StraTA: Incentivizing Agentic Reinforcement Learning with Strategic Trajectory Abstraction

NeurIPS 2026 Poster Presentation

Xiangyuan Xue

November 5, 2026

The Chinese University of Hong Kong

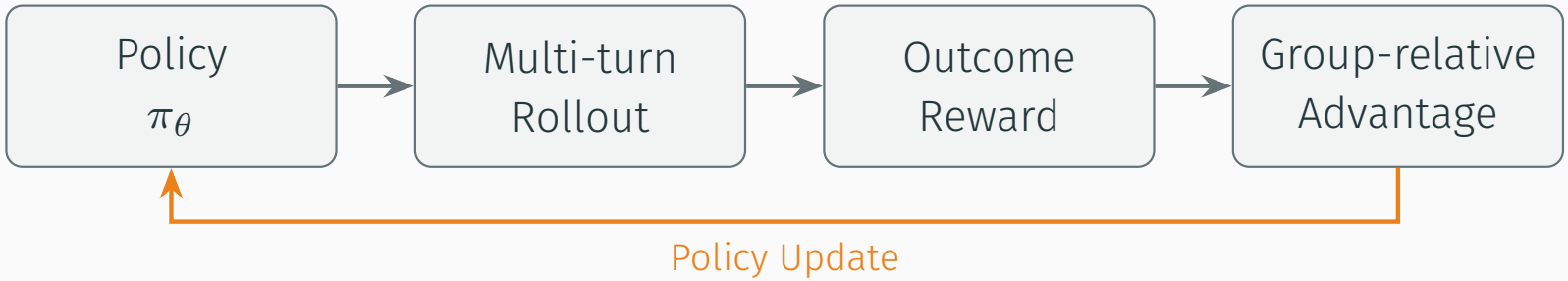
Research Background

LLM-based Agents

- An **agent** consists of a **foundation model** and a **harness**.
  - The foundation model makes decisions, while the harness provides tools, memory, and context management.
- Both components are evolving **rapidly**.
  - **Foundation models**: GPT-6-Astra, Claude-Opus-5.5, GLM-5.3, etc.
  - **Harnesses**: Codex, Claude Code, DeepSeek Harness, etc.
- Agents are enabling a wide range of **emerging applications**.
  - **Coding**: web, software, game, etc.
  - **Working**: document, chart, presentation, video, modeling, etc.

Wang, Lei, et al. "A survey on large language model based autonomous agents." arXiv preprint arXiv:2308.11432 (2023).

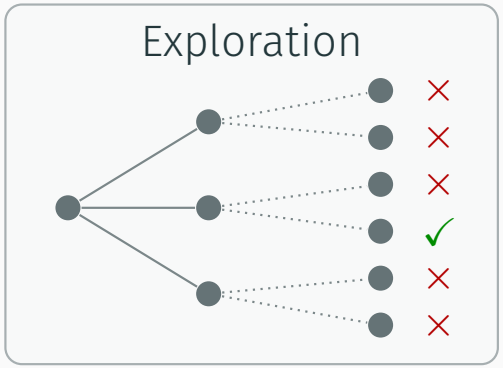
Continual Improvement of Agents

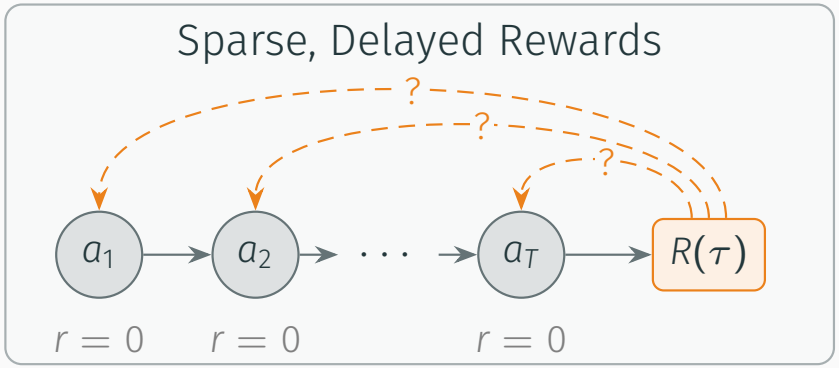


- Agents can be improved from **two complementary directions**.
  - **Model-level**: SFT, RL, OPD, etc.
  - **Harness-level**: retrieval, memory, workflow, etc.
- **Agentic RL** is one of the most promising directions.
  - The policy learns directly from **agent-environment interaction**.
  - In **GRPO-style** methods, trajectories of the same task are compared by **outcome rewards**.

