HINDSIGHT IS 20/20: ADAPTIVE CREDIT ASSIGNMENT VIA REALIZED FUTURES FOR LONG-HORIZON LLM AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Group-based reinforcement learning (RL) has emerged as an effective critic-free paradigm for training large language model (LLMs) and has recently been extended to long-horizon agentic tasks. However, sparse outcome rewards provide only weak supervision for intermediate decisions, making it difficult to determine which actions are beneficial or detrimental along successful and failed trajectories. Existing step-level group-based methods alleviate this issue through finer-grained intra-group comparisons, but their reliability depends critically on whether grouped actions are genuinely comparable and whether each group provides sufficient support for estimating expected returns. To address these limitations, we propose Hindsight-Enabled Retrospective Policy Optimization (HeRPO), a critic-free framework that elevates realized futures from return signals to posterior context for credit assignment. HeRPO extracts hindsight information from realized futures to construct hindsight-pruned grouping strategy and adaptively adjusts the balancing coefficient in generalized advantage estimation according to the effective sample support of each group. Moreover, for each executed action, HeRPO evaluates the policy both with and without privileged hindsight and converts the resulting shift in action preference into a bounded weight applied to its advantage. By jointly improving the quality of step-level credit signals and controlling the magnitude of policy updates, HeRPO enables more precise optimization for long-horizon agentic tasks. Extensive experiments on the agentic benchmarks demonstrate that HeRPO consistently outperforms other competing agentic reinforcement learning methods. Codes will be available on the GitHub.

1 INTRODUCTION

Recently, large language model (LLM) agents Jin et al. (2025); Schick et al. (2023); Yang et al. (2024b) have demonstrated remarkable progress in planning and sequential decision-making within interactive environments Wang et al. (2022); Zhou et al. (2024); Liu et al. (2024). Reinforcement learning (RL) has emerged as a central post-training paradigm for improving agents that must plan, act, and adapt over extended interactions. Among recent RL approaches, group-based optimization approaches, such as GRPO Shao et al. (2024), eliminate the need for a separate critic network, thereby avoiding the substantial computational and memory overhead typically incurred by Actor-Critic frameworks Konda & Tsitsiklis (1999); Schulman et al. (2017); Yu et al. (2025b). Instead, these methods estimate relative advantages from verifiable rewards and statistics computed within groups of sampled trajectories, making them particularly attractive for large-scale reinforcement learning of LLMs. Building on this paradigm, recent work has further explored group-based optimization in multi-turn and long-horizon agentic settings Wang et al. (2025); Jin et al. (2025); Luo et al. (2025); Yu et al. (2025a), aiming to improve policy learning over complex interaction sequences.

Despite their computational efficiency, the reliance on sparse outcome-based rewards introduces a fundamental bottleneck for agent optimization. This leads to a credit assignment challenge that it is difficult to accurately attribute a terminal reward to the intermediate decisions that are critical to the outcome (see Figure 1). Recent agentic RL methods have therefore extended GRPO-style rel-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

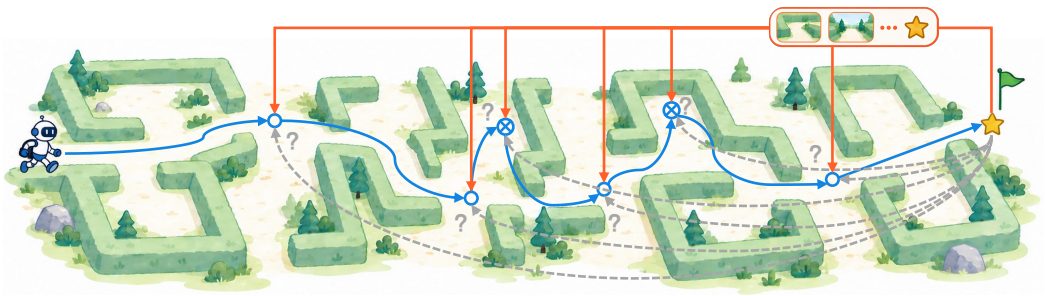


Figure 1: The executed trajectory (blue) reaches the goal, but the terminal outcome alone leaves the contribution of intermediate actions ambiguous (gray). HeRPO instead uses the realized future as privileged hindsight (orange) to retrospectively reassess them and refine credit assignment.

ative advantage estimation beyond whole trajectories by constructing step-level comparison groups over intermediate interaction actions Feng et al. (2025); He et al. (2026); Cheng et al. (2026); Dong et al. (2026); Zong et al. (2026). By aligning the granularity of intra-group comparison with environment interactions, this paradigm provides finer-grained learning signals for long-horizon agents trained under sparse outcome supervision. However, such credit assignment is driven only indirectly by comparisons, and its reliability critically depends on whether the actions within a group are genuinely comparable and whether each group provides sufficient sample support for estimating expected returns.

To address these challenges, we propose Hindsight-Enabled Retrospective Policy Optimization (HeRPO), a critic-free framework for fine-grained credit assignment in long-horizon agentic RL. The key idea of HeRPO is to treat realized futures not merely as sources of return signals, but as privileged hindsight for retrospectively assessing previous actions, as shown in Figure 1. By exploiting such hindsight throughout advantage estimation and policy optimization, HeRPO improves both the reliability of step-level credit signals and the calibration of policy updates. Concretely, HeRPO first constructs hindsight-pruned groups by retaining state-aligned actions while removes grouping relations between steps that take the same action but have inconsistent hindsight. This reduces unreliable credit transfer among actions that appear comparable at the current state but eventually lead to different futures. Based on these groups, HeRPO further introduces an adaptive generalized advantage estimation (GAE) mechanism. We define the effective sample support of each group and use it to quantify the reliability of its non-parametric value estimate and adaptively determine the GAE balance coefficient. When sufficient group samples support the estimate, HeRPO places greater emphasis on local temporal-difference credit; otherwise, it shifts toward longer-term Monte Carlo returns, enabling adaptive credit propagation across different states. Finally, HeRPO introduces hindsight self-reweighting to directly calibrate the policy update for each action. The policy model re-evaluates each executed action both with and without access to hindsight and measures the resulting shift in action preference. This posterior preference shift is then transformed into a bounded step-level weight that modulates the magnitude of the corresponding advantage, strengthening or suppressing policy updates according to the realized future of each action while preserving the original optimization direction. Notably, HeRPO introduces no additional tunable hyperparameters beyond the baseline, as its coefficients are either determined adaptively or shared with existing training configurations.

We evaluate HeRPO on long-horizon agentic benchmarks against representative agentic RL methods. Our main contributions are threefold:

- We leverage realized futures to develop hindsight-pruned grouping and effective-support adaptive GAE for reliable fine-grained credit assignment.
- We quantify the hindsight-induced preference shift to adaptively calibrate the magnitude of the policy update without altering the direction.
- Extensive experiments show that HeRPO consistently outperforms other competing methods, demonstrating its effectiveness and robustness across agentic environments and tasks.

2 Related Work

2.1 Reinforcement Learning for LLMs

Early work on reinforcement learning for LLMs primarily focused on aligning models with human preferences through RLHF Ziegler et al. (2019); Stiennon et al. (2020); Ouyang et al. (2022); Bai et al. (2022) and related preference optimization methods Rafailov et al. (2023). More recently, reinforcement learning with verifiable rewards has attracted increasing attention for improving LLM reasoning with automatically computable rewards Shao et al. (2024); Liu et al. (2025); Su et al. (2026). In particular, group-based RL methods Ahmadian et al. (2024); Shao et al. (2024); Liu et al. (2025); Yu et al. (2025b) estimate advantages from groups of responses sampled for the same query, avoiding the need for an additional value function and enabling efficient large-scale training, demonstrating strong performance on multiple areas. Different from these methods, our work focuses on improving critic-free group-based optimization by fine-grained credit assignment for long-horizon LLM agents.

2.2 Long-Horizon Agentic Reinforcement Learning

Long-horizon agentic reinforcement learning aims to optimize LLMs that interact with external environments, extending their capabilities from single-turn generation to multi-step decision-making. Recent methods address this challenge by constructing step-level comparisons across trajectories. GiGPO Feng et al. (2025) combines trajectory-level grouping with anchor-state-based micro-grouping, enabling more localized comparisons among actions taken from grouped states. GraphGPO Cheng et al. (2026) moves beyond explicit pairwise grouping by constructing a unified state-transition graph across trajectories and estimating edge-level credit based on graph distances to the goal. ARPO Dong et al. (2026) adaptively branches rollouts at uncertain tool-interaction steps with high entropy, encouraging step-level exploration of latent tool-use behaviors. Overall, these methods substantially advance long-horizon agentic RL and highlight the importance of fine-grained, step-level credit assignment. Nevertheless, most existing approaches primarily obtain step-level credit by comparing action groups across trajectories, while overlooking the effect of realized futures on assessing previous actions. HCAPO Tan et al. (2026) uses hindsight self-scoring to construct step-level advantages. HeRPO additionally prunes comparison groups, adapts GAE to group support, and uses explicit with/without-hindsight likelihood contrasts to reweight fused advantages to calibrate the policy update associated with each executed action.

3 Preliminary

Problem setup. We focus on training LLM agents that interact with an environment to complete multi-step tasks. At each time step $t = 1, 2, \ldots, T$, the agent observes a state s_t and generates a textual action a_t according to the policy $\pi_\theta(a_t \mid s_t)$, parameterized by θ. The environment then returns a scalar reward r_t and the next state s_{t+1}. A complete episode induces a trajectory $\tau = \{(s_1, a_1, r_1), \ldots, (s_T, a_T, r_T)\}$

Group-based RL. For a given task, the LLM agent samples a group of N candidate trajectories $\tau_1, \tau_2, \ldots, \tau_N$. Each trajectory τ_i receives a scalar outcome rewards $R(\tau_i)$, which reflects the overall quality or success of the final task outcome. Group-based RL methods, such as GRPO Shao et al. (2024), compute the trajectory-level advantage based on the statistics internal to the sampled group:

$$A_i^{\text{traj}} = \frac{R(\tau_i) - \mu_R}{\sigma_R}, \tag{1}$$

where μ_R and σ_R denote the mean and standard deviation of the outcome rewards within the group, respectively, and the σ_R term may optionally be omitted.

4 Hindsight-Enabled Retrospective Policy Optimization (HeRPO)

Overview of HeRPO is illustrated in Figure 2. Given a set of completed rollouts, HeRPO derives hindsight information from future trajectory subsequences and terminal outcomes. This privileged context is used to identify hindsight-inconsistent relations to prune state-aligned groups, enabling

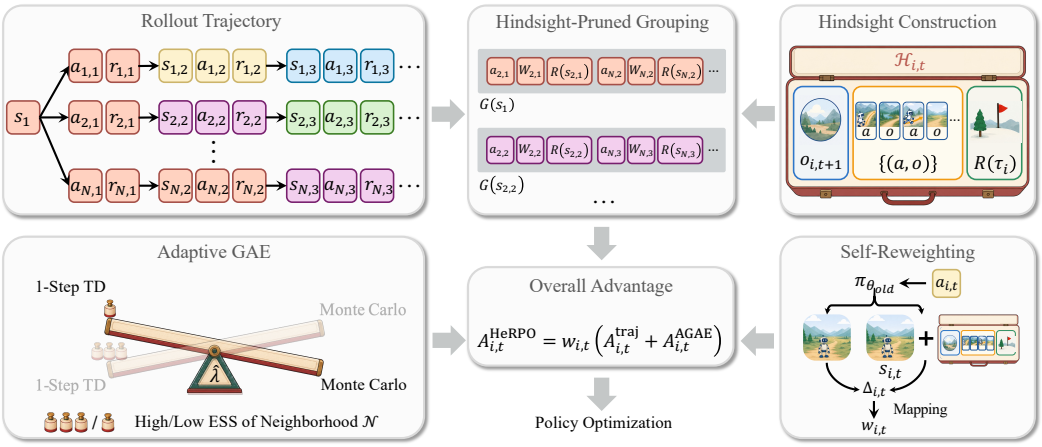


Figure 2: Overview of HeRPO. From a group of completed rollouts, HeRPO combines trajectory-level outcome comparison with a non-parametric step value estimated over dynamically consistent states. The support of this estimate controls how far credit propagates along each trajectory. The same policy then re-scores each executed action with and without privileged hindsight, producing a bounded weight that calibrates the fused advantage before policy optimization.

critic-free, adaptive step-level advantage estimation. Equipped with the hindsight information, the rollout policy retrospectively assesses each executed action, and HeRPO then leverages this reassessment to calibrate the magnitude of advantage during policy optimization.

