RACE: RECOVERY-AWARE CREDIT ESTIMATION FOR LONG-HORIZON LLM AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Group-based reinforcement learning (RL) enables critic-free post-training of large language model (LLM) agents for long-horizon tasks. However, uniform trajectory-level credit conflates useful steps with unproductive actions, while GiGPO-style exact-state grouping misses outcome evidence from semantically comparable contexts with different observations. In this work, we introduce Recovery-Aware Credit Estimation (RACE), a plug-in method that assigns action-level credit through changes in contextual recoverability. Using a frozen text encoder, RACE nonparametrically estimates contextual recoverability from outcome-labeled neighboring contexts retrieved from previously collected rollouts, without training a critic or process reward model. A leave-one-trajectory-out retrieval scheme excludes the query trajectory from its retrieved neighborhood, and the clipped change in estimated recoverability across each transition serves as an action-level credit signal added to the base group-relative advantage. The method requires no additional environment interactions and supports both binary and graded outcomes. Experiments with Qwen3-4B and Qwen3-8B on ALFWorld, WebShop, and ScienceWorld show improvements over matched GiGPO runs across all three benchmarks. With Qwen3-8B, RACE achieves absolute success-rate gains of 7.85% and 7.59% on ALFWorld's in-distribution and out-of-distribution splits, respectively, and improves the ScienceWorld graded score by 6.04 points and exact success rate by an absolute 14.75%. Further experiments show higher validation-curve AUC across all six benchmark–model settings and improved zero-shot transfer, measured by macro-averaged success and progress across BabyAI, PDDL, and TextCraft.

1 INTRODUCTION

Large language model (LLM) agents solve interactive tasks by observing an environment, taking actions, and adapting to feedback (Yao et al., 2023; Shinn et al., 2023). Reinforcement learning (RL) improves these policies using task outcomes. Group-based algorithms estimate relative advantages from multiple rollouts of the same task without training a separate critic, providing an effective approach to LLM post-training (Shao et al., 2024; Guo et al., 2025; Yu et al., 2025). In long-horizon interactions, however, an outcome may arrive after dozens of decisions (Wang et al., 2025; Feng et al., 2025). This outcome evaluates the trajectory as a whole and provides limited information about the contribution of each action. Assigning the same outcome-derived advantage throughout a trajectory can therefore reinforce detours in successful episodes and penalize useful steps in failed ones. For example, heating an object can satisfy an intermediate task requirement even when a later placement error causes the episode to fail.

Step-level grouping provides finer credit by comparing actions taken at the same state across rollouts (Feng et al., 2025; He et al., 2026). These comparisons depend on exact state recurrence. In language-agent environments, contexts with different textual observations can nevertheless reflect comparable task progress. Outcomes from trajectories containing these contexts can provide evidence about task-completion prospects beyond what is available within an exact-state group. Exact-state grouping keeps these contexts separate, leaving their outcome evidence unavailable for a shared local comparison. The resulting challenge is to obtain action-level credit from comparable rollout experience even when exact states do not recur. Figure 1 illustrates both the ambiguity of trajectory-level credit and the evidence that exact-state grouping misses.

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.

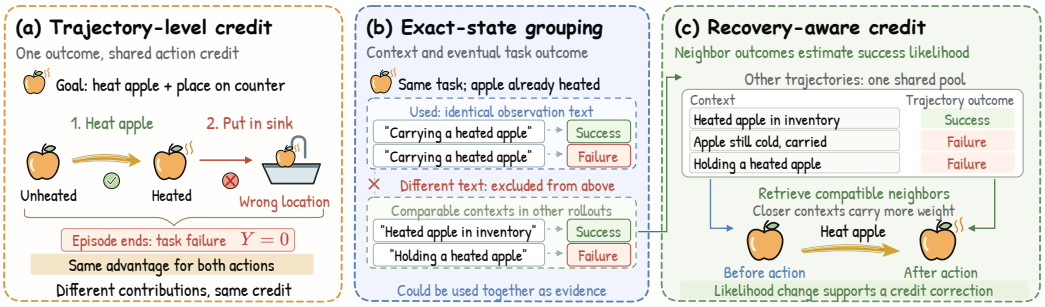


Figure 1: **Motivation for recovery-aware credit.** (a) Trajectory-level credit conflates useful and harmful actions within the same failed trajectory. (b) Exact-state grouping excludes comparable contexts with different observations from a shared local comparison. (c) RACE uses outcomes from compatible contexts in other trajectories to estimate changes in task-success likelihood across an action and correct the base advantage. Examples are schematic.

To address this problem, we propose **Recovery-Aware Credit Estimation (RACE)**, a method that evaluates actions through changes in contextual *recoverability*. Recoverability measures how likely an agent is to complete the task from its current context. RACE estimates this quantity using the terminal outcomes of other trajectories with comparable contexts. The difference in estimated recoverability across an action refines the action's credit. This connects local progress to terminal outcomes even when immediate rewards are absent, without requiring annotated process rewards.

RACE constructs context representations from the task, goal, observable state, and remaining-step bucket using a frozen text encoder. It retrieves semantically similar contexts from collected rollouts under task-compatibility rules, excluding the query trajectory. An inverse-distance weighted estimator combines the neighbors' terminal outcomes, with minimum-support requirements and shrinkage toward the batch outcome prior to regularize local estimates. The change in estimated recoverability is clipped and added to the base group-relative advantage. Weak-signal gating directs this correction to actions for which the base credit provides little differentiation. RACE integrates into the existing clipped policy objective and preserves the policy architecture and rollout budget. Its added computation uses a frozen-encoder pass and batched neighbor search, requiring neither a learned critic nor additional environment interactions.

We evaluate RACE on ALFWorld, ScienceWorld, and WebShop using Qwen3-4B and Qwen3-8B. RACE improves every reported aggregate metric over matched GiGPO training across all six benchmark–model settings. With Qwen3-8B, RACE achieves absolute success-rate gains of 7.85% and 7.59% on ALFWorld's ID and OOD splits, respectively. On ScienceWorld, the graded score improves by 6.04 points and the exact success rate by an absolute 14.75%. On WebShop, the task score improves by 2.17 points and the exact success rate by an absolute 5.50%. Further experiments show higher validation-curve AUC and improved zero-shot transfer to BabyAI, PDDL, and TextCraft. Component ablations support task-compatible cross-instance retrieval, shrinkage, and weak-signal gating.

Our contributions are as follows:

- We propose RACE for action-level credit assignment under delayed outcome supervision. By estimating changes in contextual recoverability, RACE uses compatible rollout experience to refine local credit beyond exact-state comparisons.
- We develop a nonparametric estimator based on frozen context embeddings, cross-trajectory retrieval, inverse-distance weighting, and shrinkage. Clipping and weak-signal gating integrate the resulting recovery increments into group-relative policy optimization without additional rollouts or a learned critic.
- Experiments with Qwen3-4B and Qwen3-8B on ALFWorld, ScienceWorld, and WebShop demonstrate gains over matched GiGPO training, including an absolute gain of 14.75% in ScienceWorld exact success at 8B. Training-curve analysis, zero-shot transfer, and ablations provide evidence for learning efficiency, generalization, and the contribution of the estimator's components.

2 Related Work

Reinforcement learning for language agents. Language agents alternate model-generated decisions with environment feedback, as exemplified by ReAct and Reflexion (Yao et al., 2023; Shinn et al., 2023). Recent work increasingly trains this interaction loop rather than relying only on prompting. RAGEN studies the optimization dynamics of multi-turn agent RL (Wang et al., 2025), while Agent Lightning separates agent execution from RL training and decomposes agent traces into transitions (Luo et al., 2025). Controlled studies further show that horizon length itself can destabilize policy learning even when the underlying task logic is held fixed (Kim et al., 2026). RACE targets the corresponding learning problem: assigning action-level credit from completed on-policy interactions, without altering the agent's inference-time workflow.

