Group 9: Luke Napolitano, Doruk Ozar, Vishwanath Guruvayur

We are considering a project on the fine-tuning of LLMs using LoRA and Q-LoRA to improve model safety and performance. After the success of DeepSeek-R1, it's been shown that when it comes to LLM's, optimizing parameters might surface performance gains that scaling questions have avoided. The goal is to evaluate both performance and safety outcomes in adapting these models to specific tasks. LoRA is a method for reducing the computational complexity of model training by introducing low-rank matrices into the attention layers of transformers, allowing efficient fine-tuning with fewer parameters. Q-LoRA, further optimizes the training process while preserving model quality.

By applying these methods to open-source models, we hope to assess how efficiently and effectively such models can be fine-tuned to regulatory contexts while maintaining safety standards. The research will focus on performance metrics such as task accuracy, response relevance, and model stability, alongside safety considerations, including bias mitigation, harmful output detection, and ethical concerns related to model behavior in real-world applications.

The project will involve experimenting with multiple open-source models, comparing the results of LoRA and Q-LoRA fine-tuning against traditional full-model training approaches. This will provide insights into the trade-offs between resource efficiency, performance, and safety. We hope this project can add to the growing body of LLM optimization projects being conducted.