266.5 tokens (102982.0-122174.2), with most tokens used in the eligibility determination stage compared to the ensembling and binarization stages. The median token usage in the most expensive part of the chain was 96571.5 (87497.5-106799.0), well within GPT-4o's context window of 120,000 tokens.

Error Analysis

Out of 2520 variables evaluated in total across 24 patients, 82 discrepancies were found upon programmatic grading of the application's performance against the inter-rater gold standard. The most common source of discrepancy was human error in chart review. The other most common categories of discrepancies were GPT-4o hallucination, GPT-4o misinterpretation, and disagreements in clinical judgement. Examples of each error category are provided in Table 2

and a breakdown of the variables in each error category is presented in Figure 3. Thirteen out of 82 discrepancies were in overall trial eligibility determination, leaving 69 errors in determination of individual criteria. All 26 human errors categorized were entirely in determination of individual criteria, making up 37.7% of individual criterion errors. Hallucination and misinterpretation together accounted for 26 criterion errors, or 37.7% of the total criterion errors. These two categories made up 11 out of 13 (84.6%) of the overall trial eligibility determination errors.

Table 2. Discrepancies between GPT-4o and Human determinations of individual eligibility criteria.

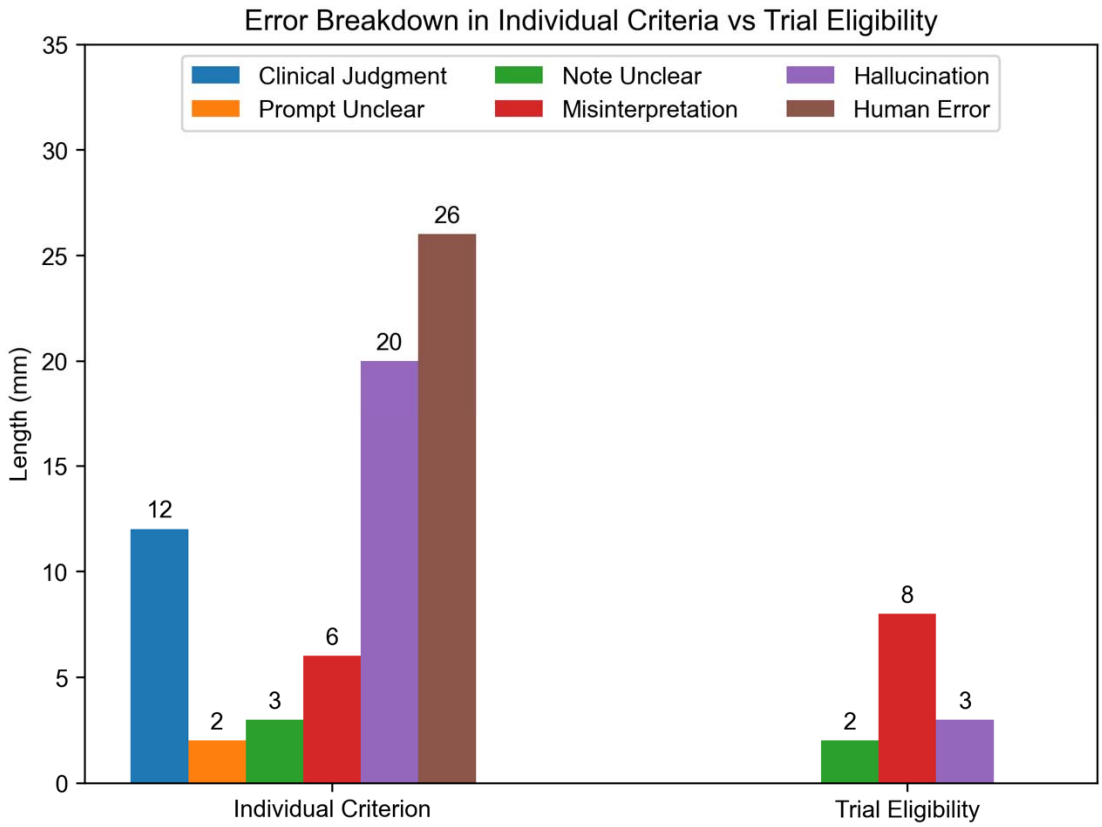
A total of 82 discrepancies were categorized after programmatic grading. The number and percentage of the total number of errors is represented. A true positive is denoted when the Criterion statement is true and the rater or model answers as such. GPT-4o denotes Generative Pre-trained Transformer 4 Omni; GI, gastrointestinal; 5-FU, 5-fluorouracil.