Fang, Jinyuan, et al. "A comprehensive survey of self-evolving ai agents: A new paradigm bridging foundation models and lifelong agentic systems." arXiv preprint arXiv:2508.07407 (2025).

Challenges of Agentic RL



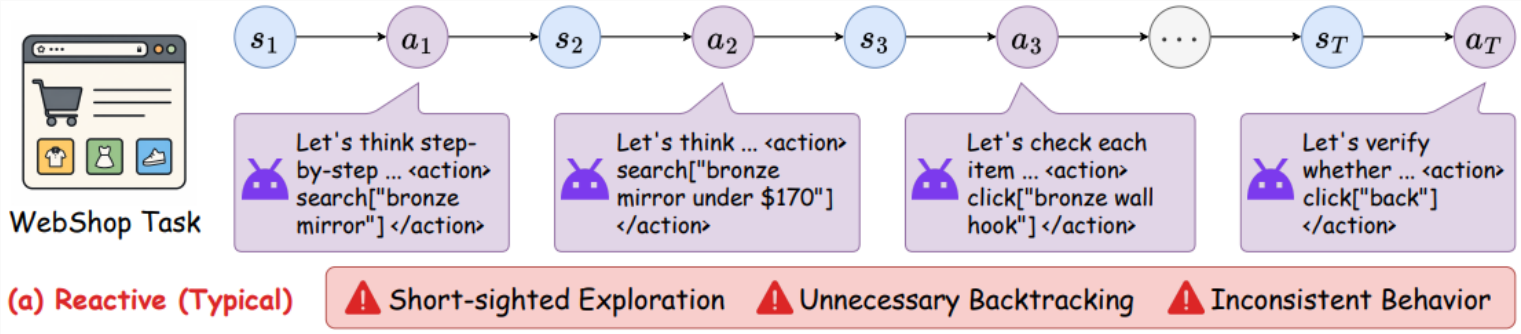


- Long-horizon agentic tasks pose **two fundamental challenges**.
  - **Exploration**: success must be discovered through **trial and error**.
  - **Sparse, delayed rewards**: **credit assignment** becomes difficult.
- Recent methods mitigate these challenges in **different ways**.
  - **RAGEN**: rollout filtering for **better exploration**.
  - **GiGPO**: step-level advantages for **better credit assignment**.

Feng, Lang, et al. "Group-in-group policy optimization for llm agent training." arXiv preprint arXiv:2505.10978 (2025).

Wang, Zihan, et al. "Ragen: Understanding self-evolution in llm agents via multi-turn reinforcement learning." arXiv preprint arXiv:2504.20073 (2025).

Limitation of Reactive Agents



- Most methods are **reactive**, without **explicit, global planning**.
- However, no planning process does **not** mean **no need to plan**.
- The agent must plan **implicitly** at **two levels** simultaneously.
  - **Episode level**: a trajectory contains many **actions**.
  - **Turn level**: an action is in fact a sequence of many **tokens**.

Zhang, Guibin, et al. "The landscape of agentic reinforcement learning for llms: A survey." arXiv preprint arXiv:2509.02547 (2025).

Research Question

- Humans typically solve complicated problems in a **plan-then-execute** manner.
  - Form a **high-level plan** from the task context.
  - Execute **local actions** under the guidance of the plan.
  - Use the feedback to improve **both planning and execution**.
- This contrast motivates a critical research question.

