Agentic Reinforcement Learning with Implicit Step Rewards

Xiaoqian Liu[1,2,3], Ke Wang[2], Yuchuan Wu[2], Fei Huang[2], Yongbin Li[2], Jianbin Jiao[1], Junge Zhang[1,3]

[1]University of Chinese Academy of Sciences [2] Tongyi Lab

[3]Institute of Automation, Chinese Academy of Sciences

ICLR

Background

■ RL for LLM

- Single-turn RL vs. Agentic RL
  - From static, single-turn tasks to **dynamic, interactive environments**;
  - From passive generators to **autonomous agents**.

- Particular challenges in agentic RL:
  - Rewards are typically sparse and delayed.
  - Trajectories are long and non-Markovian in token level.
  - Environments are non-stationary, open-ended and often with unverifiable rewards.

Motivation

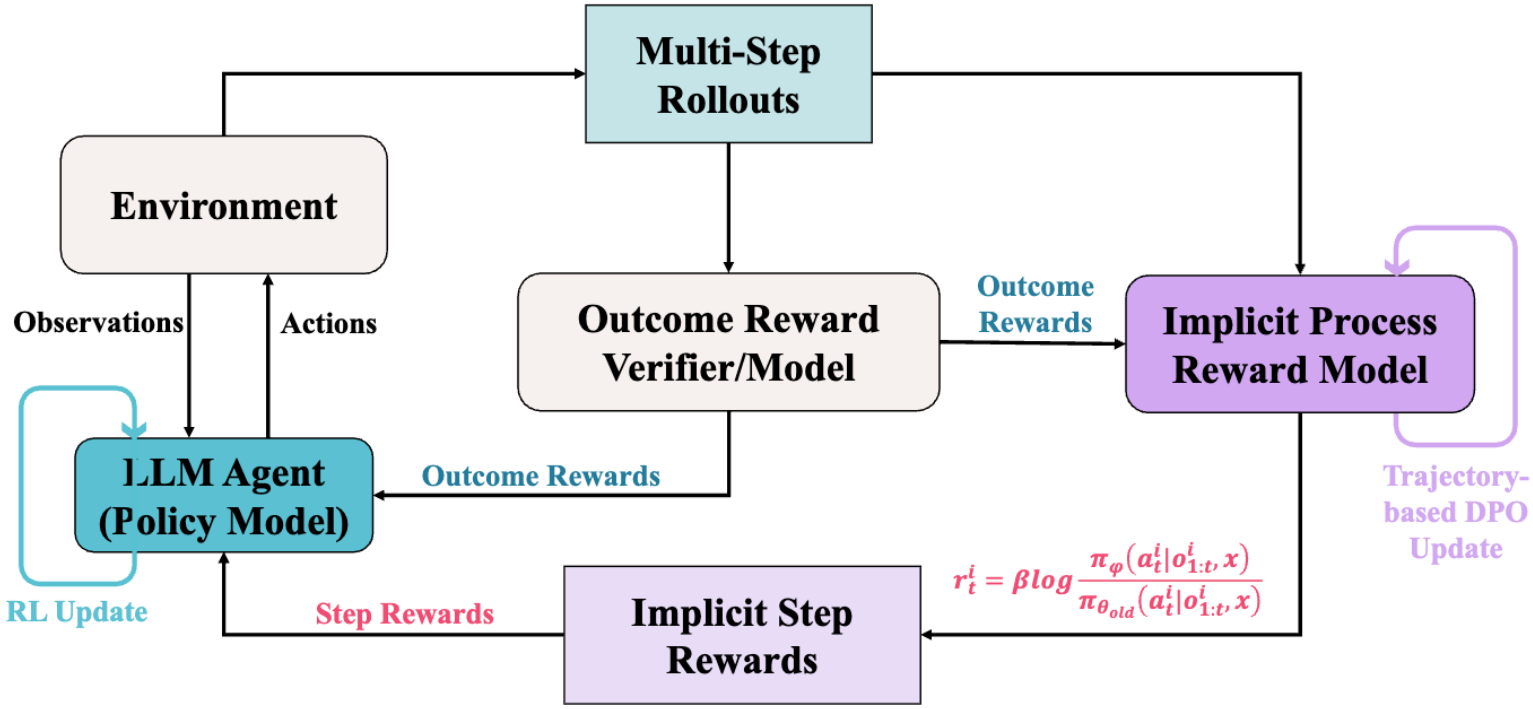
- **Credit assignment problem in agentic RL**
  - In dynamic interactive environments, relying solely on outcome rewards for policy optimization struggles to isolate effective intermediate actions.

