Citizen-Centric Artificial Intelligence Systems

Post-Trained Language Models as Agents in Sequential Games

Jim Dilkes
Supervisors: Vahid Yazdanpanah, Sebastian Stein

Background

Achieving both robust decision making and planning as well as broad generalization in autonomous agents remains a challenge.

While large language models (LLMs) have diverse world knowledge, they struggle with effective planning. This can be compensated for by using very large models generating long reasoning chains, but it is costly and slow.

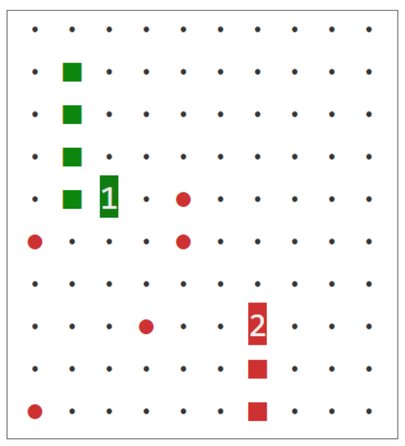
Our work explores a more efficient alternative: using reinforcement learning to directly improve decision making of LLM agents in sequential games.

Aims

Post-train an LLM to learn how to effectively plan and act in a sequential decision-making environment.

Utilise the LLM's inherent world knowledge to generalize learned strategies to previously unseen scenarios with semantically meaningful descriptions.

Framework



Snake environment

At each episode step, we prompt the LLM with instructions and a text representation of the current state (coordinates of entities and a 2D diagram of characters), then extract an action from its response.

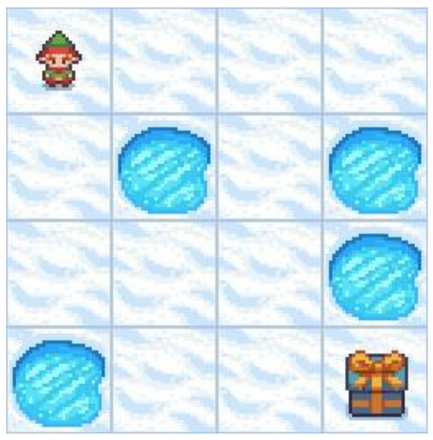
After episode termination, we calculate its advantage by comparing the *total* episode reward to other trajectories generated from the same starting state.

The LLM's parameters are updated by gradient descent on the GRPO loss [1].

We train and evaluate our agents on two environments…

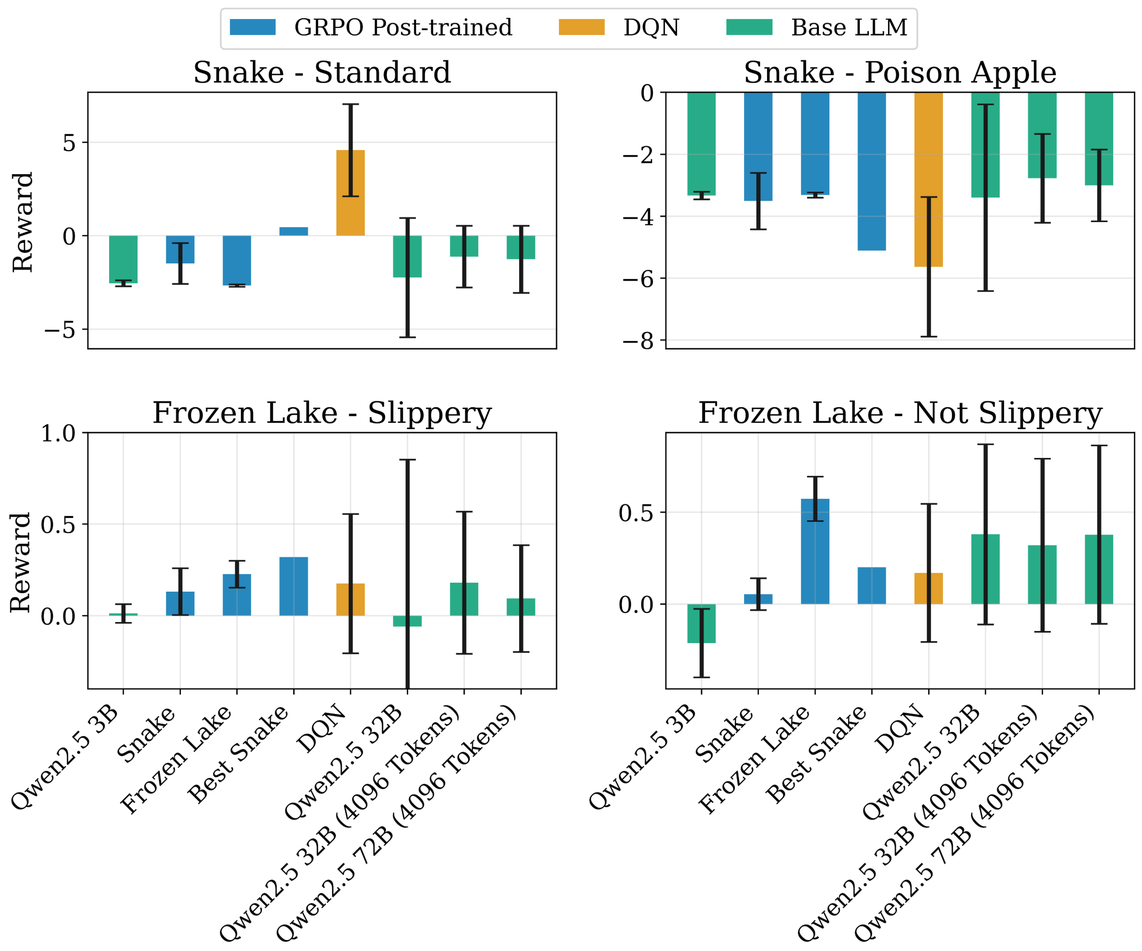
Snake: A dynamic environment where the agent must collect rewards while avoiding obstacles and a competing agent.

