Category: Trial Recruitment & Study Conduct (IIT) – Work in progress

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

S. Pepper, M. Peddapalli, M. McGuirk, I. Ratnayake, D. Pal Mudaranthakam

University of Kansas Cancer Center

1. Background

Large Language Models (LLM) have presented the ability to ingest complex and nuance unstructured human language and return answers relevant to the input; however, they suffer from a context length limitation. To surpass this context length limitation a retrieval-augmented generation (RAG) system can be implemented. By using RAG in conjunction with a LLM, a generalized natural language processing (NLP) framework can be constructed which is able to parse web pages and extract key information in a consistent format.

2. Goals

Implement an NLP system that uses LLMs and RAG to extract key eligibility criteria for clinical trials from unstructured text data on ClinicalTrials.gov. Specific focus is given to inclusion criteria related to age, Eastern Cooperative Oncology Group Score, and disease stage.

3. Solutions and Methods

The Generative Pre-Trained Transformer Quantization version of the Yi-34B LLM was used in conjunction with a vectorized embedding database that was constructed by scraping ClinicalTrials.gov using BeautifulSoup4. This vectorized database was referenced during prompting to provide RAG. Questions about clinical trial eligibility were then asked to the LLM and results were returned.

4. Outcomes

The LLM returns key values and phrases of interest related to the clinical trials. These are stored in CSV file for future use. The accuracy of these data elements has not yet been evaluated, but we will evaluate the outputs and construct a table displaying the errors and successes by data element.

5. Lessons Learned and Future Directions

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 comparing the extracted CSV of data to patient populations to assist with clinical trial navigation.