Parameter-Efficient Reinforcement Learning

Justin Qiu

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 withoutupdating 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

Will potentially focus on program synthesis or continuous self-improvement with things like math questions of increasing difficulty or some other kind of curriculum learning. Will finish some quick experiments before deciding

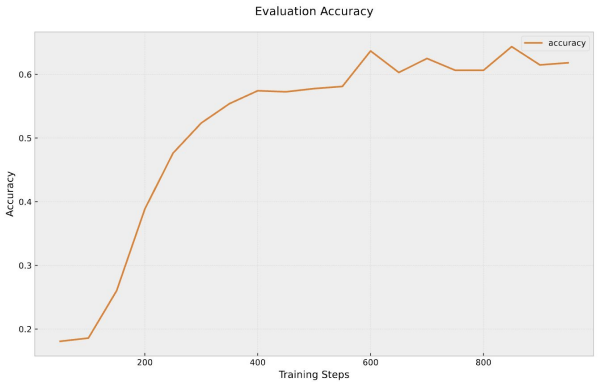
Methodology

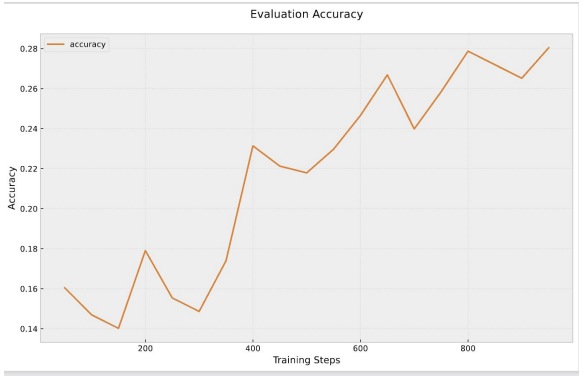
1. Modify R1 replication to use parameter-efficient methods and benchmark
    a. Discussed above
2. See if I can get the model to learn simple program synthesis through RL
3. Still debating which idea I should focus on (see previous slide)

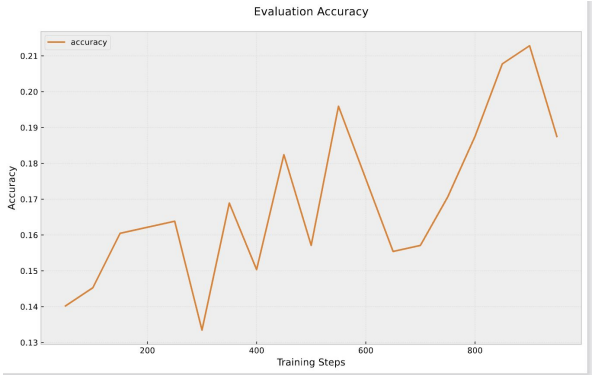
Results So Far

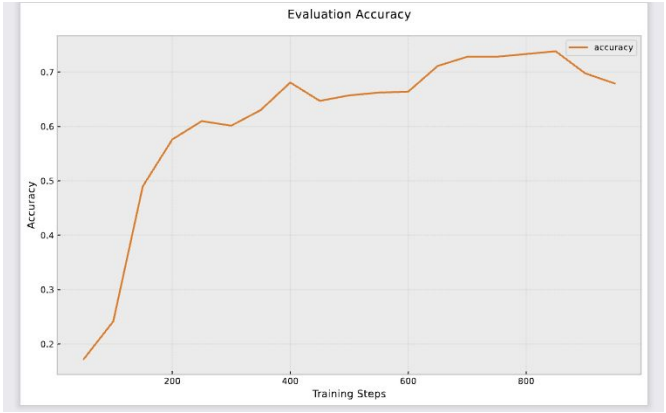
Top left: freeze all but first five; Top right: freeze all but last five

Bottom left: freeze all but last layer; Bottom right: default (no freezing)









Results so far

Freezing all but the first five layers is the only paradigm that works reasonably well

I am currently running an experiment seeing if the setup can effectively learn a simple program synthesis instance such as

f(1) = 1

f(2) = 4

f(3) = 9

etc.

Next Steps

1. Finish the simple program synthesis experiment
2. Set up LORA and run that as an experiment
3. Set up progressive layer freezing experiment
4. Overall, need to confirm overall direction of project. I feel like I’m doing experiments that aren’t super well connected and still kind of exploring, but time is short