EN.705.651.8VL.SU24 - Large Language Models: Theory and Practice
Group Final Project

Large Language Models for Matching Patients to Clinical Trials

Group Members

Nainika Devireddy, Jieming Chu, and Amin Diab

Background and Context

Clinical trials are evaluation studies that examine the effectiveness of medical interventions on human health outcomes. It is the primary method through which researchers determine the efficacy and safety of new treatments, drugs, medical devices, and diagnostic tools.

The patient recruitment process for clinical trials is a critical phase that significantly impacts the success of a study. Recruiting the right participants ensures that the study results are valid and applicable to the target population. Despite the critical nature of patient recruitment, many clinical trials face difficulties in enrolling a sufficient number of eligible participants. This challenge often leads to delays, increased costs, and sometimes failure to complete the study.

Each clinical trial has a set of eligibility criteria that specify the characteristics required for participation. These criteria can include factors such as age, gender, disease type and stage, previous treatment history, overall health status, and specific genetic markers. The criteria are designed to ensure that the study results will be relevant to the target population and that participants' safety is maximized in a controlled manner.

Identifying potential participants begins with a thorough review of medical records and databases to find individuals who meet the trial's eligibility criteria. In some cases, healthcare providers may also refer patients who they believe could benefit from or be interested in participating in a clinical trial. Health care professionals, or patients, search for federally and privately supported clinical trials through the National Institutes of Health's (NIH) ClinicalTrials.gov, a searchable registry and results database of trials conducted in the United States and around the world. ClinicalTrials.gov provides information on the participation eligibility criteria, trial purpose, conditions to be treated, and treatments to be evaluated. Currently, to verify participant eligibility, the searcher must examine each search result manually. Hence, it is a time-consuming and labor-intensive task to identify the exact studies a patient may be eligible for. To date, there are no readily available automated processes to match patients to clinical trials based on their eligibility criteria. To address these issues, innovative solutions are needed to streamline the recruitment process and enhance the matching of patients to suitable clinical trials.

Open AI's GPT-4o is the fifth generation of their generative pre-trained transformer models (GPT), known for its ability to understand and generate human-like text, as well as enhanced ability to interact with human users using voice input and output (OpenAI et al., 2023 and OpenAI. 2024). It represents a significant advancement over its predecessors, with improvements in various aspects such as performance, accuracy, and versatility.

Retrieval-Augmented-Generation (RAG) is a commonly used method designed to enhance the performance of large language models by incorporating contextual information from external data sources. This approach is particularly effective in addressing the challenge of out-of-date or incomplete information within the model's knowledge base. RAG operates by integrating an information retrieval component with a text generation component, creating a robust system capable of handling knowledge-intensive and domain-specific tasks.

The process begins with the information retrieval component, which searches for and retrieves relevant documents or data from external sources based on the input query. Once the relevant information is retrieved, it is passed along with the query to the text generation component to formulate a detailed query as input for the large language model. The large language model then generates a response that is informed by both the initial query and the supplementary information, often resulting in a more accurate and contextually appropriate output.

Embeddings play a crucial role in the RAG framework by facilitating the retrieval and generation processes. Embeddings are dense vector representations of text that capture the semantic meaning of words, phrases, or documents. In the retrieval phase, embeddings are used to represent both the query and the documents in a high-dimensional space, allowing for efficient and accurate matching of relevant information. By computing the similarity between the query embedding and the document embeddings, the system can identify the most relevant documents to supplement the language model's knowledge. This integration ensures that the generated output is not only informed by the model's pre-existing knowledge but also enriched with the latest information from the retrieved documents.

By leveraging the robustness of large language models (LLMs) and the informational context introduced by RAG, we propose using a RAG-enabled LLM architecture to optimize the identification of clinical trials a patient may be eligible for. This architecture offers the benefit of automating and expediting the patient recruitment process and ensuring that clinical trials can be conducted more effectively, robustly, and successfully.

Aim

The primary goal of this project is to evaluate different LLM approaches to identify clinical trials for which a patient is eligible for, including a RAG-enabled LLM approach. The project aims to evaluate the use of prompt-engineering and compare the use of different query methods on the task.

Data Sources

Clinical trial records were obtained from ClinicalTrials.gov using their REST API as a comma-separated-values (.csv) file. Each clinical study is uniquely identified with a National Clinical Trial Identification Number (“NCT Number”). Within each clinical study, the eligibility criteria are provided as markup text, with demarcated ‘Inclusion Criteria’ and ‘Exclusion Criteria’.

Patient records/profiles were obtained from Synthea, an open-source software used to create HIPAA-compliant synthetic patient records. Synthea outputs artificial, realistic (but not real) patient data and health records in varied formats. This project utilizes a synthetic dataset produced using Synthea on March 6, 2020, consisting of approximately 200 lifetime/longitudinal patient records. The dataset comprises 179 female breast cancer patients, 14 male breast cancer patients, and 6 patients with assorted other cancers (lung, colorectal, prostate). The records follow the Minimal Common Oncology Data Elements (mCODE) STU1 (release 1) format. mCODE is an initiative by American Society Clinical Oncology (ASCO) and the MITRE Corporation to develop a core set of structured data elements for oncology EHRs, aiming to capture research-quality data from all cancer patients to facilitate comparative effectiveness analysis and data exchange between health systems.

Safety and Ethical Considerations

Patient information is protected under the Health Insurance Portability and Accountability Act of 1996 ("HIPAA"). This research project does not use real patient data but instead uses synthetic and fictional patient data generated from Synthea, an open-source software used to create artificial patient records. This ensures that the research project will not include any HIPAA-protected information.

Methods

Pre Processing

The pre-processing phase involved preparing both clinical trial data and patient records to ensure compatibility with the large language model. For the sake of a focused evaluation approach, only clinical trials pertaining to breast cancer were used for analysis. To retrieve up-to-date clinical trial data, the ClinicalTrials.gov REST API was used to retrieve trials that match the query 'Breast Cancer'. From each study, the following attributes were extracted: NCT ID, overall progress status, start date, conditions, interventions, locations, primary completion date, date the study was posted, date the last update was made, study type, current trial phase and the eligibility criteria. The dataset is filtered to only accommodate clinical trials that are actively recruiting participants.

Stratified sampling is used to select the patients for evaluation across the three categories of patients found in the Synthea data: female, male and patients with assorted other cancers. Of the 200 patients provided by Synthea, 10 patients were selected, including eight female breast cancer patients, one male breast cancer patient and one patient with assorted other cancers (lung, colorectal, prostate).

Each patient file is presented by Synthea in a json file with nested health information. Such a structure can be challenging for an LLM to parse and decipher easily within a query. For this reason, the json file is rearranged into a patient dictionary of DataFrames based on the key:value pairs found in the file. mCODE data formats often contain links to reference pages describing the data fields found in the json file. These links were removed from the patient DataFrame for a more readable structure. It was found that this removal of hyperlinks did not negatively impact GPT-4o's ability to understand the data fields of the mCODE file.

Approaches to Querying the LLM to Match Patients to Clinical Trials

In the scope of the project, four distinct approaches were used to query the large language model, GPT-4o, in order to match patients to clinical trials based on eligibility criteria. The four approaches offer differences in input information formats, workflow architecture and query prompting. The relative performance of the LLM using each approach is described in the Results section of the paper.

Approach 1: Incorporating Full Patient Data and Clinical Trial Data in the Query

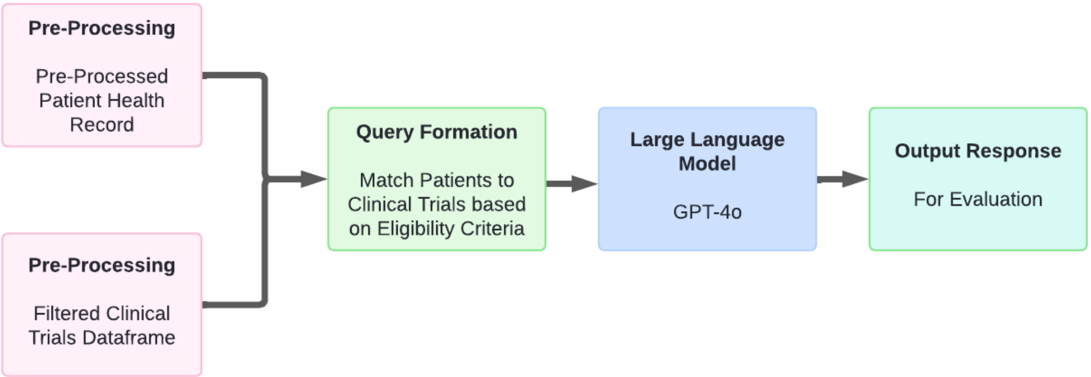
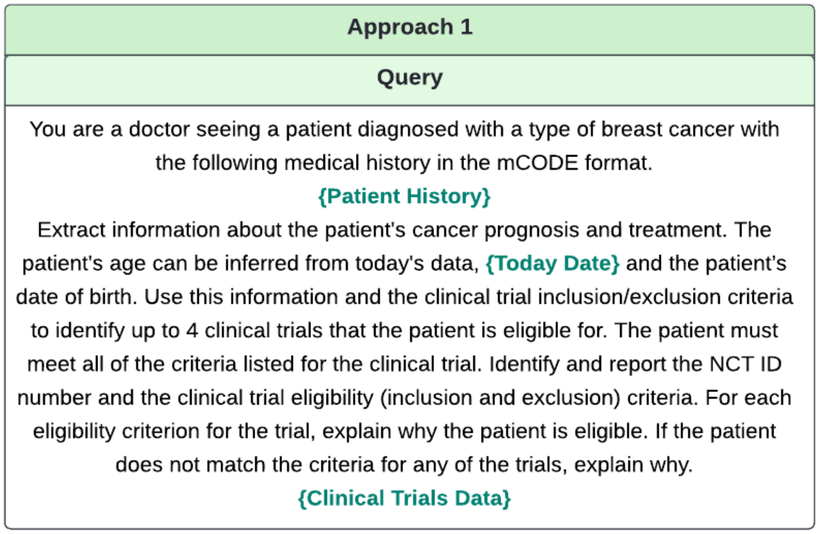

Pre-Processing
Pre-Processed Patient Health Record
Pre-Processing
Filtered Clinical Trials Dataframe
Query Formation
Match Patients to Clinical Trials based on Eligibility Criteria
Large Language Model
GPT-4o
Output Response
For Evaluation

Figure 1: *Workflow for Approach 1: Incorporating a full patient health record and complete clinical trial data in an LLM query to match the patient to eligible clinical trials*

The primary goal of Approach 1 (Figure 1) is to utilize the extensive context window of GPT-4o (128,000 tokens) by incorporating the complete pre-processed mCODE patient health record and entire clinical trial data directly into the prompt. This approach aims to maximize the relevant information available to the model, enabling more informed and accurate patient-trial matching decisions.