- Introducing process supervision signals is a direct and effective approach, but faces challenges:
  - **Manual annotation:** Costly, highly subjective, and prone to inducing reward hacking.
  - **Generative reward models:** Poor generalization across domains.
  - **Token-level rewards:** Reward variance escalates with trajectory length, resulting in unstable training.
  - **Same-state grouping:** Difficult to effectively scale to open-ended language environments.

- **Research question：**
  - How to design a credit-assignment strategy that is **label-efficient and stable**, **scales to multi-turn interactions**, and remains **robust and generalizable to (un)verifiable rewards** in open-ended environments?

Method (iStar)

- **Self-reinforcing training loop online:**
  - Policy model: generate trajectories when interacting with the environment, optimized by multi-turn RL（e.g., GRPO/RLOO/Reinforce++)；
  - Implicit process reward model (PRM)： produce implicit step rewards for each action within a trajectory, optimized by multi-turn DPO objective.



Method (iStar)

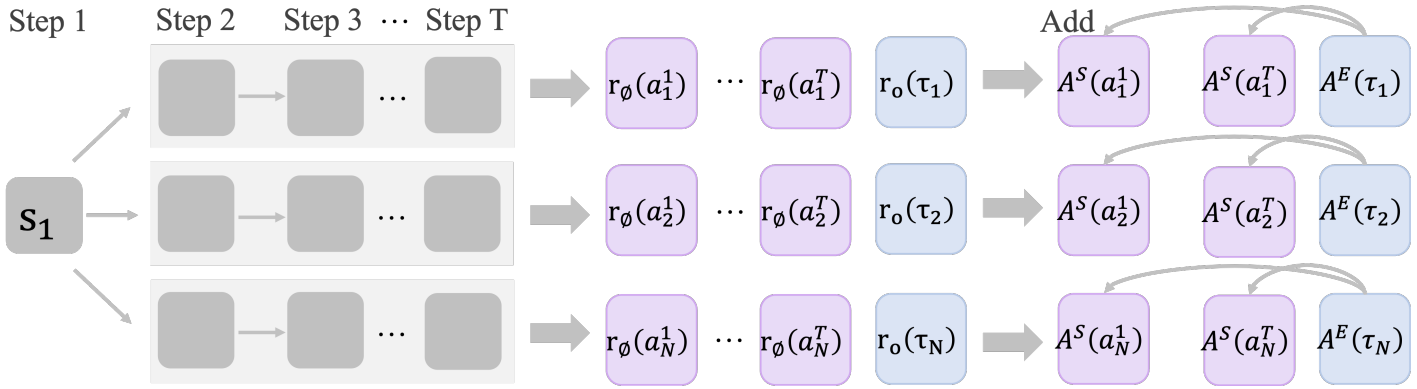
- **Optimizing implicit PRM:** based on trajectory preference and applicable to unverifiable outcome rewards.

$$\mathcal{J}_{\text{PRM}}(\phi) = -\mathbb{E}_{\substack{(\tau^+,\tau^-)\sim\pi_{\theta_{\text{old}}} \\ x\sim p(X)}} \left[\log \sigma \left(\beta \log \frac{\pi_\phi(\tau^+|x)}{\pi_{\theta_{\text{old}}}(\tau^+|x)} - \beta \log \frac{\pi_\phi(\tau^-|x)}{\pi_{\theta_{\text{old}}}(\tau^-|x)} \right) \right]$$