4.1 Hindsight-Pruned Grouping (HPG)

Although a sparse outcome reward indicates whether a trajectory ultimately succeeds, it provides little information about the contribution of an individual action. We therefore construct the hindsight information $\mathcal{H}_{i,t}$, containing the realized future and the final outcome, for the action $a_{i,t}$ by:

$$\mathcal{H}_{i,t} = \left[R(\tau_i); o_{i,t+1}; \{(a_{i,u}, o_{i,u+1})\}_{u=t+1}^{U_{i,t}}\right], \tag{2}$$

where $o_{i,t}$ denote the observation of state $s_{i,t}$, $U_{i,t} = \min(t+K, T_i)$, and K is the future-subsequence length.

We refine comparison groups to mitigate misleading credit assignment. Specifically, we prune links between identical actions with inconsistent hindsight to limit the attribution of downstream outcome differences to a shared preceding action. For a state s, the corresponding step-level group can be defined as

$$G(s) = \{(a_{i,t}, W_{i,t}, R(s_{i,t})) \mid s_{i,t} = s,\ 1 \le i \le N,\ 1 \le t \le T_i\}, \tag{3}$$

where $W_{i,t}$ indicates the normalized weight assigned to $R(s_{i,t})$ and may be either manually specified or adaptively determined. For each target step (i,t), we set the weights of candidates $(i',t') \in G(s)$ to zero to remove the grouping relation if $a_{i,t} = a_{i',t'}$ but $\mathcal{H}_{i,t} \neq \mathcal{H}_{i',t'}$, then let $G_{i,t}(s)$ denote the remaining candidates with positive weights, renormalized to satisfy the constraint $\sum_{W \in G_{i,t}(s)} W = 1$. $R(s_{i,t})$ denotes the discounted episodic return in GiGPO Feng et al. (2025). The difference in final returns may stem from subsequent decisions and should not be attributed indiscriminately to the common preceding actions of the trajectory. Credit should be more focused on decisions that cause trajectory divergence

4.2 Adaptive GAE (AGAE)

The retained group members provide a non-parametric substitute for a learned critic. Specifically, we estimate the value of state $s_{i,t}$ by aggregating the weighted returns within the group G:

$$V_{i,t} = \sum_{W,R \in G_{i,t}(s)} WR. \tag{4}$$

To precisely assign credit to individual steps, we employ Generalized Advantage Estimation (GAE) Schulman et al. (2015) to construct a refined advantage signal:

$$\delta_{i,t} = r_{i,t} + \gamma V_{i,t+1} - V_{i,t},\ A_{i,t}^{\mathrm{GAE}} = \delta_{i,t} + \gamma\lambda A_{i,t+1}^{\mathrm{GAE}}, \tag{5}$$

where $A_{i,T_i+1}^{\mathrm{GAE}} = V_{i,T_i+1} = 0$ is set as the termination condition.

The reliability of the value estimate depends on both the number of retained samples and the concentration of their weights. We use effective sample support as a proxy for the amount of evidence contributing to each value estimate:

$$E_{i,t} = \frac{1}{\sum_{W \in G(s_{i,t})} W^2} - 1. \tag{6}$$

This quantity reflects the number of contributing samples and the concentration of their weights. Consequently, a large $E_{i,t}$ indicates that the value estimate is supported by a broad set of samples rather than dominated by a few influential ones, whereas a small $E_{i,t}$ indicates the opposite.

We then map effective sample support to an adaptive balance coefficient:

$$\hat{\lambda}_{i,t} = 1 - \mathrm{clip}\left(\frac{E_{i,t}}{N}, 0, 1\right), \tag{7}$$

which dynamically balances reliance on the one-step temporal-difference estimate and the Monte Carlo return. The resulting adaptive GAE is defined as

$$A_{i,t}^{\mathrm{AGAE}} = \delta_{i,t} + \gamma\hat{\lambda}_{i,t} A_{i,t+1}^{\mathrm{AGAE}}. \tag{8}$$

In this study, we assign equal normalized weights to all retained members, *i.e.*, $W = \frac{1}{|G_{i,t}(s)|}, \forall W \in G_{i,t}(s)$. Under this weighting scheme, the effective sample support simplifies to $E_{i,t} = |G_{i,t}(s)| - 1$. Thus, the adaptive coefficient $\hat{\lambda}$ is negatively correlated with the number of group sample. When the value estimate is supported by a large amount of group samples, $\hat{\lambda}$ approaches 0, placing more trust in the local one-step temporal-difference estimate. Conversely, when only a few group members remain after filtering, $\hat{\lambda}$ approaches 1, making the estimate closer to the Monte Carlo return.

4.3 Hindsight Self-Reweighting (HSR)

To modulate the policy update, HeRPO employs the rollout policy $\pi_{\theta_{\mathrm{old}}}$ and evaluates the same executed action with a token length of $L_{i,t}$ under both the standard context and the hindsight-augmented context. We then produce their log-likelihood difference as the step-level posterior preference shift $\Delta_{i,t}$:

$$\Delta_{i,t} = \frac{1}{L_{i,t}} \sum_{q=1}^{L_{i,t}} \log \frac{\pi_{\theta_{\mathrm{old}}}(a_{i,t,q} \mid s_{i,t}, a_{i,t,<q}, \mathcal{H}_{i,t})}{\pi_{\theta_{\mathrm{old}}}(a_{i,t,q} \mid s_{i,t}, a_{i,t,<q})}, \tag{9}$$

where q is the token index, and the scalar $\Delta_{i,t}$ is detached from gradient computation. A positive $\Delta_{i,t}$ indicates that hindsight makes the executed action more plausible, whereas a negative one denotes posterior disagreement. HeRPO converts this preference shift into a step-level weight by:

$$w_{i,t} = \mathrm{clip}\Big(\exp\Big(\mathrm{sgn}(A_i^{\mathrm{traj}} + A_{i,t}^{\mathrm{AGAE}})\Delta_{i,t}\Big), 1-\epsilon, 1+\epsilon\Big), \tag{10}$$

where sgn denotes the sign function, and we reuse the clipping coefficient ($0 \le \epsilon < 1$) of the policy-optimization objective to bound the weight. When $A_i^{\mathrm{traj}} > 0$, hindsight endorsement strengthens the update, whereas disagreement weakens it. When $A_i^{\mathrm{traj}} < 0$, this behavior is reversed. Since $w_{i,t}$ remains positive and bounded, it adaptively modulates the magnitude of each step-level policy update without independently reversing its optimization direction.

4.4 Policy Optimization

HeRPO combines the trajectory-level task signal with the adaptive step-level credit estimate. For each executed action, we form the overall advantage as:

$$A_{i,t}^{\mathrm{HeRPO}} = w_{i,t}(A_i^{\mathrm{traj}} + A_{i,t}^{\mathrm{AGAE}}). \tag{11}$$

The trajectory- and step-level components therefore determine the update direction jointly, and the weight derived from the hindsight self-reweighting calibrates their magnitude. Then the clipped policy optimization objective of HeRPO is:

$$\begin{aligned}\mathcal{J}_{\mathrm{HeRPO}}(\theta) =& \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\frac{1}{T_i}\sum_{t=1}^{T_i}\min\left(\rho_\theta(a_{i,t})A_{i,t}^{\mathrm{HeRPO}}, \mathrm{clip}(\rho_\theta(a_{i,t}), 1-\epsilon, 1+\epsilon)\, A_{i,t}^{\mathrm{HeRPO}}\right)\right] \\ & - \beta\mathbb{D}_{KL}(\pi_\theta \| \pi_{\theta_{\mathrm{ref}}}),\end{aligned} \tag{12}$$

where $\rho_\theta(a_{i,t}) = \frac{\pi_\theta(a_{i,t}|s_{i,t})}{\pi_{\theta_{\mathrm{old}}}(a_{i,t}|s_{i,t})}$ is the importance sampling ratio and β controls the strength of the KL penalty.

5 Experiments

5.1 Experimental Setup

Benchmarks. We evaluate HeRPO on three challenging long-horizon agentic benchmarks, including ALFWorld Shridhar et al. (2021), WebShop Yao et al. (2022), and Sokoban Schrader (2018). ALFWorld evaluates text-based embodied decision making across six household task families: Pick, Look, Clean, Heat, Cool, and Pick2, comprising 3,827 task instances in total. We report the success rate for each task family as well as the overall success rate (%). WebShop is a large-scale web-interactive environment designed to evaluate grounded web interaction for product search and purchase, containing over 1.1 million products and approximately 12,000 user instructions. We report both the task score and success rate (%). Sokoban Schrader (2018) is a visual puzzle environment designed to evaluate spatial reasoning and long-horizon planning, in which an agent must push all boxes onto designated target locations while avoiding irreversible deadlocks. We procedurally generate 6×6 rooms containing a single box and provide the board states as RGB observations. Each episode is limited to at most 15 environment steps, and we report the task success rate (%).

Compared methods. We compare HeRPO against two closed-source models, GPT-4o OpenAI (2024) and Gemini-2.5-Pro Comanici et al. (2025); two prompting methods ReAct Yao et al. (2023) and Reflexion Shinn et al. (2023); and seven reinforcement-learning methods PPO Schulman et al. (2017), RLOO Ahmadian et al. (2024), GRPO Shao et al. (2024), EMPG Wang et al. (2026), GiGPO Feng et al. (2025), HCAPO Tan et al. (2026), and GraphGPO Cheng et al. (2026). To ensure a consistent and standardized comparison, we directly adopt the results of competing methods reported in the original HCAPO Tan et al. (2026) and GraphGPO Cheng et al. (2026) papers.

Implementation Details. Following prior works Feng et al. (2025); Cheng et al. (2026), we use Qwen2.5-1.5B/7B-Instruct Yang et al. (2024a) as the base LLMs for ALFWorld and WebShop, and Qwen2.5-VL-3B-Instruct for the visually grounded Sokoban environment. All RL-based methods share the same training framework (*i.e.*, verl-agent) to ensure fair comparisons. The rollout group size N for group-based RL methods is set to 8, and the discount factor γ is set to 0.95 following Feng et al. (2025). For HeRPO, the future-subsequence length $K = 2$ is set equal to the history length, and the clipping bound ϵ is shared with that used in the policy-optimization objective. Consequently, HeRPO introduces no additional tunable hyperparameters compared with GiGPO. Detailed training configurations are elaborated in Appendix B.

5.2 Experimental Results

5.2.1 Comparison on Agentic Benchmarks

Tables 1 and 2 compare HeRPO with prompting and RL-based methods on ALFWorld, WebShop, and Sokoban. Prompting methods improve over the untrained backbones but remain substantially behind RL-trained agents, highlighting the difficulty of learning effective long-horizon interaction from prompting alone. GRPO improves the average success rate on ALFWorld to 72.8% and 77.6% for the 1.5B and 7B models, respectively, while also yielding substantial gains on WebShop and Sokoban. GiGPO further improves performance by introducing the state-aligned grouping strategy, demonstrating the benefit of fine-grained credit assignment for training LLM-based agents. HeRPO consistently achieves the best performance across all three benchmarks and at all model scales.