Category	**Criterion**	**Correctness of GPT-4o and Human Answers**	**Explanation**	**N (%)**
Prompt Unclear	The patient has a diagnosis of pancreatic adenocarcinoma.	GPT-4o False Positive, Human True Negative	GPT-4o states there is a diagnosis of "pancreatic adenocarcinoma", but it is signet cell based, not pancreatic ductal adenocarcinoma. This was an error of prompt engineering.	2 (2.4)
Note Unclear	The patient has a history of other malignancy within 2 years that does not include malignancies with a low rate of metastatic spread.	GPT-4o False Positive, Human True Negative	Patient had renal cell carcinoma, but it was resected in 2015. The model had no way of knowing the date of the note. By study design, all patients were all seen within the last year, so we had information the model was unable to	5 (6.1)

			access in the deidentified note.	
Clinical Judgment	The patient has a history of GI condition which could impair absorption or ability to ingest study drug (i.e. evidence of GI obstruction).	GPT-4o False Positive, Human True Negative	Patient is status post duodenal stent placement. GPT-4o did not connect duodenal stent placement mentioned in the note to the small bowel obstruction.	12 (14.6)
Misinterpretation	The patient has had disease progression on 5-FU for metastatic or unresectable pancreatic ductal adenocarcinoma.	GPT-4o False Negative, Human True Positive	GPT-4o was unable to identify that the patient received 5-FU for metastatic disease and experienced disease progression despite the treatment.	14 (17.0)
	The patient is eligible for trial NCT04543071.	GPT-4o False Positive, Human True Negative	GPT-4o screened the patient into this trial. However, the patient must be naïve to treatment for metastatic disease for this trial.	
Hallucination	The patient's hemoglobin is less than 9.0 g/dL.	GPT-4o False Positive, Human True Negative	The note states in the laboratory values section that the hemoglobin is 10.2.	23 (28.0)
Human Error	The patient has had prior radiotherapy for pancreatic cancer.	GPT-4o True Positive, Human False Negative	The note states that the patient had radiotherapy, and this was missed by the raters.	26 (31.7)

Figure 3. Error Breakdown

Errors were categorized into six error types and separated into errors in determining individual eligibility criteria and errors in determining overall trial eligibility.


Error Breakdown in Individual Criteria vs Trial Eligibility
Clinical Judgment
Prompt Unclear
Note Unclear
Misinterpretation
Hallucination
Human Error
Length (mm)
35
30
25
20
15
10
5
0
12
2
3
6
20
26
2
8
3
Individual Criterion
Trial Eligibility

Discussion

The application of large language models to the problem of clinical trials screening has been widely discussed. Findings have shown promise in isolation in extraction of individual eligibility criteria, understanding preprocessed clinical summaries and case reports, and evaluation of structured eligibility data.[7–11] Limitations in context window length and maintenance of patient privacy are largely responsible for the lack of studies using data more representative of the electronic medical record. However, bridging this gap is important for evaluating the performance of LLMs in scenarios resembling the electronic medical record. To this end, our application fills this niche and shows high sensitivity and specificity in selecting patients for appropriate clinical trials in a fraction of the time and at fraction of the cost of existing screening mechanisms.

One strength of this study stems from our ability to use imperfect and unstructured data as input into an algorithmic decision. Our findings highlight that the imperfection of free-text clinical notes only slightly hinders the performance of a general-use model when compared to the performance in other studies on more preprocessed data.[7–11] Another strength was the simultaneous evaluation of GPT-4o's ability to extract individual eligibility criteria as well as overall trial eligibility, providing a proxy for explainability alongside insight into the strengths and limitations of the technology. This data may be useful to surgeons in finding alternate trials for which patients are eligible or reconsidering eligibility if any of the ”missed” criteria represent clinical conditions that may resolve. GPT-4o's near perfect performance on criteria such as age and resectability of disease suggested that it could reliably glean these details from a free-text clinical note. This suggests that LLM based tools may also be useful in determining resectability, which can also guide operative and clinical management.

On the other hand, the model generated some errors which preclude immediate widespread use in the clinical setting. For example, one patient was incorrectly deemed eligible for four trials with an eligibility criterion requiring that patients be naïve to treatment for metastatic disease. Sure enough, the individual criterion assessing whether the patient had experienced progression of metastatic disease on chemotherapy was also incorrect, suggesting that the model had also struggled to pick up on a piece of crucial information needed to score correctly on overall eligibility. We also found that the model hallucinated in ways that could not be explained. GPT-4o's sensitivity for detecting metformin on a patient's medication list was surprisingly low. Out of five patients taking metformin, GPT-4o was only able to detect one, in contrast to its perfect record on Hepatitis B status or its near-perfect performance on tumor staging, both which appear in a much less standardized form than metformin in typical oncology notes.

Several limitations were encountered in this study. First, the pool of patients enrolled in trials at a single center for pancreatic cancer was relatively small. While representative of previously enrolled patients at our center, the cohort may not be representative of the true population of patients with pancreatic cancer, possibly limiting the generalizability of our results. Future directions will invariably involve validation in larger datasets as well as prospective comparison in resource use, time, and accuracy to current screening processes. By design, the trial-agnostic structure of our application is a convenient foundation on which to expand our dataset to more clinical trials to achieve maximal generalizability of the tool. A limitation inherent to this field is the black box nature of generative AI obscuring explainability, which we attempted to mitigate with chain of thought prompting to shed light on the information that the model could extract in isolation alongside demonstration of higher-level reasoning.