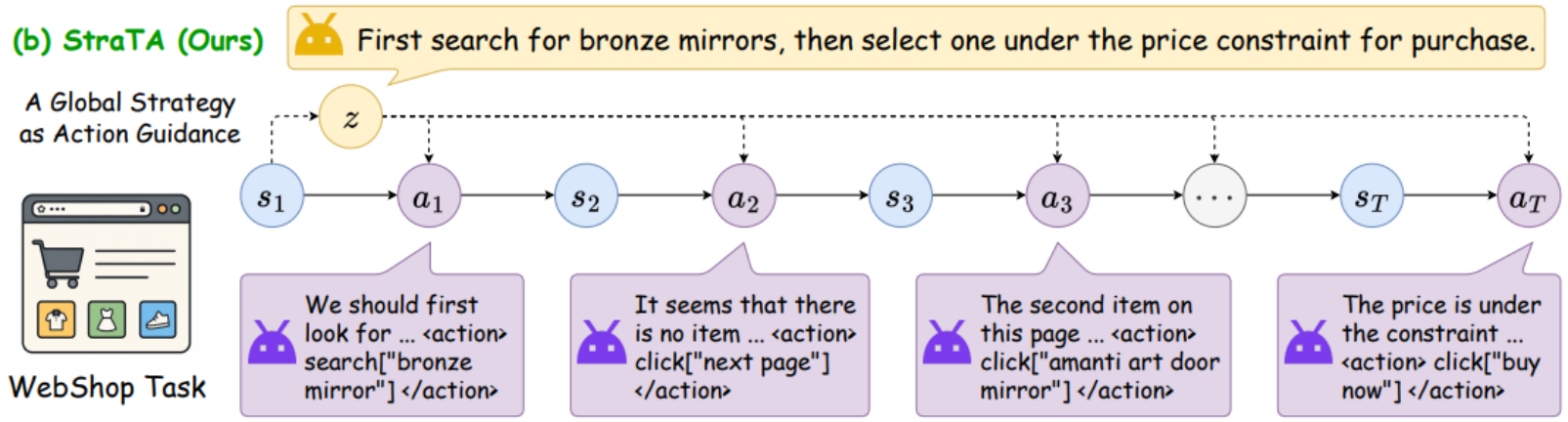
Research Question

Can agentic RL benefit from integrating long-horizon decision-making with **explicit, learnable global planning**?

Erdogan, Lutfi Eren, et al. "Plan-and-act: Improving planning of agents for long-horizon tasks." arXiv preprint arXiv:2503.09572 (2025).

StraTA Framework

Strategy-guided Task Execution

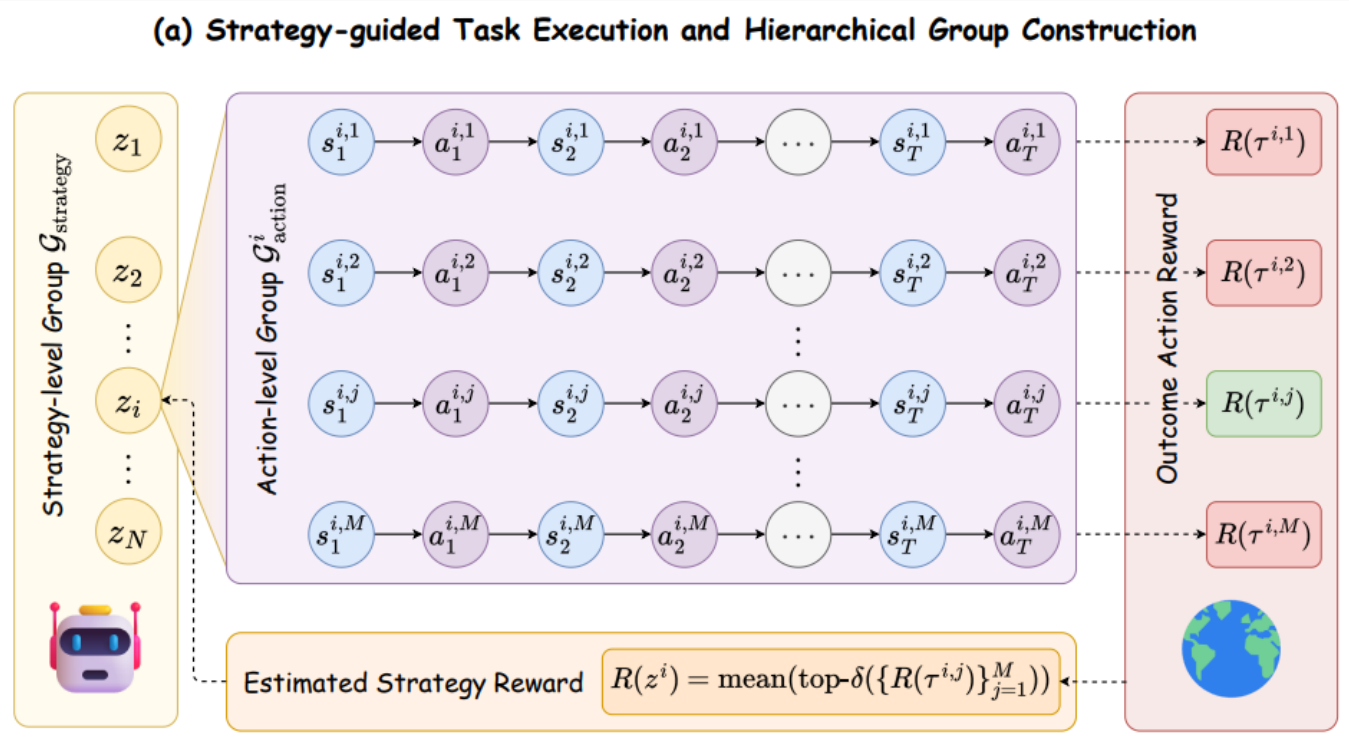


- We propose **Strategic Trajectory Abstraction (StraTA)**.
- A single policy first generates a **compact plan**, and then conditions **every action** on it

$$z \sim \pi_\theta(\cdot \mid s_1), \quad a_t \sim \pi_\theta(\cdot \mid z, s_t).$$

- Decompose the learning objective into **planning** and **execution**.