Table 1: Performance on ALFWorld and WebShop. For most results, we report the mean ± standard deviation over three runs with different random seeds. The best and second-best results are highlighted in **bold** and underlined, respectively.

Type	Method	ALFWorld							WebShop	
		Pick	Look	Clean	Heat	Cool	Pick2	All	Score	Succ.
Closed-Source Models										
Prompting	GPT-4o	75.3	60.8	31.2	56.7	21.6	49.8	48.0	31.8	23.7
Prompting	Gemini-2.5-Pro	92.8	63.3	62.1	69.0	26.6	58.7	60.3	42.5	35.9
Qwen2.5-1.5B-Instruct										
Prompting	Qwen2.5	5.9	5.5	3.3	9.7	4.2	0.0	4.1	23.1	5.2
Prompting	ReAct	17.4	20.5	15.7	6.2	7.7	2.0	12.8	40.1	11.3
Prompting	Reflexion	35.3	22.2	21.7	13.6	19.4	3.7	21.8	55.8	21.9
RL Training	PPO	64.8±3.5	40.5±6.9	57.1±4.9	60.6±6.6	46.4±4.0	47.4±1.9	54.4±3.1	73.8±3.0	51.5±2.9
RL Training	RLOO	88.3±3.0	52.8±8.6	71.0±5.9	62.8±8.7	66.4±5.5	56.9±4.7	69.7±2.5	73.9±5.6	52.1±6.7
RL Training	GRPO	85.3±1.5	53.7±8.0	84.5±6.8	78.2±7.9	59.7±5.0	53.5±5.6	72.8±3.6	75.8±3.5	56.8±3.8
RL Training	EMPG	85.5	33.5	78.9	76.2	74.7	69.1	73.7	80.4	60.8
RL Training	GiGPO	94.4±5.9	67.5±4.6	94.8±3.8	94.4±7.8	79.8±4.7	76.4±5.4	86.7±1.7	83.1±1.6	65.0±3.2
RL Training	HCAPO	88.6±7.0	75.0±0.0	97.6±1.8	90.7±6.9	84.2±0.0	74.2±6.9	87.0±4.1	83.8±0.7	68.5±1.0
RL Training	GraphGPO	95.2±1.6	**100.0±0.0**	85.3±2.6	85.7±5.8	**96.3±2.6**	93.7±2.2	92.7±1.3	89.3±1.5	78.7±3.9
RL Training	**HeRPO (Ours)**	**100.0±0.0**	83.3±7.2	**100.0±0.0**	**98.4±2.7**	86.7±2.9	**94.7±0.0**	**95.8±0.5**	**91.6±1.6**	**80.7±1.2**
Qwen2.5-7B-Instruct										
RL Training	Qwen2.5	33.4	21.6	19.3	6.9	2.8	3.2	14.8	26.4	7.8
RL Training	ReAct	48.5	35.4	34.3	13.2	18.2	17.6	31.2	46.2	19.5
RL Training	Reflexion	62.0	41.6	44.9	30.9	36.3	23.8	42.7	58.1	28.8
RL Training	PPO (critic)	92.3±4.0	64.0±8.4	92.5±2.4	89.5±7.0	80.3±2.0	68.8±8.3	80.4±2.7	81.4±3.1	68.7±5.1
RL Training	RLOO	87.6±4.3	78.2±8.3	87.3±5.8	81.3±7.6	71.9±5.2	48.9±8.4	75.5±4.6	80.3±3.2	65.7±4.0
RL Training	GRPO	90.8±5.1	66.1±6.7	89.3±5.4	74.7±6.9	72.5±5.4	64.7±7.3	77.6±5.2	79.3±2.8	66.1±3.7
RL Training	EMPG	92.9	75.2	74.8	86.3	73.7	65.3	78.5	81.0	69.3
RL Training	GiGPO	97.7±1.6	82.7±7.9	**98.8±1.6**	83.7±7.2	89.3±8.2	79.2±6.6	90.8±1.3	84.4±2.9	72.8±3.2
RL Training	HCAPO	99.1±1.3	90.3±2.0	97.3±1.9	81.8±8.8	90.8±6.6	81.9±10.0	91.4±2.3	85.1±1.3	73.8±2.8
RL Training	GraphGPO	**100.0±0.0**	**100.0±0.0**	91.4±1.5	92.9±5.8	**94.4±0.0**	92.1±2.2	95.3±1.1	86.9±0.7	80.3±1.3
RL Training	**HeRPO (Ours)**	**100.0±0.0**	94.3±5.6	97.5±2.1	**100.0±0.0**	90.3±5.0	**100.0±0.0**	**97.4±0.5**	**92.2±1.8**	**83.1±1.5**

Compared with the strongest competing method (*i.e.*, GraphGPO), HeRPO improves the success rate on ALFWorld and WebShop by $3.1\%/2.1\%$ and $2.0\%/2.8\%$ with the 1.5B/7B backbones, respectively. HeRPO also yields a consistent performance gain of 1.5% on Sokoban. These results demonstrate our HeRPO's superiority across diverse interaction dynamics and task structures.

5.2.2 Ablation Study

Table 2: Performance on Sokoban. For most results, we report the mean ± standard deviation of the success rate (%) over three runs with different random seeds. The best and second-best results are highlighted in **bold** and underlined, respectively.

Type	Method	Sokoban [6×6]
Prompting	Qwen2.5-VL-3B-Instruct	11.7
RL Training	GRPO	67.1±4.7
RL Training	GiGPO	81.0±3.6
RL Training	GraphGPO	87.0±0.7
RL Training	HeRPO	**88.5**±1.2

We evaluate the contribution of each component in our HeRPO using the Qwen2.5-1.5B-Instruct model, with the results reported in Table 3. The results reveal that: (1) each component individually improves the overall metrics over the baseline on both benchmarks, demonstrating its effectiveness; (2) combining multiple components further improves performance, indicating that they provide complementary benefits for robust credit assignment across different tasks; and (3) jointly incorporating all three components consistently achieves the best overall performance on both benchmarks, demonstrating their synergistic effect and motivating the full HeRPO framework.

5.2.3 Adaptive vs. Fixed λ

To demonstrate the superiority of our adaptive GAE over the vanilla GAE for agentic RL, we replace the adaptive coefficient with fixed λ values of 0.1, 0.3, 0.5, 0.7, and 0.9, and evaluate all variants on ALFWorld using Qwen2.5-1.5B-Instruct. As shown in Figure 3(a), our adaptive strategy achieves performance competitive with the best fixed-λ setting evaluated. Performance declines as the fixed λ increases beyond 0.3, indicating sensitivity to the choice of λ in this setting. By adjusting λ

Table 3: Ablation study of HPG (Hindsight-Pruned Grouping), AGAE (Adaptive GAE), and HSR (Hindsight Self-Reweighting) with Qwen2.5-1.5B-Instruct as the base model. The best and second-best results are highlighted in **bold** and underlined, respectively.

Method	ALFWorld							WebShop	
	Pick	Look	Clean	Heat	Cool	Pick2	All	Score	Succ.
GiGPO (Baseline)	94.4±5.9	67.5±4.6	94.8±3.8	94.4±7.8	79.8±4.7	76.4±5.4	86.7±1.7	83.1±1.6	65.0±3.2
+ HSR	95.6±2.1	62.5±0.0	94.2±2.7	92.9±3.4	85.0±0.0	81.6±11.2	89.1±1.1	83.9±0.3	68.0±2.1
+ HPG	91.9±3.5	88.9±0.0	89.7±8.0	98.3±2.9	84.8±2.6	88.9±5.6	90.6±2.8	85.4±2.9	70.6±4.3
+ AGAE	96.8±3.2	**93.8±6.3**	98.8±2.1	89.3±9.3	85.8±6.2	88.3±5.8	93.0±1.6	87.2±0.8	74.2±1.4
+ HPG + AGAE	95.7±1.9	87.5±6.3	98.8±2.1	**100.0±0.0**	85.0±5.0	91.7±2.9	93.5±2.0	88.0±0.4	77.1±2.3
+ HPG + AGAE + HSR (HeRPO)	**100.0±0.0**	83.3±7.2	**100.0±0.0**	98.4±2.7	**86.7±2.9**	**94.7±0.0**	**95.8±0.5**	**91.6±1.6**	**80.7±1.2**



(a) Adaptive vs. Fixed λ (b) Comparison of Hindsight Self-Reweighting Granularities

Figure 3: (a) Performance comparison of the adaptive (ours) and fixed GAE coefficients on ALFWorld using Qwen2.5-1.5B-Instruct. (b) Training dynamics under different granularities of hindsight self-reweighting on ALFWorld using Qwen2.5-1.5B-Instruct. Curves report the mean overall success rate over training.

according to the effective group support, our method maintains strong performance without requiring a separate hyperparameter search for λ.

5.2.4 Discussion on the Granularity of Hindsight Self-Reweighting

To isolate the effect of weighting granularity in hindsight self-reweighting, we disable the other two components of HeRPO and retain only the baseline augmented with hindsight self-reweighting. We compare trajectory-, step-, and token-level weighting on ALFWorld using Qwen2.5-1.5B-Instruct. Under trajectory- and step-level weighting, all tokens within the same trajectory or interaction step, respectively, share the same advantage weight w. As shown in Figure 3(b), step-level weighting achieves the strongest performance throughout most of the later stages of training. Trajectory-level weighting remains stable but converges to a suboptimal performance, suggesting that sharing the weight across an entire trajectory yields overly coarse credit assignment. In contrast, token-level weighting may amplify local language-modeling noise and introduce inconsistent credit within a single action, causing model collapse. This behavior indicates unstable optimization when credit is assigned at an excessively fine granularity. These results are also consistent with the nature of agentic tasks, where environment feedback is naturally associated with executed actions rather than individual tokens. The action step therefore provides a more appropriate granularity for credit assignment, motivating the step-level weighting adopted in HeRPO.

5.2.5 Scalability of Hindsight Self-Reweighting

We further investigate whether hindsight self-reweighting (HSR) generalizes beyond the GiGPO baseline to other agentic RL methods. To this end, we additionally integrate HSR into GRPO Shao et al. (2024) and GraphGPO Cheng et al. (2026), using Qwen2.5-1.5B-Instruct and following the same training protocol on ALFWorld and WebShop. As shown in Table 4, incorporating HSR consistently improves the overall performance of all three baseline methods across both benchmarks, with particularly pronounced gains for GRPO, which lacks fine-grained credit assignment. These results demonstrate the generalizability of HSR across various agentic RL methods and highlight its potential as a plug-and-play module for improving credit assignment.

Table 4: Scalability of hindsight self-reweighting (HSR) across GRPO, GiGPO, and GraphGPO on ALFWorld and WebShop using Qwen2.5-1.5B-Instruct. Results are reported as mean ± standard deviation over three random seeds. Shaded rows denote HSR-augmented variants.