- **Theoretical analysis:**
  - The learning objective of implicit PRM is equivalent to a BT model with a step-wise reward function.

$$\begin{aligned} \mathbb{P}(\tau_1 \succ \tau_2) &= \sigma \left(\sum_{t=1}^{T_1} \beta \log \frac{\pi^*_\phi(a^1_t \mid o^1_{1:t}, x)}{\pi_{\theta_{\text{old}}}(a^1_t \mid o^1_{1:t}, x)} - \sum_{t=1}^{T_2} \beta \log \frac{\pi^*_\phi(a^2_t \mid o^2_{1:t}, x)}{\pi_{\theta_{\text{old}}}(a^2_t \mid o^2_{1:t}, x)} \right) \\ &= \sigma \left(\sum_{t=1}^{T_1} r^*_\phi(o^1_{1:t}, a^1_t) - \sum_{t=1}^{T_2} r^*_\phi(o^2_{1:t}, a^2_t) \right) \end{aligned}$$

- **Policy learning:** trajectory-level & step-level advantages for credit assignment in multi-turn RL.



Experiments

■ Performance on benchmarks.

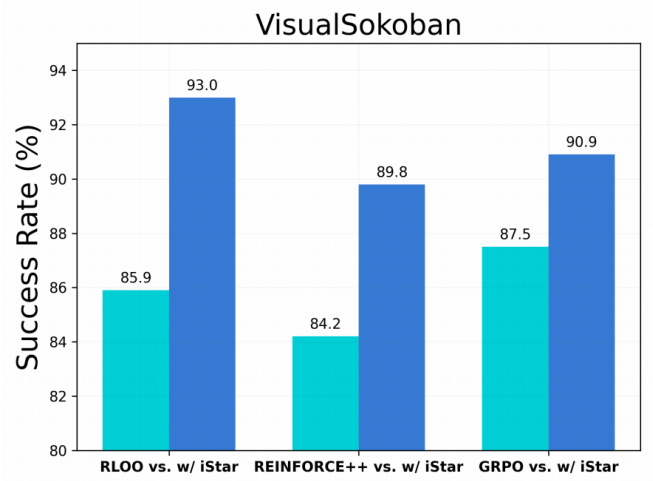
Table 1: **Performance on WebShop and VisualSokoban**. Qwen2.5-7B-Instruct and Qwen2.5-VL-7B-Instruct serve as the base models for the policy model in WebShop and VisualSokoban, respectively. Note that Deepseek-R1 and PPO training do not currently support multi-modal scenarios, and PRIME is only applicable to tasks with binary outcome rewards. Results are averaged over three random seeds.

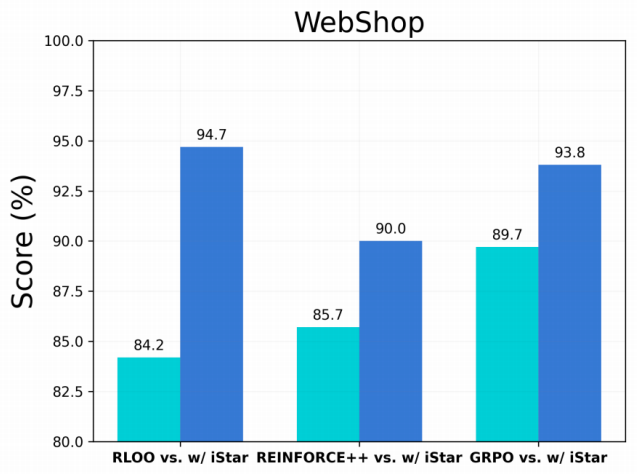
Method	WebShop		VisualSokoban
	Success	Score	Success
Prompting frontier LLMs (ReAct)			
GPT-5	37.5	66.1	16.6
Gemini-2.5-Pro	30.5	38.4	16.0
DeepSeek-R1	29.3	39.8	-
Claude-Sonnet-4-Thinking	35.2	62.0	19.1
Base Model (ReAct)	21.5	47.3	14.1
+ PPO	78.2 ± 4.5	86.6 ± 1.1	-
+ GRPO	80.1 ± 1.7	89.3 ± 2.8	85.6 ± 2.8
+ RLOO	77.4 ± 1.1	87.6 ± 4.7	86.3 ± 0.6
+ REINFORCE++	77.0 ± 3.9	85.8 ± 0.1	81.4 ± 8.8
+ PRIME (Cui et al., 2025)	81.5 ± 1.8	91.3 ± 0.6	-
+ GiGPO (Feng et al., 2025)	84.1 ± 3.9	91.2 ± 1.5	85.9 ± 2.6
+ RLOO w/ iStar	**86.5** ± 2.8	**93.6** ± 1.0	**91.7** ± 1.2