Hierarchical Group Construction



- For each task, StraTA samples N **plans**, and M **executions** under each plan, forming a plan group and N action groups.
- Actions inherit **outcome rewards**, and plans take the **top-δ mean**

$$R(a_t^{i,j}) = R(\tau^{i,j}), \quad R(z^i) = \text{mean}\left(\text{top}_\delta\left(\{R(\tau^{i,j})\}_{j=1}^M\right)\right).$$

Group-relative Policy Optimization

- The advantage is computed within the corresponding group

$$A(x) = \frac{R(x) - \textbf{mean}\left(\{R(x')\}_{x' \in \mathcal{G}}\right)}{\textbf{std}\left(\{R(x')\}_{x' \in \mathcal{G}}\right)}, \quad \mathcal{G} \in \{\mathcal{G}_{\text{plan}}, \mathcal{G}^1_{\text{action}}, \ldots, \mathcal{G}^N_{\text{action}}\}.$$

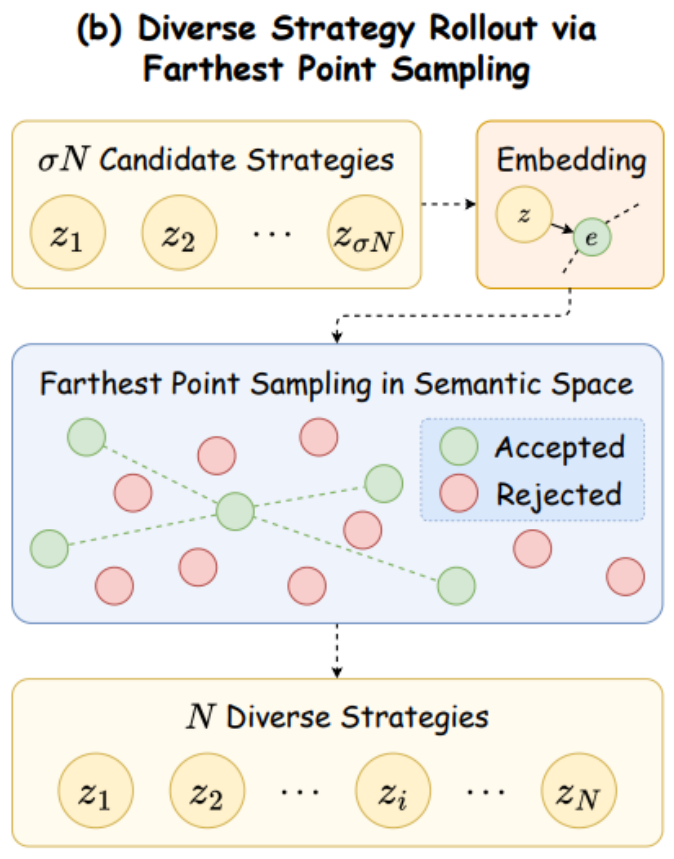
- GRPO's objective optimizes the response with its advantage

$$\mathcal{L}(x, A; \theta) = \frac{1}{|x|} \sum_{k=1}^{|x|} \textbf{min}\left(\rho_{x,k}(\theta)A, \textbf{clip}\left(\rho_{x,k}(\theta), 1-\epsilon, 1+\epsilon\right) A\right).$$