Type	Method	ALFWorld							WebShop	
		Pick	Look	Clean	Heat	Cool	Pick2	All	Score	Succ.
RL Training	GRPO	85.3±1.5	53.7±8.0	84.5±6.8	78.2±7.9	59.7±5.0	53.5±5.6	72.8±3.6	75.8±3.5	56.8±3.8
RL Training	GRPO + HSR	93.1±5.3	64.2±8.9	80.6±7.0	78.3±8.0	77.8±4.8	47.9±3.6	76.1±1.5	79.5±5.3	63.5±4.8
RL Training	GiGPO	94.4±5.9	67.5±4.6	94.8±3.8	94.4±7.8	79.8±4.7	76.4±5.4	86.7±1.7	83.1±1.6	65.0±3.2
RL Training	GiGPO + HSR	95.6±2.1	62.5±0.0	94.2±2.7	92.9±3.4	85.0±0.0	81.6±11.2	89.1±1.1	83.9±0.3	68.0±2.1
RL Training	GraphGPO	95.2±1.6	100.0±0.0	85.3±2.6	85.7±5.8	96.3±2.6	93.7±2.2	92.7±1.3	89.3±1.5	78.7±3.9
RL Training	GraphGPO + HSR	99.1±1.5	91.7±0.0	98.5±2.6	97.8±3.8	81.2±1.8	91.2±3.0	93.8±0.8	89.5±0.5	79.9±1.8

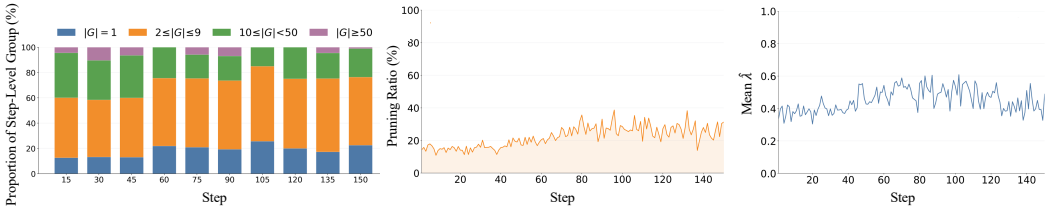


Figure 4: Training dynamics of HeRPO on ALFWorld. Left: proportion of step-level groups across four retained-support ranges. Middle: fraction of state-aligned candidates removed by hindsight-pruned grouping. Right: mean adaptive coefficient $\hat{\lambda}$ used in adaptive GAE.

5.2.6 Training Dynamics

We analyze the training dynamics of HeRPO on ALFWorld to examine how retained step-level group, hindsight pruning, and the adaptive coefficient $\hat{\lambda}$ evolve during training. As shown in Figure 4 (left), the retained step-level groups are consistently dominated by small-to-moderate support (*i.e.*, $2 \leq |G| \leq 9$), accounting for over half of the samples throughout training. $|G| = 1$ groups constitute only 12.6%-25.7% of the samples, while groups with $|G| \geq 50$ remain rare. Despite short-term fluctuations, the pruning ratio in Figure 4 (middle) exhibits a clear overall increase from the early to the later stages of training, indicating that hindsight pruning remains active as the policy evolves. Figure 4 (right) further shows that the mean adaptive coefficient $\hat{\lambda}$ varies substantially rather than converging to a fixed value. This variation is consistent with the design of adaptive GAE, which continuously adjusts $\hat{\lambda}$ according to changes in the effective support of the retained groups. Together, these results verify that both hindsight pruning and adaptive coefficient $\hat{\lambda}$ remain dynamic throughout training.

6 Conclusion

In this work, we introduced HeRPO, a critic-free framework for fine-grained credit assignment in long-horizon agentic RL. The key idea of HeRPO is to elevate realized futures from mere return signals to privileged hindsight for retrospectively assessing intermediate actions. By leveraging this hindsight, HeRPO constructs more reliable comparison action groups, adaptively balances local and long-term credit according to effective sample support, and calibrates the magnitude of policy updates through hindsight-induced shifts in action preference. HeRPO improves both the reliability of step-level credit signals and the precision of policy optimization under sparse outcome supervision. Extensive experiments across long-horizon agentic benchmarks demonstrate that HeRPO consistently outperforms representative agentic RL methods, highlighting the effectiveness of incorporating realized futures into fine-grained credit assignment. HeRPO nevertheless has two limitations. First, its hindsight-based assessments are derived from the policy itself, and their quality is therefore bounded by the model's underlying ability. Second, when the future feedback is ambiguous, the resulting credit signals may become less informative. Future work will explore extending HeRPO to more complex and diverse agentic environments, as well as developing more robust mechanisms for calculating credit signals under limited conditions.

AI USE STATEMENT

Generative AI tools were used to generate and refine the visual elements in the figures and to improve the clarity, grammar, and readability of the manuscript. All AI-assisted visual and textual content was carefully reviewed and revised by the authors to ensure accuracy and consistency. The research questions, methodology, experiments, analyses, scientific claims, and conclusions were developed and validated by the authors. The authors take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

We provide detailed information to facilitate the reproduction of our results. The complete HeRPO training procedure is summarized in Algorithm 1, including trajectory collection, hindsight construction, hindsight-pruned grouping, adaptive advantage estimation, hindsight self-reweighting, and policy optimization. The mathematical definitions of all components and their corresponding hyperparameters are specified in the main text.

Detailed implementation and training configurations are provided in the appendix, including the backbone models, hardware, learning rates, rollout group size, discount factor, clipping coefficient, KL coefficient, training duration, maximum interaction horizons, and prompt and response length limits for each environment. We additionally describe the reward schemes, batch construction, rollout sampling strategy, and evaluation metrics used for ALFWorld, WebShop, Sokoban, and SciWorld.

To make the agent–environment interaction reproducible, we provide the complete prompt templates used for all main experimental environments, together with the construction of the privileged hindsight context. We also report component ablations, rollout-group-size ablations, hindsight-construction ablations, and training dynamics to clarify the contribution and behavior of individual design choices.

For comparisons with prior methods, we follow the corresponding open-source training frameworks and experimental protocols whenever applicable, and explicitly state that baseline results are adopted from previously published work. We report the mean and standard deviation over three runs with different random seeds (0, 1, and 2) for most main experimental results. The source code and configurations required to reproduce HeRPO will be released upon publication.

REFERENCES

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting REINFORCE-style optimization for learning from human feedback in LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12248–12267, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.662.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Xin Cheng, Shuo He, Lang Feng, Haiyang Xu, Ming Yan, Lei Feng, and Bo An. Beyond trajectory-level attribution: Graph-based credit assignment for agentic reinforcement learning. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*. PMLR, 2026.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. doi: 10.48550/arXiv.2507.06261. URL https://arxiv.org/abs/2507.06261.

Guanting Dong, Hangyu Mao, Kai Ma, Licheng Bao, Yifei Chen, Zhongyuan Wang, Zhongxia Chen, Jiazhen Du, Huiyang Wang, Fuzheng Zhang, Guorui Zhou, Yutao Zhu, Ji-Rong Wen, and Zhicheng Dou. Agentic reinforced policy optimization. In *International Conference on Learning Representations*, 2026. URL https://iclr.cc/virtual/2026/poster/10009306.

Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. Group-in-group policy optimization for LLM agent training. In *Advances in Neural Information Processing Systems*, volume 38, pp. 46375–46408, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/420c9f777c0b4f78d515e53cf74d58b2-Abstract-Conference.html.

Shuo He, Lang Feng, Qi Wei, Xin Cheng, Lei Feng, and Bo An. Hierarchy-of-groups policy optimization for long-horizon agentic tasks. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=T8Dev99qnz.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Dong Wang, Hamed Zamani, and Jiawei Han. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*, 2025.

Vijay Konda and John Tsitsiklis. Actor-critic algorithms. *Advances in neural information processing systems*, 12, 1999.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. Agentbench: Evaluating llms as agents. In *International Conference on Learning Representations*, 2024.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding R1-Zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*, 2025.

Xufang Luo, Yuge Zhang, Zhiyuan He, Zilong Wang, Siyun Zhao, Dongsheng Li, Luna K. Qiu, and Yuqing Yang. Agent lightning: Train any ai agents with reinforcement learning. *arXiv preprint arXiv:2508.03680*, 2025.

OpenAI. GPT-4o system card. Technical report, OpenAI, August 2024. URL https://openai.com/index/gpt-4o-system-card/.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744, 2022.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pp. 53728–53741, 2023.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems*, volume 36, pp. 68539–68551, 2023.

Max-Philipp B. Schrader. gym-sokoban: Sokoban environment for openai gym. https://github.com/mpSchrader/gym-sokoban, 2018. Version 0.0.6.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*, 2015.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. doi: 10.48550/arXiv.2402.03300. URL https://arxiv.org/abs/2402.03300.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-0377. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/1b44b878bb782e6954cd888628510e90-Abstract-Conference.html.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. ALFWorld: Aligning text and embodied environments for interactive learning. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=0IOX0YcCdTn.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F. Christiano. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*, volume 33, pp. 3008–3021, 2020.

Yi Su, Dian Yu, Linfeng Song, Juntao Li, Haitao Mi, Zhaopeng Tu, Min Zhang, and Dong Yu. Crossing the reward bridge: Expanding reinforcement learning with verifiable rewards across diverse domains. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3872–3892, 2026.

Hui-Ze Tan, Xiao-Wen Yang, Hao Chen, Jie-Jing Shao, Yi Wen, Yuteng Shen, Weihong Luo, Xiku Du, Lan-Zhe Guo, and Yu-Feng Li. Hindsight credit assignment for long-horizon LLM agents. *arXiv preprint arXiv:2603.08754*, 2026. doi: 10.48550/arXiv.2603.08754. URL https://arxiv.org/abs/2603.08754.

Jiawei Wang, Jiacai Liu, Yuqian Fu, Yingru Li, Xintao Wang, Yuan Lin, Yu Yue, Lin Zhang, Yang Wang, and Ke Wang. Harnessing uncertainty: Entropy-modulated policy gradients for long-horizon LLM agents. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*. PMLR, 2026.

Ruoyao Wang, Peter Jansen, Marc-Alexandre Côté, and Prithviraj Ammanabrolu. ScienceWorld: Is your agent smarter than a 5th grader? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11279–11298. Association for Computational Linguistics, 2022.

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, and Manling Li. Ragen: Understanding self-evolution in llm agents via multi-turn reinforcement learning. *arXiv preprint arXiv:2504.20073*, 2025.

Zhiheng Xi, Jixuan Huang, Chenyang Liao, Baodai Huang, Honglin Guo, Jiaqi Liu, Rui Zheng, Junjie Ye, Jiazheng Zhang, Wenxiang Chen, Wei He, Yiwen Ding, Guanyu Li, Zehui Chen, Zhengyin Du, Xuesong Yao, Yufei Xu, Jiecao Chen, Tao Gui, Zuxuan Wu, Qi Zhang, Xuanjing Huang, and Yu-Gang Jiang. Agentgym-rl: Training llm agents for long-horizon decision making through multi-turn reinforcement learning. *arXiv preprint arXiv:2509.08755*, 2025. URL https://arxiv.org/abs/2509.08755.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024a.

John Yang, Carlos E. Jimenez, Alexander Wettig, Kilian Lieret, Shunyu Yao, Karthik Narasimhan, and Ofir Press. Swe-agent: Agent-computer interfaces enable automated software engineering. In *Advances in Neural Information Processing Systems*, volume 37, pp. 50528–50652, 2024b.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. WebShop: Towards scalable real-world web interaction with grounded language agents. In *Advances in Neural Information Processing Systems*, volume 35, 2022. doi: 10.52202/068431-1508. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/82ad13ec01f9fe44c01cb91814fd7b8c-Abstract-Conference.html.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

Hongli Yu, Tinghong Chen, Jiangtao Feng, Jiangjie Chen, Weinan Dai, Qiying Yu, Ya-Qin Zhang, Wei-Ying Ma, Jingjing Liu, Mingxuan Wang, and Hao Zhou. Memagent: Reshaping long-context llm with multi-conv rl-based memory agent. *arXiv preprint arXiv:2507.02259*, 2025a.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Juncai Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Ru Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Yonghui Wu, and Mingxuan Wang. DAPO: An open-source LLM reinforcement learning system at scale. In *Advances in Neural Information Processing Systems*, volume 38, pp. 113222–113244. Curran Associates, Inc., 2025b. doi: 10.52202/085713-3775.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. Webarena: A realistic web environment for building autonomous agents. In *International Conference on Learning Representations*, 2024.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

Zefang Zong, Dingwei Chen, Yang Li, Qi Yi, Bo Zhou, Chengming Li, Bo Qian, Peng Chen, and Jie Jiang. AT^{2}PO: Agentic turn-based policy optimization via tree search. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, pp. 24120–24143. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1106. URL https://aclanthology.org/2026.acl-long.1106/.

