Phase I: Model-Guided Reflection Trajectory Generation

MCTS Search

Revision Trajectory (τ^r)

The agent is stuck in an **infinite loop** of trying to execute an **invalid action**, which is not helpful in solving the task.

○ Action in Initial Trajectory (τ^i)

● Action in Bad Trajectory (τ^b)

● Action in Good Trajectory (τ^g)

◉ Transition Point

Reward = 0   Reward = 0.7

Phase II: Iterative Self-Training with Revision Trajectories

Revision Trajectory (τ^r)

Good Trajectory (τ^g)

$t = t'$

$\alpha < \text{Reward}\,(\tau^g) = \text{Reward}\,(\tau^r)$

loss function

$$L(\theta) = \mathbb{E}_{(\tau^g, u) \sim \mathscr{D}_{\text{Good}}} \left[\log \pi_\theta(\tau^g \mid u) \right] +$$

$$\mathbb{E}_{(\tau^r, u) \sim \mathscr{D}_{\text{Revision}}} \left[\log \pi_\theta(\text{rs}, \tau^g_{(t > t')} \mid u, \tau^b_{(t \leq t')}) \right]$$

Iterative Supervised Fine-Tuning