- StraTA's objective **jointly optimizes** plans and actions

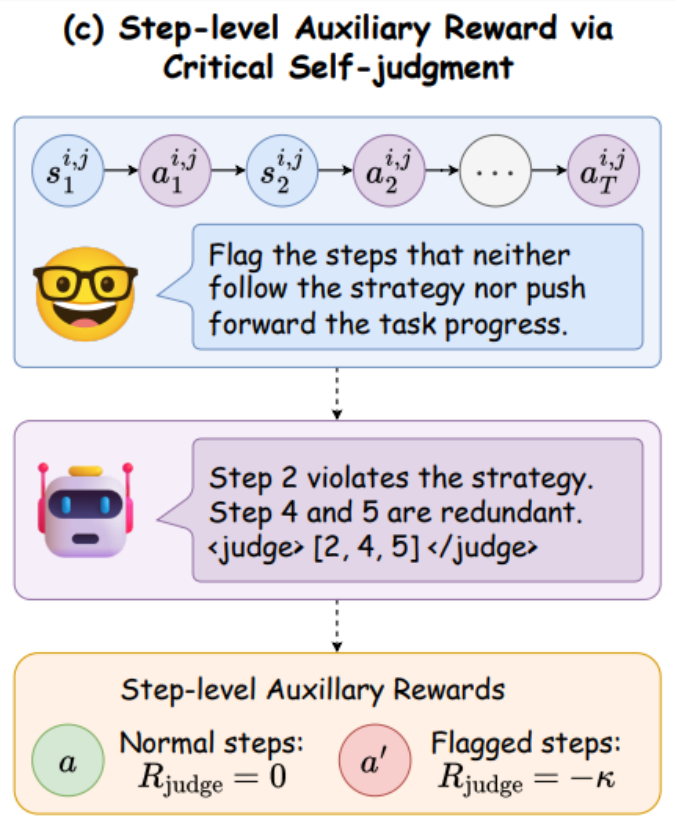
$$J_{\text{StraTA}}(\theta) = \mathbb{E}\left[\sum_{i=1}^{N} \mathcal{L}\left(z^i, A(z^i); \theta\right) + \sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{t=1}^{T_{i,j}} \mathcal{L}\left(a_t^{i,j}, A(a_t^{i,j}); \theta\right) - \beta D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}})\right].$$

Diverse Strategy Rollout



- Independently sampled plans are often **highly similar**, even at high temperatures.
- StraTA **oversamples** σN candidate plans and maps them into vectors in a **semantic space**.
- **Farthest point sampling** selects N plans that are maximally dissimilar to each other.
- The resulting **semantically diverse** plans can effectively encourage exploration.

Step-level Auxiliary Reward



- All actions **share the outcome reward**, which cannot ensure that each action **follows the plan**.
- After each trajectory ends, the policy **retrospectively judges** the whole trajectory.
- Steps that **neither follow the plan nor advance the task** receive a penalty of $-\kappa$.
- Due to the errors of **LLM-as-a-judge**, the penalty κ should be kept **small**.

Experimental Results

Experimental Setup

- We evaluate StraTA on **three interactive benchmarks**.
  - **ALFWorld**: completing household tasks.
  - **WebShop**: searching for and purchasing products.
  - **SciWorld**: conducting scientific experiments.
- We train models at **two scales**.
  - **Qwen2.5-1.5B-Instruct**: showing the **improvement** from RL training.
  - **Qwen2.5-7B-Instruct**: showing the **scaling effect** on a larger model.
- We compare StraTA with two groups of baselines.
  - **Prompting-based**: GPT, Claude, Gemini, Vanilla, and ReAct.
  - **Training-based**: PPO, RLOO, GRPO, GiGPO, and AgentGym-RL.
- The training configuration largely follows **GiGPO**.
  - **Training**: 150 steps with 16 tasks per batch on 8 NVIDIA H100 GPUs.
  - **Rollout**: $N = 4$, $M = 8$, i.e., **32 trajectories per task**.
  - **Hyperparameters**: $\sigma = 8$, $\delta = 0.5$, $\kappa = 0.1$.