Group-relative optimization and fine-grained credit. GRPO introduced critic-free, within-group advantage estimation for verifiable reasoning (Shao et al., 2024), and subsequent systems have scaled this paradigm through improved sampling and optimization (Guo et al., 2025; Yu et al., 2025). For interactive agents, GiGPO supplements episode groups with local groups at exactly repeated anchor states (Feng et al., 2025), while turn-level GRPO uses explicit intermediate rewards to construct per-turn advantages (Zeng et al., 2025). More recent methods derive finer supervision from trajectory structure: SALT builds graphs over same-prompt trajectories (Li et al., 2026), STAPO uses normalized entropy to identify trajectory-neglect steps (Qi et al., 2026), and ELPO searches for the first irrecoverable tool-use error (Liang et al., 2026b). RACE estimates action-level changes in recoverability using outcome evidence from semantically comparable contexts in other trajectories, without annotated intermediate rewards or additional diagnostic rollouts.

Temporal credit and non-parametric estimation. Temporal credit assignment broadly asks how an action influences a delayed outcome (Pignatelli et al., 2024); learned counterfactual estimators such as COCOA address this question with an explicit outcome model (Meulemans et al., 2023). RACE instead estimates a local recovery potential directly from rollout data. This design is related to nearest-neighbor value estimation in Neural Episodic Control (Pritzel et al., 2017) and retrieval-based probability estimation in kNN-LM (Khandelwal et al., 2020), but changes both the target and its role: neighbors are structured interaction contexts, their labels are trajectory outcomes, and their weighted estimate is used only to correct the training advantage. Leave-one-trajectory-out filtering excludes the query trajectory, minimum-support constraints determine where corrections apply, and shrinkage regularizes local estimates. Retrieval is used during training.

3 Method

RACE estimates contextual *recoverability*—the expected terminal outcome of comparable continuations—from collected rollouts. As shown in Figure 2, it embeds rollout contexts, retrieves compatible contexts from other trajectories, and estimates recoverability from their outcome labels. The change across an action supplies a bounded correction to the base group-relative advantage, broadcast to its action tokens. This requires neither a learned critic nor additional environment interactions.

3.1 Setup and Credit-Assignment Gap

For each task in a batch, the behavior policy $\pi_{\theta_{\text{old}}}$ samples a group of trajectories, yielding N trajectories in total. Let $[N] = \{1, \ldots, N\}$ and x_i denote the task instance for trajectory i:

$$\tau_i = \left\{(h_{i,t}, a_{i,t})\right\}_{t=0}^{T_i-1}, \qquad i \in [N], Y_i \in [0,1], \tag{1}$$

Here, t indexes environment decisions, $h_{i,t}$ is the observable context before action $a_{i,t}$, and T_i is the episode length. The terminal outcome Y_i is mapped to $[0,1]$ and may represent binary success or graded completion.

Let $A_{i,t}^{\text{base}}$ denote the scalar action-level advantage produced by a critic-free group estimator. GiGPO, for example, combines an episode-level relative advantage with a local comparison among actions taken at repeated anchor states (Feng et al., 2025). Such comparisons are informative when

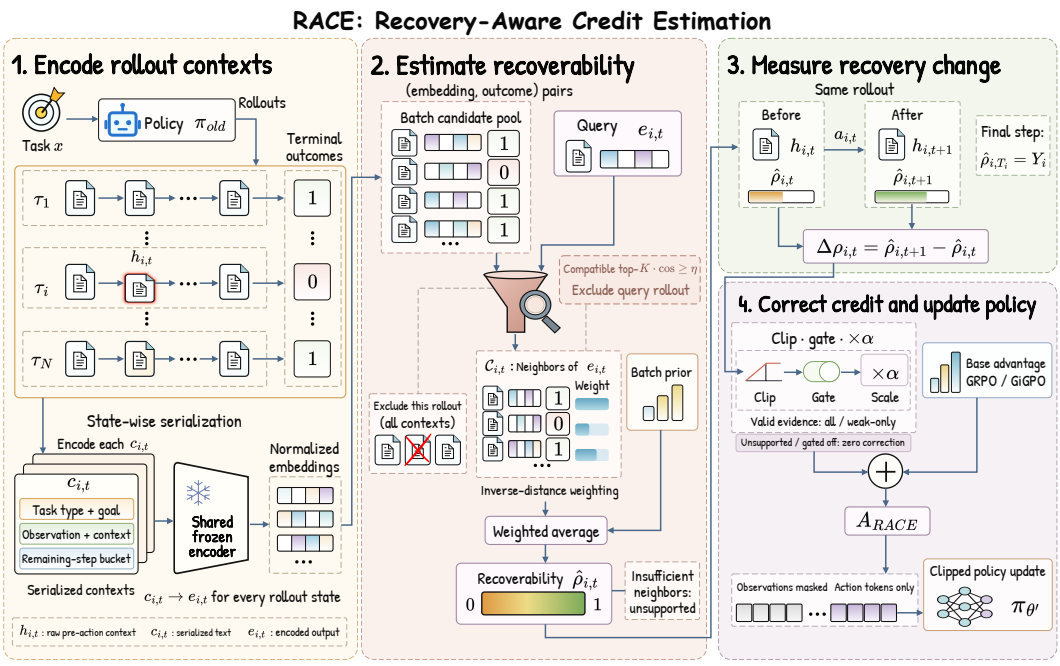


Figure 2: Overview of RACE. Grouped policy rollouts provide both the base advantage and the experience used for recovery estimation. Each interaction context is serialized with its task, goal, current observable state, and remaining-step bucket, then embedded by a frozen text encoder. Compatible contexts from other trajectories supply outcome-labeled neighbors for a shrinkage-weighted recoverability estimate. RACE adds the clipped change in recoverability across an action to the base advantage; the existing actor-loss mask then restricts policy optimization to generated action tokens.

exact states recur, whereas semantically related contexts need not be identical. Moreover, assigning the same delayed outcome throughout an episode cannot distinguish an action that improves the remaining path to success from one whose effect is later repaired. RACE supplies this missing transition-level distinction while leaving $A^{\text{base}}_{i,t}$ and its rollout construction unchanged.

3.2 Recoverability from Rollout Experience

Context representation. Language-agent environments are partially observed. RACE therefore represents each decision using rollout observations and benchmark metadata, without assuming access to a fully observed Markov state:

$$c_{i,t} = \text{Serialize}\big(q_{i,t}, g_i, b_{i,t}, \widetilde{o}_{i,t}\big), \tag{2}$$

where $q_{i,t}$ denotes the task type or interaction stage, g_i the goal, $b_{i,t}$ the remaining-step bucket, and $\widetilde{o}_{i,t}$ the state summary. The explicit goal field may be empty when goal-relative features encode it. RACE uses a benchmark-provided summary when available, including goal-relative features where applicable; otherwise it uses the recorded anchor observation. A frozen text encoder f_ϕ with fixed parameters ϕ mean-pools non-padding token representations and produces the unit-norm embedding

$$e_{i,t} = \frac{f_\phi(c_{i,t})}{\|f_\phi(c_{i,t})\|_2}. \tag{3}$$

The embedding map remains fixed during policy optimization, without an additional value-learning objective.