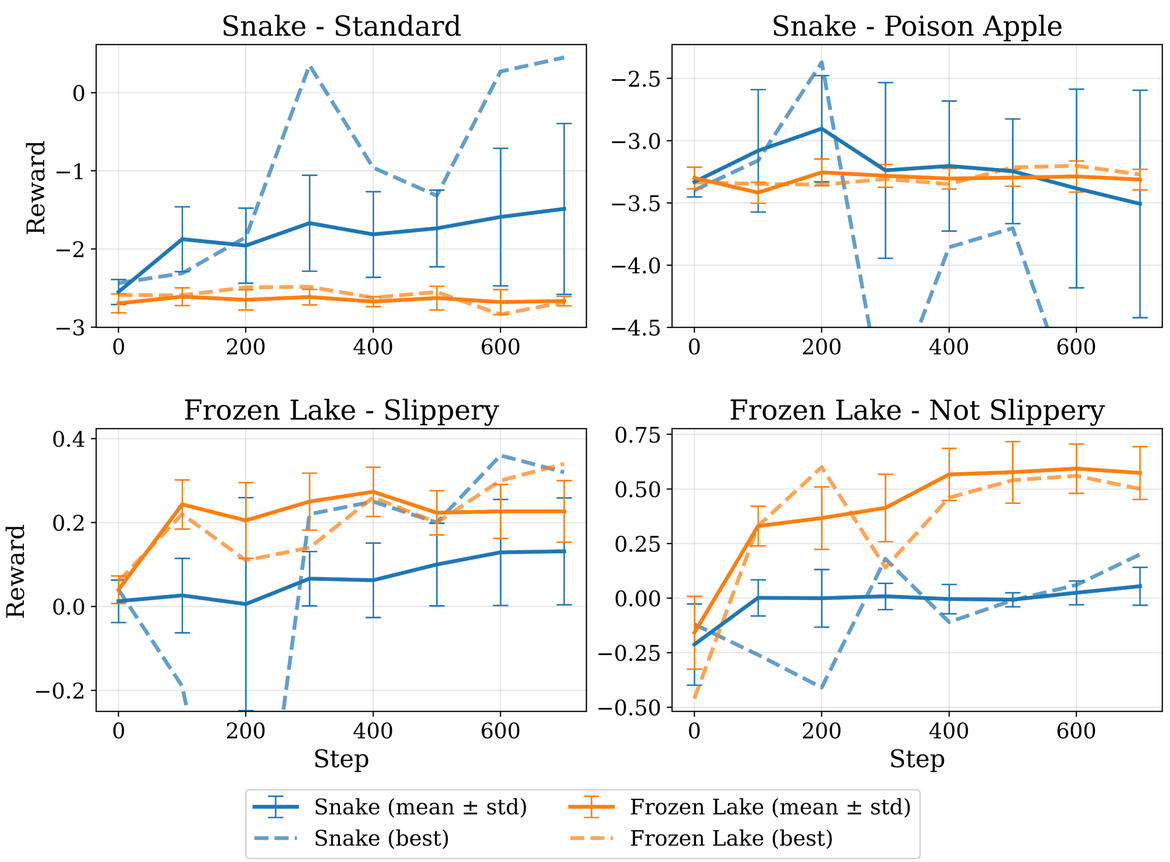
Frozen Lake: A stochastic environment requiring navigation across a grid with probabilistic movement.


Frozen Lake environment [2]

Preliminary Results

- Post-training LLMs with GRPO improves performance on sequential games
- Small LLMs (3B parameters) post-trained with GRPO outperform larger base models





- A baseline Snake agent using a Deep Q-Network significantly outperforms the post-trained agent on Snake
- The best post-trained models demonstrate better cross-task generalisation capabilities than the DQN agent

Future work will explore more complex interactive scenarios, such as those where human instruction or inter-agent communication shape the agents aims and understanding of the world.

[1] Zhihong et. Al. *DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models* (2024)

[2] Towers et.al. *Gymnasium: A Standard Interface for Reinforcement Learning Environments* (2024), image: gymnasium.farama.org

UK Research and Innovation
University of Southampton
j.dilkes@soton.ac.uk