Background

- Deepseek R1 demonstrated the potential of pure RL for finetuning pretrained LLMs
- They update all weights of their model, which can be computationally expensive
- Other methods of training models like supervised finetuning have been shown to retain most of their performance without having to update all parameters

Objective

- We want to explore whether we can achieve similar RL performance without updating all of the model weights
- Some methods to explore include LoRA-like methods, freezing all but the first few or last few layers, or greedily selecting layers to update weights on during training

Methodology

- Start with modifying replication of R1 to only train final layers. Currently planning to also use GRPO and a similar reward function as in the paper, with formatting rewards and correctness rewards