Leveraging a RAG empowered LLM for Natural Language Processing on ClinicalTrials.gov

Sam Pepper, MS, Matthew McGuirk, Mahitha Peddapalli, Dinesh Pal Mudaranthakam, PhD
The University of Kansas Medical Center, Kansas City, KS
Department of Biostatistics & Data Science, University of Kansas Medical Center

Introduction

Clinical trials play a crucial role in advancing medical treatments. However, a significant challenge for clinical trials is efficiently identifying suitable participants. Eligibility to participant in a trial relies on inclusion and exclusion criteria encompassing factors such as age, gender, medical history, disease stage, and prior treatments. Extracting inclusion and exclusion criteria from clinical trial descriptions, often presented in unstructured text format on websites like ClinicalTrials.gov, can be a laborious and time-consuming process but is necessary to determine patient eligibility.

By leveraging Natural Language Processing (NLP) techniques, specifically the combination of a Large Language Model (LLM) and Retrieval Augmented Generation (RAG), we can automate the process of extracting inclusion and exclusion criteria from unstructured text and store it in a structured format. This automation allows for the construction of a simple eligibility database and allows researchers and clinicians to efficiently compare eligibility criteria with the electronic medical record (EMR) to determine patients that may be eligible for participation in specific clinical trials.

We created a system consisting of a RAG empowered LLM to extract key clinical trial eligibility criteria from unstructured text data on ClinicalTrials.gov. Our targeted criteria to extract was inclusion criteria related to age, Eastern Cooperative Oncology Group Score (ECOG), and disease stage.

Methods and Materials

For the LLM we utilized the quantized Nous-Hermes-2-Yi-34B, a Generalized Post-Training Quantization (GPTQ) model which could be run on a local computer. To construct the RAG System, we scraped webpages from ClinicalTrials.gov and categorized them by Trial ID. The retrieved text was then segmented into smaller chunks of 1000 characters in length with 150 characters of overlap. These text chunks were then transformed into vectorized representations and stored in a database. During LLM prompting, we queried the database of text chunks specific to the Trial ID being analyzed using cosine similarity as a metric for measuring the similarity between the text chunk and our desired inclusion criteria. By comparing similarities between the desired inclusion criteria (ex. "inclusion criteria age") with the stored vectors, we returned the top three most similar text chunks. The retrieved text chunks, along with a question about what the clinical trial eligibility were included in the LLM prompt which requested that the model return the desired eligibility criteria. The LLM's response was then saved to a CSV and the process continued until all desired criteria were extracted across all desired trial ids. These responses were then manually evaluated for accuracy and scored as correct or incorrect.

Trial ID	Age Criteria	Age Min	Age Max	ECOG Criteria	ECOG Min	ECOG Max
NCT04221451	The age inclusion eligibility criteria for clinical trial NCT04221451 is: Ages Eligible for Study: 2 Years and older (Child, Adult, Older Adult).	2	NA	The ECOG inclusion eligibility criteria for clinical trial NCT04221451 is not explicitly mentioned in the provided information.	NA	NA

Example of a Response

N = 36	Age Criteria	Age Min	Age Max	ECOG Criteria	ECOG Min	ECOG Max
Pass	36	33	32	33	32	33
Fail	0	3	4	3	4	3
Success	100%	92%	89%	92%	89%	92%

Accuracy of Results

Results

The LLM and RAG system was able to return key inclusion criteria phrases and values accurately succeeding with a rate of 89% or better across the six elements extracted. This indicates LLMs may be a promising tools for NLP related to clinical trials. We were only able to evaluate 36 clinical trials due to time constraints however we plan on increasing our sample size as well as the inclusion criteria being extracted.

LLMs require large amounts of processing power to operate and as such this is a limitation when running one locally. Local LLMs will not be as powerful as cloud-based ones. Also, LLMs are subject to hallucination and continued effort should be made to reduce this impact. Future ideas include introducing more powerful LLMs and using results produced from this study to train a Low-Rank Adaptation (LoRA) model to improve performance.

Research reported in this publication was supported by the National Cancer Institute Cancer Center Support Grant P30 CA168524 and used BISR core