ALFWorld and WebShop Results

Method	ALFWorld							WebShop	
	Pick	Look	Clean	Heat	Cool	Pick2	All	Score	Succ
Closed-Source Model									
GPT-5.1	80.0	53.8	22.2	62.5	44.0	66.7	55.7	31.2	22.2
Claude-4-Sonnet	80.0	76.9	63.0	75.0	72.0	70.8	72.9	23.7	18.4
Gemini-2.5-Flash	71.4	30.8	40.7	18.8	20.0	25.0	38.6	4.7	4.4
Qwen2.5-1.5B-Instruct									
Vanilla	$5.7_{\pm 0.0}$	$25.6_{\pm 4.4}$	$1.2_{\pm 2.1}$	$0.0_{\pm 0.0}$	$0.0_{\pm 0.0}$	$2.8_{\pm 2.4}$	$4.5_{\pm 0.4}$	$20.2_{\pm 0.7}$	$2.9_{\pm 0.8}$
ReAct	$30.5_{\pm 1.3}$	$12.8_{\pm 9.6}$	$7.4_{\pm 3.0}$	$18.8_{\pm 8.8}$	$17.3_{\pm 1.9}$	$12.5_{\pm 6.8}$	$17.6_{\pm 2.0}$	$16.2_{\pm 0.3}$	$9.1_{\pm 0.4}$
PPO	$86.7_{\pm 4.4}$	$61.5_{\pm 7.7}$	$93.8_{\pm 2.1}$	$97.9_{\pm 3.6}$	$84.0_{\pm 0.0}$	$80.6_{\pm 2.4}$	$85.5_{\pm 2.2}$	$78.2_{\pm 0.4}$	$52.8_{\pm 0.2}$
RLOO†	$88.3_{\pm 3.0}$	$52.8_{\pm 8.6}$	$71.0_{\pm 5.9}$	$62.8_{\pm 8.7}$	$66.4_{\pm 5.5}$	$56.9_{\pm 4.7}$	$69.7_{\pm 2.5}$	$73.9_{\pm 5.6}$	$52.1_{\pm 6.7}$
GRPO	$80.0_{\pm 0.0}$	$61.5_{\pm 0.0}$	$81.5_{\pm 3.7}$	$95.8_{\pm 7.2}$	$62.7_{\pm 4.6}$	$69.4_{\pm 4.8}$	$75.5_{\pm 2.9}$	$80.0_{\pm 0.5}$	$62.2_{\pm 1.2}$
GiGPO†	$94.4_{\pm 5.9}$	$67.5_{\pm 4.6}$	$94.8_{\pm 3.8}$	$94.4_{\pm 7.8}$	$79.8_{\pm 4.7}$	$76.4_{\pm 5.4}$	$86.7_{\pm 1.7}$	$83.1_{\pm 1.6}$	$65.0_{\pm 3.2}$
StraTA (Ours)	$89.5_{\pm 1.6}$	$87.2_{\pm 4.4}$	$92.6_{\pm 0.0}$	$93.8_{\pm 0.0}$	$90.7_{\pm 4.6}$	$90.3_{\pm 2.4}$	$90.7_{\pm 1.9}$	$91.1_{\pm 0.3}$	$82.5_{\pm 0.6}$
Qwen2.5-7B-Instruct									
Vanilla	$50.5_{\pm 9.1}$	$43.6_{\pm 4.4}$	$27.2_{\pm 5.7}$	$14.6_{\pm 7.2}$	$16.0_{\pm 4.0}$	$5.6_{\pm 4.8}$	$27.4_{\pm 1.5}$	$31.2_{\pm 1.0}$	$5.3_{\pm 1.5}$
ReAct	$56.2_{\pm 3.6}$	$43.6_{\pm 3.6}$	$49.4_{\pm 4.6}$	$43.8_{\pm 5.1}$	$37.3_{\pm 6.8}$	$18.1_{\pm 5.2}$	$42.4_{\pm 1.7}$	$35.7_{\pm 0.2}$	$27.3_{\pm 0.2}$
PPO	$96.2_{\pm 1.7}$	$71.8_{\pm 8.9}$	$96.3_{\pm 0.0}$	$91.7_{\pm 3.6}$	$70.7_{\pm 6.1}$	$87.5_{\pm 0.0}$	$87.4_{\pm 1.1}$	$82.2_{\pm 0.2}$	$68.9_{\pm 1.0}$
RLOO†	$87.6_{\pm 4.3}$	$78.2_{\pm 8.3}$	$87.3_{\pm 5.8}$	$81.3_{\pm 7.6}$	$71.9_{\pm 5.2}$	$48.9_{\pm 8.4}$	$75.5_{\pm 4.6}$	$80.3_{\pm 3.2}$	$65.7_{\pm 4.0}$
GRPO	$98.1_{\pm 2.7}$	$71.8_{\pm 9.6}$	$85.2_{\pm 3.0}$	$58.3_{\pm 7.8}$	$73.3_{\pm 5.0}$	$69.4_{\pm 3.9}$	$79.3_{\pm 1.5}$	$84.1_{\pm 0.7}$	$71.5_{\pm 0.7}$
GiGPO†	$97.7_{\pm 1.6}$	$82.7_{\pm 7.9}$	$98.8_{\pm 1.6}$	$83.7_{\pm 7.2}$	$89.3_{\pm 8.2}$	$79.2_{\pm 6.6}$	$90.8_{\pm 1.3}$	$84.4_{\pm 2.9}$	$72.8_{\pm 3.2}$
StraTA (Ours)	$94.3_{\pm 0.0}$	$92.3_{\pm 0.0}$	$97.5_{\pm 2.1}$	$100.0_{\pm 0.0}$	$93.3_{\pm 4.6}$	$81.9_{\pm 2.4}$	$93.1_{\pm 0.8}$	$91.2_{\pm 0.3}$	$84.2_{\pm 0.3}$

