EE/CprE/SE 491 WEEKLY REPORT 6

10/17/2024 – 10/24/2024

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

Our work this week was originally going to focus on setting up our group VM with an ideal collaborative version of the work we have done so far to setup an LLM for our project. Upon discovering our VM was not made to the requested specifications, we pivoted to a focus on continuing to understand the inner workings of LLMs and experimenting with fine-tuning and training our models. No significant changes were made to our project.

Past Week accomplishments

- Narrowing out dataset coverage
- Confirming functions of VM from etg
- Fine tuning LLMs continuation of last week

Pending Issues

- N/A

Individual Contributions

Name	Individual Contributions	Hours this week	Hours cumulative
Abrahim Toutoungi	- Explored vm - Tried making datasets with chatgpt - Worked on implementing fine tuning in google colab using qlora	6	32
Gabriel Carlson	- Installed dependencies for RAG on vm - Updated VM's software	7	31

	- Researched using QLorA and peft to fine-tune LLMs - Fine-tuned the Llama 3.1 8B model using unsloth and medical conversational dataset - Documented testing of tuned LLM with basic medical queries		
Halle Northway	- Exploring VM, contacting ETG about resolving issues - Experimented with constraints for finetuning LLM responses and different datasets - Researching different datasets, particularly about emergency response situtations	7	33
Brianna Norman	- Exploring VM, issues pending - Making LLM from start for better understanding - Working through implementing fine tuning - Creating conversational datasets to train on	10	32
Ellery Sabado	- Tried to FineTune through sloth and conda and Ollama - Create a simple conversational dataset - Used that conversational dataset and used on previous sentimental analysis FineTuning (Used huggingFace models)	6	31
Emma Zatkalik	- Trying to get fine tuning to work locally on my computer in VScode - Reformatting dataset on medical questions and answers that we can use for training our model - Finetuning in Google Colab - Peft, LoRA, bits and bytes, huggingface datasets, llama3.1 8B - Pushing fine-tuned model to huggingface and using it	9	33

Comments and extended discussion (optional)

N/A

Plans for upcoming week

- Install necessary applications to VM
- Setup LLM on VM once re-provisioned
- Compare results from fine-tuning and compile a collaborative LLM

Summary of weekly advisor meeting

Next Steps:

- If an LLM returns a resource, look at where the resource was found. LLM should cite sources for RAG?
    - Industry panel: How do you know what the llm is giving is true or credible
- Create mini server for VM for models so we arent running multiple at the same time
- Find cases that you like in existing data sets, copy them and merge them with other data sets
- Tensorboard or wndb
    - Take what is happening internally and output to graph for fine-tuned
- **Work on getting an LLM fully fine tuned on one dataset**
- Week of 18th of November
    - Work on presentation
- Week of 2nd
    - Start practicing presenting
- Week of 9th
    - Presentation