The prompt (Figure 2) includes the two detailed DataFrames within the input body of query text provided to the LLM. A detailed query was constructed, prompting the model to extract information about the patient's cancer prognosis and treatment from the mCODE-formatted medical history. Subsequently, the model was instructed to use the extracted information to identify up to four suitable clinical trials for the patient. The model was required to detail the NCT ID and eligibility criteria of each study provided in the response and explain why or why not the patient met the criteria for each trial.



Approach 1

Query

You are a doctor seeing a patient diagnosed with a type of breast cancer with the following medical history in the mCODE format.

{Patient History}

Extract information about the patient's cancer prognosis and treatment. The patient's age can be inferred from today's data, **{Today Date}** and the patient's date of birth. Use this information and the clinical trial inclusion/exclusion criteria to identify up to 4 clinical trials that the patient is eligible for. The patient must meet all of the criteria listed for the clinical trial. Identify and report the NCT ID number and the clinical trial eligibility (inclusion and exclusion) criteria. For each eligibility criterion for the trial, explain why the patient is eligible. If the patient does not match the criteria for any of the trials, explain why.

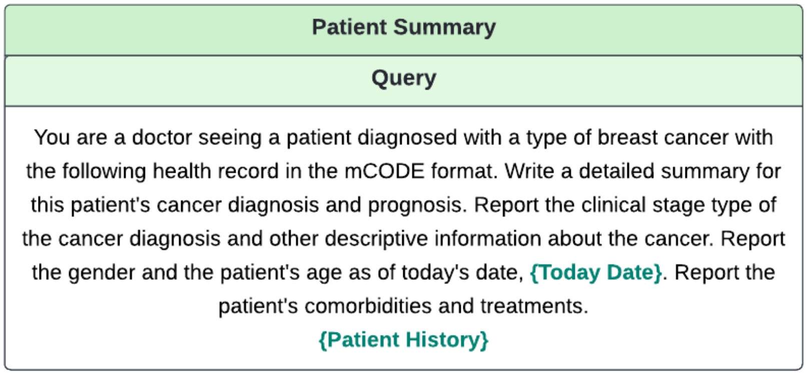
{Clinical Trials Data}

Figure 2: *LLM Query for Approach 1*

The following approaches (2, 3 and 4) each incorporate the concept of providing the LLM with a summary of the patient information, as opposed to the complete pre-processed patient health record in the query. The intermediate step below describes how the summary of the patient record is generated using GPT-4o.

Prompting the LLM to Summarize the Patient Health Record

In order to consolidate the key information from the patient health record, the LLM is tasked with summarizing key information relevant to cancer diagnosis and prognosis, including comorbidities and treatments involved (Figure 3). This intermediate step mitigates the need to input irrelevant patient information found in the patient health record into the query and provide focused patient details to improve the performance of the model.



Patient Summary

Query

You are a doctor seeing a patient diagnosed with a type of breast cancer with the following health record in the mCODE format. Write a detailed summary for this patient's cancer diagnosis and prognosis. Report the clinical stage type of the cancer diagnosis and other descriptive information about the cancer. Report the gender and the patient's age as of today's date, **{Today Date}**. Report the patient's comorbidities and treatments.
{Patient History}

Figure 3: *LLM Query for Generating a Patient Summary*

Approach 2: Incorporating Patient Summary and Clinical Trial Data in the Query

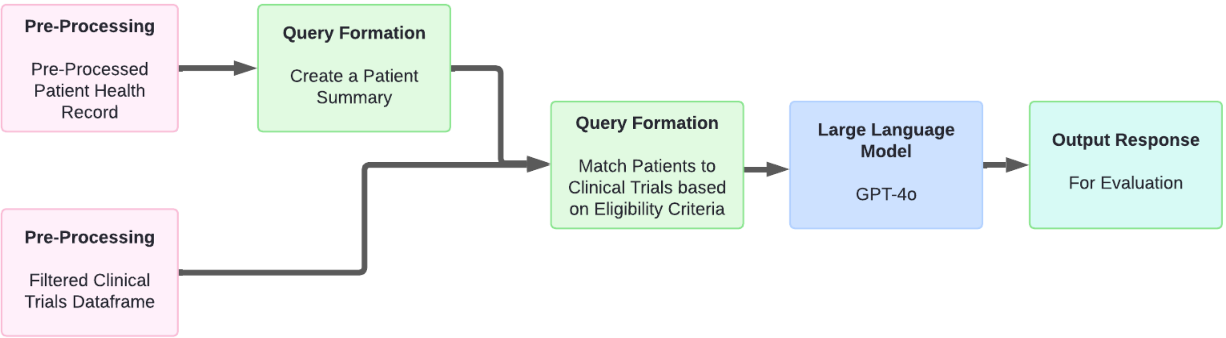
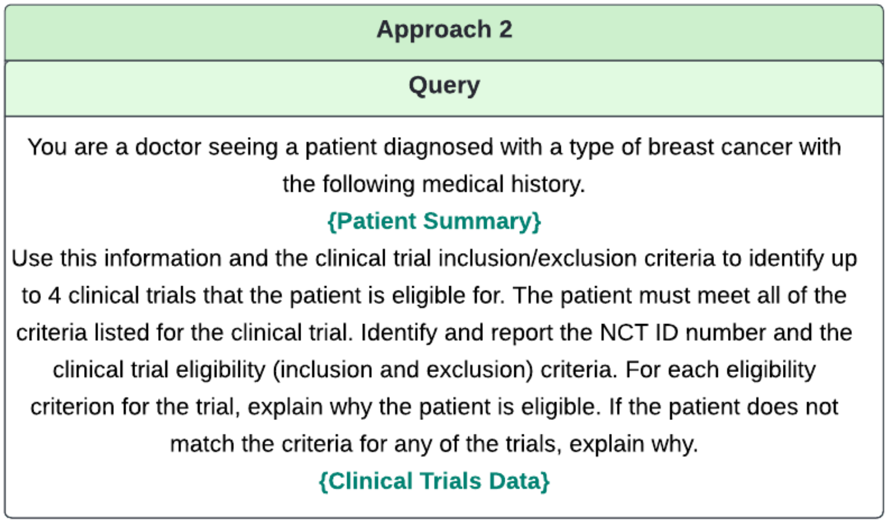


Figure 4: *Workflow for Approach 2: Incorporating a patient summary and the complete clinical trial data in an LLM query to match the patient to eligible clinical trials*

Approach 2 (Figure 4) simplifies the input query by incorporating an LLM-generated summary of the patient health record along with the entire clinical trial data into the prompt given to the model (Figure 5). By only providing key patient information, this approach aims to streamline the data provided to the LLM, making it more manageable for the model while still enabling accurate patient-trial matching decisions.

The model was instructed to use this summarized information to identify up to four suitable clinical trials for the patient. Similar to Approach 1, the model was required to detail the NCT ID and eligibility criteria of each study provided in the response and explain why or why not the patient met the criteria for each trial.



Approach 2

Query

You are a doctor seeing a patient diagnosed with a type of breast cancer with the following medical history.

{Patient Summary}

Use this information and the clinical trial inclusion/exclusion criteria to identify up to 4 clinical trials that the patient is eligible for. The patient must meet all of the criteria listed for the clinical trial. Identify and report the NCT ID number and the clinical trial eligibility (inclusion and exclusion) criteria. For each eligibility criterion for the trial, explain why the patient is eligible. If the patient does not match the criteria for any of the trials, explain why.

{Clinical Trials Data}

Figure 5: *LLM Query for Approach 2*

Approach 3: Incorporating Patient Summary and Clinical Trial Data in a RAG-Enabled LLM Query

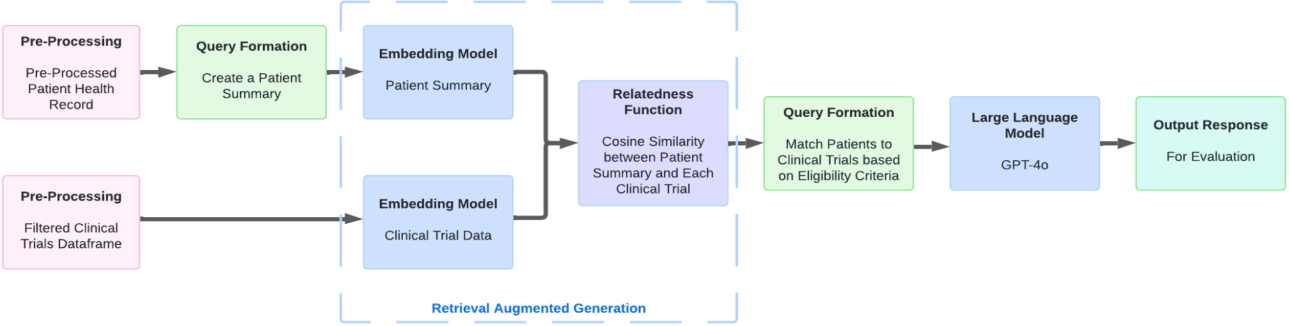
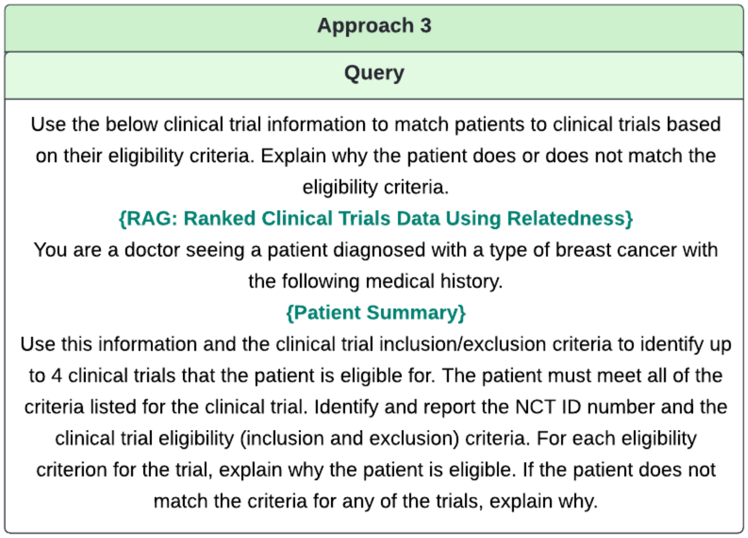