A Pseudo Code

Algorithm 1 summarizes the pipeline of HeRPO. The privileged hindsight is constructed only after an episode terminates and is used retrospectively to estimate credit and calibrate the subsequent policy update, which will not be used during validation.

Algorithm 1 Hindsight-Enabled Retrospective Policy Optimization (HeRPO)

Require: Initial policy π_θ, task distribution $\mathcal{D}$, group size N, discount γ, clipping bound ϵ, KL coefficient β, maximum step T

for each training iteration **do**
  Update the old rollout policy: $\pi_{\theta_{\text{old}}} \leftarrow \pi_\theta$
  Sample a task $x \sim \mathcal{D}$ and collect N complete trajectories $\{\tau_i\}_{i=1}^{N}$ using $\pi_{\theta_{\text{old}}}$
  Compute A_i^{traj} for each trajectory τ_i by Eq. (1)
  for each executed action step (i, t) **do**
    ▷ **Hindsight-Pruned Grouping:**
    Construct $\mathcal{H}_{i,t}$ by Eq. (2)
    Form $G(s_{i,t})$ and Estimate $V_{i,t}$ by Eq. (4)
    ▷ **Adaptive GAE:**
    Compute $E_{i,t}$ and $\hat{\lambda}_{i,t}$ by Eq. (6) and Eq. (7), respectively
    Estimate $A_{i,t}^{\text{AGAE}} \leftarrow \delta_{i,t} + \gamma\hat{\lambda}_{i,t} A_{i,t+1}^{\text{AGAE}}$
    ▷ **Hindsight Self-Reweighting:**
    Compute the preference shift $\Delta_{i,t}$ by Eq. (9) and then map it into $w_{i,t}$ by Eq. (10)
    $A_{i,t}^{\text{HeRPO}} \leftarrow w_{i,t}(A_i^{\text{traj}} + A_{i,t}^{\text{AGAE}})$
  end for
  Update θ by maximizing the objective in Eq. (12)
end for

B Implementation Details

B.1 Training Configuration

We implement HeRPO and conduct experiments on the ALFWorld, WebShop, and Sokoban environments using the verl-agent Feng et al. (2025) training framework. For experiments with Qwen2.5-1.5B-Instruct and Qwen2.5-VL-3B-Instruct, training is conducted on two NVIDIA H20 GPUs, each with 96 GB of memory. For the larger Qwen2.5-7B-Instruct model, we use four NVIDIA H20 GPUs, each with 96 GB of memory. Unless otherwise specified, all experiments are trained for 150 epochs.

For all model scales, we use a clipping ratio ϵ of 0.2, and a KL-divergence loss coefficient of 0.01. The learning rate is set to 1×10^{-6} for the actor and set to 1×10^{-5} for the critic, where the critic is included only in PPO. We sample training rollouts with temperature 1.0 and evaluate with temperature 0.4.

B.2 Prompt Templates

The prompt templates used for ALFWorld, WebShop, and Sokoban are shown in Figure 5, Figure 6, and Figure 7, respectively. All prompts are constructed using Python-style string formatting, with placeholders enclosed in {} denoting semantic slots that are dynamically instantiated at each interaction step.

Specifically, {task_description} specifies the task objective; {step_count} and {current_step} identify the interaction progress; {history_length} and {action_history} provide up to two recent action-observation records; {current_observation} contains the latest environment response; and {admissible_actions} enumerates the actions allowed at the current state. For Sokoban, the environment serves the {image} marker as a placeholder to represent the visual observation. Reasoning and the selected environment action are parsed from the required <think></think> and <action></action> tags, respectively.

Prompt Template for ALFWorld

You are an expert agent operating in the ALFRED Embodied Environment. Your task is to: {task_description}
Prior to this step, you have already taken {step_count} step(s). Below are the most recent {history_length} observations and the corresponding actions you took: {action_history}
You are now at step {current_step} and your current observation is: {current_observation}
Your admissible actions of the current situation are: [{admissible_actions}].
Now it's your turn to take an action.
You should first reason step-by-step about the current situation. This reasoning process MUST be enclosed within <think> </think> tags.
Once you've finished your reasoning, you should choose an admissible action for the current step and present it within <action> </action> tags.

Figure 5: Prompt template used for ALFWorld.

Prompt Template for WebShop

You are an expert autonomous agent operating in the WebShop e-commerce environment.
Your task is to: {task_description}.
Prior to this step, you have already taken {step_count} step(s). Below are the most recent {history_length} observations and the corresponding actions you took: {action_history}
You are now at step {current_step} and your current observation is: {current_observation}.
Your admissible actions of the current situation are: [{available_actions}].
Now it's your turn to take one action for the current step.
You should first reason step-by-step about the current situation, then think carefully which admissible action best advances the shopping goal. This reasoning process MUST be enclosed within <think> </think> tags.
Once you've finished your reasoning, you should choose an admissible action for the current step and present it within <action> </action> tags.

Figure 6: Prompt template used for WebShop.

Prompt Template for Sokoban

You are an expert agent operating in the Sokoban environment. Your goal is to push all the boxes onto the target spots. Once all boxes are on the targets, you win!
 # Rules
You can only push boxes. You can't pull them, so plan ahead to avoid getting stuck.
You can't walk through or push boxes into walls.
To avoid traps, do not push boxes into corners or against walls where they can't be moved again.
 # Visual Elements in the Image:
Character: A small, green alien-like figure with two antennae and black eyes. It represents you.
Box: A yellow crate marked with an orange "X" across its front. It is the box you need to push.
Target: A black tile outlined in red, with a small red diamond shape in the center. It marks the destination where a box should be pushed.
 # Current Step
Your current observation is shown in the image: {image}
Your admissible actions are ["up", "down", "left", "right"].
Now it's your turn to make a move (choose ONE action only for the current step).
You should first reason step-by-step about the current situation—observe the positions of boxes and targets, plan a path to push a box toward a target, and avoid traps like corners or walls. This reasoning process MUST be enclosed within <think> </think> tags.
Once you've finished your reasoning, you should choose an admissible action for the current step and present it within <action> </action> tags.

Figure 7: Prompt template used for Sokoban.

As shown in Figure 8, HeRPO constructs the hindsight from the terminal episodic outcome and realized futures. Up to $K = 2$ future actions with corresponding observations are included. At the end of a trajectory, the future-action section is omitted.

Hindsight Construction

=== Reference Context ===

The following information contains the final outcome, the immediate next observation, and future actions with corresponding observations for reference.

[Outcome]

final reward = {final_reward}

[Immediate Next Observation]

observation= {next_observation}

[Future Actions and Observations]

step+{num}:
  action: {future_action}
  observation: {future_observation}

Figure 8: Privileged hindsight information used in our HeRPO.

B.3 ALFWORLD

ALFWorld episodes contain at most 50 environment steps. We set the maximum prompt and response lengths to 2,048 and 512 tokens, respectively. The 1.5B and 7B models use mini-batch sizes of 256 and 64, respectively. At each interaction step, the environment provides the current textual observation together with the set of admissible actions. We adopt a rule-based reward scheme that assigns a reward of 10 upon successful task completion and 0 otherwise. Invalid actions generated by the agent incur an additional penalty of -0.1. Each training batch contains 16 task instances, from each of which we sample $N = 8$ trajectories starting from the same initial state, yielding 128 environment trajectories per rollout batch. We evaluate the agent based on task success across the six ALFWorld task families, as well as overall success.

B.4 WEBSHOP

WebShop episodes contain at most 15 environment steps. Because product pages and action spaces generally induce longer contexts, we increase the maximum prompt length to 4,096 tokens while retaining a maximum response length of 512 tokens. Both backbone models use a mini-batch size of 64. We use the environment-provided task score as the episodic outcome signal, assigning a reward of 10 for successful task completion and 0 otherwise. Invalid actions incur an additional penalty of -0.1. Each training batch contains 16 task instances, from each of which we sample $N = 8$ trajectories starting from the same initial state, yielding 128 environment trajectories per rollout batch. We report both the average task score and the exact task success rate.

B.5 SOKOBAN

For the supplementary visual Sokoban setting, we use Qwen2.5-VL-3B-Instruct with RGB observations of the game board. Episodes are generated on 6×6 rooms containing a single box and are limited to at most 15 environment steps. We set the maximum prompt and response lengths to 1,024 and 512 tokens, respectively, and use a mini-batch size of 64. We adopt a rule-based reward scheme that assigns a reward of 10 upon successful task completion and 0 otherwise. Invalid actions generated by the agent incur an additional penalty of -0.1. For each rollout batch, we sample $N = 8$ trajectories for each of 32 independent task instances, resulting in 256 trajectories in total. We evaluate performance using the environment-provided terminal success signal and report the validation success rate.

C ADDITIONAL EXPERIMENTS

C.1 COMPARISON ON SCIWORLD

SciWorld is a text-based environment for scientific task solving Wang et al. (2022), designed to evaluate whether agents can apply elementary scientific knowledge through grounded experimentation rather than answering questions from static text alone. At each interaction step, the agent receives a

Table 5: Overall success rate (%) comparison on SciWorld. The best and second-best results are highlighted in **bold** and underlined, respectively.

Type	Model	SciWorld
Proprietary	GPT-4o	21.0
Proprietary	Qwen-Max	13.5
Proprietary	Gemini-2.5-Flash	21.0
Proprietary	OpenAI o4-mini	29.5
Proprietary	OpenAI o3	41.5
Proprietary	Gemini-2.5-Pro	12.5
Models $\geq$100B	Qwen3-235B-A22B	23.5
Models $\geq$100B	DeepSeek-V3-0324	0.5
Models $\geq$100B	DeepSeek-R1-0528	4.5
Models $<$100B	Qwen2.5-72B-Instruct	9.5
Models $<$100B	Qwen3-32B	14.0
Models $<$100B	Llama-3.1-70B-Instruct	25.0
Qwen2.5-7B-Instruct		
RL Training	AgentGym-RL-7B	50.5
RL Training	ScalingInter-7B	<u>57.0</u>
RL Training	HeRPO (Ours)	**61.5**

Table 6: Performance comparison of HeRPO under different rollout group sizes on ALFWorld using Qwen2.5-1.5B-Instruct as the backbone. Results are reported as mean $\pm$ standard deviation over three random seeds.

Group Size	ALFWorld						
	Pick	Look	Clean	Heat	Cool	Pick2	All
4	98.0±1.7	66.7±7.2	98.7±2.2	100.0±0.0	83.3±2.9	86.0±6.1	92.4±0.9
6	100.0±0.0	75.0±0.0	98.7±2.2	95.2±0.0	88.3±7.6	89.5±5.3	94.0±0.5
8 (Ours)	100.0±0.0	83.3±7.2	100.0±0.0	98.4±2.7	86.7±2.9	94.7±0.0	95.8±0.5

textual observation, selects an executable action to inspect or manipulate the simulated environment, and continues interacting until it completes the target experiment or reaches the interaction limit.

