EE/CprE/SE 491 WEEKLY REPORT 5

10/10/2024 – 10/17/2024

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

This week's goal was to explore more into the topic of fine-tuned LLMs. We also were to collect a more official list of datasets that we will use for our model and continue to work on overarching project tasks, like mock-ups, more research on LLMs, getting a VM, and collecting a requirements list of needed libraries or packages. No significant changes were made to our project during this week.

Past Week accomplishments

- Received VM from ETG
- Experimented with the fine-tuning approach to LLMs
- Researched sentiment analysis in more depth
- Worked on a basic frontend

Pending Issues

- N/A

Individual Contributions

Name	Individual Contributions	Hours this week	Hours cumulative
Abrahim Toutoungi	- Worked on frontend sample - Researched about cleaning up datasets - Looked into resources acquired last week - Worked on lightning talk 3	6	26

	- Looked into NLTK sentiment analysis		
Gabriel Carlson	- Created requirements list and zip for deployment on VM - Researched using FastAPI and langserve to serve chains/runnables on RESTful API - Worked on debugging langserve conversational retrieval chain with huggingface models	6	24
Halle Northway	- Got group project VM approved and created - Created fine-tuning LLM sample using Llama3 and mental-health-datasets github repo		26
Brianna Norman	- Experimented with fine-tuning LLM locally in vscode using LoRA - Looked into implementing conversational datasets with a RAG	5	22
Ellery Sabado	- Research more about Fine-tuning - Learned more about LoRA and how it works with a finetuning model. - Implemented a fine-tuning model through huggingface, LoRA, and a conversational dataset(ex. IMDB reviews)	6	25

Emma Zatkalik	- Researched about finetuned LLMs
	- Got a simple finetuned LLM working on Google Colab
	  - Unsloth, Llama3.1 8B, SFTTrainer, huggingface
	  - peft, bits and bytes, QLora
	  - Looked at more style based datasets
		6	24

Comments and extended discussion (optional)
N/A

Plans for upcoming week
- Continue implementing fine-tuning
- Access the VM and do the initial setup
- Collaborate with each other on the VM
- Think more about the design of the UI

- Gather metrics for LLMs (time to train, how long prompts take on different OS, how our experiments compare to the VM performance.)

Summary of weekly advisor meeting

Next steps

- Putting our RAG experiments together
    - Figure out documents we want to use for RAG
    - Send to Amhed by Monday Night
- Lightning talk
    - Send to Ahmed for review
- Start messing around with fine-tuning
    - LLM will adjust responses to match dataset responses and presentation style (not giving new information)
    - Recommends using conversational datasets
    - won't get as good of results as RAG, will be slower but the response will be more
    tailored to that dataset's response
- Request a VM
    - GPU
    - Ubuntu 20.04 or 22.04
    - Atleast 512 or 1TB storage
    - 16 or 32 gb Ram
    - GUI
- Q Laura - Optimization and minimization technique
    - We will be using 8 bit or 4 bit
    - Depends on the model which one will be better
- IF USING GOOGLE COLLAB
    - DONT SLEEP COMPUTER
    - DONT USE SAFARI
    - DONT REFRESH

Thinking Ahead

- Future dependancies/libraries to use
- bitsandbytes
- Accelerate
- Peft