Pre-Processing
Pre-Processed Patient Health Record
Query Formation
Create a Patient Summary
Embedding Model
Patient Summary
Pre-Processing
Filtered Clinical Trials Dataframe
Embedding Model
Clinical Trial Data
Relatedness Function
Cosine Similarity between Patient Summary and Each Clinical Trial
Retrieval Augmented Generation
Query Formation
Match Patients to Clinical Trials based on Eligibility Criteria
Large Language Model
GPT-4o
Output Response
For Evaluation

Figure 6: *Workflow for Approach 3: Incorporating a patient summary and ranked clinical trial data in an RAG-enabled LLM query to match the patient to eligible clinical trials*

Approach 3 (Figure 6) utilizes Retrieval-Augmented Generation to enhance the model's access to relevant clinical trial information. This approach aims to refine the input data by computing word embeddings for the eligibility criteria of each clinical trial and the patient summary, allowing for precise retrieval of the most relevant trials.

RAG computes a word embedding for the text describing the eligibility criteria of each clinical trial, as well as for the queried patient summary. Using the cosine similarity function on these word embeddings, RAG retrieves clinical trials that are most closely related to the patient summary. The most relevant clinical trials are then provided to the model in a prompt (Figure 7) to refine and enhance the information available to the model, enabling more accurate patient-trial matching decisions.



Approach 3

Query

Use the below clinical trial information to match patients to clinical trials based on their eligibility criteria. Explain why the patient does or does not match the eligibility criteria.

{RAG: Ranked Clinical Trials Data Using Relatedness}

You are a doctor seeing a patient diagnosed with a type of breast cancer with the following medical history.

{Patient Summary}

Use this information and the clinical trial inclusion/exclusion criteria to identify up to 4 clinical trials that the patient is eligible for. The patient must meet all of the criteria listed for the clinical trial. Identify and report the NCT ID number and the clinical trial eligibility (inclusion and exclusion) criteria. For each eligibility criterion for the trial, explain why the patient is eligible. If the patient does not match the criteria for any of the trials, explain why.

Figure 7: *LLM Query for Approach 3*

Approach 4: Incorporating Patient Summary and Clinical Trial Data in a RAG-Enabled LLM Query with a Modified Similarity Metric

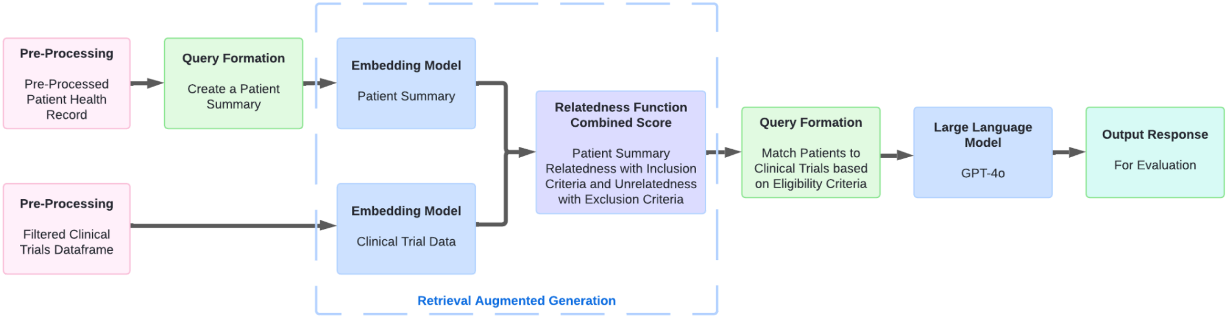
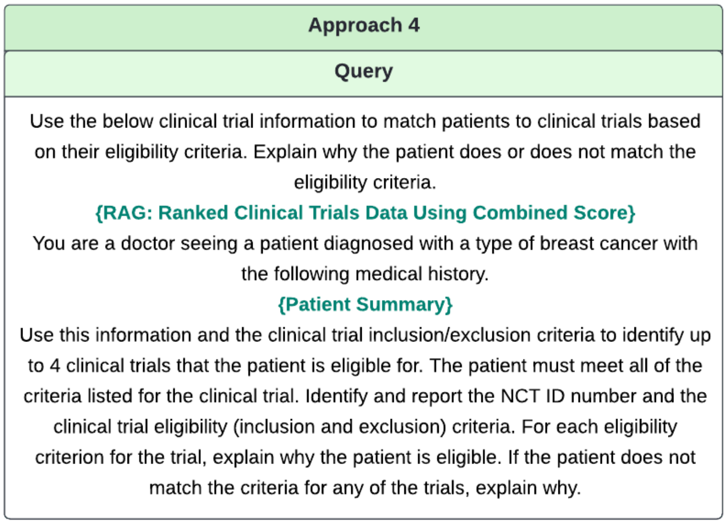

Pre-Processing
Pre-Processed Patient Health Record
Query Formation
Create a Patient Summary
Embedding Model
Patient Summary
Pre-Processing
Filtered Clinical Trials Dataframe
Embedding Model
Clinical Trial Data
Relatedness Function Combined Score
Patient Summary Relatedness with Inclusion Criteria and Unrelatedness with Exclusion Criteria
Retrieval Augmented Generation
Query Formation
Match Patients to Clinical Trials based on Eligibility Criteria
Large Language Model
GPT-4o
Output Response
For Evaluation

Figure 8: *Workflow for Approach 4: Incorporating a patient summary and ranked clinical trial data using a modified similarity metric in an RAG-enabled LLM query to match the patient to eligible clinical trials*

Approach 4 (Figure 8) refines the similarity metric used in Approach 3 in an attempt for more accurate patient-trial matching. While Approach 3 utilized a single similarity metric between the eligibility criteria and the patient summary, Approach 4 adopts a modified similarity metric that differentiates between the inclusion and exclusion criteria for each clinical trial.

Approach 4 computes separate word embeddings for the inclusion criteria text and the exclusion criteria text. A combined relatedness score is then calculated by assessing the relatedness between the inclusion criteria and the patient summary, as well as the unrelatedness between the exclusion criteria and the patient summary. This combined related score is used to rank clinical trials for relevance to the patient. The most relevant clinical trials are provided to the model in a prompt (Figure 9) to refine and enhance the information available to the model. This approach aims to ensure that the inclusion criteria are relevant to the patient, while the exclusion criteria are unrelated, thereby improving the precision of patient-trial matching decisions.



Approach 4

Query

Use the below clinical trial information to match patients to clinical trials based on their eligibility criteria. Explain why the patient does or does not match the eligibility criteria.

{RAG: Ranked Clinical Trials Data Using Combined Score}

You are a doctor seeing a patient diagnosed with a type of breast cancer with the following medical history.

{Patient Summary}

Use this information and the clinical trial inclusion/exclusion criteria to identify up to 4 clinical trials that the patient is eligible for. The patient must meet all of the criteria listed for the clinical trial. Identify and report the NCT ID number and the clinical trial eligibility (inclusion and exclusion) criteria. For each eligibility criterion for the trial, explain why the patient is eligible. If the patient does not match the criteria for any of the trials, explain why.

Figure 9: *LLM Query for Approach 4*

Evaluation Methods

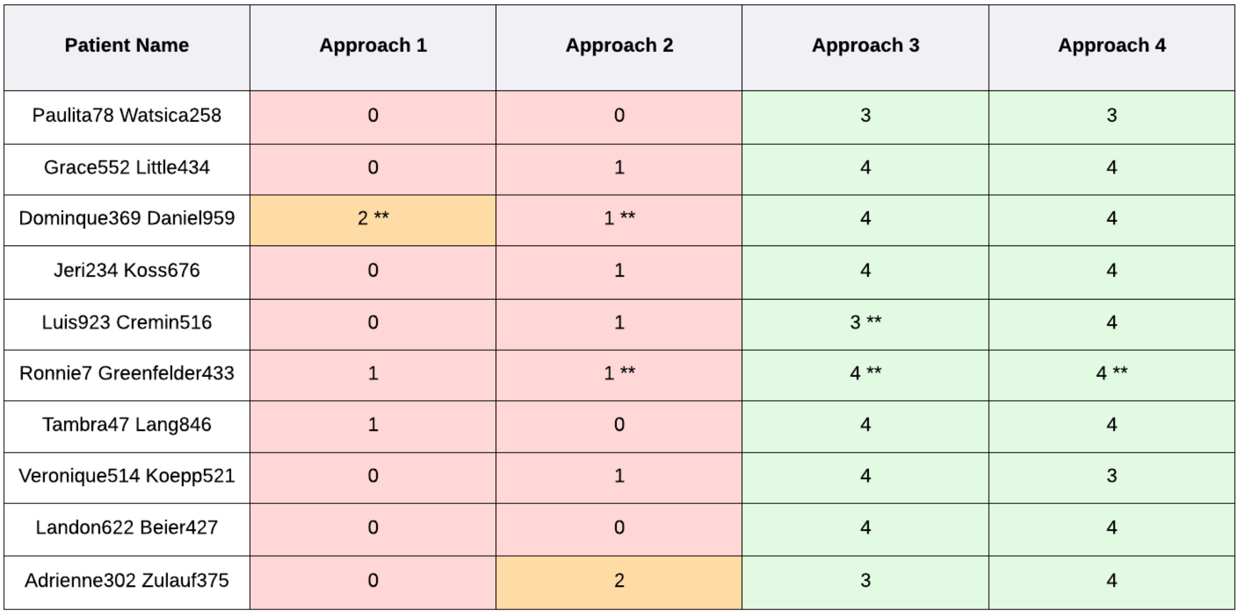
In order to evaluate the relative performance of the four approaches described, ten selected patients were matched to clinical trials using each of the four respective approaches. Each LLM response was evaluated by three team members using a range of qualitative and quantitative metrics.

