Project Overview

Background

- **Deepseek R1** demonstrated the potential of pure RL for fine-tuning pretrained LLMs
- **All weights updated** in their model can be computationally expensive
- **Other training methods** like supervised fine-tuning retain most performance without updating all parameters

Objective

- Explore achieving similar RL performance **without updating all model weights**
- Investigate methods such as:
  - LoRA-like approaches
  - Freezing all but the first few or last few layers
  - Greedily selecting layers to update during training

Methodology

1. **Modify R1 Replication**:
   - Train only final layers
   - Use GRPO and similar reward functions (formatting and correctness rewards)
2. **Start with a Small Base Model**:
   - Example: Qwen 0.5B for ease of use
3. **Experiment with RL Algorithms**:
   - Compare DPO, PPO, GRPO, etc.
4. **Utilize Common Datasets**:
   - Countdown tasks for R1-like models
   - Explore curriculum learning with progressively harder math datasets

Progress Report

- **R1 Replication**:
  - Running on Colab
  - Exploring GPU rental options if needed
- **Familiarization**:
  - TRL and VERL frameworks
- **Current Tasks**:
  - Modifying `GRPOTrainer` from TRL to only modify final layers

Next Steps

1. **Complete Modification**:
   - Finish modifying `GRPOTrainer` to update only final layers
   - Benchmark the modified model
2. **Benchmarking**:
   - Evaluate on tasks like AIME, MMLU, etc.
3. **Future Experiments**:
   - Implement curriculum learning with harder math datasets