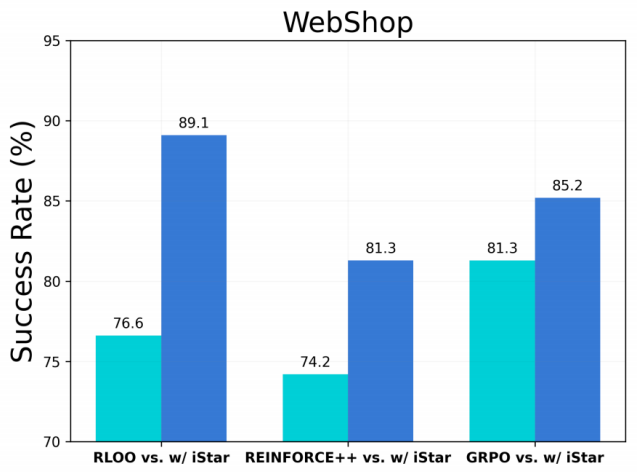
Table 2: **Performance on Sotopia.** 'Self-Chat" refers to dialogues where the model under evaluation interacts with itself, while "GPT-4o-as-Partner" denotes interactions between the model and GPT-4o. "Goal" refers to the goal completion rate (on a scale of 0-10). The "Hard" subset comprises test scenarios in SOTOPIA that require advanced reasoning capabilities, and "All" represents the complete test set. Results are averaged over three random seeds.

Method	Self-Chat		GPT-4o-as-Partner	
	Goal (Hard)	Goal (All)	Goal (Hard)	Goal (All)
Prompting frontier LLMs (ReAct)				
GPT-5	7.21	8.95	**7.70**	**8.90**
Gemini-2.5-Pro	6.74	8.27	7.43	8.41
DeepSeek-R1	6.98	8.56	7.30	8.44
Claude-Sonnet-4-Thinking	6.39	8.64	7.02	8.62
Qwen2.5-7B-Instruct (ReAct)	5.56	6.77	5.51	7.30
+ PPO	6.63 ± 0.24	8.25 ± 0.09	6.27 ± 0.14	8.07 ± 0.08
+ GRPO	6.97 ± 0.24	8.31 ± 0.06	6.42 ± 0.31	7.84 ± 0.06
+ RLOO	5.70 ± 0.16	7.13 ± 0.02	6.09 ± 0.13	7.77 ± 0.03
+ REINFORCE++	6.17 ± 0.30	7.87 ± 0.09	6.38 ± 0.05	7.93 ± 0.09
+ GRPO w/ iStar	**7.11** ± 0.19	**8.42** ± 0.03	**6.76** ± 0.18	**8.36** ± 0.03
Llama3.1-8B-Instruct (ReAct)	5.89	6.95	5.82	7.43
+ PPO	7.76 ± 0.14	9.05 ± 0.03	6.64 ± 0.03	8.14 ± 0.01
+ GRPO	7.92 ± 0.08	9.12 ± 0.02	6.68 ± 0.03	8.14 ± 0.02
+ RLOO	6.48 ± 0.15	8.33 ± 0.03	6.51 ± 0.14	8.02 ± 0.06
+ REINFORCE++	7.84 ± 0.14	9.06 ± 0.04	6.38 ± 0.23	7.99 ± 0.10
+ GRPO w/ iStar	**8.06** ± 0.11	**9.20** ± 0.03	**7.16** ± 0.14	**8.45** ± 0.03

■ Compatible with various RL algorithms.