A quantitative ranking (0 - 4) was used to quantify the performance of the model as each model was tasked with identifying up to four clinical trials for each patient. This numerical assessment will provide a standardized method for evaluating different aspects of the model's outputs, including accuracy, relevance, and completeness. The accuracy was based on a) whether the LLM's response correctly reported the clinical trial's eligibility criteria, b) whether any part of the patient history has been misrepresented in the model's output and c) whether the model was overly strict in evaluating the eligibility of the patient to a trial. By that reasoning, an LLM response that is free from the aforementioned errors is awarded a full ranking score of 4. These rankings will facilitate a clear comparison between the model's performance and human expert evaluations, highlighting areas of strength and opportunities for improvement.

A qualitative analysis was also used to evaluate the LLM's responses. A detailed approach was used to identify points of hallucination, incorrect reasoning, lack of medical terminology, etc. In Supplement S2, the LLM's output under the four different approaches for patient, Paulita78 Watsica258, are presented in detail. Assessment of the error types and the details of the error are presented.

Results

Evaluation results, as shown in Table 1, provided quantitative rankings for the four approaches applied to the ten selected patients. Approaches 3 and 4 demonstrated the best performance, utilizing the Retrieval-Augmented Generation (RAG) method to enhance trial matching. The improvement in performance is a statistically significant one as is confirmed by a t-test on the scores seen in Table 1. A t-test on the scores of Approach 1 and Approach 3 showed a p-value of 1.7 e-7, and a t-test on the scores of Approach 2 and Approach 3 showed a p-value of 9.1 e-8, thereby confirming that the improvement in performance seen with Approach 3 is a statistically significant one compared to performance of each of Approaches 1 and 2.



Patient Name	Approach 1	Approach 2	Approach 3	Approach 4
Paulita78 Watsica258	0	0	3	3
Grace552 Little434	0	1	4	4
Dominque369 Daniel959	2 **	1 **	4	4
Jeri234 Koss676	0	1	4	4
Luis923 Cremin516	0	1	3 **	4
Ronnie7 Greenfelder433	1	1 **	4 **	4 **
Tambra47 Lang846	1	0	4	4
Veronique514 Koepp521	0	1	4	3
Landon622 Beier427	0	0	4	4
Adrienne302 Zulauf375	0	2	3	4

Table 1: *Quantitative Ranking Results in Evaluating the Four Approaches to Match Patients to Clinical Trials. Red denotes that the approach exhibited a score below 50% (i.e. below 2 out of 4), orange denotes a 50% score and green denotes a score above 50%.*

Approach 4, which incorporated a modified ranking mechanism, offered a relatively modest improvement (in terms of aggregate ranking scores) compared with the performance of Approach 3. Future evaluations with more patients may shed further light on the statistical significance of this additional improvement in performance since a t-test on the scores for

Approaches 3 and 4, as seen in Table 1, yields a p-value of 0.59, hinting the difference in performance between Approach 3 and Approach 4 is not likely to be a statistically significant one. However, from a qualitative evaluation of Approaches 3 and 4, slight differences were noted in the exclusion criteria of the clinical trials reported by the model. Approach 3 returned clinical trials for which the patient is not eligible for and accurately states why. Meanwhile, Approach 4 returned clinical trials with exclusion criteria for which the patient summary does not explicitly confirm eligibility. Without the explicit confirmation of the information, the model accurately states that the patient may not be eligible due to insufficient information. While both approaches remain accurate in their responses, the two approaches have unique evaluations of the exclusion criteria. Nevertheless, the observations made from using Approach 4 highlight the potential for other enhanced measures to RAG that can be implemented to bring further improvements in performance.

Approach 1, which relied on full patient files and full clinical trial data without additional summarization or scoring, performed with the least accuracy. This approach struggled with the sheer volume and complexity of the raw data, leading to difficulties in accurately parsing and matching patient information to clinical trial criteria. The lack of summarization meant that the model had to process an overwhelming amount of detailed information, which often resulted in less reliable and less precise matching outcomes.

A closer examination of specific cases further illustrates these findings. A patient diagnosed with colon cancer, Grace552 Little434, was evaluated by all four approaches. Only Approaches 2, 3, and 4 correctly acknowledged that she does not meet the criteria for the breast cancer trials, due to being diagnosed with an entirely different cancer type, demonstrating the approach's ability to accurately discriminate between trials on detailed patient summaries and structured data evaluation. However, Approach 2, despite identifying the mismatch in cancer type, failed to address ineligibility or eligibility of the patient based on other criteria. Approach 1 failed to recognize this mismatch and repeatedly stated that the patient is diagnosed with breast cancer, likely due to its inability to effectively parse and integrate the complex data without summarization.

Another patient, Landon622 Beier427, diagnosed with male breast cancer, was also evaluated. Only Approaches 2, 3, and 4 accurately identified his eligibility for male breast cancer trials and correctly listed appropriate trials. However, Approach 2, despite correctly assessing the patient as male, failed to address ineligibility or eligibility of the patient based on other criteria. Approach 1 again struggled, failing to recognize Landon's specific condition and thus not providing accurate trial matches. This underscores the importance of structured data processing and the use of patient summaries in enhancing the model's accuracy.

Supplement 1 includes the qualitative evaluation of the LLM responses. This qualitative assessment further illuminated the differences in performance among the approaches. Approaches 3 and 4 produced the most accurate and contextually relevant matches, consistently identifying suitable clinical trials for the patients based on detailed and comprehensive criteria matching. Each eligibility criterion was explicitly addressed in the response, with the model explaining why the patient met or did not meet specific requirements. The responses of Approached 3 and 4 also correctly acknowledged when the patient summary had insufficient information to determine eligibility. This transparency in reasoning not only improved the trustworthiness of the results but also provided valuable insights for researchers reviewing the recommendations.

Approach 1's qualitative evaluation revealed its shortcomings in handling the complexity and volume of the data. The model often failed to parse all relevant information accurately, leading to mismatches or missed eligibility criteria. This lack of precision was particularly evident in cases with complex medical histories or multiple comorbidities, where the model struggled to integrate all necessary data points without the aid of summarization or structured evaluation. Despite being asked to, Approaches 1 and 2 did not address each of the eligibility criteria of the clinical trials mentioned in the response.

Approaches 1 and 2 repeatedly exhibited the following errors: a) incorrectly reporting the clinical trial's eligibility criteria, and b) misrepresentation of the patient history. For example, approaches 1 and 2 often misinterpret age inclusion criteria, stating a certain age range that is different from that found in the clinical trial's eligibility criteria. Approaches 1 and 2 also disregarded information on the patient's treatment history. Additionally, limitations of the Synthea dataset include not having the HER-2 status of oncology patients, a common protein that is associated with the growth of cancer cells. Often, approaches 1 and 2 hallucinated the HER-2 status or inappropriately ignored such eligibility criteria.

On rare occasions, Approaches 3 and 4 exhibited tendencies to be overly strict in evaluating the eligibility criteria. For example, one of the clinical trials required the patient to be fluent in spoken and written Spanish. The model stated the patient as 'Ineligible as primary language is English' for this trial. While it is a fair assumption to make that the patient does not speak other languages apart from their primary language, the model did not have sufficient information to make this judgment. Reporting that the criteria was not met due to insufficient information may have been a more appropriate response.

Furthermore, in evaluation of the medical accuracy of the models' responses, it was found that the models were capable of making inferences on the patient's state of health. In a particular case of patient Paulita78 Watsica25, a 56 year old female patient, all four approaches identified clinical trials that include or exclude patients based on menopausal status. Approaches 1 and 2 ignored describing the menopausal status requirement altogether and thereby potentially implied that the patient is premenopausal at age 56. Approaches 3 and 4, on the other hand, explicitly stated that the 56 year old patient is post-menopausal and not of child-bearing age (See Supplements S1 and S2). This observation shows that the model inherently makes certain medical assumptions based on statistical patterns in the data it has been trained on, without the patient record explicitly stating information that supports those assumptions. While these assumptions can be statistically accurate on a broad scale, they may not account for individual patient nuances and specific clinical contexts (Ceylan et al., 2015 & Swanner, 2023). Reporting that the statistical mean natural menopausal age is below age 55 for women, with potential cases of outliers, could have been a more appropriate response. Such generalizations can be detrimental if blindly relied upon, as they might overlook unique patient factors or rare conditions. The model's behavior could reflect an underlying bias in the model training corpus. It is therefore essential that healthcare professionals critically evaluate the model's outputs and integrate their clinical judgment to ensure safe and effective patient care.

Discussion

Presenting the large language model with raw mCODE electronic health record (EHR) data to identify matching clinical trials presents a challenging task due to the complexity and detail of mCODE data, which includes various structured elements such as a patient's medical history, diagnoses, and treatments. Parsing this raw data requires the model to understand and interpret a wide array of medical terminologies and formats, making it significantly more demanding than handling plain text summaries. In prior studies, researchers often used short-form plain text patient summary data as input (Koopman, 2016); for example, Jin et al. (2024) utilized GPT-4 to process these concise summaries, which distill essential information into a straightforward, narrative form. In contrast, raw mCODE EHR data encompasses detailed attributes like age, gender, disease type, stage, and treatment history, necessitating advanced data extraction and contextual understanding to accurately match patients with clinical trials. While handling such detailed and structured data would greatly enhance the model's applicability in real-world clinical settings, it also requires more sophisticated natural language processing capabilities and deeper integration with medical ontologies and standards.