We follow the open-source AgentGym-RL framework Xi et al. (2025) and adopt its released data split, where training and test sets contain 2,120 and 200 task instances, and use Qwen2.5-7B-Instruct as the backbone model. SciWorld can be categorized into five subsets according to their task objectives: Measure, Test-Conductivity, Find, Chemistry-Mix, and Lifespan. We report the overall task success rate (%) on the test set. Each episode allows at most 20 interaction steps, with a discount factor of $\gamma = 0.95$. We set the maximum prompt and response lengths to 1,024 and 4,096 tokens, respectively, while limiting each interaction to at most 200 newly generated tokens and the overall model context length to 8,192 tokens. Each training batch contains 16 task instances, with $N = 8$ trajectories sampled per instance, resulting in 128 trajectories per rollout batch. We use an actor learning rate of 1×10^{-6} and a clipping coefficient of 0.2. All models are trained for 250 optimization steps on eight NVIDIA H20 GPUs.

As shown in Table 5, HeRPO achieves the best overall performance, substantially outperforming both proprietary and open-source models. Under the same RL training setup with Qwen2.5-7B-Instruct as the backbone, HeRPO also consistently surpasses AgentGym-RL-7B and ScalingInter-7B. These results further demonstrate the effectiveness and generalizability of our step-level credit assignment mechanism, extending its benefits beyond the embodied, web-interaction, and visual-planning environments studied in the main experiments to interactive scientific experimentation.

Table 7: Performance comparison of different hindsight construction strategies in HeRPO on ALFWorld using Qwen2.5-1.5B-Instruct as the backbone. Results are reported as mean ± standard deviation over three random seeds.

Setting	ALFWorld						
	Pick	Look	Clean	Heat	Cool	Pick2	All
w/o Outcome	100.0±0.0	70.8±7.2	100.0±0.0	98.4±2.7	85.6±4.1	91.2±3.0	94.0±1.2
w/o Future Act./Obs.	98.0±1.7	88.0±12.5	97.4±2.2	98.4±2.7	85.3±5.5	98.3±2.9	95.3±1.4
HeRPO (Ours)	100.0±0.0	83.3±7.2	100.0±0.0	98.4±2.7	86.7±2.9	94.7±0.0	95.8±0.5

C.2 Ablation on Rollout Group Size

We investigate the effect of the rollout group size on ALFWorld. As shown in Table 6, increasing the rollout group size consistently improves the overall performance, suggesting that a larger group provides more diverse trajectories for within-group comparison and enables more reliable relative credit assignment.

C.3 Ablation on Hindsight Construction

We further investigate the effect of different hindsight construction strategies on ALFWorld. As shown in Table 7, constructing hindsight with all signals achieves the strongest overall performance. Removing either future action/observation signal or outcome signal leads to performance degradation, indicating that they provide complementary information for reassessing previous actions. In particular, outcome information provides an important signal for distinguishing the consequences of different decisions.

D Case Study

Follow, we present two complete case study trajectories of an agent trained by HeRPO on ALFWorld and WebShop, including the prompt and the response at each step.

D.1 ALFWorld

The task is to "put a hot plate in diningtable." The six-step trajectory first locates and acquires a plate, then uses the microwave to satisfy the required state transformation before navigating to the target receptacle.

Step 1: Locate a candidate object

Prompt.

```
You are an expert agent operating in the ALFRED Embodied Environment.
Your current observation is: -= Welcome to TextWorld, ALFRED! =-

You are in the middle of a room. Looking quickly around you, you see a cabinet 11, a cabinet 10, a
      cabinet 9, a cabinet 8, a cabinet 7, a cabinet 6, a cabinet 5, a cabinet 4, a cabinet 3, a cabinet
      2, a cabinet 1, a coffeemachine 1, a countertop 2, a countertop 1, a diningtable 2, a diningtable 1,
       a drawer 3, a drawer 2, a drawer 1, a fridge 1, a garbagecan 1, a microwave 1, a sidetable 1, a
      sinkbasin 1, a stoveburner 4, a stoveburner 3, a stoveburner 2, a stoveburner 1, and a toaster 1.

Your task is to: put a hot plate in diningtable.
Your admissible actions of the current situation are: ['go to cabinet 1'
 'go to cabinet 10'
 'go to cabinet 11'
 'go to cabinet 2'
 'go to cabinet 3'
 'go to cabinet 4'
 'go to cabinet 5'
 'go to cabinet 6'
 'go to cabinet 7'
 'go to cabinet 8'
 'go to cabinet 9'
 'go to coffeemachine 1'
 'go to countertop 1'
 'go to countertop 2'
 'go to diningtable 1'
```

```
 'go to diningtable 2'
 'go to drawer 1'
 'go to drawer 2'
 'go to drawer 3'
 'go to fridge 1'
 'go to garbagecan 1'
 'go to microwave 1'
 'go to sidetable 1'
 'go to sinkbasin 1'
 'go to stoveburner 1'
 'go to stoveburner 2'
 'go to stoveburner 3'
 'go to stoveburner 4'
 'go to toaster 1'
 'inventory'
 'look'].

Now it's your turn to take an action.
You should first reason step-by-step about the current situation. This reasoning process MUST be enclosed
      within <think> </think> tags.
Once you've finished your reasoning, you should choose an admissible action for current step and present
     it within <action> </action> tags.
```

Response.

<think> The task is to put a hot plate on the diningtable. I need to find a plate, heat it, and then place it on the diningtable. </think>

<action>go to countertop 1</action>

Step 2: Acquire the plate

Prompt.

```
You are an expert agent operating in the ALFRED Embodied Environment. Your task is to: put a hot plate in
      diningtable.
Prior to this step, you have already taken 1 step(s). Below are the most recent 1 observaitons and the
     corresponding actions you took: [Observation 1: '-= Welcome to TextWorld, ALFRED! =-

You are in the middle of a room. Looking quickly around you, you see a cabinet 11, a cabinet 10, a
     cabinet 9, a cabinet 8, a cabinet 7, a cabinet 6, a cabinet 5, a cabinet 4, a cabinet 3, a cabinet
     2, a cabinet 1, a coffeemachine 1, a countertop 2, a countertop 1, a diningtable 2, a diningtable 1,
      a drawer 3, a drawer 2, a drawer 1, a fridge 1, a garbagecan 1, a microwave 1, a sidetable 1, a
     sinkbasin 1, a stoveburner 4, a stoveburner 3, a stoveburner 2, a stoveburner 1, and a toaster 1.

Your task is to: put a hot plate in diningtable.', Action 1: 'go to countertop 1']
You are now at step 2 and your current observation is: You arrive at countertop 1. On the countertop 1,
     you see a pan 1, a plate 2, a soapbottle 2, and a tomato 1.
Your admissible actions of the current situation are: ['examine countertop 1'
 'go to cabinet 1'
 'go to cabinet 10'
 'go to cabinet 11'
 'go to cabinet 2'
 'go to cabinet 3'
 'go to cabinet 4'
 'go to cabinet 5'
 'go to cabinet 6'
 'go to cabinet 7'
 'go to cabinet 8'
 'go to cabinet 9'
 'go to coffeemachine 1'
 'go to countertop 2'
 'go to diningtable 1'
 'go to diningtable 2'
 'go to drawer 1'
 'go to drawer 2'
 'go to drawer 3'
 'go to fridge 1'
 'go to garbagecan 1'
 'go to microwave 1'
 'go to sidetable 1'
 'go to sinkbasin 1'
 'go to stoveburner 1'
 'go to stoveburner 2'
 'go to stoveburner 3'
 'go to stoveburner 4'
 'go to toaster 1'
 'inventory'
 'look'
 'take pan 1 from countertop 1'
 'take plate 2 from countertop 1'
 'take soapbottle 2 from countertop 1'
 'take tomato 1 from countertop 1'].

Now it's your turn to take an action.
You should first reason step-by-step about the current situation. This reasoning process MUST be enclosed
      within <think> </think> tags.
Once you've finished your reasoning, you should choose an admissible action for current step and present
     it within <action> </action> tags.
```

Response.

<think> The task is to put a hot plate in diningtable. I have already observed the countertop 1 where I can find a plate 2. I need to take the plate and put it on the diningtable. </think>

<action>take plate 2 from countertop 1</action>

Step 3: Navigate to the heating appliance

Prompt.

```
You are an expert agent operating in the ALFRED Embodied Environment. Your task is to: put a hot plate in
      diningtable.
Prior to this step, you have already taken 2 step(s). Below are the most recent 2 observaitons and the
      corresponding actions you took: [Observation 1: '-= Welcome to TextWorld, ALFRED! =-

You are in the middle of a room. Looking quickly around you, you see a cabinet 11, a cabinet 10, a
      cabinet 9, a cabinet 8, a cabinet 7, a cabinet 6, a cabinet 5, a cabinet 4, a cabinet 3, a cabinet
      2, a cabinet 1, a coffeemachine 1, a countertop 2, a countertop 1, a diningtable 2, a diningtable 1,
       a drawer 3, a drawer 2, a drawer 1, a fridge 1, a garbagecan 1, a microwave 1, a sidetable 1, a
      sinkbasin 1, a stoveburner 4, a stoveburner 3, a stoveburner 2, a stoveburner 1, and a toaster 1.

Your task is to: put a hot plate in diningtable.', Action 1: 'go to countertop 1']
[Observation 2: 'You arrive at countertop 1. On the countertop 1, you see a pan 1, a plate 2, a
      soapbottle 2, and a tomato 1.', Action 2: 'take plate 2 from countertop 1']
You are now at step 3 and your current observation is: You pick up the plate 2 from the countertop 1.
Your admissible actions of the current situation are: ['examine countertop 1'
 'examine plate 2'
 'go to cabinet 1'
 'go to cabinet 10'
 'go to cabinet 11'
 'go to cabinet 2'
 'go to cabinet 3'
 'go to cabinet 4'
 'go to cabinet 5'
 'go to cabinet 6'
 'go to cabinet 7'
 'go to cabinet 8'
 'go to cabinet 9'
 'go to coffeemachine 1'
 'go to countertop 2'
 'go to diningtable 1'
 'go to diningtable 2'
 'go to drawer 1'
 'go to drawer 2'
 'go to drawer 3'
 'go to fridge 1'
 'go to garbagecan 1'
 'go to microwave 1'
 'go to sidetable 1'
 'go to sinkbasin 1'
 'go to stoveburner 1'
 'go to stoveburner 2'
 'go to stoveburner 3'
 'go to stoveburner 4'
 'go to toaster 1'
 'inventory'
 'look'
 'move plate 2 to countertop 1'].

Now it's your turn to take an action.
You should first reason step-by-step about the current situation. This reasoning process MUST be enclosed
      within <think> </think> tags.
Once you've finished your reasoning, you should choose an admissible action for current step and present
      it within <action> </action> tags.
```

Response.

<think> The task is to put a hot plate in the diningtable. I have already taken two steps to get the plate from the countertop. Now, I need to heat the plate and then place it on the diningtable. </think>

<action>go to microwave 1</action>

Step 4: Transform the object state

Prompt.

```
You are an expert agent operating in the ALFRED Embodied Environment. Your task is to: put a hot plate in
      diningtable.
Prior to this step, you have already taken 3 step(s). Below are the most recent 2 observaitons and the
      corresponding actions you took: [Observation 2: 'You arrive at countertop 1. On the countertop 1,
      you see a pan 1, a plate 2, a soapbottle 2, and a tomato 1.', Action 2: 'take plate 2 from
      countertop 1']
[Observation 3: 'You pick up the plate 2 from the countertop 1.', Action 3: 'go to microwave 1']
You are now at step 4 and your current observation is: You arrive at microwave 1. The microwave 1 is
      closed.
Your admissible actions of the current situation are: ['examine microwave 1'
```

```
 'examine plate 2'
 'go to cabinet 1'
 'go to cabinet 10'
 'go to cabinet 11'
 'go to cabinet 2'
 'go to cabinet 3'
 'go to cabinet 4'
 'go to cabinet 5'
 'go to cabinet 6'
 'go to cabinet 7'
 'go to cabinet 8'
 'go to cabinet 9'
 'go to coffeemachine 1'
 'go to countertop 1'
 'go to countertop 2'
 'go to diningtable 1'
 'go to diningtable 2'
 'go to drawer 1'
 'go to drawer 2'
 'go to drawer 3'
 'go to fridge 1'
 'go to garbagecan 1'
 'go to sidetable 1'
 'go to sinkbasin 1'
 'go to stoveburner 1'
 'go to stoveburner 2'
 'go to stoveburner 3'
 'go to stoveburner 4'
 'go to toaster 1'
 'heat plate 2 with microwave 1'
 'inventory'
 'look'
 'open microwave 1'].

Now it's your turn to take an action.
You should first reason step-by-step about the current situation. This reasoning process MUST be enclosed
      within <think> </think> tags.
Once you've finished your reasoning, you should choose an admissible action for current step and present
     it within <action> </action> tags.
```

Response.