Eligible episodic neighbors. RACE retrieves evidence only from contexts that satisfy an explicit compatibility rule. Let $\mathcal{P}$ be the candidate context pool, where (j,u) indexes decision u of trajectory j. The predicate $\Gamma((i,t),(j,u)) \in \{0,1\}$ requires the same environment and benchmark-specific compatibility. With cosine similarity threshold η, the eligible set for query (i,t) is

$$\mathcal{C}_{i,t} = \Big\{(j,u) \in \mathcal{P} : \Gamma((i,t),(j,u)) = 1,\ \tau_j \neq \tau_i,\ e_{i,t}^\top e_{j,u} \geq \eta\Big\}. \tag{4}$$

The condition $\tau_j \neq \tau_i$ excludes the query trajectory from the retrieved neighborhood. The compatibility rule may also exclude the same task identifier to avoid neighborhoods dominated by replicated rollouts. We retain up to K eligible contexts with the highest cosine similarity as $\mathcal{N}_{i,t}$. The pool contains current-batch records and may include a bounded memory of recent rollout records, subject to the same eligibility rules.

Shrinkage-weighted recoverability. Neighbor (j, u) contributes its trajectory outcome Y_j with the inverse-distance weight

$$w_{i,t}^{j,u} = \frac{1}{\|e_{i,t} - e_{j,u}\|_2^2 + \epsilon_{\text{ker}}}, \tag{5}$$

where the subscript identifies the query, the superscript identifies the neighbor, and $\epsilon_{\text{ker}} > 0$ stabilizes the kernel. Following episodic nearest-neighbor estimation (Pritzel et al., 2017), closer contexts receive greater weight without fitting a parametric value function. For a neighborhood containing at least M records, with $1 \leq M \leq K$, we estimate

$$\widehat{\rho}_{i,t} = \frac{\sum_{(j,u)\in\mathcal{N}_{i,t}} w_{i,t}^{j,u} Y_j + \lambda\rho_0}{\sum_{(j,u)\in\mathcal{N}_{i,t}} w_{i,t}^{j,u} + \lambda}, \tag{6}$$

where ρ_0 is the mean outcome over distinct trajectories in the current batch and $\lambda \geq 0$ is the prior strength. The estimate shrinks local outcome evidence toward ρ_0, with stronger shrinkage when the total neighbor weight is small relative to λ.

3.3 Recovery Increment and Policy Update

Action-level recovery increment. RACE evaluates action $a_{i,t}$ by comparing recoverability before and after the transition. For non-terminal actions, the correction is

$$\Delta\rho_{i,t} = \text{clip}(\widehat{\rho}_{i,t+1} - \widehat{\rho}_{i,t}, -c_\rho, c_\rho)\,. \tag{7}$$

For the final action, we anchor the post-action potential to the observed outcome by setting $\widehat{\rho}_{i,T_i} = Y_i$. Thus, a positive increment identifies an action after which comparable continuations terminate with higher outcomes, while a negative increment identifies the converse. Clipping by $c_\rho \geq 0$ bounds the influence of individual retrieval estimates.

Let $\bar{\chi}_{i,t}$ indicate that both required potentials are supported. RACE forms

$$A_{i,t}^{\text{RACE}} = A_{i,t}^{\text{base}} + \alpha\,\bar{\chi}_{i,t} g_{i,t} \Delta\rho_{i,t}, \tag{8}$$

where $\alpha \geq 0$ controls the correction scale. The gate $g_{i,t}$ specifies where recovery evidence complements the base estimator. Applying RACE to every supported transition sets $g_{i,t} = 1$. The weak-signal variant instead uses

$$g_{i,t} = \mathbb{I}\big[\text{Std}_{j\in\mathcal{G}(x_i)}(Y_j) \leq \epsilon_g \;\vee\; |A_{i,t}^{\text{base}}| \leq \epsilon_A\big]\,, \tag{9}$$

where $\mathcal{G}(x_i)$ contains the indices of distinct sampled trajectories for task x_i, and $\mathbb{I}[\cdot]$ is the indicator function. The nonnegative thresholds ϵ_g and ϵ_A specify negligible outcome variation and base-advantage magnitude. The gate targets outcome-homogeneous groups or actions with negligible base credit and may additionally exclude all-success groups. It does not alter the recoverability estimator.

Policy optimization. RACE broadcasts the scalar correction over each action's response span. The base actor-loss mask selects generated action tokens, excluding observations from the policy loss. For a PPO-style objective, we use the same clipped surrogate:

$$\mathcal{J}_{\text{RACE}}(\theta) = \mathbb{E}_{i,t,k}\Big[\min\big(r_{i,t,k}(\theta)A_{i,t}^{\text{RACE}}, \text{clip}(r_{i,t,k}(\theta), 1-\epsilon, 1+\epsilon)A_{i,t}^{\text{RACE}}\big)\Big], \tag{10}$$

where k indexes action tokens and ϵ is the PPO clipping parameter. For token $a_{i,t,k}$ and its prefix $a_{i,t,<k}$, the importance ratio is $r_{i,t,k}(\theta) = \pi_\theta(a_{i,t,k} \mid h_{i,t}, a_{i,t,<k})/\pi_{\theta_{\text{old}}}(a_{i,t,k} \mid h_{i,t}, a_{i,t,<k})$. The behavior parameters θ_{old} remain fixed while θ is optimized. The expectation uses rollout action tokens with the base loss aggregation. Retrieval is computed outside the policy graph, and all KL or entropy terms of the base algorithm are retained. RACE changes the credit supplied to the optimizer without changing the policy architecture, rollout budget, or clipped policy objective.

Interpretation. For a trajectory with supported potentials, the raw recovery differences telescope before clipping and gating:

$$\sum_{t=0}^{T_i-1} \left(\widehat{\rho}_{i,t+1} - \widehat{\rho}_{i,t}\right) = Y_i - \widehat{\rho}_{i,0}. \tag{11}$$

The identity expresses the outcome-relative correction as a sum of transition-level recoverability increments. Gating selects the transitions to correct, and clipping bounds each increment.

The complete estimator and update algorithms appear in Appendix A.

4 Experiments

4.1 Experimental Setup

Benchmarks and evaluation. We evaluate on ALFWorld (Shridhar et al., 2021), ScienceWorld (Wang et al., 2022), and WebShop (Yao et al., 2022), covering household instruction following, scientific procedures, and goal-conditioned shopping. Evaluation uses ALFWorld's 140 ID and 134 OOD tasks, 1,024 held-out ScienceWorld episodes, and WebShop's 500 test goals. We report task-completion success rates on all three benchmarks, alongside ScienceWorld's graded score and WebShop's task score for partial progress. We evaluate fixed final checkpoints and report means and sample standard deviations across independently trained checkpoints, after averaging evaluation seeds within each checkpoint.

Baselines and comparison protocol. We compare RACE with GiGPO (Feng et al., 2025) under matched initialization, training instances, rollout budgets, policy objectives, and decoding settings. Table 1 also reports published results for Agent-R1 with GRPO, PPO, REINFORCE, and RLOO (Cheng et al., 2025), R^3L (Shi et al., 2026), Constant-context (Xie et al., 2026), BEACON (Wang et al., 2026), PlanPO (Liang et al., 2026a), and EnvHarness+GRPO (Huang et al., 2026), as well as ReAct (Yao et al., 2023; Xiong et al., 2025) and zero-shot teacher results (Chen et al., 2026).

Implementation. For our experiments, we implement RACE within the GiGPO training framework using Qwen3-4B and Qwen3-8B (Yang et al., 2025). For retrieval, serialized rollout contexts are embedded by the frozen `all-MiniLM-L6-v2` encoder[1] (Wang et al., 2020). Changes in estimated recoverability provide the credit correction used in policy optimization. Training and retrieval configurations for the main experiments are summarized in Appendix B.

4.2 Main Results

We evaluate RACE against matched GiGPO runs on ALFWorld, ScienceWorld, and WebShop (Table 1, Parts I–III). The table also includes published results from other systems.