The LLM model's performance was limited by the quality of the Synthea data. Limitations in the Synthea dataset, such as the absence of genetic information in patient health records, were noted. These gaps can significantly affect the accuracy of clinical trial matching, as genetic markers often play a crucial role in determining patient eligibility for a trial. It was seen in Approaches 3 and 4 that the LLM was able to correctly identify that such information was missing and would be useful for matching patients to the appropriate trials. Addressing these limitations by incorporating more comprehensive data sources could enhance the model's performance and reliability.

Synthea data, while well-structured, does not perfectly mimic the variability and complexity found in real-world electronic health records. Real patient data is often unstructured, posing challenges for direct application of the model. In order to comply with safety and ethical considerations, real patient data was not used in the scope of this project. For real-world implementation, consideration must be given to how these real-world data challenges can be addressed, potentially through advanced data preprocessing techniques or adaptive learning algorithms.

Prompt engineering played a significant role in improving the quality of the model's responses. By instructing the model to provide reasons or explanations for each criterion, higher quality answers were produced, and the reasoning process was demystified. Splitting instructions into separate sentences and carefully ordering the prompts, with patient summaries first or instructions first, also proved beneficial. Emphasizing specific information in the patient summary and ensuring all criteria were met further refined the outputs.

Prompting the model for multiple clinical trial recommendations facilitated a better understanding of its reasoning processes. When asked to provide several options, the model demonstrated a more comprehensive approach to matching patients with clinical trials, thus offering a broader perspective on its decision-making capabilities. This approach also allowed for the identification of potential inconsistencies or areas where the model could be improved.

It was noted that the specificity of the LLM model's response correlated with the specificity and comprehensiveness of the data input regarding the patient health record; more specific patient summaries yielded more specific model responses. Hence, the intermediate step of prompting the LLM to summarize the patient record was a limiting factor in the success of the model. In the prompt, the model is asked to specifically focus on the age/gender of the patient, cancer diagnosis and prognosis, treatments and comorbidities. Adding specificity in the prompting resulted in the LLM being able to recall the key information, producing a concise and focused summary of the patient's health record. Additionally, using keywords matching the health record structure and data labels, such as 'clinical staging' pointed the LLM in the direction of the information. This method helped in distilling the essential details from the comprehensive data, thereby enhancing the accuracy and relevance of the model's responses.

With GPT-4o's impressive context window of 128,000 tokens, maximizing the context window did not always correlate with improved answer quality, as seen in Approaches 1 and 2. While more extensive textual input provided the model with additional information and context, it also increased the complexity of processing and led to less coherent responses. This underscores the need for a balanced approach in context window management, ensuring sufficient information is given in the prompt without overwhelming the model.

Improvements and Future Considerations

In future research, comparing the model with other LLMs, such as BioGPT or LLAMA, is recommended to benchmark its performance and identify areas for improvement. BioGPT, with its specialized training on biomedical literature, could provide insights into how domain-specific training impacts the effectiveness of clinical trial matching. Similarly, LLAMA, known for its efficiency and scalability, could offer alternative approaches or algorithms that might enhance the current model. These comparisons could reveal strengths and weaknesses, guiding further refinements in model architecture and querying processes.

Employing more robust evaluation methods to ensure a comprehensive assessment of the model's performance may improve understanding of the four approaches. This could involve multi-dimensional evaluations that not only focus on accuracy but also consider factors such as recall, precision, and F1 scores. Additionally, real-world validation with clinical experts and pilot implementations in healthcare settings could provide practical feedback on the model's effectiveness and usability. Robust evaluations would help in understanding the model's practical utility and readiness for deployment.

Integrating the ClinicalTrials API as a tool for the LLM could significantly enhance real-time querying capabilities based on patient conditions, without limiting the model to only breast cancer clinical trials. This integration would allow the model to access up-to-date clinical trial data dynamically, ensuring that recommendations are current and relevant. Real-time access to clinical trial databases would streamline the patient matching process, making it more efficient and reducing the manual workload on healthcare providers.

Techniques such as Reinforcement Learning from Human Feedback (RLHF) could be used to refine the model's decision-making processes. By incorporating feedback from human experts, the model can continuously learn and improve its accuracy in matching patients to appropriate clinical trials. This iterative learning approach ensures that the model evolves based on practical insights and experiences from the field.

Example-based prompting was proposed as a method to guide the model's responses more effectively. By providing specific examples or templates, the model can better understand the expected output format and the critical information needed for clinical trial matching. This can improve the clarity and relevance of the model's responses, ensuring that it focuses on the most pertinent details. However, with the vastness of medical terminology and complex patient cases, there are several approaches that can be used to match patients to clinical trials that may not be easily learned in a few examples.

Feature extraction techniques could be used to crystallize the features that are most relevant to the search process as per the clinical trial data, and use those features with specific patient health record data to provide the model with more precise information on eligibility criteria. Currently, the system attempts to maintain as much information as possible from the health records to have a comprehensive matching system. Furthermore, a scoring system could also be implemented to prompt the LLM to rank the relevance of clinical trials based on how well they match the patient's profile, providing a more structured and quantifiable approach to trial matching.

Overall, these improvements could potentially enhance the model's accuracy, efficiency, and practical applicability in clinical settings, ultimately improving patient recruitment for clinical trials.

Conclusion

This project has demonstrated the efficacy of four approaches to using LLMs to match patients with clinical trials. It was found that the implementation of the Retrieval-Augmented Generation (RAG) method significantly enhances the precision and accuracy of matching patients to clinical trials when utilizing large language models. By incorporating external data sources and a structured relatedness scoring mechanism, the RAG-enabled approaches produced the most reliable and contextually relevant matches. These findings indicate the potential of similarity-based RAG in streamlining the patient recruitment process, thereby improving the efficiency and success rates of clinical trials. However, the potential for further improvement was identified. Significant insights were gained from this study regarding the impact of quality input patient data and the importance of prompt engineering.

The sensitivity of the model to input data was observed to be critical, indicating that more accurate and comprehensive patient information could enhance matching outcomes. The use of prompt engineering played a notable role in refining the responses, underscoring the necessity for carefully crafted queries to guide the model effectively.

While the current implementation of RAG proved successful, continuous improvement in data processing techniques, more robust evaluation methods, and advanced prompt engineering strategies will be essential for more consistent and accurate results in clinical trial matching. The lessons learned from this study will guide future developments, ensuring that the model remains a valuable tool in clinical research and patient recruitment.

References

- OpenAI et. al. (2023). GPT-4 Technical Report. https://arxiv.org/abs/2303.08774
- OpenAI. (2024). Hello GPT-4o. https://openai.com/index/hello-gpt-4o/
- Luo et. al. (2022). BioGPT: Generative Pre-trained Transformer for Biomedical Text Generation and Mining. https://arxiv.org/abs/2210.10341
- Clinicaltrials.gov.
- NIH: User Guide for Clinicaltrials.gov Website: https://clinicaltrials.gov/submit-studies/prs-help/user-guide#intro
- Jin, Q., Wang, Z., Floudas, C. S., Chen, F., Gong, C., Bracken-Clarke, D., ... & Lu, Z. (2023). Matching patients to clinical trials with large language models. ArXiv.arXiv:2307.15051v4. Accessed on August 2, 2024. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10418514/
- Koopman, B., & Zuccon, G. (2016, July). A test collection for matching patients to clinical trials. In Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval (pp. 669-672).
- Msv, J. (2023, July 21). Tutorial: Build a Q&A bot for Academy Awards based on ChatGPT. The New Stack. https://thenewstack.io/tutorial-build-a-qa-bot-for-academy-awards-based-on-chatgpt/
- Stack OverFlow, Accessed on August 4th 2024. https://stackoverflow.com/questions/78415818/how-to-get-full-results-with-clinicaltrials-gov-api-in-python
- Roberts, K., Demner-Fushman, D., Voorhees, E.M., Bedrick, S. & Hersh, W.R. Overview of the TREC 2021 Clinical Trials Track. in Proceedings of the Thirtieth Text REtrieval Conference (TREC 2021) (2021).
- Walonoski, J., Kramer, M., Nichols, J., Quina, A., Moesel, C., Hall, D., ... & McLachlan, S. (2018). Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. Journal of the American Medical Informatics Association, 25(3), 230-238.
- Ceylan, B., & Özerdoğan, N. (2015). Factors affecting age of onset of menopause and determination of quality of life in menopause. Turkish journal of obstetrics and gynecology, 12(1), 43–49. https://doi.org/10.4274/tjod.79836
- Swanner, K. D., & Richmond, L. B. (2023). A 65-Year-Old Woman With No Menopause History: A Case Report. Cureus, 15(9), e44792. https://doi.org/10.7759/cureus.44792

Supplements

EN.705.651.8VL.SU24 - Large Language Models: Theory and Practice
Group Final Project

S1: LLM Output Highlights

Approach 1: Incorporating Full Patient Data and Clinical Trial Data in the Query

Synthetic Patient Name: Paulita78 Watsica258

```
Based on the provided data, the patient is a female born on
March 10, 1968. As of today's date, August 1, 2024, she is 56
years old. The patient has been diagnosed with malignant
neoplasm of the breast (breast cancer), as indicated in the
'Condition' section of the mCODE data.


To find suitable clinical trials for the patient, we need to
consider her age, diagnosis, and other relevant medical history.
Here are four clinical trials that the patient might be eligible
for, based on the inclusion criteria related to age, gender, and
diagnosis:


### Clinical Trial 1: NCT03147430

- **Eligibility Criteria**: Women who receive a standard of care
for breast cancer.

- **Reason for Eligibility**:

  - **Gender**: The trial specifically mentions women, which
matches the patient's gender.

  - **Diagnosis**: The trial is for patients receiving standard
care for breast cancer, which is likely applicable to this
patient given her diagnosis.


### Clinical Trial 2: NCT04638725

- **Eligibility Criteria**: HER2-positive breast cancer
patients.

- **Reason for Eligibility**:
  - **Diagnosis**: If the patient's breast cancer is
HER2-positive (this information needs to be confirmed), she
would be eligible based on her diagnosis.


### Clinical Trial 3: NCT05989776

- **Eligibility Criteria**: Patients over 18 and under 75 years
of age with breast cancer.

- **Reason for Eligibility**:

  - **Age**: The patient is 56 years old, which falls within the
specified age range.

  - **Diagnosis**: The patient has been diagnosed with breast
cancer.


### Clinical Trial 4: NCT04506476

- **Eligibility Criteria**: Capacity for consent, diagnosed with
breast cancer, and undergoing or completed treatment.

- **Reason for Eligibility**:

  - **Consent**: Assuming the patient can provide consent.

  - **Diagnosis and Treatment**: The patient has breast cancer
and is likely undergoing treatment based on her medical history.


For each of these trials, further details from the patient's
medical records (such as specific type of breast cancer, current
treatment details, and overall health status) would be necessary
to confirm eligibility. Additionally, logistical considerations
such as the patient's ability to travel to the trial locations
and her interest in participating should also be taken into
account.


If more specific information about the patient's type of breast
cancer (e.g., HER2 status) or other health conditions is
available, it would help refine the selection of appropriate
clinical trials.
```

