RAGEN: Understanding Self-Evolution in LLM Agents via Multi-Turn Reinforcement Learning

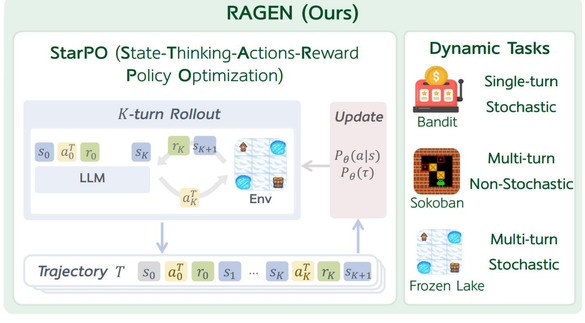
Zihan Wang*, Kangrui Wang*, Qineng Wang*, Pingyue Zhang*, Linjie Li*, Zhengyuan Yang, Xing Jin, Kefan Yu, Minh Nhat Nguyen, Licheng Liu, Eli Gottlieb, Yiping Lu, Kyunghyun Cho, Jiajun Wu, Li Fei-Fei, Lijuan Wang, Yejin Choi, Manling Li

"Zero" Style Training: LLMs learn to self-evolve with minimal supervision

Pretraining | Cold-Start Stage | RL with verifiable rewards

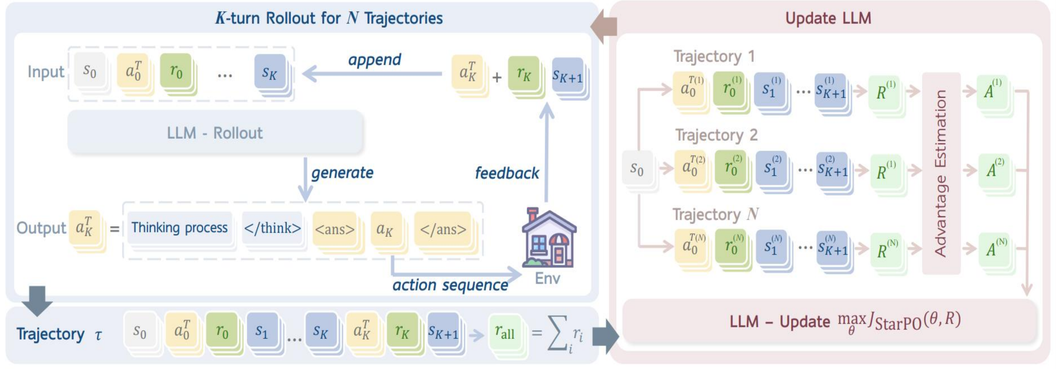
- No need for human demonstration in SFT
- No need for human preference data across domains
- No need to train reward models

Extending to Real-World: static → dynamic environments



- Prompt is finite; environments can have infinite states
- LLMs can interact to get feedback, not just respond
- Diversity comes for free: randomness, history, context
- More challenges: partial observation, credit assignment, compounding errors

Defining Agent RL Training: MDP as sequence prediction



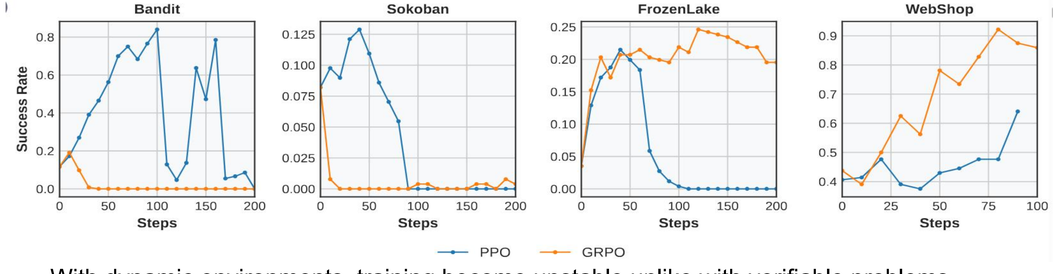
Optimizing: a single answer → *entire multi-turn interaction trajectory*

$$J_{\text{step}}(\theta) = \mathbb{E}_{s\sim\mathcal{D},a\sim\pi_\theta(\cdot|s)}[R(s,a)]$$

$$J_{\text{StarPO}}(\theta) = \mathbb{E}_{\mathcal{M},\tau\sim\pi_\theta}[R(\tau)]$$

- *Rollout generation*
- *Structured Output*
- *Action Execution*
- *Feedback Loop*
- *Trajectory evaluation*

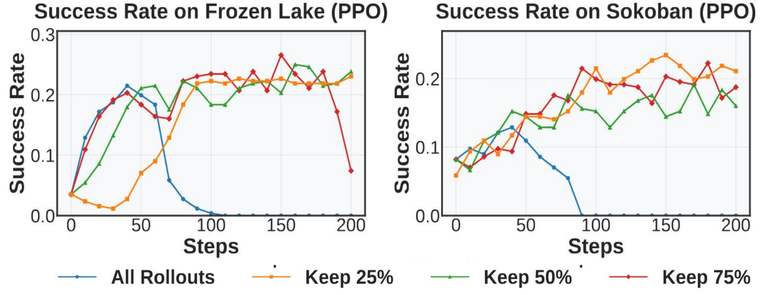
Can LLM Agents self-evolve with dynamic environments?



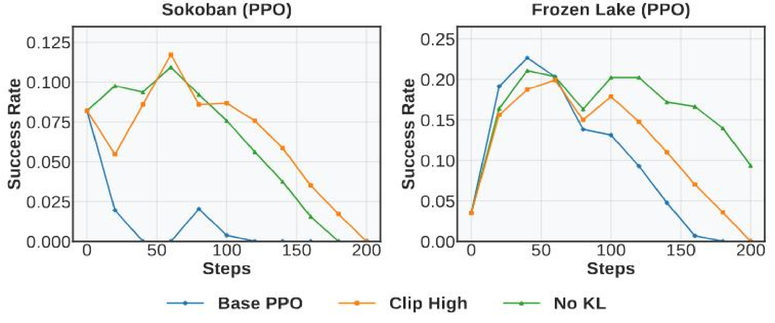
- With dynamic environments, training become unstable unlike with verifiable problems
- LLMs stuck in Echo Trap where they repeatedly reason about states after RL training
- Potential reasons: model is unfamiliar with environment, reward has high variance, state change can be random - model tend to repeat reasoning patterns for less challenging problems and reasoning chains become limited after training

Guiding models to forbid being "trapped"

- Models can only generate diverse reasoning in part of cases
- Measuring reasoning diversity is difficult, so we measure outcome diversity by measuring outcome reward variance
- Keeping diverse trajectories help model learning better



- Model learn on diverse and long sequences in agent settings
- Optimization gradient also become less certain and variance become larger
- Stabilization methods of single-turn RL works well under agent settings



Limitations and Future Work

- Agent learns well, while reasoning occurs conditionally on semantic-rich environments
- Stochastic environments still hinder agent learning and lead to conservative policies
- Future work: more modalities, credit assignment…