EE/CprE/SE 491 WEEKLY REPORT 4

10/3/2024 – 10/10/2024

Group number: 35

Project title: Universal Response Engine: LLMs for Good

Client &/Advisor: Ahmed Nazar and Mohamed Selim

Team Members/Role:

Abrahim Toutoungi - Stakeholder Liaison

Gabriel Carlson - Communications Manager

Halle Northway - Meeting Coordinator

Brianna Norman - Project Deliverables Manager

Ellery Sabado - Timeline Coordinator

Emma Zatkalik - Assignment Manager

Weekly Summary

The group was tasked with implementing a RAG (Retrieval-Augmented Generation) to improve the accuracy of our implemented LLMs. We were to use llama3.1 or mistral, langchain, huggingface embeddings, and FAISS. The dataset used for the RAG LLM should be applicable to our project and ideally produce a response that is accurate and uses the data we fed to it. No significant changes were made to our project during this week.

Past Week accomplishments

- Have RAG (retrieval-augmented generation) implementations for Chroma and FAISS
- Added datasets previously researched to RAG vector store
- Began testing Ollama (llama3) to see if it fits our needs
- Reached out to mental health organization (Red Shirt Foundation) and got resources from them

Pending Issues

- Ensuring datasets contain the desired information while staying within the project scope.
- We are not subject matter experts on the topics we are collecting datasets for (see: medical assistance). How to ensure that given sources are reliable?

Individual Contributions

Name	Individual Contributions	Hours this week	Hours cumulative
Abrahim Toutoungi	- Implemented RAG using Ollama, huggingface, and langchain. - Tried chroma and FAISS	7	20

	- Used a first aid pdf locally and asked LLM about it. - Researched more about making LLMs - Spoke with The Red Shirt organization to get better resources relating to suicide prevention		
Gabriel Carlson	- Worked on implementing RAG using Llama, ChromaDB, and langchain - Added additional embedded documents like medical dataset for conversational retrieval chain - Tested conversational chain - Updated the Gantt chart - Spoke with The Red Shirt organization to get better resources relating to suicide prevention	7	18
Halle Northway	- Working off RAG implemented last week, created a new RAG using HuggingFace, Ollama LLM model Llama3, Langchain - Used RAG to read from local pdf, txt, and csv files and provide relevant responses - Requested VM	8	19
Brianna Norman	- Researched more about RAG usage - Looked through team datasets - Altered implemented RAG from previous week	3	17
Ellery Sabado	- Created a RAG with hugginface, Ollama, Fasis, and langChain to read JSON and PDFs. (Also tried ChromaDB) - Tried to implement the Coverational retrieval chain to the RAG. - Used the dataset that we researched last week and used it in the RAG	7	19
Emma Zatkalik	- Explored more datasets on huggingface, specifically csv and json formats for medical chats - Implemented RAG LLM on Google Colab and VScode - Got txt, csv, json, and pdf datasets working, some local, some from huggingface	6	18

Comments and extended discussion (optional)

N/A

Plans for upcoming week

- Compiling our RAG implementations into a unified RAG file for testing
- Preparing for lightning talk next thursday
- Request VM for hosting with requirements:
    - GPU for LLMs
    - Ubuntu 20.04 or 22.04
    - Atleast 512Gb of storage
    - Atleast 16Gb of RAM
    - GUI
- Research Q Laura for optimization/minimization of LLMs
- Stretch Goal: Research fine-tuning LLMs

Summary of weekly advisor meeting

Notes

- Datasets look good

Tasks

- Implementing RAGs for the datasets we found
- See if we can extract data from the dataset
- Then we are going to try fine-tuning the LLMs
- Report back on the results and see how it went
- Then work on embedding pictures/videos
- Ahmed uses Google Collab to train large datasets
- Use conversational retrieval langchain
- Stick to Mistral and LLama How do people use videos or pictures with LLMS
- Process non-text source as is, using an embedding model that supports that specific source
    - Clip from OpenAI takes images and embeds them
    - Guava from Intel takes videos and embeds them
    - Using these can be expensive, or you need a subscription to use
- Take the input and do the work that the normal models do
    - Take the image and encode it to base64
    - Take audio and encode/transform to encoding using FFT
    - Take location data and encode it to text
    - Can take a video and transcribe the audio to text, then process the text descriptions for chunks to a set of frames