Gains across benchmarks and model scales. RACE improves all aggregate metrics over matched GiGPO runs at both model scales. On ALFWorld (Table 1, Part I), Qwen3-8B achieves absolute ID and OOD success-rate gains of 7.85% and 7.59%, reaching 90.95% and 88.93%, respectively. Qwen3-4B reaches 86.31% ID and 81.97% OOD success. On ScienceWorld (Part II), graded score and exact success increase by 6.04 points and an absolute 14.75% at 8B, and by 2.36 points and 5.60% at 4B, respectively. On WebShop (Part III), task score and exact success increase by 2.17 points and an absolute 5.50% at 8B, and by 4.85 points and 8.47% at 4B. On both ScienceWorld and WebShop, the larger gains in exact success indicate improved task completion alongside partial progress.

Training efficiency and cross-environment transfer. RACE achieves higher horizon-normalized validation AUC and final validation performance in all six benchmark–model settings (Figure 3(a)). Appendix E provides the full validation and on-policy learning curves. We also evaluate the 8B policies trained on each environment on BabyAI, PDDL, and TextCraft without additional training

[1] Checkpoint: `https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2`.

Table 1: Main results on a 0–100 scale (higher is better). Shading denotes RACE; bold marks the higher mean within each matched comparison. Values are means ± sample standard deviations across trained checkpoints. Published baselines retain their reported evaluation protocols. †: prompting (teachers: zero-shot); other methods are environment-trained. Constant-context uses SFT+RL. "/": published seen/unseen results; "–": unreported. R^3L reports task score rather than exact success.

Part I: ALFWorld

Backbone	Method	ID success ↑	OOD success ↑
Qwen3-32B	Teacher† (Chen et al., 2026)	–	50.00
Llama-3.1-70B	ReAct† (Xiong et al., 2025)	78.60	73.90
GPT-4o	ReAct† (Xiong et al., 2025)	78.60	83.60
Qwen3-4B	Agent-R1 (GRPO) (Cheng et al., 2025)	81.29	74.58
	Agent-R1 (PPO) (Cheng et al., 2025)	76.42	72.38
	Agent-R1 (REINFORCE) (Cheng et al., 2025)	73.84	69.57
	Agent-R1 (RLOO) (Cheng et al., 2025)	79.08	73.46
	Constant-context (Xie et al., 2026)	76.40 ±1.50	81.30 ±1.90
	GiGPO (reproduced)	81.07 ±2.86	81.47 ±2.99
	RACE	**86.31** ±2.19	**81.97** ±0.18
Qwen3-8B	Constant-context (Xie et al., 2026)	83.60 ±1.80	89.60 ±0.40
	EnvHarness+GRPO (Huang et al., 2026)	87.90	88.80
	GiGPO (reproduced)	83.10 ±6.06	81.34 ±1.76
	RACE	**90.95** ±2.02	**88.93** ±9.67

Part II: ScienceWorld

Backbone	Method	Graded score ↑	Success ↑
Qwen3-32B	Teacher† (Chen et al., 2026)	28.80	12.00
GPT-4o	ReAct† (Xiong et al., 2025)	60.00 /56.00	59.80 /57.80
Qwen2.5-1.5B	BEACON (Wang et al., 2026)	58.90	45.30
Qwen3-4B	R^3L (Shi et al., 2026)	39.80	–
	Constant-context (Xie et al., 2026)	72.40 /75.10	63.10 /65.40
	GiGPO (reproduced)	71.61 ±0.49	42.77 ±4.84
	RACE	**73.97** ±3.30	**48.37** ±7.01
Qwen3-8B	GiGPO (reproduced)	74.96 ±4.31	49.64 ±0.85
	RACE	**81.00** ±4.10	**64.39** ±4.60

Part III: WebShop

Backbone	Method	Task score ↑	Success ↑
Qwen3-32B	Teacher† (Chen et al., 2026)	52.80	51.00
Llama-3.1-70B	ReAct† (Xiong et al., 2025)	59.40	–
GPT-4o	ReAct† (Xiong et al., 2025)	63.50	–
PaLM-540B	ReAct† (Yao et al., 2023)	66.60	40.00
Qwen2.5-1.5B	BEACON (Wang et al., 2026)	86.10	75.60
	PlanPO (Liang et al., 2026a)	86.80 ±1.50	77.20 ±5.60
Qwen3-4B	Agent-R1 (GRPO) (Cheng et al., 2025)	65.83	44.20
	Agent-R1 (PPO) (Cheng et al., 2025)	70.18	46.00
	Agent-R1 (REINFORCE) (Cheng et al., 2025)	63.41	41.80
	Agent-R1 (RLOO) (Cheng et al., 2025)	68.02	45.10
	R^3L (Shi et al., 2026)	74.60	–
	GiGPO (reproduced)	83.47 ±4.65	68.18 ±7.74
	RACE	**88.32** ±1.10	**76.64** ±0.60
Qwen3-8B	EnvHarness+GRPO (Huang et al., 2026)	79.20	67.40
	GiGPO (reproduced)	85.60 ±1.24	73.07 ±3.77
	RACE	**87.77** ±1.89	**78.57** ±4.10

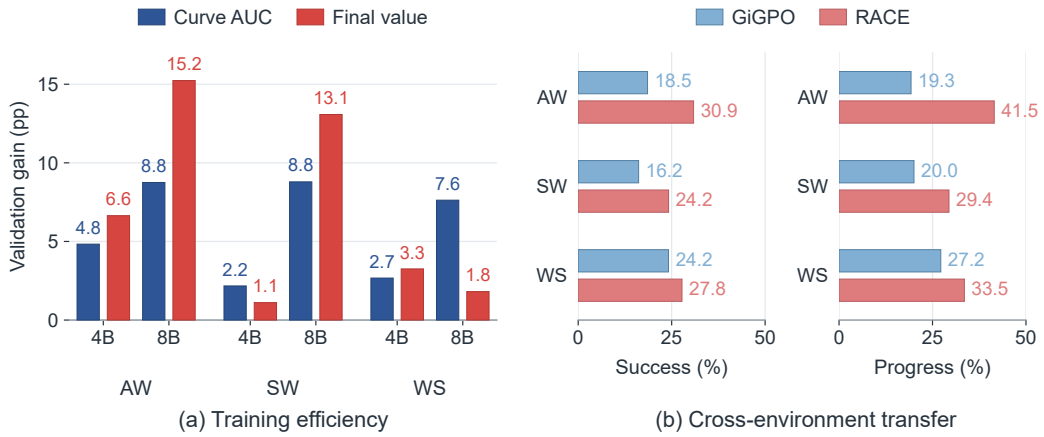


Figure 3: Training and cross-environment transfer with RACE. (a) Absolute gains (%) over matched GiGPO runs in horizon-normalized validation AUC and final validation performance. (b) Zero-shot transfer of 8B policies to BabyAI, PDDL, and TextCraft. Success and progress are averaged equally over the three target environments and then across checkpoints. Blue and red horizontal bars denote GiGPO and RACE, respectively. AW = ALFWorld, SW = ScienceWorld, WS = WebShop; pp denotes absolute differences between percentage values.

Table 2: ALFWorld ablations with Qwen3-4B after 100 training updates. Success rates are percentages; scores use the original evaluation scale. Higher is better for all metrics. Shading denotes RACE; bold marks the highest value in each column, including ties.

Configuration	ID success	ID score	OOD success	OOD score
GiGPO	63.57	3.655	58.96	2.505
RACE	**82.14**	**5.350**	**81.34**	**5.436**
w/o task-type grouping	**82.14**	5.332	78.36	4.794
within-task retrieval only	77.14	4.298	72.39	3.437
w/o shrinkage	75.00	4.346	70.15	3.483
w/o weak-signal gate	76.43	4.443	59.70	2.775

(Figure 3(b)). Averaged equally over the three targets, absolute success gains are 12.31%, 8.02%, and 3.61% for ALFWorld-, ScienceWorld-, and WebShop-trained models, respectively; absolute progress gains are 22.25%, 9.36%, and 6.31%. Appendix D gives the evaluation protocol and task-level results.