While we did not evaluate GPT-4o's performance against human performance in a high-fidelity simulation of the screening process, our findings suggest that its performance on individual criteria can rival that of non-expert screeners and provide a complementary set of strengths in support of a largely human process. Additionally, the simplicity of our application suggests that the tool could be connected directly to an electronic health record, bringing aspirations toward systematic clinical trial screening even closer into view. While efforts to train specific models in search of a performance boost are needed in this relatively nascent field, it is reassuring to know that our results concur with findings that state that general use models can display impressive performance "out of the box" without specialization beyond application-specific prompting.[16] Future work exploring validation on different models, both open-source and commercially available, would be easily implemented given the application's model-agnostic nature and would explore generalizability, increase feasibility for integration into existing electronic health record systems, and elucidate the upper limit of performance on this important clinical problem.

Conclusion

Our findings highlight that the imperfection of free-text clinical notes only slightly hinders the performance of a general-use model when compared to previous performance on more preprocessed data. Use of this tool in clinical trials screening could complement human existing screening efforts. This study represents the frontier of pitting of emerging large language model technology against the historically unruly terrain of the electronic medical record.

References

1. Kitterman DR, Cheng SK, Dilts DM, Orwoll ES. The Prevalence and Economic Impact of Low-Enrolling Clinical Studies at an Academic Medical Center. *Acad Med J Assoc Am Med Coll*. 2011;86(11):1360-1366. doi:10.1097/ACM.0b013e3182306440

2. Woo M. An AI boost for clinical trials. *Nature*. 2019;573(7775):S100-S102. doi:10.1038/d41586-019-02871-3

3. Penberthy LT, Dahman BA, Petkov VI, DeShazo JP. Effort Required in Eligibility Screening for Clinical Trials. *J Oncol Pract*. 2012;8(6):365-370. doi:10.1200/JOP.2012.000646

4. Sateren WB, Trimble EL, Abrams J, et al. How Sociodemographics, Presence of Oncology Specialists, and Hospital Cancer Programs Affect Accrual to Cancer Treatment Trials. *J Clin Oncol*. Published online September 21, 2016. doi:10.1200/JCO.2002.08.056

5. Unger JM, Shulman LN, Facktor MA, Nelson H, Fleury ME. National Estimates of the Participation of Patients With Cancer in Clinical Research Studies Based on Commission on Cancer Accreditation Data. *J Clin Oncol*. Published online April 2, 2024. doi:10.1200/JCO.23.01030

6. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature*. 2023;620(7972):172-180. doi:10.1038/s41586-023-06291-2

7. Jin Q, Wang Z, Floudas CS, et al. Matching Patients to Clinical Trials with Large Language Models. Published online 2023. doi:10.48550/ARXIV.2307.15051

8. Hamer DM den, Schoor P, Polak TB, Kapitan D. Improving Patient Pre-screening for Clinical Trials: Assisting Physicians with Large Language Models. Published online June 29, 2023. doi:10.48550/arXiv.2304.07396

9. Wong C, Zhang S, Gu Y, et al. Scaling Clinical Trial Matching Using Large Language Models: A Case Study in Oncology. In: *Proceedings of the 8th Machine Learning for Healthcare*

Conference. PMLR; 2023:846-862. Accessed September 16, 2024. https://proceedings.mlr.press/v219/wong23a.html

10. Beattie J, Neufeld S, Yang D, et al. Utilizing Large Language Models for Enhanced Clinical Trial Matching: A Study on Automation in Patient Screening. *Cureus*. 16(5):e60044. doi:10.7759/cureus.60044

11. Unlu O, Shin J, Mailly CJ, et al. Retrieval-Augmented Generation–Enabled GPT-4 for Clinical Trial Screening. *NEJM AI*. 2024;1(7). doi:10.1056/AIoa2400181

12. Norgeot B, Quer G, Beaulieu-Jones BK, et al. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. *Nat Med*. 2020;26(9):1320-1324. doi:10.1038/s41591-020-1041-y

13. Rights (OCR) O for C. Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. September 7, 2012. Accessed September 18, 2024. https://www.hhs.gov/hipaa/for-professionals/special-topics/de-identification/index.html

14. Hello GPT-4o. Accessed September 18, 2024. https://openai.com/index/hello-gpt-4o/

15. Introduction | □□□ LangChain. Accessed September 18, 2024. https://python.langchain.com/docs/introduction/

16. Nori H, Lee YT, Zhang S, et al. Can Generalist Foundation Models Outcompete Special-Purpose Tuning? Case Study in Medicine. Published online 2023. doi:10.48550/ARXIV.2311.16452