EN.705.651.8VL.SU24 - Large Language Models: Theory and Practice
Group Final Project

Approach 2: Incorporating Patient Summary and Clinical Trial Data in the Query

Synthetic Patient Name: Paulita78 Watsica258

```
Based on the provided clinical trial data and the patient's
medical history, let's evaluate Mrs. Watsica's eligibility for
some of the listed clinical trials. We will focus on trials
related to breast cancer, as that is her primary diagnosis.


### Trial 1: NCT03147430

**Eligibility Criteria:**

- Women who receive a standard of care for breast cancer.


**Patient's Eligibility:**

- Mrs. Watsica has been diagnosed with breast cancer and has
received standard care including radiation therapy. Therefore,
she meets this criterion.


### Trial 2: NCT04638725 (SIGHER)

**Eligibility Criteria:**

- HER2-positive breast cancer patients.


**Patient's Eligibility:**

- The patient's medical history does not specify that her breast
cancer is HER2-positive. Without this information, we cannot
confirm her eligibility for this trial.


### Trial 3: NCT05989776 (EVAPREF)

**Eligibility Criteria:**

- Patients over 18 and under 75 years of age.

- Diagnosed with breast cancer.
```

Patient's Eligibility:

- Mrs. Watsica is 56 years old, which fits within the age range specified.

- She has been diagnosed with breast cancer, meeting the second criterion.

Trial 4: NCT05781633

Eligibility Criteria:

- Signed informed consent.

- Female patients with histologically confirmed breast cancer.

Patient's Eligibility:

- Assuming Mrs. Watsica is willing to sign informed consent, she meets the first criterion.

- She has a diagnosis of malignant neoplasm of the breast, which is histologically confirmed breast cancer, thus meeting the second criterion.

Conclusion:

Mrs. Watsica is eligible for the following clinical trials based on the information provided:

- **NCT03147430**: She meets the criterion of receiving standard care for breast cancer.

- **NCT05989776 (EVAPREF)**: She fits the age range and has a diagnosis of breast cancer.

- **NCT05781633**: She can provide informed consent and has histologically confirmed breast cancer.

She does not appear eligible for NCT04638725 (SIGHER) as there is no information about her breast cancer being HER2-positive.

Further testing would be required to determine HER2 status if considering this trial.

Approach 3: Incorporating Patient Summary and Clinical Trial Data in a RAG-Enabled LLM Query

Synthetic Patient Name: Paulita78 Watsica258

Based on the provided patient summary and the clinical trial eligibility criteria, Mrs. Paulita78 Watsica258 does not match the criteria for any of the clinical trials listed. Here is a detailed explanation of why she does not meet the eligibility criteria for each trial:

1. **NCT04360330**

   - **Inclusion Criteria:**

     - Female, ≥ 50 years of age. (Eligible)

     - Oncotype or MammaPrint diagnosis results are required prior to the start of treatment. (Not specified in patient summary)

     - Histologically confirmed invasive breast cancer. (Eligible)

     - Clinical stage T1N0M0. (Eligible)

     - Receptor status: ER/PR positive and HER2 negative. (Not specified in patient summary)

     - Unifocal breast cancer. (Not specified in patient summary)

     - ECOG 0, 1. (Not specified in patient summary)

     - Ability to undergo MRI. (Not specified in patient summary)

     - WOCBP must agree to use adequate contraception or agree to undergo sexual abstinence. (Not applicable as she is post-menopausal)

    - Ability to understand the investigational nature, potential risks and benefits of the research study and willingness to sign the written informed consent and HIPAA document(s). (Eligible)

  - **Exclusion Criteria:**

    - Patients without histologically confirmed invasive breast cancer. (Eligible)

    - Patients without Oncotype or MammaPrint diagnosis results at the start of treatment. (Not specified in patient summary)

    - Patients with metastatic disease. (Eligible)

    - ECOG 2, 3, 4. (Not specified in patient summary)

    - Patients that are unable to undergo MRI. (Not specified in patient summary)

    - Prior history of radiation to the chest. (Ineligible due to radiation therapy in 2020)

    - History of collagenous disease. (Eligible)

    - Any serious medical or psychiatric illness/condition likely to interfere or limit compliance with study requirements/treatment. (Eligible)

    - Diagnosis of another primary malignancy within the last 5 years. (Eligible)

    - Patients unable to consent, who are pregnant or nursing, or are prisoners. (Eligible)

2. **NCT04965909**

  - **Inclusion Criteria:**

    - Woman aged between 18 - 65 years. (Eligible)

    - Diagnosis of stage 0 - III breast cancer. (Eligible)

    - Primary treatment completed at least 3 months ago. (Eligible)

    - Pain related to primary treatment in the last 6 months. (Not specified in patient summary)

    - Access to Internet and an electronic device. (Not specified in patient summary)

    - Ability to communicate fluently in Spanish. (Ineligible as primary language is English)

    - Approval of participation in the study by the coordinator of the health team. (Not specified in patient summary)

  - **Exclusion Criteria:**

    - Another previous type of cancer or breast cancer recurrence in a period of less than 1 year. (Eligible)

    - Medical diagnosis of a neurological, autoimmune or cardiovascular disease. (Eligible)

    - Pathology associated with a contraindication to physical exercise. (Eligible)

    - Diagnosis of serious psychiatric or neurologic disorders. (Eligible)

3. **NCT05033925**

  - **Inclusion Criteria:**

    - Female, at least 18 years old. (Eligible)

    - Welfare scale 0, 1, and 2 (ECOG - WHO). (Not specified in patient summary)

    - Patients with stage IV breast cancer. (Ineligible as patient has stage 1A)

    - Patients who are willing to participate in the test and sign an informed consent. (Eligible)

    - Patients who undergo breast cancer chemotherapy with luminal B sub-type or triple negative with anthracycline base. (Not specified in patient summary)

    - Patients who are willing and able to fill out a questionnaire. (Eligible)

    - Patients who are willing and able to comply with the test protocols during the test. (Eligible)

  - **Exclusion Criteria:**

    - Unable to meet the test protocol. (Eligible)

    - Patients with liver and kidney disorders. (Eligible)

    - Patients with other diseases/disorders that are meaningful according to the investigators. (Eligible)

    - Patients with cancer that has metastasized to the brain. (Eligible)

    - Pregnant women and breastfeeding mothers. (Eligible)

    - Patients with the ejection fraction smaller-than or equal to 55%. (Not specified in patient summary)

4. **NCT05891561**

  - **Inclusion Criteria:**

    - Histologically diagnosed unilateral primary invasive breast cancer patient with pT1, pN0, and M0 disease. (Eligible)

    - HER2-positive. (Not specified in patient summary)

    - Complete clinical pathological information. (Not specified in patient summary)

    - ECOG 0-1. (Not specified in patient summary)

    - Currently not pregnant or breast-feeding. (Eligible)

    - Fine organ function. (Not specified in patient summary)

    - Have good compliance with planned treatment, understand the study process and sign a written informed consent. (Eligible)

  - **Exclusion Criteria:**

    - Bilateral or metastatic breast cancer. (Eligible)

    - Receiving neoadjuvant treatment. (Not specified in patient summary)

    - Other malignancies within 5 years. (Eligible)

    - Severe systemic infections or other serious illnesses. (Eligible)

- HIV infection, active hepatitis B or C infection. (Eligible)

- Known allergy to or intolerance to a therapeutic drug or its excipients. (Eligible)

- Prior history of chemotherapy, endocrine therapy, biotherapy, or radiation therapy for any reason. (Ineligible due to prior radiation therapy)

- Enrollment of another investigational study within 4 weeks prior to initial administration of the investigational treatment. (Not specified in patient summary)

- Receiving live vaccine within 30 days prior to initial administration of the investigational treatment. (Not specified in patient summary)

- History of mental illness or drug abuse that may affect compliance with the trial requirements. (Eligible)

- The researchers determine that the patients were not suitable for the study. (Eligible)

Given the detailed criteria and the patient's medical history, Mrs. Watsica does not fully meet the eligibility requirements for any of the clinical trials listed. The primary reasons include prior radiation therapy, language requirements, and specific diagnostic or treatment history not aligning with the trial criteria.

