Introduction

In this lab, we explored the concept of QLoRA (Quantized Low-Rank Adaptation), a technique that allows efficient fine-tuning of large language models (LLMs) by combining 4-bit quantization and LoRA (Low-Rank Adaptation). This is particularly useful when working with limited GPU memory (e.g., 16GB), enabling us to train 7B models effectively.

Experiment Summary

We fine-tuned a quantized model (e.g., `bloomz-560m` or `Meta-Llama-3-8B`) using the PEFT library and Hugging Face's `awesome-chatgpt-prompts` dataset. The training involved using 4-bit quantization (NF4) and LoRA configuration with `r=16`, `alpha=16`, and dropout to avoid overfitting.

Observations

- The **pretrained model** responded well but lacked task-specific behavior.
- After **fine-tuning**, responses aligned more closely with the dataset — motivational, structured, and aligned with coaching tone.
- Only 50 examples and 5 epochs were enough to noticeably shift the model behavior.

Failure Cases

- Some prompts led to **repetitive outputs** or **hallucinations**, especially vague prompts.
- The model struggled with factual or logic-based queries unrelated to motivational tasks — which is expected due to the training data nature.

Key Learnings

- QLoRA significantly reduces memory usage and allows fine-tuning of large models on commodity hardware.
- LoRA parameters (`r`, `alpha`, `target_modules`) impact the model behavior and must be adjusted based on architecture.
- Proper prompt engineering and dataset alignment are crucial for successful fine-tuning.