EE/CprE/SE 491 WEEKLY REPORT 2

9/19/2024 – 9/26/2024

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 overall objective of this week was to continue learning more about large language models and how to implement them. More specifically, how to run them locally on our computers and utilize Python to adjust certain settings on the model, some settings include temperature, which represents how creative the model should be with its response, and top_p, which specifies what percentile the model should be choosing its answers from, and repetition_penalty, which represents how much the model should avoid using the same words over and over again in its response. No significant changes were made to the project. The scope has been explained to be flexible with our preferences.

Past Week accomplishments

- Continued LLM Research
  - Introducing how RAG plays a role
- LLM Experimentation
  - Running LLMs locally
- Use Case Brainstorming
  - Narrowing the scope of our project
- User Interface Brainstorming
  - Using other models for reference, but focusing on the our goal of being user-friendly and for good

Pending Issues

- Continuing to figure out how to run an LLM programmatically
- Continuing to learn more about LLMs (narrow in on a design)

Individual Contributions

Name	Individual Contributions	Hours this week	Hours cumulative
Abrahim Toutoungi	- Kept researching LLM stuff - Played around with some figma to model front end - Read about sentiment analysis - Tried running LLM using hugging face	5	9
Garbiel Carlson	- Installed and Researched: - Langchain - Langserve - Structured Output - Conversational Retrieval Chain (uses RAG) - Tested HuggingFaceEndpoint using langchain_huggingface - Started testing a Conversational Retrieval Chain using huggingface's API and manually input context, prompts, and embeddings.	4	8
Halle Northway	- Continuing LLM research, especially on how it is planned into application development cycle - Ran meta-llama model with LangChain LLM framework and experimented with its responses - Started brainstorming use cases for project	3	7
Brianna Norman	- Continued LLMs research, worked a bit in LLM studio with llama 3.1, getting into HuggingFace - Studied organizational logic for resource pages, considered how to best present information to users - Looked into RAG	4	8
Ellery Sabado	- Ran Ollama, specifically llama3.1, on my device - Trying to use HuggingFace to run an LLM on my device - Research more about LLMs, RAG, and LangChain	4	8
Emma Zatkalik	- Continual LLM research - Trying to run LLM (llama3.1) with python - Reading about RAG	4	8

	- Observing how changing LLM parameters change its response		

Comments and extended discussion (optional)
N/A

Plans for upcoming week

- Collecting good datasets that are applicable to our project
    - IEEE -> GitHub repositories -> hugging face -> kaggle
- Making at least one RAG
- Continue testing with langchain

Summary of weekly advisor meeting

In our meeting, we began by recapping everyone's week and what they have been focusing on, any issues they encountered, and any other additional information each member wanted to add. Then, after our recaps, Ahmed answered questions and walked us through how document embedding and retrieval works using langchain. We went through the steps from acquiring the document, processing the document into text files, embedding the generated text files, and adding them to the vector store. We also discussed how these vector stores are used when prompting the LLM. We discussed some of the main focuses of the project functionality and what datasets we wanted to include. Lastly, we talked about our todo items for this upcoming week.