PROGRESS REPORT

Group 9: Doruk Ozar, Luke Napolitano, Vishwanath Guruvayur

So far, the majority of our work has focused on conducting an extensive literature review on **LoRA**, **Q-LoRA**, and other advanced **parameter-efficient fine-tuning (PEFT)** methods, to understand what they are and their applications in LLM fine-tuning. We have also explored performance and safety concerns associated with LLMs, like hallucination and content moderation, along with mitigation strategies to ensure responsible AI usage. The fine-tuning lecture in class was very insightful, sparking several ideas about how to approach our next steps.

We initially identified benchmark datasets in multiple domains, including **Finance**, **Mathematics**, and **Conversational Domains**, and evaluated these to determine which dataset best aligned with our goals for model performance and safety evaluation. After selecting the OpenMathInstruct-2 Math dataset, we preprocessed it to ensure that it is clean, structured, and properly labeled for use in fine-tuning experiments. Additionally, we have chosen newer versions of open-source LLMs, including **LLaMA**, **Quen**, and **Gemma**, starting with LlaMA for experimentation. We are currently running **baseline performance evaluations** to establish initial benchmarks. These baselines will help us assess how well the models perform on the selected task before any fine-tuning is done.

Our next steps involve implementing **LoRA**, **Q-LoRA**, and other PEFT techniques into the models. We plan to conduct **initial training runs** on a subset of the data to assess training efficiency, with the expectation that LoRA and Q-LoRA will improve resource efficiency compared to traditional full-model fine-tuning methods. Once the initial training runs are complete, we will expand fine-tuning to larger datasets and further optimize hyperparameters to enhance both model performance and safety alignment. The evaluation will include key metrics such as **accuracy**, **stability**, and **bias mitigation**. We have been doing a lit review of how we would need specific evaluation metrics for domain-specific data. For example, while studying about HELM, we figured that it might not be the best option for the Mathematics dataset and we should focus on symbolic accuracy and error rate in general.

Our final aim is to compare the performance of the fine-tuned models with traditional full-model fine-tuning approaches to analyze trade-offs between **resource efficiency** and **model quality**. Finally, we will document our findings, summarize the challenges encountered during the experimentation process, and create visualizations and comparative metrics to include in the final project report. This project has already helped us understand the methods of further polishing the transformer architecture by adding adaptive layers and how research is being done to make the training process more computationally efficient and affordable. **Hands-on experimentation with these methods** and comparing them on specific domain data should help us understand the core idea behind these methods better and how we can decide on the architecture of LLMs in real-world applications.