4.3 Ablation Study

We ablate four RACE components with Qwen3-4B on ALFWorld, using the same 100-update training budget for all configurations (Table 2). Each variant removes task-type grouping, restricts retrieval to the current task instance, removes empirical-Bayes shrinkage, or applies the correction without the weak-signal gate.

Effect of each component. Full RACE achieves the best or tied-best result on all four aggregate metrics. Its absolute OOD success-rate gains over the variants without task-type grouping and restricted to within-task retrieval are 2.98% and 8.95%, respectively. These comparisons support task-compatible cross-instance retrieval. Full RACE also achieves absolute OOD success-rate gains of 11.19% and 21.64% over the variants without shrinkage and without weak-signal gating, respectively. Figure 4(a) shows OOD success across six task families; panel (b) summarizes the ID and OOD success differences between full RACE and each ablation. Appendix F.2 gives the task-level values.

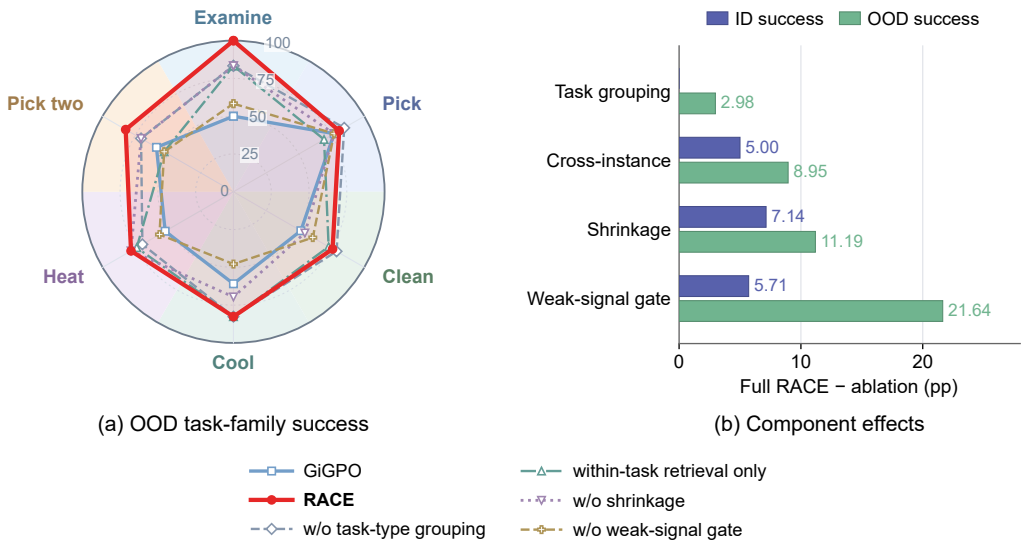


Figure 4: ALFWorld ablations with Qwen3-4B after 100 training updates. (a) OOD success across six task families on a full 0–100% radial scale. (b) Absolute ID and OOD success-rate gains (%) of full RACE over each ablation, computed as full RACE minus the ablation using Table 2. "Cross-instance" denotes the comparison with within-task retrieval only. Task-level values appear in Appendix F.2.

Table 3: Time-averaged RACE diagnostics on ScienceWorld. Correlation is measured at the trajectory level; relative correction is mean absolute correction divided by mean absolute base advantage.

Diagnostic	Qwen3-4B	Qwen3-8B
Recoverability–outcome correlation	0.60	0.53
Supported steps (%)	98.9	98.6
Relative correction (%)	5.7	5.5

4.4 Recoverability Diagnostics

On ScienceWorld, recoverability estimates correlate positively with trajectory outcomes at both model scales (Table 3). Eligible neighbors support over 98% of action steps. The mean absolute correction is 5.7% of the mean absolute base advantage at 4B and 5.5% at 8B. Appendix F.1 provides the full diagnostic curves.

5 Conclusion

We introduced RACE, a plug-in credit estimator for long-horizon language-agent reinforcement learning. RACE estimates contextual recoverability from compatible, outcome-labeled contexts in collected on-policy rollouts and adds the clipped change across each transition to the base group-relative advantage. This provides action-level credit from delayed outcomes without training a critic or collecting additional environment interactions. With Qwen3-4B and Qwen3-8B, RACE improves all matched aggregate metrics over GiGPO on ALFWorld, ScienceWorld, and WebShop, covering exact completion and graded progress. RACE also achieves higher validation AUCs and improves zero-shot transfer to BabyAI, PDDL, and TextCraft. Component ablations support task-compatible cross-instance retrieval, shrinkage, and weak-signal gating. ScienceWorld diagnostics show positive recoverability–outcome correlations and neighbor support for over 98% of action steps. These findings support contextual recoverability as a source of fine-grained credit for critic-free group-based reinforcement learning.

AI USE STATEMENT

In this work, we used generative AI tools for language editing and polishing, including grammar checking and improving the readability of the manuscript, and for generating auxiliary code related to environment setup, dependency installation, and configuration. We did not use generative AI tools for research ideation, method implementation, experimental design, result interpretation, mathematical claims or proofs, or other core research tasks.

We have reviewed all AI-assisted work. Language edits were manually verified to preserve the intended technical meaning, and AI-generated auxiliary code was reviewed and tested before use. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies credit assignment for LLM agents using established benchmark environments: ALFWorld, WebShop, and ScienceWorld, with transfer evaluation on BabyAI, PDDL, and TextCraft. All agent interactions take place within these benchmark environments. The study does not involve recruiting human participants or executing real-world purchases or physical actions.

REPRODUCIBILITY STATEMENT

Section 3 specifies RACE, with estimator and policy-update pseudocode in Appendix A. Appendix B documents the models, retrieval encoder, training configurations, benchmark splits, and evaluation procedures, including the benchmark-specific settings in Table 4. The zero-shot transfer protocol for BabyAI, PDDL, and TextCraft is detailed in Appendix D.

REFERENCES

Chishui Chen, Yaoyou Fan, Te Sun, Yi Yang, Chenghao Sun, Delin Mao, Hongbo Qiao, Zuowei Zhang, Junxi Wang, Chenxing Sun, Yangen Hu, Lu Pan, Xuyang Liu, and Linfeng Zhang. Look ahead before you distill: Future trajectory validation of teacher guidance for agentic on-policy distillation. *arXiv preprint arXiv:2608.01953*, 2026. URL https://arxiv.org/abs/2608.01953.

Mingyue Cheng, Shuo Yu, Daoyu Wang, Qingchuan Li, Xiaoyu Tao, Jie Ouyang, Yucong Luo, Yitong Zhou, Qi Liu, and Enhong Chen. Agent-R1: A unified and modular framework for agentic reinforcement learning. *arXiv preprint arXiv:2511.14460*, 2025. doi: 10.48550/arXiv.2511.14460. URL https://arxiv.org/abs/2511.14460.

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, Main Conference, pp. 46375–46408, 2025. doi: 10.52202/085713-1544. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/420c9f777c0b4f78d515e53cf74d58b2-Abstract-Conference.html.

Daya Guo, Dejian Yang, Haowei Zhang, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645:633–638, 2025. doi: 10.1038/s41586-025-09422-z. URL https://www.nature.com/articles/s41586-025-09422-z.

Shuo He, Lang Feng, Qi Wei, Xin Cheng, Lei Feng, and Bo An. Hierarchy-of-groups policy optimization for long-horizon agentic tasks, 2026. URL https://arxiv.org/abs/2602.22817.