Synthetic Patient Name: Paulita78 Watsica258

Approach 4: Incorporating Patient Summary and Clinical Trial Data in a RAG-Enabled LLM Query with a Modified Similarity Metric

Based on the provided patient summary and the clinical trial eligibility criteria, Mrs. Paulita78 Watsica258 does not match the criteria for any of the clinical trials listed. Here is a detailed explanation of why she does not meet the eligibility criteria for each trial:

NCT04360330

Inclusion Criteria:

1. Female, ≥ 50 years of age. **(Eligible)**

2. Oncotype or MammaPrint diagnosis results are required prior to the start of treatment. **(Not specified in patient summary)**

3. Histologically confirmed invasive breast cancer. **(Eligible)**

4. Clinical stage T1N0M0. **(Eligible)**

5. Receptor status: ER/PR positive and HER2 negative. **(Not specified in patient summary)**

6. Unifocal breast cancer. **(Not specified in patient summary)**

7. ECOG 0, 1. **(Not specified in patient summary)**

8. Ability to undergo MRI. **(Not specified in patient summary)**

9. WOCBP must agree to use adequate contraception or agree to undergo sexual abstinence. **(Not applicable as she is postmenopausal)**

10. Ability to understand the investigational nature, potential risks and benefits of the research study and willingness to sign the written informed consent and HIPAA document(s). **(Eligible)**

Exclusion Criteria:

1. Patients without histologically confirmed invasive breast cancer. **(Eligible)**

2. Patients without Oncotype or MammaPrint diagnosis results at the start of treatment. **(Not specified in patient summary)**

3. Patients with metastatic disease. **(Eligible)**

4. ECOG 2, 3, 4. **(Not specified in patient summary)**

5. Patients that are unable to undergo MRI. **(Not specified in patient summary)**

6. Prior history of radiation to the chest. **(Not eligible due to radiation therapy in 2020)**

7. History of collagenous disease. **(Eligible)**

8. Any serious medical or psychiatric illness/condition likely to interfere or limit compliance with study requirements/treatment. **(Eligible)**

9. Diagnosis of another primary malignancy within the last 5 years. **(Eligible)**

10. Patients unable to consent, who are pregnant or nursing, or are prisoners. **(Eligible)**

NCT05889390

Inclusion Criteria:

1. At least 18 years of age. **(Eligible)**

2. Female patient. **(Eligible)**

3. Life expectancy ≥ 6 months. **(Eligible)**

4. De novo histological/cytological diagnosis of HER2-negative breast tumor involving one breast. **(Not specified in patient summary)**

5. Diagnosis of breast tumor ≤ 40 days. **(Not eligible, diagnosed in 2020)**

6. Locally advanced stage disease (stage II and III) requiring neoadjuvant treatment. **(Not eligible, Stage 1A)**

7. ECOG status: 0-2. **(Not specified in patient summary)**

8. Suitable for and designated by the investigator for neoadjuvant therapy. **(Not specified in patient summary)**

9. Willingness to participate in the trial and signed the informed consent form for the protocol. **(Eligible)**

Exclusion Criteria:

1. Patient is ≤ 18 years of age. **(Eligible)**

2. Tumor of both breasts. **(Eligible)**

3. Diagnosis of breast tumor > 40 days. **(Not eligible, diagnosed in 2020)**

4. HER2 positive breast tumor. **(Not specified in patient summary)**

5. Has already received some anticancer therapy. **(Not eligible, received radiation therapy)**

6. Any previous cancer requiring anti-tumor treatment within 5 years prior to selection. **(Eligible)**

7. Co-existing serious diseases. **(Eligible)**

NCT04481932

Inclusion Criteria:

1. Female between 18 and 70 years old. **(Eligible)**

2. Histologically confirmed as invasive breast cancer. **(Eligible)**

3. ECOG PS 0-1. **(Not specified in patient summary)**

4. The expected survival time is not less than 12 weeks. **(Eligible)**

5. Standard immunohistochemical HER2-positive breast cancer patients. **(Not specified in patient summary)**

6. The status of hormone receptors (ER and PR) can be known. **(Not specified in patient summary)**

7. Clinical examination or imaging examination of primary lesion >2cm. **(Not specified in patient summary)**

8. Patients who are operable and have not received any previous anti-tumor therapy. **(Not eligible, received radiation therapy)**

9. Echocardiography indicated left ventricular ejection fraction (LVEF)≥55%. **(Not specified in patient summary)**

10. Adequate organ and bone marrow function. **(Not specified in patient summary)**

11. I have agreed and signed the informed consent. **(Eligible)**

Exclusion Criteria:

1. Have received any previous anti-tumor treatment for primary invasive breast cancer. **(Not eligible, received radiation therapy)**

2. Previous (<10 years) or other malignant tumors. **(Eligible)**

3. Metastatic breast cancer (M1), bilateral or ipsilateral multifocal breast cancer. **(Eligible)**

4. Uncontrolled hypertension, systolic blood pressure > 150 MMHG and/or diastolic blood pressure > 100 MMHG. **(Not specified in patient summary)**

5. Receive other anti-tumor treatments within 4 weeks before enrollment. **(Eligible)**

6. Inability to swallow, intestinal obstruction or other factors affecting the use and absorption of medication. **(Eligible)**

7. Persons with allergic constitution or known history of allergy to the drug components of the program. **(Eligible)**

8. The patient has a severe concomitant disease or other conditions that the researcher considers inappropriate for the patient to participate in the study. **(Eligible)**

9. Non-surgically sterilized female patients of childbearing age must have negative serum or urine HCG tests within 14 days prior to study inclusion. **(Not applicable as she is postmenopausal)**

10. Other circumstances deemed inappropriate for inclusion by the researcher. **(Eligible)**

NCT04913064

Inclusion Criteria:

1. Postmenopausal women who have increased risk of breast cancer and high BMI ≥ 30 kg/m^2. **(Eligible)**

2. The study population will consist of women with a relative risk of developing breast cancer that is at least > 2 x that of

the general population for their age group based on any of the following: **(Not specified in patient summary)**

3. Body mass index (BMI) $\geq$ 30 (Asian BMI $\geq$ 25 kg/m^2). **(Eligible)**

4. Postmenopausal, defined as continuous absence of menstruation for 12+ months. **(Eligible)**

5. Bilateral mammogram within the 12 months prior to study enrollment that is read as not suspicious for breast cancer. **(Not specified in patient summary)**

6. Serum creatinine of 1.5 X upper limit of institutional norm or less. **(Not specified in patient summary)**

7. Total bilirubin of 1.5 X upper limit of institutional norm or less. **(Not specified in patient summary)**

8. ALT and AST of less than 2 X upper limit of institutional norm or less. **(Not specified in patient summary)**

9. Hemoglobin of 9.0 gm/dL or more. **(Not specified in patient summary)**

10. Platelets of 100,000/mm^3 or more. **(Not specified in patient summary)**

11. Total white blood cell (WBC) of 3500/mm^3. **(Not specified in patient summary)**

12. Absolute neutrophil count (ANC) of 1500/mm^3 or more. **(Not specified in patient summary)**

13. Must be willing to have about 40-50 ml of blood drawn at 0 and 3 months. **(Eligible)**

14. Must be able to swallow pills. **(Eligible)**

15. This study will assess WBM powder in postmenopausal women of 21 years of age or older. **(Eligible)**

16. Participants will have ECOG performance status 0 or 1. **(Not specified in patient summary)**

17. The effects of WBM on the developing fetus are unknown. Pregnant and women with childbearing potential are not eligible to participate in this study of postmenopausal women at high risk of breast cancer. **(Eligible)**

18. All subjects must have the ability to understand and the willingness to sign a written informed consent. **(Eligible)**

19. Prior therapies, including chemoprevention or adjuvant therapies, must have been completed for 6 or more months prior to study entry. **(Not eligible, received radiation therapy in 2020)**

Exclusion Criteria:

1. Active malignancy within the past 5 years with the exception of non-melanoma skin cancer or carcinoma in situ of the cervix. **(Eligible)**

2. Ongoing chemotherapy, radiation therapy, or other cancer-related treatment. **(Not eligible, received radiation therapy in 2020)**

3. History of a bleeding tendency or current use of Coumadin or other anticoagulants. **(Eligible)**

4. Concurrent use of hormone-modifying medications including systemic hormone replacement therapy. **(Not specified in patient summary)**

5. Concurrent use of immunosuppressant medications. **(Eligible)**

6. Uncontrolled intercurrent illness. **(Eligible)**

7. Chronic use of any herbal or dietary supplement containing mushrooms within the 3 months prior to entry on the study. **(Not specified in patient summary)**

8. Treatment with other investigational agents. **(Eligible)**

9. Premenopausal status. **(Eligible)**

10. Known sensitivity or allergic reactions attributed to compounds of similar chemical or biologic composition to WBM. **(Eligible)**

11. Subjects on a standing regimen of full dose aspirin (≥ 325 mg/day), NSAIDs or NSAID-containing products. **(Not specified in patient summary)**

12. Any other condition that would, in the Investigator's judgment, contraindicate the patient's participation in the

clinical study due to safety concerns with clinical study procedures. **(Eligible)**

13. Subjects, who in the opinion of the investigator, may not be able to comply with the safety monitoring requirements of the study. **(Eligible)**

Summary

Mrs. Watsica does not meet the eligibility criteria for any of the clinical trials due to various reasons such as prior radiation therapy, lack of specific diagnostic results (e.g., Oncotype or MammaPrint), and the timing of her diagnosis. Therefore, she is not eligible for any of the listed clinical trials.

EN.705.651.8VL.SU24 - Large Language Models: Theory and Practice
Group Final Project

S2: Human Qualitative Evaluation of LLM Responses

Human Expert Evaluation of LLM Response for Approach 1: Incorporating Full Patient Data and Clinical Trial Data in the Query

Synthetic Patient Name: Paulita78 Watseka 258 (Walonoski, et al., 2018).(See Suppl. S1)

LLM Generated Trial #1: NCT03147430

LLM's Assessment: Include

Human Expert Consensus Assessment: Exclude

- Error 1: Reasoning Error.
- Error details: Identification of a trial that is a non-treatment study. (discuss, we didn't prompt the LLM to look for treatment trials either so maybe it is not technically erring). The trial's name is called Early Detection of Breast Cancer in Women With Suspicious Mammograms. The trial's description also states that it is a non-treatment study and therefore won't benefit the patient who has malignant neoplasm of the breast. The LLM fails to infer that the patient should look for a treatment trial.

Synthetic Patient Name: Paulita78 Watsica258 (Walonoski, et al., 2018).

LLM Generated Trial #2: NCT04638725

LLM's Assessment: Potentially Include

Human Expert Consensus Assessment: Exclude

- Error 1: Recall Error
- Error details:The study description is very specific about not including patients who are not HER2 positive. The trial's name is called Identification of Genetic Determinants for Treatment Resistance/Sensitivity and/or Toxicity in Adjuvant Setting for HER2 Positive Breast Cancer (SIGHER). The trial's description states that this is a multicenter, non-randomized, prospective cohort study. The purpose of the study is to identify constitutional genetic factors associated with histological response, resistance or sensibility to treatment in human epidermal growth factor receptor 2 (HER2)-positive breast cancer. 9000 patients will be enrolled in this study. Blood samples will be collected after informed consent and inclusion in the study.Patients will be treated and followed according to the standards of their treating center. They will be followed every six months for five years. The synthetic patient’s mCODE data does not indicate anywhere the HER2 status.