<think> The task is to put a hot plate in the diningtable. I have already taken the plate from the countertop and am now at the microwave 1. The microwave is closed. I need to heat the plate using the microwave. </think>

<action>heat plate 2 with microwave 1</action>

Step 5: Navigate to the target receptacle

Prompt.

```
You are an expert agent operating in the ALFRED Embodied Environment. Your task is to: put a hot plate in
      diningtable.
Prior to this step, you have already taken 4 step(s). Below are the most recent 2 observaitons and the
     corresponding actions you took: [Observation 3: 'You pick up the plate 2 from the countertop 1.',
     Action 3: 'go to microwave 1']
[Observation 4: 'You arrive at microwave 1. The microwave 1 is closed.', Action 4: 'heat plate 2 with
     microwave 1']
You are now at step 5 and your current observation is: You heat the plate 2 using the microwave 1.
Your admissible actions of the current situation are: ['examine microwave 1'
 'examine plate 2'
 'go to cabinet 1'
 'go to cabinet 10'
 'go to cabinet 11'
 'go to cabinet 2'
 'go to cabinet 3'
 'go to cabinet 4'
 'go to cabinet 5'
 'go to cabinet 6'
 'go to cabinet 7'
 'go to cabinet 8'
 'go to cabinet 9'
 'go to coffeemachine 1'
 'go to countertop 1'
 'go to countertop 2'
 'go to diningtable 1'
 'go to diningtable 2'
 'go to drawer 1'
 'go to drawer 2'
 'go to drawer 3'
 'go to fridge 1'
 'go to garbagecan 1'
 'go to sidetable 1'
 'go to sinkbasin 1'
 'go to stoveburner 1'
 'go to stoveburner 2'
 'go to stoveburner 3'
```

```
 'go to stoveburner 4'
 'go to toaster 1'
 'heat plate 2 with microwave 1'
 'inventory'
 'look'
 'open microwave 1'].

Now it's your turn to take an action.
You should first reason step-by-step about the current situation. This reasoning process MUST be enclosed
      within <think> </think> tags.
Once you've finished your reasoning, you should choose an admissible action for current step and present
     it within <action> </action> tags.
```

Response.

<think> The previous actions involved heating the plate 2 using the microwave 1. Now, we need to put the hot plate in the diningtable 1. </think>

<action>go to diningtable 1</action>

Step 6: Complete the placement

Prompt.

```
You are an expert agent operating in the ALFRED Embodied Environment. Your task is to: put a hot plate in
      diningtable.
Prior to this step, you have already taken 5 step(s). Below are the most recent 2 observaitons and the
     corresponding actions you took: [Observation 4: 'You arrive at microwave 1. The microwave 1 is
     closed.', Action 4: 'heat plate 2 with microwave 1']
[Observation 5: 'You heat the plate 2 using the microwave 1.', Action 5: 'go to diningtable 1']
You are now at step 6 and your current observation is: You arrive at diningtable 1. On the diningtable 1,
      you see a bowl 1, a cup 2, and a plate 1.
Your admissible actions of the current situation are: ['examine diningtable 1'
 'examine plate 2'
 'go to cabinet 1'
 'go to cabinet 10'
 'go to cabinet 11'
 'go to cabinet 2'
 'go to cabinet 3'
 'go to cabinet 4'
 'go to cabinet 5'
 'go to cabinet 6'
 'go to cabinet 7'
 'go to cabinet 8'
 'go to cabinet 9'
 'go to coffeemachine 1'
 'go to countertop 1'
 'go to countertop 2'
 'go to diningtable 2'
 'go to drawer 1'
 'go to drawer 2'
 'go to drawer 3'
 'go to fridge 1'
 'go to garbagecan 1'
 'go to microwave 1'
 'go to sidetable 1'
 'go to sinkbasin 1'
 'go to stoveburner 1'
 'go to stoveburner 2'
 'go to stoveburner 3'
 'go to stoveburner 4'
 'go to toaster 1'
 'inventory'
 'look'
 'move plate 2 to diningtable 1'].

Now it's your turn to take an action.
You should first reason step-by-step about the current situation. This reasoning process MUST be enclosed
      within <think> </think> tags.
Once you've finished your reasoning, you should choose an admissible action for current step and present
     it within <action> </action> tags.
```

Response.

<think> The task is to put a hot plate in the diningtable. I have already prepared a hot plate using the microwave and have it ready. The plate is located on the diningtable 1. I need to move the plate to the diningtable 1 to place it there. </think>

<action>move plate 2 to diningtable 1</action>

D.2 WEBSHOP

The user requests men's boots with lace closure and rubber soles, in the color "boulder," size 15, and priced below $140. The five-step trajectory searches with the principal constraints, selects a product within the price range, configures the requested attributes, and completes the purchase.

Step 1: Form a constraint-aware query

Prompt.

```
You are an expert autonomous agent operating in the WebShop e-commerce environment.
Your task is to: Find me men's boots with lace closure, rubber outsole, rubber sole with color: boulder,
     and size: 15, and price lower than 140.00 dollars.
Your current observation is: 'Search'.
Your admissible actions of the current situation are:
[
'search[<your query>]',
'click[search]',
].

Now it's your turn to take one action for the current step.
You should first reason step-by-step about the current situation, then think carefully which admissible
     action best advances the shopping goal. This reasoning process MUST be enclosed within <think> </
     think> tags.
Once you've finished your reasoning, you should choose an admissible action for current step and present
     it within <action> </action> tags.
```

Response.

<think> To find the men's boots with the specified criteria, I need to search for them in the webshop e-commerce environment. The search query should include the color ("boulder"), size (15), and price (lower than 140.00). This will help me filter the results and find the desired items. </think>

<action>search[boulder,men'sboots,15,140.00]</action>

Step 2: Select a product within budget

Prompt.

```
You are an expert autonomous agent operating in the WebShop e-commerce environment.
Your task is to: Find me men's boots with lace closure, rubber outsole, rubber sole with color: boulder,
     and size: 15, and price lower than 140.00 dollars.
Prior to this step, you have already taken 1 step(s). Below are the most recent 1 observations and the
     corresponding actions you took: [Observation 1: ''Search'', Action 1: 'search[boulder, men's boots,
     15, 140.00]']
You are now at step 2 and your current observation is: 'Back to Search' [SEP] 'Page 1 (Total results: 50)
     ' [SEP] 'Next >' [SEP] 'B07L6DV555' [SEP] 'Merrell Men's Primer Mid LTR Fashion Boot' [SEP] '$112.91
      to $134.33' [SEP] 'B01EX1BWE4' [SEP] 'Milwaukee Leather MBM9075 Men's Black 6-inch Plain Toe Dual
     Zipper Lock Leather Boots' [SEP] '$99.82 to $134.99' [SEP] 'B01JJR8GJG' [SEP] 'Cat Footwear Men's
     Carnaby Canvas Fashion Sneaker' [SEP] '$79.95' [SEP] 'B01LXTLLVG' [SEP] 'PUMA V1.08 Tricks Top
     Trainer Mens Soccer Sneakers/Boots-Grey-5.5' [SEP] '$55.5' [SEP] 'B00ITCBD3Y' [SEP] 'Laredo Womens
     Spellbound Studded Square Toe Dress Boots Mid Calf Low Heel 1-2" - Black' [SEP] '$132.0 to $150.26'
     [SEP] 'B09QXF3V3X' [SEP] 'DEUVOUM Summer Trend Mesh Shoes Men's Sports Shoes Solid Color Lace-Up
     Sneakers Fashion All-Match Walking Shoes Outdoor Hiking Shoes Non-Slip Shock-Absorbing Casual Sports
      Shoes' [SEP] '$100.0' [SEP] 'B01MSP2B0I' [SEP] 'Wrangler Men's Retro Slim Fit Straight Leg Jean,
     Black, 42W x 32L' [SEP] '$63.05' [SEP] 'B09Q37JQZ6' [SEP] 'Women's Sexy Swimsuit One Piece High Neck
      Halter Bikini Floral Stiching See Through Monokini Tummy Control Beachwear' [SEP] '$10.99 to $18
     .99' [SEP] 'B06Y52T6KC' [SEP] 'Ellie Shoes Women's 414-Wonder Boot' [SEP] '$59.99 to $107.99' [SEP]
     'B07JZ5HVN6' [SEP] 'Women's Ladies Sneakers,Limsea Fashion Classic Boots Ankle Short Leather Shoes'
     [SEP] '$6.63 to $8.37'.
Your admissible actions of the current situation are:
[
'click[back to search]',
'click[next >]',
'click[b07l6dv555]',
'click[b01ex1bwe4]',
'click[b01jjr8gjg]',
'click[b01lxtllvg]',
'click[b00itcbd3y]',
'click[b09qxf3v3x]',
'click[b01msp2b0i]',
'click[b09q37jqz6]',
'click[b06y52t6kc]',
'click[b07jz5hvn6]',
].

Now it's your turn to take one action for the current step.
You should first reason step-by-step about the current situation, then think carefully which admissible
     action best advances the shopping goal. This reasoning process MUST be enclosed within <think> </
     think> tags.
Once you've finished your reasoning, you should choose an admissible action for current step and present
     it within <action> </action> tags.
```

Response.