Chengsong Huang, Zifeng Wang, Rujun Han, Jun Yan, Yanfei Chen, Zoey CuiZhu, Ke Jiang, Peng Xia, Han Yu, Yufan Zhuang, Yifei Ming, Jiaqi Pan, Bhavana Dalvi Mishra, Jiaxin Huang, Burak Gokturk, Tomas Pfister, and Chen-Yu Lee. EnvHarness: Awakening static worlds for agent learning. *arXiv preprint arXiv:2608.19880*, 2026. URL https://arxiv.org/abs/2608.19880.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Generalization through memorization: Nearest neighbor language models. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HklBjCEKvH.

Sunghwan Kim, Junhee Cho, Beong-woo Kwak, Taeyoon Kwon, Liang Wang, Nan Yang, Xingxing Zhang, Furu Wei, and Jinyoung Yeo. On training large language models for long-horizon tasks: An empirical study of horizon length. In *International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=PnHfrCMKtp.

Jiazheng Li, Yawei Wang, Qiaojing Yan, Yijun Tian, Zhichao Xu, Huan Song, Panpan Xu, and Lin Lee Cheong. SALT: Step-level advantage assignment for long-horizon agents via trajectory graph. In *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 4709–4725. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-eacl.247. URL https://aclanthology.org/2026.findings-eacl.247/.

Dayang Liang, Liyuan He, Xuan Feng, Shuxin Li, Bo An, and Yunlong Liu. PlanPO: Group planning-aware policy optimization for multi-turn agentic LLMs. *arXiv preprint arXiv:2608.17289*, 2026a. doi: 10.48550/arXiv.2608.17289. URL https://arxiv.org/abs/2608.17289.

Qiao Liang, Yuke Zhu, Chao Ge, Lei Yang, Ying Shen, Bo Zheng, and Sheng Guo. Learning from the irrecoverable: Error-localized policy optimization for tool-integrated LLM reasoning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, pp. 11008–11028. Association for Computational Linguistics, 2026b. doi: 10.18653/v1/2026.acl-long.504. URL https://aclanthology.org/2026.acl-long.504/.

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. doi: 10.48550/arXiv.2508.03680. URL https://arxiv.org/abs/2508.03680.

Chang Ma, Junlei Zhang, Zhihao Zhu, Cheng Yang, Yujiu Yang, Yaohui Jin, Zhenzhong Lan, Lingpeng Kong, and Junxian He. AgentBoard: An analytical evaluation board of multi-turn LLM agents. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://arxiv.org/abs/2401.13178.

Alexander Meulemans, Simon Schug, Seijin Kobayashi, Nathaniel Daw, and Gregory Wayne. Would i have gotten that reward? long-term credit assignment by counterfactual contribution analysis. In *Advances in Neural Information Processing Systems*, volume 36, pp. 68685–68735, 2023. doi: 10.52202/075280-3006. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/d8bd445c2abe1343cce0e14b361b2fb3-Abstract-Conference.html.

Eduardo Pignatelli, Johan Ferret, Hado van Hasselt, Matthieu Geist, Thomas Mesnard, Olivier Pietquin, and Laura Toni. A survey of temporal credit assignment in deep reinforcement learning. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=bNtr6SLgZf.

Alexander Pritzel, Benigno Uria, Sriram Srinivasan, Adrià Puigdomènech Badia, Oriol Vinyals, Demis Hassabis, Daan Wierstra, and Charles Blundell. Neural episodic control. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2827–2836, 2017. URL https://proceedings.mlr.press/v70/pritzel17a.html.

Qiuyi Qi, Tian Liang, Mutian Bao, Jinjian Zhang, Dongnan Liu, Wei Zhou, Linjian Mo, Ming Kong, Jie Liu, Feng Zhang, and Qiang Zhu. STAPO: Selective trajectory-aware policy optimization for LLM agent training. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, pp. 28371–28392. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1308. URL https://aclanthology.org/2026.acl-long.1308/.

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.

Weijie Shi, Yanxi Chen, Zexi Li, Xuchen Pan, Yuchang Sun, Jiajie Xu, Xiaofang Zhou, and Yaliang Li. R^3L: Reflect-then-retry reinforcement learning with language-guided exploration, pivotal credit, and positive amplification. *arXiv preprint arXiv:2601.03715*, 2026. doi: 10.48550/arXiv.2601.03715. URL https://arxiv.org/abs/2601.03715.

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 Cote, 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.

Ruoyao Wang, Peter Jansen, Marc-Alexandre Cote, 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. doi: 10.18653/v1/2022.emnlp-main.775. URL https://aclanthology.org/2022.emnlp-main.775/.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems*, volume 33, pp. 5776–5788, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Zihan Wang, Kangrui Wang, Qineng Wang, Pingyue Zhang, Linjie Li, Zhengyuan Yang, Kefan Yu, Minh Nhat Nguyen, et al. RAGEN: Understanding self-evolution in LLM agents via multi-turn reinforcement learning. *arXiv preprint arXiv:2504.20073*, 2025. doi: 10.48550/arXiv.2504.20073. URL https://arxiv.org/abs/2504.20073.

Zixuan Wang, Yuchen Yan, Hongxing Li, Teng Pan, Dingming Li, Ruiqing Zhang, Weiming Lu, Jun Xiao, Yueting Zhuang, and Yongliang Shen. Milestone-guided policy learning for long-horizon language agents. *arXiv preprint arXiv:2605.06078*, 2026. doi: 10.48550/arXiv.2605.06078. URL https://arxiv.org/abs/2605.06078.

Zhiheng Xi, Yiwen Ding, Wenxiang Chen, Boyang Hong, Honglin Guo, Junzhe Wang, Xin Guo, Dingwen Yang, Chenyang Liao, Wei He, Songyang Gao, Lu Chen, Rui Zheng, Yicheng Zou, Tao Gui, Qi Zhang, Xipeng Qiu, Xuanjing Huang, Zuxuan Wu, and Yu-Gang Jiang. AgentGym: Evaluating and training large language model-based agents across diverse environments. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pp. 27914–27961. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.1355. URL https://aclanthology.org/2025.acl-long.1355/.

Haoyang Xie, Xinyuan Wang, Yancheng Wang, Puda Zhao, and Feng Ju. From history to state: Constant-context skill learning for LLM agents. *arXiv preprint arXiv:2605.05413*, 2026. URL https://arxiv.org/abs/2605.05413.

Weimin Xiong, Yifan Song, Qingxiu Dong, Bingchan Zhao, Feifan Song, Xun Wang, and Sujian Li. MPO: Boosting LLM agents with meta plan optimization. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 3914–3935. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.210. URL https://aclanthology.org/2025.findings-emnlp.210/.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

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. URL https://arxiv.org/abs/2207.01206.

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://arxiv.org/abs/2210.03629.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, et al. DAPO: An open-source LLM reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025. doi: 10.48550/arXiv.2503.14476. URL https://arxiv.org/abs/2503.14476.

Siliang Zeng, Quan Wei, William Brown, Oana Frunza, Yuriy Nevmyvaka, and Mingyi Hong. Reinforcing multi-turn reasoning in LLM agents via turn-level credit assignment. *arXiv preprint arXiv:2505.11821v1*, 2025. doi: 10.48550/arXiv.2505.11821. URL https://arxiv.org/abs/2505.11821v1.

A RACE PSEUDOCODE AND IMPLEMENTATION DETAILS

Algorithms 1 and 2 describe recoverability estimation and its integration into a policy update. A record $q = (i, t)$ identifies context t in trajectory i, whose length is T_i and normalized outcome is $Y_i \in [0, 1]$. For a batch $\mathcal{B}$ of N trajectories, the context records are $\mathcal{Q}(\mathcal{B}) = \{(i, t) : i \in [N], 0 \leq t < T_i\}$, where $[N] = \{1, \ldots, N\}$. The optional memory $\mathcal{D}$ contains historical context records. The training protocol specifies the outcome map Ω, compatibility predicate Γ, nonnegative similarity kernel κ, prior rule $\mathcal{R}$, and application gate Gate. For unit-norm embeddings, the inverse-distance kernel in the main text is $\kappa(s) = 1/(2 - 2s + \varepsilon_{\mathrm{ker}})$, where s is cosine similarity.

