Leveraging LLMs and RAG for Medical Report Parsing

Project Management and Software Development
for Medical Applications

General Info

Contact Persons: André Mourato, Miroslav Březík
Contact Email: mourato@virtonomy.io, brezik@virtonomy.io

Project Abstract

This project aims to explore the use of Large Language Models (LLMs) for parsing medical reports to extract key clinical information accurately. It will investigate the potential of Retrieval-Augmented Generation (RAG) to enhance the LLM's parsing capabilities by incorporating external, up-to-date medical literature and report parsing. The project will test and compare the effectiveness of using RAG versus a pure LLM approach to determine if RAG significantly improves the model's accuracy and reliability in handling complex medical data. Additionally, the project will explore RAG's application in natural language querying of a NoSQL database for efficient data retrieval. By focusing on domain-specific fine-tuning, query optimization, and explainable AI, this project seeks to streamline information extraction and database interaction for healthcare professionals.

Background and Motivation

Virtonomy GmbH is developing the first web platform for conducting fully data-driven clinical trials with medical devices using virtual patients. Our system is based on clinical scans (CT, MRI), pathological data, and medical device data. With the ability to perform anatomical studies and simulations on large virtual patient cohorts, we enable implant manufacturers to perform faster, safer, and innovative design iterations of their devices. To create these virtual patients, their metadata and comprehensive clinical history are of utmost importance. This involves processing a vast amount of data, including complex medical reports linked to each patient. Therefore, an efficient way to parse these medical reports is crucial to manage, analyze, and utilize the wealth of information that comes to us. Implementing an optimized parsing system will significantly enhance our ability to simulate accurate patient scenarios and streamline clinical trials using our virtual patient platform.

Student's Tasks Description

- Data preparation: the student will gain access to reports from hundreds of patients, understand the key information that needs to be extracted, and annotate several of them accordingly.
- Large Language Model: use a pre-trained LLM (e.g., GPT-4, ClinicalBERT) and annotated medical reports to parse new reports according to the requirements.
- RAG Integration: implement RAG to augment the LLM's parsing capabilities and improve the accuracy of complex or rare medical cases using external medical databases.
- RAG vs. Pure LLM Evaluation: design and conduct experiments to compare the performance of the RAG-enhanced LLM against the LLM alone.
- (Optional) Database Querying: Develop and test a natural language interface for querying a NoSQL database using both the RAG-enhanced LLM and the pure LLM to explore differences in data retrieval accuracy and efficiency.

Technical Prerequisites

- Solid mathematical background.
- Proficiency in Python.
- Good understanding of Deep Learning.
- Familiarity with LLMs (RAG is a plus)
- Basic understanding of Git.

References

- Zabir Al Nazi and Wei Peng (2024). "Large Language Models in Healthcare and Medical Domain: A Review." Informatics, 11(3), 57
- Sun, Y., Wang, S., Li, Y., Feng, S., Tian, H., Wu, H., et al. (2020). "ERNIE 2.0: A Continual Pre-Training Framework for Language Understanding." AAAI 34: 8968–8975.
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., et al. (2020). "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks." arXiv