<think> To find the men's boots with the specified criteria, I need to filter the search results based on the given conditions. The conditions are: men's boots with lace closure, rubber outsole, rubber sole with color: boulder, size: 15, and price lower than 140.00 dollars. I will filter the search results accordingly. </think>

<action>click[b07l6dv555]</action>

Step 3: Select the requested color

Prompt.

```
You are an expert autonomous agent operating in the WebShop e-commerce environment.
Your task is to: Find me men's boots with lace closure, rubber outsole, rubber sole with color: boulder,
     and size: 15, and price lower than 140.00 dollars.
Prior to this step, you have already taken 2 step(s). Below are the most recent 2 observations and the
     corresponding actions you took: [Observation 1: ''Search'', Action 1: 'search[boulder, men's boots,
     15, 140.00]']
[Observation 2: ''Back to Search' [SEP] 'Page 1 (Total results: 50)' [SEP] 'Next >' [SEP] 'B07L6DV555' [
     SEP] 'Merrell Men's Primer Mid LTR Fashion Boot' [SEP] '$112.91 to $134.33' [SEP] 'B01EX1BWE4' [SEP]
      'Milwaukee Leather MBM9075 Men's Black 6-inch Plain Toe Dual Zipper Lock Leather Boots' [SEP] '$99
     .82 to $134.99' [SEP] 'B01JJR8GJG' [SEP] 'Cat Footwear Men's Carnaby Canvas Fashion Sneaker' [SEP] '
     $79.95' [SEP] 'B01LXTLLVG' [SEP] 'PUMA V1.08 Tricks Top Trainer Mens Soccer Sneakers/Boots-Grey-5.5'
      [SEP] '$55.5' [SEP] 'B00ITCBD3Y' [SEP] 'Laredo Womens Spellbound Studded Square Toe Dress Boots Mid
      Calf Low Heel 1-2" - Black' [SEP] '$132.0 to $150.26' [SEP] 'B09QXF3V3X' [SEP] 'DEUVOUM Summer
     Trend Mesh Shoes Men's Sports Shoes Solid Color Lace-Up Sneakers Fashion All-Match Walking Shoes
     Outdoor Hiking Shoes Non-Slip Shock-Absorbing Casual Sports Shoes' [SEP] '$100.0' [SEP] 'B01MSP2B0I'
      [SEP] 'Wrangler Men's Retro Slim Fit Straight Leg Jean, Black, 42W x 32L' [SEP] '$63.05' [SEP] '
     B09Q37JQZ6' [SEP] 'Women's Sexy Swimsuit One Piece High Neck Halter Bikini Floral Stiching See
     Through Monokini Tummy Control Beachwear' [SEP] '$10.99 to $18.99' [SEP] 'B06Y52T6KC' [SEP] 'Ellie
     Shoes Women's 414-Wonder Boot' [SEP] '$59.99 to $107.99' [SEP] 'B07JZ5HVN6' [SEP] 'Women's Ladies
     Sneakers,Limsea Fashion Classic Boots Ankle Short Leather Shoes' [SEP] '$6.63 to $8.37'', Action 2:
     'click[b07l6dv555]']
You are now at step 3 and your current observation is: 'Back to Search' [SEP] '< Prev' [SEP] 'size' [SEP]
      '7' [SEP] '7.5' [SEP] '9' [SEP] '9.5' [SEP] '10' [SEP] '10.5' [SEP] '11' [SEP] '11.5' [SEP] '12' [
     SEP] '14' [SEP] '15' [SEP] 'color' [SEP] 'boulder' [SEP] 'black' [SEP] 'Merrell Men's Primer Mid LTR
      Fashion Boot' [SEP] 'Price: $112.91 to $134.33' [SEP] 'Rating: N.A.' [SEP] 'Description' [SEP] '
     Features' [SEP] 'Reviews' [SEP] 'Buy Now'.
Your admissible actions of the current situation are:
[
'click[back to search]',
'click[< prev]',
'click[description]',
'click[features]',
'click[reviews]',
'click[buy now]',
'click[7]',
'click[7.5]',
'click[9]',
'click[9.5]',
'click[10]',
'click[10.5]',
'click[11]',
'click[11.5]',
'click[12]',
'click[14]',
'click[15]',
'click[boulder]',
'click[black]',
].

Now it's your turn to take one action for the current step.
You should first reason step-by-step about the current situation, then think carefully which admissible
     action best advances the shopping goal. This reasoning process MUST be enclosed within <think> </
     think> tags.
Once you've finished your reasoning, you should choose an admissible action for current step and present
     it within <action> </action> tags.
```

Response.

<think> To find the desired men's boots with specific criteria, I need to filter the search results based on the given conditions. The current observation indicates that I am on the product page for Merrell Men's Primer Mid LTR Fashion Boot with size 15 and color boulder. However, I need to find other products that match these criteria. </think>

<action>click[boulder]</action>

Step 4: Select the requested size

Prompt.

```
You are an expert autonomous agent operating in the WebShop e-commerce environment.
Your task is to: Find me men's boots with lace closure, rubber outsole, rubber sole with color: boulder,
     and size: 15, and price lower than 140.00 dollars.
Prior to this step, you have already taken 3 step(s). Below are the most recent 2 observations and the
     corresponding actions you took: [Observation 2: ''Back to Search' [SEP] 'Page 1 (Total results: 50)'
      [SEP] 'Next >' [SEP] 'B07L6DV555' [SEP] 'Merrell Men's Primer Mid LTR Fashion Boot' [SEP] '$112.91
     to $134.33' [SEP] 'B01EX1BWE4' [SEP] 'Milwaukee Leather MBM9075 Men's Black 6-inch Plain Toe Dual
     Zipper Lock Leather Boots' [SEP] '$99.82 to $134.99' [SEP] 'B01JJR8GJG' [SEP] 'Cat Footwear Men's
     Carnaby Canvas Fashion Sneaker' [SEP] '$79.95' [SEP] 'B01LXTLLVG' [SEP] 'PUMA V1.08 Tricks Top
     Trainer Mens Soccer Sneakers/Boots-Grey-5.5' [SEP] '$55.5' [SEP] 'B00ITCBD3Y' [SEP] 'Laredo Womens
     Spellbound Studded Square Toe Dress Boots Mid Calf Low Heel 1-2" - Black' [SEP] '$132.0 to $150.26'
     [SEP] 'B09QXF3V3X' [SEP] 'DEUVOUM Summer Trend Mesh Shoes Men's Sports Shoes Solid Color Lace-Up
     Sneakers Fashion All-Match Walking Shoes Outdoor Hiking Shoes Non-Slip Shock-Absorbing Casual Sports
      Shoes' [SEP] '$100.0' [SEP] 'B01MSP2B0I' [SEP] 'Wrangler Men's Retro Slim Fit Straight Leg Jean,
     Black, 42W x 32L' [SEP] '$63.05' [SEP] 'B09Q37JQZ6' [SEP] 'Women's Sexy Swimsuit One Piece High Neck
```

```
      Halter Bikini Floral Stiching See Through Monokini Tummy Control Beachwear' [SEP] '$10.99 to $18
     .99' [SEP] 'B06Y52T6KC' [SEP] 'Ellie Shoes Women's 414-Wonder Boot' [SEP] '$59.99 to $107.99' [SEP]
     'B07JZ5HVN6' [SEP] 'Women's Ladies Sneakers,Limsea Fashion Classic Boots Ankle Short Leather Shoes'
     [SEP] '$6.63 to $8.37'', Action 2: 'click[b0716dv555]']
[Observation 3: ''Back to Search' [SEP] '< Prev' [SEP] 'size' [SEP] '7' [SEP] '7.5' [SEP] '9' [SEP] '9.5'
      [SEP] '10' [SEP] '10.5' [SEP] '11' [SEP] '11.5' [SEP] '12' [SEP] '14' [SEP] '15' [SEP] 'color' [SEP
     ] 'boulder' [SEP] 'black' [SEP] 'Merrell Men's Primer Mid LTR Fashion Boot' [SEP] 'Price: $112.91 to
      $134.33' [SEP] 'Rating: N.A.' [SEP] 'Description' [SEP] 'Features' [SEP] 'Reviews' [SEP] 'Buy Now
     '', Action 3: 'click[boulder]']
You are now at step 4 and your current observation is: 'Back to Search' [SEP] '< Prev' [SEP] 'size' [SEP]
      '7' [SEP] '7.5' [SEP] '9' [SEP] '9.5' [SEP] '10' [SEP] '10.5' [SEP] '11' [SEP] '11.5' [SEP] '12' [
     SEP] '14' [SEP] '15' [SEP] 'color' [SEP] 'boulder' [SEP] 'black' [SEP] 'Merrell Men's Primer Mid LTR
      Fashion Boot' [SEP] 'Price: $112.91 to $134.33' [SEP] 'Rating: N.A.' [SEP] 'Description' [SEP] '
     Features' [SEP] 'Reviews' [SEP] 'Buy Now'.
Your admissible actions of the current situation are:
[
'click[back to search]',
'click[< prev]',
'click[description]',
'click[features]',
'click[reviews]',
'click[buy now]',
'click[7]',
'click[7.5]',
'click[9]',
'click[9.5]',
'click[10]',
'click[10.5]',
'click[11]',
'click[11.5]',
'click[12]',
'click[14]',
'click[15]',
'click[boulder]',
'click[black]',
].

Now it's your turn to take one action for the current step.
You should first reason step-by-step about the current situation, then think carefully which admissible
     action best advances the shopping goal. This reasoning process MUST be enclosed within <think> </
     think> tags.
Once you've finished your reasoning, you should choose an admissible action for current step and present
     it within <action> </action> tags.
```

Response.

<think> To find the men's boots with the specified criteria, I need to filter the results based on the given conditions. Starting with the current observation, I should check if there are any results matching all the criteria. If not, I should move to the next step to refine the search. </think>

<action>click[15]</action>

Step 5: Complete the purchase

Prompt.

```
You are an expert autonomous agent operating in the WebShop e-commerce environment.
Your task is to: Find me men's boots with lace closure, rubber outsole, rubber sole with color: boulder,
     and size: 15, and price lower than 140.00 dollars.
Prior to this step, you have already taken 4 step(s). Below are the most recent 2 observations and the
     corresponding actions you took: [Observation 3: ''Back to Search' [SEP] '< Prev' [SEP] 'size' [SEP]
     '7' [SEP] '7.5' [SEP] '9' [SEP] '9.5' [SEP] '10' [SEP] '10.5' [SEP] '11' [SEP] '11.5' [SEP] '12' [
     SEP] '14' [SEP] '15' [SEP] 'color' [SEP] 'boulder' [SEP] 'black' [SEP] 'Merrell Men's Primer Mid LTR
      Fashion Boot' [SEP] 'Price: $112.91 to $134.33' [SEP] 'Rating: N.A.' [SEP] 'Description' [SEP] '
     Features' [SEP] 'Reviews' [SEP] 'Buy Now'', Action 3: 'click[boulder]']
[Observation 4: ''Back to Search' [SEP] '< Prev' [SEP] 'size' [SEP] '7' [SEP] '7.5' [SEP] '9' [SEP] '9.5'
      [SEP] '10' [SEP] '10.5' [SEP] '11' [SEP] '11.5' [SEP] '12' [SEP] '14' [SEP] '15' [SEP] 'color' [SEP
     ] 'boulder' [SEP] 'black' [SEP] 'Merrell Men's Primer Mid LTR Fashion Boot' [SEP] 'Price: $112.91 to
      $134.33' [SEP] 'Rating: N.A.' [SEP] 'Description' [SEP] 'Features' [SEP] 'Reviews' [SEP] 'Buy Now
     '', Action 4: 'click[15]']
You are now at step 5 and your current observation is: 'Back to Search' [SEP] '< Prev' [SEP] 'size' [SEP]
      '7' [SEP] '7.5' [SEP] '9' [SEP] '9.5' [SEP] '10' [SEP] '10.5' [SEP] '11' [SEP] '11.5' [SEP] '12' [
     SEP] '14' [SEP] '15' [SEP] 'color' [SEP] 'boulder' [SEP] 'black' [SEP] 'Merrell Men's Primer Mid LTR
      Fashion Boot' [SEP] 'Price: $112.91 to $134.33' [SEP] 'Rating: N.A.' [SEP] 'Description' [SEP] '
     Features' [SEP] 'Reviews' [SEP] 'Buy Now'.
Your admissible actions of the current situation are:
[
'click[back to search]',
'click[< prev]',
'click[description]',
'click[features]',
'click[reviews]',
'click[buy now]',
'click[7]',
'click[7.5]',
'click[9]',
'click[9.5]',
'click[10]',
```

```
'click[10.5]',
'click[11]',
'click[11.5]',
'click[12]',
'click[14]',
'click[15]',
'click[boulder]',
'click[black]',
].

Now it's your turn to take one action for the current step.
You should first reason step-by-step about the current situation, then think carefully which admissible
     action best advances the shopping goal. This reasoning process MUST be enclosed within <think> </
     think> tags.
Once you've finished your reasoning, you should choose an admissible action for current step and present
     it within <action> </action> tags.
```

Response.

<think> To find the desired men's boots with specific criteria, I need to filter the available options based on the given conditions. The current observation shows that I am at the "size" step with the size "15" and the color "boulder". Now, I need to find a product that matches all the given criteria. </think>

<action>click[buy now]</action>