A.1 RECOVERABILITY ESTIMATOR

Algorithm 1 Outcome-weighted recoverability estimation

Require: Query records $\mathcal{Q}$; candidate records $\mathcal{P}$; normalized embeddings $\{e_q\}$; trajectory outcomes $\{Y_i\}$; compatibility predicate Γ; neighbor cap K; minimum support M; similarity floor η; nonnegative kernel κ; shrinkage strength λ; prior $\rho_0(q)$

Ensure: Recoverability $\widehat{\rho}_q$ and support indicator χ_q for every $q \in \mathcal{Q}$

for all $q = (i, t) \in \mathcal{Q}$ **do**
  $\mathcal{C}_q \leftarrow \{r = (j, u) \in \mathcal{P} : \Gamma(q, r) = 1,\ j \neq i,\ e_q^\top e_r \geq \eta\}$ ▷ leave one trajectory out
  $\mathcal{N}_q \leftarrow \textsc{TopK}(\mathcal{C}_q, r \mapsto e_q^\top e_r, K)$
  if $|\mathcal{N}_q| < M$ **then**
    $\widehat{\rho}_q \leftarrow \perp$; $\chi_q \leftarrow 0$ ▷ do not extrapolate without support
  else
    for all $r = (j, u) \in \mathcal{N}_q$ **do**
      $w_{qr} \leftarrow \kappa(e_q^\top e_r)$
    end for
    $$\widehat{\rho}_q \leftarrow \frac{\sum_{r=(j,u)\in\mathcal{N}_q} w_{qr} Y_j + \lambda \rho_0(q)}{\sum_{r\in\mathcal{N}_q} w_{qr} + \lambda}$$
    $\chi_q \leftarrow 1$
  end if
end for
return $\{(\widehat{\rho}_q, \chi_q) : q \in \mathcal{Q}\}$

Estimator invariants. The query trajectory is excluded from its retrieved neighborhood, and an unsupported pre- or post-action estimate produces zero correction. The terminal action uses the observed outcome rather than a retrieved successor. Outcome priors and gates are computed over distinct trajectories rather than step records.

Computational complexity. Let B be the number of context records in an update and d the embedding dimension. Within compatibility bucket b, let $\mathcal{Q}_b$ and $\mathcal{P}_b$ denote the query and candidate records. Excluding encoder inference, batched similarity computation costs $O(\sum_b |\mathcal{Q}_b||\mathcal{P}_b|d)$, with at most $O(BK)$ storage for selected neighbors. Context strings are deduplicated before encoding. The estimator reuses the base estimator's rollouts without additional policy or environment calls.

A.2 POLICY-UPDATE INTEGRATION

Algorithm 2 RACE credit estimation and policy update

Require: Old policy $\pi_{\theta_{\text{old}}}$; grouped task batch $\mathcal{X}$; base group estimator; frozen encoder f_ϕ; optional bounded memory $\mathcal{D}$; outcome map Ω; compatibility predicate Γ; kernel κ; prior rule $\mathcal{R}$; gate rule Gate; RACE parameters $K, M, \eta, \lambda, \alpha, c_\rho$

Ensure: Updated policy parameters θ

Sample grouped trajectories $\mathcal{B} = \{\tau_i, Z_i\}_{i=1}^N$ from $\pi_{\theta_{\text{old}}}$ on $\mathcal{X}$
Normalize trajectory outcomes: $Y_i \leftarrow \Omega(Z_i) \in [0, 1]$
Compute base scalar advantages $\{A_{i,t}^{\text{base}}\}$ with the chosen group estimator
for all $q = (i, t) \in \mathcal{Q}(\mathcal{B})$ **do**
    $c_q \leftarrow$ SERIALIZE(task type or stage, goal, remaining-step bucket, state summary)
end for
Encode each unique context once and normalize: $e_q \leftarrow f_\phi(c_q)/\|f_\phi(c_q)\|_2$
$\mathcal{P} \leftarrow \mathcal{Q}(\mathcal{B}) \cup \mathcal{D}$ ▷ $\mathcal{D} = \varnothing$ for batch-only retrieval
Compute $\rho_0(q) \leftarrow \mathcal{R}(q, \{Y_i\}_{\tau_i \in \mathcal{B}})$ over distinct trajectories
$(\widehat{\rho}, \chi) \leftarrow$ ESTIMATERECOVERABILITY$(\mathcal{Q}(\mathcal{B}), \mathcal{P}, e, Y, \Gamma, K, M, \eta, \kappa, \lambda, \rho_0)$
for all $q = (i, t) \in \mathcal{Q}(\mathcal{B})$ **do**
    if $t = T_i - 1$ **then**
        $\rho_{i,t}^+ \leftarrow Y_i$; $\chi_{i,t}^+ \leftarrow 1$ ▷ anchor the terminal post-action context
    else
        $\rho_{i,t}^+ \leftarrow \widehat{\rho}_{i,t+1}$; $\chi_{i,t}^+ \leftarrow \chi_{i,t+1}$
    end if
    $\bar{\chi}_{i,t} \leftarrow \chi_{i,t}\chi_{i,t}^+$
    if $\bar{\chi}_{i,t} = 1$ **then**
        $\Delta\rho_{i,t} \leftarrow$ CLIP$(\rho_{i,t}^+ - \widehat{\rho}_{i,t}, -c_\rho, c_\rho)$
    else
        $\Delta\rho_{i,t} \leftarrow 0$
    end if
    $g_{i,t} \leftarrow \text{Gate}\big(A_{i,t}^{\text{base}}, \{Y_j : j \in \mathcal{G}(x_i)\}\big) \in \{0, 1\}$
    $\delta A_{i,t} \leftarrow \alpha\, \bar{\chi}_{i,t} g_{i,t} \Delta\rho_{i,t}$
    Broadcast $\delta A_{i,t}$ only to generated action tokens using the existing response mask
    $A_{i,t}^{\text{RACE}} \leftarrow A_{i,t}^{\text{base}} + \delta A_{i,t}$
end for
If buffer retrieval is enabled, append $\mathcal{Q}(\mathcal{B})$ to $\mathcal{D}$ only after all estimates are computed
Update θ with the unchanged clipped policy objective using A^{RACE}
return θ

Policy-update integration. RACE writes the scalar correction to the stored response span, and the actor loss mask excludes observation tokens. With bounded memory, the current batch is appended after all estimates have been computed, keeping its records separate from historical candidates during the update.

B EXPERIMENTAL PROTOCOLS

B.1 SHARED TRAINING AND EVALUATION SETTINGS

Controlled comparisons. Matched GiGPO and RACE runs share the pretrained checkpoint, training instances, rollout groups, policy objective, decoding configuration, and interaction budget. Each update contains 16 task instances with eight trajectories per instance. We evaluate the predefined final checkpoint of each run. Evaluation results are averaged over seeds within each checkpoint; we then report the mean and sample standard deviation across independently trained checkpoints.

Retrieval encoder. All RACE configurations use the frozen `sentence-transformers/all-MiniLM-L6-v2` encoder (Wang et al., 2020). Serialized contexts are truncated to 256 tokens on ALFWorld and WebShop and 512 tokens on

ScienceWorld. We mean-pool the final hidden states over non-padding positions and ℓ_2-normalize the resulting 384-dimensional vectors for cosine retrieval. Encoding runs on CPU.

Training and retrieval configurations. The policy is optimized with AdamW at a learning rate of 10^{-6} and a PPO clipping range of 0.2. RACE uses minimum support $M = 3$, increment clip $c_\rho = 0.5$, and inverse-distance kernel stabilizer $\varepsilon_{\text{ker}} = 0.001$. Table 4 summarizes benchmark-specific settings for the main experiments.

