University of New Haven

TAGLIATELA COLLEGE OF ENGINEERING

MASTER OF SCIENCE IN DATA SCIENCE

Natural Language Processing (DSCI-6004-01)
Final Project Presentation

-By :
Rohan Reddy Marri &
Devakinandan Pentyala

Project Overview

- Develop a **Retrieval-Augmented Generation (RAG)** system to answer **biomedical research** questions using scientific articles.
- Combine information retrieval from **PubMed abstracts** with open-source LLMs

i. Llama-3
ii. Mistral-7B
iii. Phi-3

- Retrieve relevant biomedical passages, provide them as context to the LLM, and generate **evidence-based, citation-supported answers**.
- Focus on improving **factual accuracy**, **reasoning quality**, and **faithfulness** to retrieved evidence.
- Evaluate and compare the performance of multiple LLMs in terms of **accuracy**, **clarity**, and **hallucination control**.
- Advance **trustworthy AI** tools for biomedical research and knowledge discovery.

Project Objectives

➢ Objective:

- Develop a domain specific Retrieval Augmented Generation(RAG) pipeline to efficiently retrieve and summarize biomedical research information for factual question answering.

➢ Approach:

- Integrate PubMed abstracts and biomedical literature as the retrieval corpus.
- Build a custom retriever pipeline that ground model responses in verified scientific evidence.
- Implement & compare 3 open-source LLMs – Llama-3,Mistral-7B and Phi-3 within the same RAG framework.

➢ Evaluation Dimensions:

- Factual Accuracy : Measure corrected of biomedical facts against gold-standard datasets.
- Reasoning Ability:Assess model's ability to infer logical and evidence-based conclusions.
- Hallucination Rate: Identify and reduce unsupported or fabricated responses.

➢ Optimization Focus:

- Fine-tune and optimize smaller LLMs for factual biomedical QA to achieve high accuracy with low computational cost.

➢ Expected Outcomes:

- Deliver a reliable, interpretable RAG system for scientific text understanding.
- Provide comparative insights into model reliability, reasoning patterns, and trustworthiness in biomedical applications.

Why is this project worth doing?

- Biomedical literature is growing exponentially, making it difficult for researchers to quickly find relevant and high-quality evidence.
- Traditional search tools often return incomplete or non-contextual information, slowing down research progress.
- Large Language Models (LLMs) frequently generate inaccurate or hallucinated responses when handling factual biomedical questions.
- Retrieval-Augmented Generation (RAG) systems ground responses in verified scientific evidence, improving factual accuracy and trustworthiness.
- This project supports the development of reliable, evidence-based AI tools that enhance information retrieval in healthcare and biomedical research.
- Contributes to open, transparent AI research aimed at improving knowledge accessibility for scientists, clinicians, and policymakers.

Review of related work

- Retrieval-Augmented Generation (RAG) combines information retrieval and language generation to produce fact-grounded answers (link).
- REALM (Guu et al., 2020) introduced retrieval during pretraining, allowing models to access external knowledge efficiently.(link)
- Dense Passage Retrieval (DPR) by Karpukhin et al. (2020) improved retrieval precision using dual-encoder dense embeddings.(link)
- Fusion-in-Decoder (FiD) architecture (Izacard & Grave, 2021) enhanced reasoning by encoding multiple retrieved passages separately before generation.(link)
- Modern RAG systems balance retrieval accuracy, generation quality, and computational efficiency for scalable QA tasks.(link)
- PubMed Abstracts: Comprehensive biomedical research literature from PubMed.(link)
- Llama 3 (by Meta AI): A large-language-model family (8B–405B parameters), instruction-tuned for reasoning and generation.(link)
- Licensing & openness note: Llama-3 is released with weights accessible, but its license has been critiqued for not meeting full “open-source” OSI definition.(link)
- Mistral-7B — High-performance open-weight model optimized for inference efficiency.(link)

Approach / System Architecture

1. Data Source: PubMed abstracts or Kaggle 'PubMed 200k RCT' dataset.

2. Preprocessing: Text cleaning and chunking (512-token chunks).

3. Embedding: Using sentence-transformers (MiniLM or Bio-Sentence-BERT).

4. Vector Store: FAISS or ChromaDB for document retrieval.

5. RAG Inference: Retrieve top-k docs and feed into LLM to generate evidence-based answer

6. Models: Llama-3-8B, Mistral-7B, Phi-3-mini.

7.Outcome:Reliable, evidence-grounded biomedical question answering system with reduced hallucination and improved factual accuracy.

Deliverables

- RAG System Code (GitHub)
- Proposal Slides
- Ten or more Biomedical Questions Dataset
- Evaluation Report comparing Llama-3, Mistral-7B, Phi-3
- Research-style Paper or Report
- Expected Outcome: Open-source biomedical RAG assistant with detailed model comparison.

Evaluation Methodology

- Factual Accuracy: Measures the correctness of biomedical information generated by the model against verified sources.
- Faithfulness: Evaluates how well the generated answer aligns with the retrieved evidence; ensures model doesn't invent facts.
- Hallucination Rate: Tracks the frequency of unsupported or fabricated claims in model outputs.
- Clarity & Coherence: Assesses human readability and logical flow of the generated answers.
- Retrieval Precision: Measures the relevance of top-k retrieved documents to the user query; higher precision improves grounding and reduces hallucinations.
- Comparative Evaluation: Outputs of Llama-3-8B, Mistral-7B, and Phi-3-mini are compared for each question across all metrics. Helps identify the best model for factual, trustworthy, and readable biomedical QA.

THANK YOU !!

University of New Haven

TAGLIATELA COLLEGE OF ENGINEERING