Weekly Report: LLM + RAG Model Project

Edward Yang

April 17, 2025

Week Overview: 070425 - 130425

- **Project Name:** LLM + RAG Model Development for Disease Diagnosis
- **Goals:**
  - **Goal 1:** Create Document preprocessing functions and Embedding Model for the Disease Dataset
  - **Goal 2:** Use vector search for retrieving relevant documents
  - **Goal 3:** Integrate the RAG model with the LLM for generating responses
  - **Goal 4:** Test the model with sample queries
  - **Goal 5:** Compare the model's output accuracy against the Non-RAG model (LLM-only). This can be done by creating a fine-tuning model to evaluate diagnostic accuracy of the two models, trained on the hierarchical ICD-10 Disease Dataset, where partial matches at different levels of specificity have different clinical value (similar to the method in the Medfound paper).
  - **Goal 6:** Create a report on the model's performance and potential improvements

Progress

- **Completed Tasks:**
  - **Task 1:** Created data_loader.py and other tools for loading and preprocessing the dataset.
  - **Task 2:** Started on the embedding model, using the Sentence-BERT model.
  - **Task 3:** Used ChromaDB as a vector database for storing the embeddings.
- **Ongoing Tasks:**
  - **Task 1:** Creating semantic search pipeline
  - **Task 2:** Implementing LLM for generating responses
  - **Task 3:** Fine-tuning the model for evaluating diagnostic accuracy
  - **Task 4:** comparing model accuracy against non-RAG model

Challenges

- **Challenge 1:** Description and potential solutions
- **Challenge 2:** Description and potential solutions

Next Steps

- **Planned Tasks for Next Week:**
  - **Task 1:** Description
  - **Task 2:** Description
  - **Task 3:** Description
- **Milestones to Achieve:**
  - **Milestone 1:** Description
  - **Milestone 2:** Description

Additional Notes

Include any additional notes, observations, or comments.