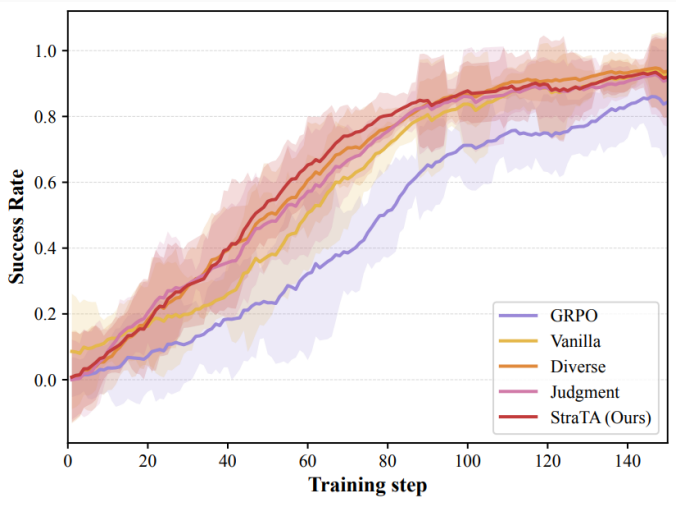
- Training-based methods **significantly outperform** prompting-based ones, with **GiGPO** being the best.
- StraTA outperforms GiGPO by **2.3%** on ALFWorld and **11.4%** on WebShop based on Qwen2.5-7B-Instruct.

SciWorld Results

Method	Category	SciWorld					
		Measure	Test-Cond	Find	Chem-Mix	Lifespan	All
Closed-Source Model							
GPT-5.1	Prompting-Based	32.1	46.5	41.8	69.6	38.3	43.0
Claude-4-Sonnet	Prompting-Based	45.4	54.7	60.3	45.6	90.0	57.4
Gemini-2.5-Flash	Prompting-Based	40.7	51.3	53.7	33.0	50.0	49.6
Qwen2.5-7B-Instruct							
AgentGym-RL†	Training-Based	24.5	59.0	65.9	0.0	66.7	50.5
ScalingInter†	Training-Based	34.0	55.4	88.6	0.0	73.3	57.0
Vanilla	Prompting-Based	$2.1_{\pm 1.5}$	$12.3_{\pm 1.3}$	$21.2_{\pm 1.6}$	$9.6_{\pm 3.3}$	$25.0_{\pm 2.4}$	$13.9_{\pm 0.8}$
ReAct	Prompting-Based	$12.2_{\pm 2.1}$	$14.4_{\pm 1.2}$	$33.6_{\pm 2.7}$	$9.6_{\pm 2.8}$	$28.9_{\pm 3.4}$	$20.0_{\pm 0.3}$
PPO	Training-Based	$1.8_{\pm 1.3}$	$57.5_{\pm 0.2}$	$60.9_{\pm 0.8}$	$6.9_{\pm 1.9}$	$97.8_{\pm 3.1}$	$51.4_{\pm 0.3}$
GRPO	Training-Based	$31.2_{\pm 4.8}$	$49.8_{\pm 2.9}$	$32.7_{\pm 3.6}$	$14.7_{\pm 5.4}$	$53.3_{\pm 5.0}$	$41.8_{\pm 2.2}$
StraTA (Ours)	Training-Based	$55.7_{\pm 3.2}$	$54.4_{\pm 0.7}$	$78.6_{\pm 1.1}$	$16.3_{\pm 1.2}$	$100.0_{\pm 0.0}$	$63.5_{\pm 0.7}$

- Training-based methods **significantly outperform** prompting-based ones on the same backbone, with **ScalingInter** as the best, on par with Claude-4-Sonnet.
- StraTA achieves the **highest overall score**, exceeding ScalingInter by **6.5%** and Claude-4-Sonnet by **6.1%** points.

