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

Project Proposal

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.

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 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. Microsoft's BioGPT is a specialized variant of the GPT architecture tailored for biomedical applications. Pre-trained on 15 million PubMed abstracts, it is designed to leverage the vast amount of textual data available in the biomedical domain, such as research papers, clinical reports, and other scientific literature, to improve the generation and understanding of biomedical text. While similar to GPT4 in architecture, BioGPT consists of more domain-specific training that may be deemed useful in matching patients to clinical trials (Luo et. al, 2022).

Retrieval-Augmented-Generation (RAG) is a commonly used method to introduce contextual information to a large language model, combating out-of-date information. RAG acts as an information retrieval component from external data sources to supplement a text generator model to facilitate knowledge-intensive and domain-specific tasks.

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 develop a RAG-enabled LLM architecture to identify clinical trials for which a patient is eligible for. This identification is to be carried out using a dataset of clinical trial records extracted from ClinicalTrials.gov.

Data Source

Clinical trial records will be downloaded from ClinicalTrials.gov using their REST API as a comma-separated-values (.csv) file. Each clinical study is uniquely identified with an National Clinical Trial Identification Number ("NCT Number"). Within each clinical study, the eligibility criteria is provided as markup text.

Patient records/profiles will be generated 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.

Data Set

Data fields that identify the clinical study (NCT Number, Title) and describe the eligibility requirements (inclusion and exclusion criteria, min/max age, sex/gender, etc) will be included in the dataset. Only clinical trials that are actively recruiting or will recruit in the near-future are included in the dataset. For the sake of computational and temporal constraints, clinical studies belonging to a specific set of conditions/diseases will be included in the dataset.

Methods

We aim to develop a RAG-enabled chained agentive LLM architecture to match patients to clinical trials via eligibility criteria. This architecture will use RAG to preload contextual information and use prompt chaining to perform clinical trial matching. The contextual information will include the database of recruiting clinical trial records and its associated eligibility requirements, along with NIH’s Unified Medical Language System documentation. Synthetic patient profiles will be taken from Synthea, an open-source software used to create artificial patient records. The artificial records will be prompt-chained and passed through the RAG-enabled LLM architecture to match patients to available clinical trial records.

We also aim to explore the incorporation of traditional ML models in the workflow. The RAG-enabled LLM architecture will be compared to GPT 4o and Bio-GPT as standalone baselines. The baselines models will directly parse the information from clinicaltrials.gov data via the restAPI. We, as domain experts, will also manually match patients to clinical trials using the clinicaltrials.gov user interface to act as the ground-truth for the models.

Evaluations

Evaluations will focus on comparing the correctness of the developed model to that of human experts. We will compare the performance of human experts to that of the developed architecture. Human experts will evaluate the correctness of the LLM model’s results. We also aim to discuss the cost-effectiveness and real-world value of the methods, using metrics such as time taken, labor costs, and computational costs.

Safety and Ethical Considerations

Patient information is protected under the Health Insurance Portability and Accountability Act of 1996 (“HIPAA”). This research project will not use real patient data but will instead use 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 relevant information.

Potential Data Sources

NIH: Glossary of Medical Research
https://www.nlm.nih.gov/research/umls/index.html

NIH: User Guide for Clinicaltrials.gov Website
https://clinicaltrials.gov/submit-studies/prs-help/user-guide#intro

SyntheticMass: Synthetic Patient Records
https://synthea.mitre.org/downloads

Similar Projects

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10418514/

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.