Proposal: From PPO to Preference-Based Optimization in RL-GPT

Team#: Minjae Kim (20253101), Jaewon Cho (20254595), Jeehye Na (20254242)

We propose to reproduce and extend the RL-GPT framework, a method that combines Large Language Models (LLMs) with reinforcement learning for long-horizon embodied control. RL-GPT consists of two interacting agents:

- a **Slow Agent** that decomposes high-level tasks into sub-actions, and
- a **Fast Agent** that executes each sub-action via code or RL.

The original work trains the RL path with **PPO** using two scalar rewards: a CLIP-based semantic alignment reward and a distance-based shaping reward. However, PPO relies on hand-tuned rewards and a value model, often leading to instability, reward hacking, and limited sample reuse.

To address these issues, we will replace PPO with **Direct Preference Optimization (DPO)** and **Group Relative Preference Optimization (GRPO)**—algorithms that learn directly from **relative trajectory preferences** rather than numeric returns.

Approach

Step 1. PPO baseline: Reproduce RL-GPT's PPO training on small MineDojo tasks (e.g., harvesting, crafting) to establish a baseline.

Step 2. Preference dataset: Collect multiple rollouts per sub-action under identical conditions. Compute composite CLIP + Distance scores and normalize within each group to derive pairwise (DPO) or group-wise (GRPO) preferences.

Step 3. Preference-based training: Train policies using DPO and GRPO objectives, which remove value estimation and optimize trajectory likelihoods with KL regularization.

Step 4. Empirical analysis: Compare PPO, DPO, and GRPO in terms of: Sample Efficiency, Stability, Alignment

Computational Considerations

Experiments will focus on small MineDojo or grid-based environments to remain GPU-feasible.

Significance

This work bridges **LLM alignment methods (DPO/GRPO)** and **hierarchical RL**, offering a preference-based alternative to reward-driven training that enhances **robustness, efficiency, and interpretability**.