Experiments

■ Improving sample efficiency and training stability during multi-turn RL.

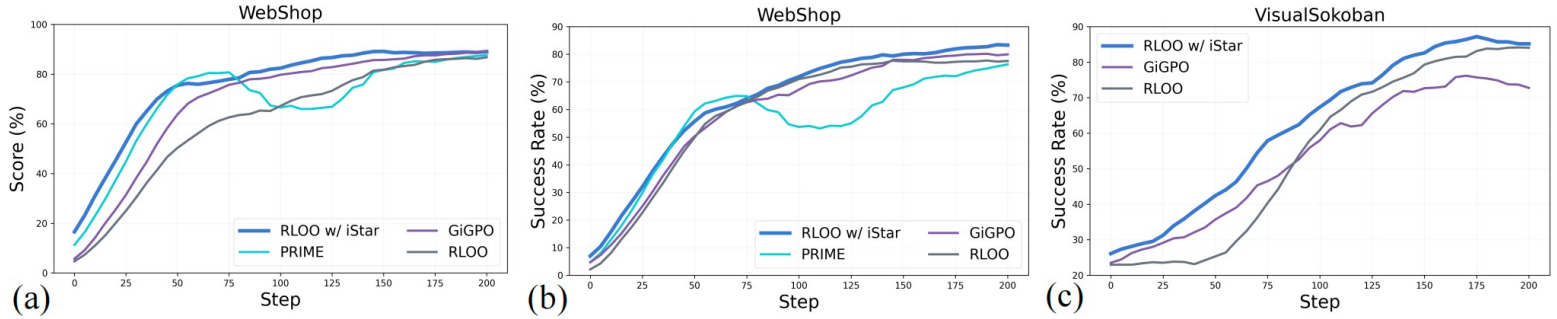


Figure 4: **Validation performance (10-step moving) during RL in WebShop and VisualSokoban.** Note that PRIME can only be applied to tasks with binary outcome rewards.

■ Enhance exploration efficiency.

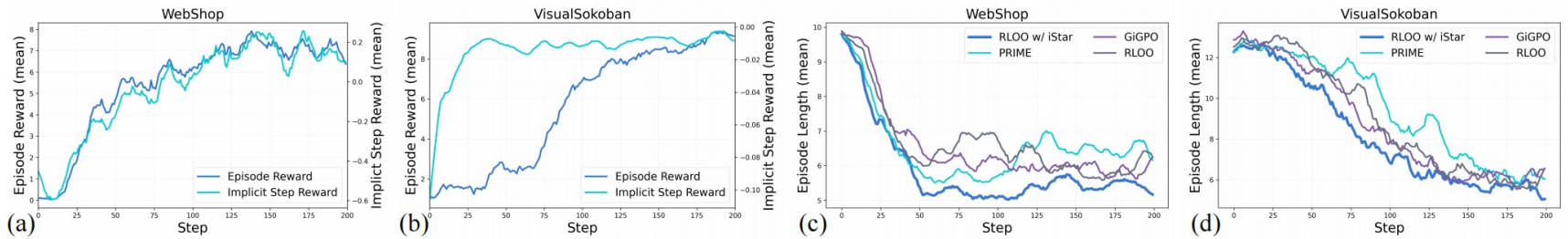


Figure 5: **Training dynamics (10-step moving) of iStar in WebShop and VisualSokoban**. (a)-(b): Dynamics of the episode and implicit step rewards during RL training by our method. (c)-(d): The episode length versus training step compared to baselines.

Code

ICLR

Thanks!

E-mail: liuxiaoqian23@mails.ucas.ac.cn