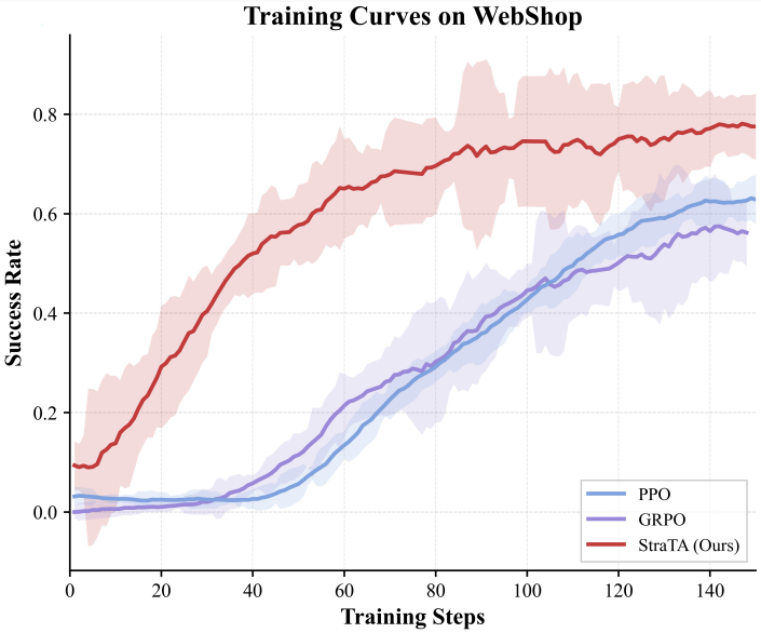
Ablation Study

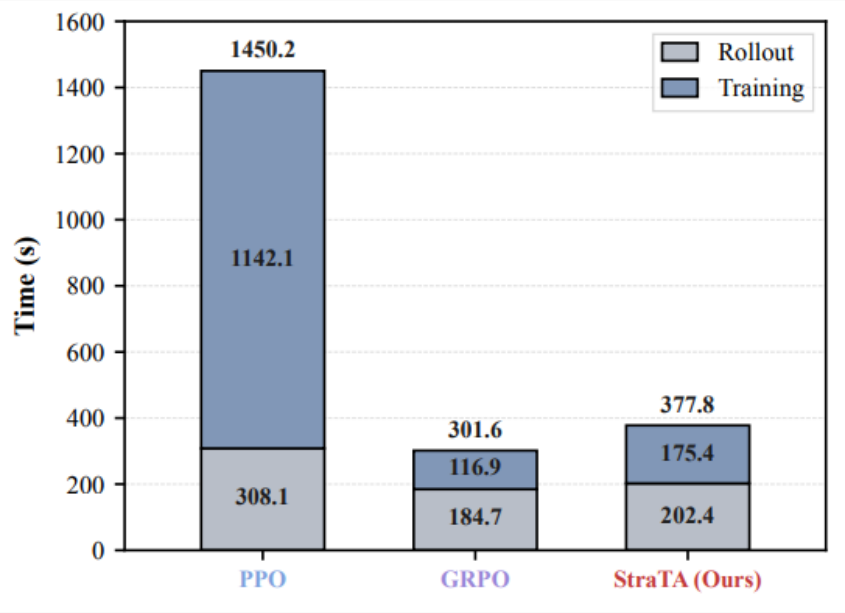


Variant	ALFWorld	WebShop
Vanilla	79.0	64.0
Diversity	87.9	64.6
Alignment	81.9	66.7
StraTA (Ours)	**88.6**	**73.4**

- We use **Qwen2.5-3B-Instruct** and only **4 tasks per batch**, which makes the differences more visible.
- **Vanilla** already performs well, showing the benefit of **explicit planning** and **joint optimization**.
- **Diversity** and **Alignment** each contribute to the performance, with **different effects** on different tasks.
- Combining both components yields the **best results**.

Training Efficiency and Cost





- StraTA **converges much faster** than PPO and GRPO.
- StraTA takes about **25% more time** per step than GRPO, and roughly **a quarter** of the time required by PPO.
- In general, StraTA achieves **significant performance gains** with only a **small extra training cost**.

Future Work

From Fixed Strategies to Adaptive Refinement

- **Limitation**: the plan is generated once and **fixed** thereafter.
  - Critical information may emerge **only through exploration**.
  - The agent should **revise its plan** when observations contradict it.
- **Challenge**: there is no good way to **assign credit** to revisions.
  - The outcome depends on both the **revised plan** and its **execution**.
  - A single trajectory cannot tell what would happen **without revision**.
- **Direction**: explore **tree-structured**, **counterfactual** rollouts.
  - Branches from a **shared state** enable **precise credit assignment**.
  - The infrastructure must support **forking** environment states.

Towards Reliable Step-level Rewards

- **Limitation**: self-judgment comes from the **policy under training**.
  - Its errors **shift** as the policy evolves, so κ must remain **small**.
  - Agreement with GPT-5.5 rises from **72.6%** to **90.1%**, but is **imperfect**.
- **Challenge**: LLM judges are **noisy**, **biased**, and **exploitable**.
  - They cannot match the **accuracy** of rule-based **verifiers**.
  - The policy easily learns **patterns** that **hack** the judge.
- **Direction**: **rubric-based RL** with **explicit, task-specific criteria**.
  - Judgments become more **consistent**, **interpretable**, and **robust**.
  - It also adds **training overhead**, where **latency** is the bottleneck.

Thank You for Listening!