Synthetic Patient Name: Paulita78 Watsica258 (Walonoski, et al., 2018).

LLM Generated Trial #3: NCT05989776

LLM's Assessment: Include

Human Expert Consensus Assessment: Exclude

- Error 1: Hallucination
- Error details: LLM claims patient is 56 and falls within the inclusion criteria age range of 18-75. But the actual eligibility criteria says that patients should be between the ages of 18-40.
- Error 2: Reasoning + Medical Knowledge Inaccuracy
- Error details: The trial's name is called Increasing Access to Fertility Preservation for Women With Breast Cancer (EVAPREF). The trial's summary shows that it is a trial mainly interested in increasing fertility preservation. The patient is 56 years old and her menopausal status as well as fertility age needs to be assessed and cannot be based on assumption.

Synthetic Patient Name: Paulita78 Watsica258 (Walonoski, et al., 2018).

LLM Generated Trial #4: NCT04506476

LLM's Assessment: Include

Human Expert Consensus Assessment: Exclude

- Error 1: Recall Error
- Error details:Trial name is Trial Evaluating the Benefit of a Fitness Tracker Based Workout During Adjuvant Radiotherapy of Breast Cancer. LLM misses the most important inclusion criteria of "Indication for adjuvant radiotherapy of breast cancer after breast-conserving surgery or Ablatio mammae". The synthetic patient is prescribed Megavoltage radiation therapy using photons (procedure), not RT, and therefore should be excluded.

Human Expert Evaluation of LLM Response for Approach 2 Incorporating Patient Summary and Clinical Trial Data in the Query

Synthetic Patient Name: Paulita78 Watsica258 (Walonoski, et al., 2018).

LLM Generated Trial #4: NCT05781633

LLM's Assessment: Include

Human Expert Consensus Assessment: Exclude

- Error 1: Recall Error
- Error details: Trial's name is the Efficacy and Safety of Eutideron, Etoposide, and Bevacizumab in Patients With Brain Metastases From Breast Cancer. The synthetic

patient did not receive Eutideron, Etoposide, or Bevacizumab. The synthetic patient does not have brain metastases from breast cancer.

Human Expert Evaluation of LLM Response for Approach 3 Incorporating Patient Summary and Clinical Trial Data in a RAG-Enabled LLM Query

The LLM correctly indicated that all four of the trial numbers generated do not match the patient's profile. Further research is needed to determine whether the LLM's claim that none of the trials pulled from the clinicaltrials.gov database match the patient is correct.

Synthetic Patient Name: Paulita78 Watsica258 (Walonoski, et al., 2018).(See Suppl. S1)

LLM Generated Trial#1: NCT03147430

LLM's Assessment: Exclude

Human Expert Consensus Assessment: Exclude

- LLM Performance Improvement 1: Better recall
- Improvement details: the LLM correctly and exhaustively generated the inclusion and exclusion criteria of the trial and appended its comment to the end of each criterion. In cases where information in the inclusion criteria is not specified in the patient summary, the LLM correctly stated as such. Improvement details: the LLM correctly understood from the patient record that the patient underwent radiation therapy in 2020 and that this makes the patient excluded from the trial because the exclusion criteria specifies that prior history of radiation to the chest is an exclusion criterion.
- LLM Performance Error 1: Medical Knowledge Inaccuracy
- Error details: the LLM incorrectly inferred from the patient record that the patient at 56 years old is already postmenopausal and therefore does not satisfy the inclusion criterion *WOCBP must agree to use adequate contraception or agree to undergo sexual abstinence*.

Human Expert Evaluation of LLM Response for Approach 4 Incorporating Patient Summary and Clinical Trial Data in a RAG-Enabled LLM Query with a Modified Similarity Metric

The LLM correctly indicated that all four of the trial numbers generated do not match the patient's profile. Further research is needed to determine whether the LLM's claim that none of the trials pulled from the clinicaltrials.gov database match the patient is correct. Compared with Approach 3, the LLM generated 2 different trials in Approach 4 and determined them to not be matches.

Synthetic Patient Name: Paulita78 Watsica258 (Walonoski, et al., 2018).(See Suppl. S1)

LLM Generated Trial#1: NCT03147430

LLM's Assessment: Exclude

Human Expert Consensus Assessment: Exclude

- LLM Performance Improvement 1: Better reasoning
- Improvement details: the LLM correctly and exhaustively generated the inclusion and exclusion criteria of the trial and appended its comment to the end of each criterion. In cases where information in the inclusion criteria is not specified in the patient summary, the LLM correctly stated as such. Improvement details: the LLM correctly understood from the patient record that the patient underwent radiation therapy in 2020 and that this makes the patient excluded from the trial because the exclusion criteria specifies that prior history of radiation to the chest is an exclusion criterion.
- LLM Performance Error 1: Medical Knowledge Inaccuracy
- Error details: the LLM incorrectly inferred from the patient record that the patient at 56 years old is already postmenopausal and therefore does not satisfy the inclusion criterion *WOCBP must agree to use adequate contraception or agree to undergo sexual abstinence*.

Synthetic Patient Name: Paulita78 Watsica258 (Walonoski, et al., 2018).(See Suppl. S1)

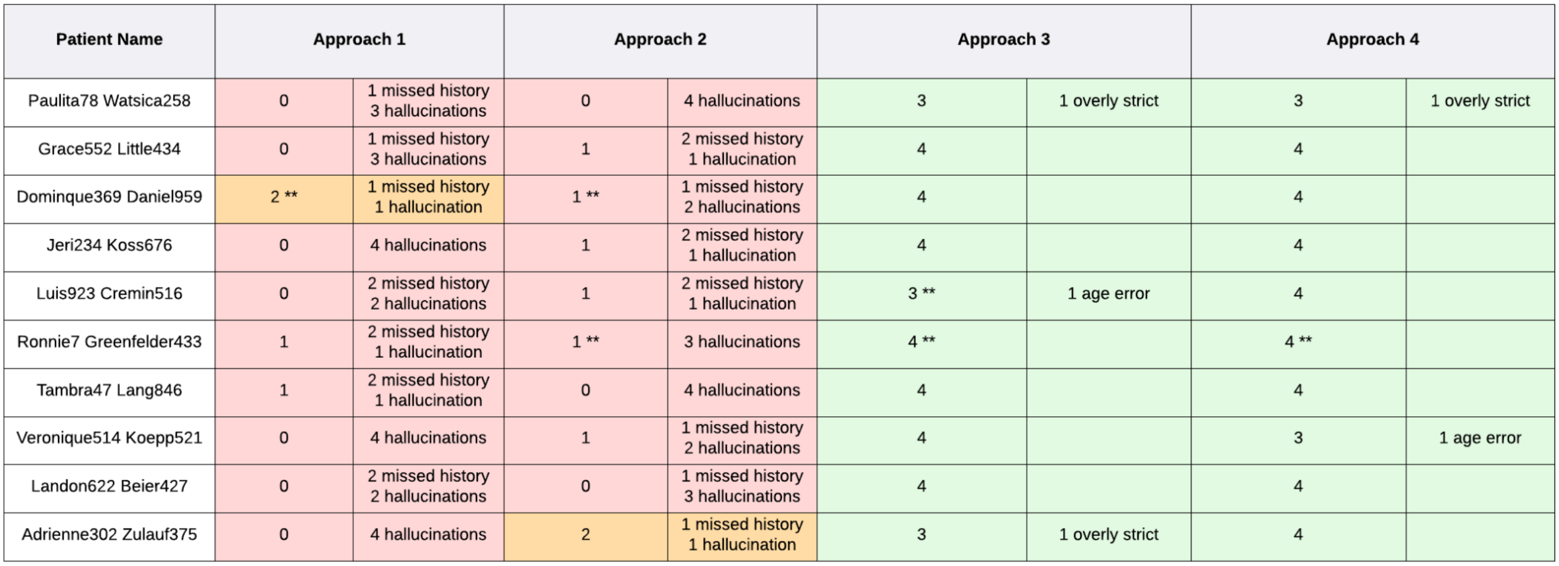
LLM Generated Trial#1: NCT04913064

LLM's Assessment: Exclude

Human Expert Consensus Assessment: Exclude

- LLM Performance Error 1 : Reasoning Error
- Error details: The patient's medical record summary states clearly that the patient's race is White and does not present any information regarding the patient's last measured weight, height, or BMI. The model, however, reasoned that that the patient matches this inclusion criteria.

S3: Table



Patient Name	Approach 1		Approach 2		Approach 3		Approach 4	
Paulita78 Watsica258	0	1 missed history 3 hallucinations	0	4 hallucinations	3	1 overly strict	3	1 overly strict
Grace552 Little434	0	1 missed history 3 hallucinations	1	2 missed history 1 hallucination	4		4	
Dominque369 Daniel959	2 **	1 missed history 1 hallucination	1 **	1 missed history 2 hallucinations	4		4	
Jeri234 Koss676	0	4 hallucinations	1	2 missed history 1 hallucination	4		4	
Luis923 Cremin516	0	2 missed history 2 hallucinations	1	2 missed history 1 hallucination	3 **	1 age error	4	
Ronnie7 Greenfelder433	1	2 missed history 1 hallucination	1 **	3 hallucinations	4 **		4 **	
Tambra47 Lang846	1	2 missed history 1 hallucination	0	4 hallucinations	4		4	
Veronique514 Koepp521	0	4 hallucinations	1	1 missed history 2 hallucinations	4		3	1 age error
Landon622 Beier427	0	2 missed history 2 hallucinations	0	1 missed history 3 hallucinations	4		4	
Adrienne302 Zulauf375	0	4 hallucinations	2	1 missed history 1 hallucination	3	1 overly strict	4	

Table 2: *Quantitative Ranking Results in Evaluating the Four Approaches to Match Patients to Clinical Trials. Brief descriptions of the nature of the errors made by the model are included. Red denotes that the approach exhibited a score below 50% (i.e. below 2 out of 4), orange denotes a 50% score and green denotes a score above 50%.*