Table 4: Training and RACE configurations for the main experiments with Qwen3-4B and Qwen3-8B.

Setting	ALFWorld	ScienceWorld	WebShop
Evaluated checkpoint (updates)	150	300	150
Interaction horizon	50	30	15
Prompt / response token limits	2048 / 512	2048 / 256	4096 / 512
Actor KL-loss coefficient	Off	0.001	Off
Entropy coefficient	0.001	0	0.001
Neighbor cap K	16	32	4B: 16; 8B: 64
Similarity threshold η	0.65	0.75	0.65
Shrinkage strength λ	2.0	0.5	0.5
Correction scale α	0.5	1.25	0.5
Retrieval pool	Current batch	Batch + memory (4,096)	Current batch
Task-type grouping	Task family	Disabled	Page stage
Leave-one-task-group-out	No	Yes	Yes
Correction gate	Weak-signal steps	All supported steps	Weak-signal steps

Reward-side KL is disabled for all three benchmarks. ScienceWorld additionally retrieves from a bounded memory of 4,096 prior-update context records. WebShop excludes all-success groups from the correction gate.

B.2 Benchmark-Specific Protocols

ALFWorld. Policies are trained for 150 updates with a maximum interaction horizon of 50 steps. Final checkpoints are evaluated on the complete standard ID and OOD splits, which contain 140 and 134 tasks, respectively. The primary metric is episode success. We additionally report success for six task families: examining an object under a light source (Examine), pick-and-place (Pick), cleaning and placing (Clean), cooling and placing (Cool), heating and placing (Heat), and placing two objects (Pick2).

ScienceWorld. Policies are trained for 300 updates with a 30-step interaction horizon. Evaluation covers 23 task types; seven types whose simulator-clock requirements exceed this horizon are excluded. Each final checkpoint is evaluated on 1,024 held-out episodes from two deterministic evaluation seeds with 512 episodes each. We report the official graded score for partial progress and exact task-completion success, both on a 0–100 scale.

WebShop. Policies are trained for 150 updates with a 15-step interaction horizon. Final evaluation uses the standard 500-goal held-out set with identical goals and decoding settings for the matched methods. RACE retrieves up to $K = 16$ neighbors for Qwen3-4B and $K = 64$ for Qwen3-8B using a goal-relative state representation while leaving the policy input unchanged. We report the official normalized task score and exact episode success. Each checkpoint is evaluated under seeds 0, 1, and 2 using sampling with temperature 0.4 and a maximum of 512 generated tokens per action. At 4B, we evaluate three independently trained checkpoints per method at update 150. We average the three evaluation seeds within each checkpoint, then report the mean and sample standard deviation across checkpoints.

C Fine-Grained Benchmark Results

C.1 ALFWorld

Table 5 reports ALFWorld success by task family. At 8B, RACE achieves absolute success-rate gains of 26.84% on ID Cool and 19.70% on OOD Clean.

Table 5: ALFWorld success rates (%) by task family on the ID and OOD splits. Overall reports success over the complete split. Values are means ± sample standard deviations across independently trained checkpoints. Bold marks the higher mean within each matched backbone and split; shaded columns denote RACE.

(a) In-distribution

Task family	Qwen3-4B		Qwen3-8B	
	GiGPO	RACE	GiGPO	RACE
Examine	65.07 ±9.13	**82.64** ±2.95	77.15 ±23.08	**86.94** ±0.00
Pick	93.01 ±5.32	**95.70** ±6.08	94.09 ±6.84	**94.62** ±4.56
Clean	95.15 ±5.18	**98.19** ±2.56	95.75 ±6.02	**100.00** ±0.00
Cool	**69.40** ±5.26	64.20 ±6.29	60.51 ±0.98	**87.35** ±5.31
Heat	89.58 ±8.84	**90.62** ±4.42	**81.25** ±8.84	77.08 ±14.73
Pick2	67.26 ±17.68	**81.55** ±5.89	80.95 ±0.00	**91.07** ±2.53
Overall	81.07 ±2.86	**86.31** ±2.19	83.10 ±6.06	**90.95** ±2.02

(b) Out-of-distribution

Task family	Qwen3-4B		Qwen3-8B	
	GiGPO	RACE	GiGPO	RACE
Examine	84.72 ±5.89	**91.67** ±11.79	**95.83** ±1.96	88.89 ±15.71
Pick	**83.33** ±0.00	65.38 ±7.25	77.56 ±4.53	**79.49** ±18.13
Clean	80.30 ±6.43	**83.33** ±7.86	74.24 ±0.71	**93.94** ±7.14
Cool	84.00 ±3.12	**88.58** ±3.27	81.04 ±1.07	**90.90** ±0.00
Heat	73.61 ±13.72	**79.89** ±2.95	84.07 ±0.90	**88.85** ±7.82
Pick2	86.27 ±8.32	**92.16** ±5.55	87.25 ±6.93	**91.18** ±12.48
Overall	81.47 ±2.99	**81.97** ±0.18	81.34 ±1.76	**88.93** ±9.67

C.2 ScienceWorld

We examine category-level performance on ScienceWorld by grouping its 23 task types into seven semantic categories (Table 6). For Qwen3-8B, RACE increases the overall graded score from 74.96 to 81.00 and exact task-completion success from 49.64% to 64.39%. The largest category-level gains are in Matter and Measurement, at 15.93 and 12.07 points, respectively.

Table 6: ScienceWorld results on a 0–100 scale. Task-family rows report macro-averaged graded scores over task types; Overall reports the episode-level graded score, and Exact success reports task-completion success (%). Each checkpoint is evaluated on 1,024 episodes. Values are means ± sample standard deviations across independently trained checkpoints. Bold marks the higher mean within each matched backbone; shaded columns denote RACE.

Task family	Qwen3-4B		Qwen3-8B	
	GiGPO	RACE	GiGPO	RACE
Matter	44.24 ±18.55	**50.11** ±14.46	51.40 ±22.98	**67.33** ±3.49
Measurement	60.42 ±3.64	**66.94** ±6.28	63.52 ±7.61	**75.59** ±11.02
Electricity	**72.01** ±6.76	70.72 ±10.83	**67.33** ±9.16	64.61 ±0.88
Classification	98.39 ±1.00	**99.06** ±0.52	99.64 ±0.41	**99.73** ±0.35
Chemistry	32.25 ±11.58	**37.29** ±16.75	**48.85** ±5.33	42.50 ±15.17
Lifespan	77.23 ±3.51	**84.53** ±10.10	86.71 ±7.02	**94.42** ±5.53
Life stages	**30.05** ±16.86	24.92 ±15.92	20.55 ±12.00	**22.67** ±20.60
Overall	71.61 ±0.49	**73.97** ±3.30	74.96 ±4.31	**81.00** ±4.10
Exact success	42.77 ±4.84	**48.37** ±7.01	49.64 ±0.85	**64.39** ±4.60

D Cross-Environment Transfer

D.1 Evaluation Protocol and Aggregate Results

We evaluate matched Qwen3-8B models in zero-shot transfer to BabyAI and PDDL from AgentBoard (Ma et al., 2024) and TextCraft from AgentGym (Xi et al., 2025). The test sets contain 112, 60, and 100 episodes, respectively. We use the official task assets and environment wrappers, disable Qwen3's thinking mode, and decode greedily with at most 100 new tokens per interaction. Episode horizons are 30 steps for BabyAI and PDDL and 20 steps for TextCraft. We report success and the official subgoal-based progress measure on a 0–100 scale. Task-macro results weight the three target environments equally.

Across all three training sources, RACE improves both task-macro success and task-macro progress over matched GiGPO models. Table 7 gives the underlying means for BabyAI, PDDL, and TextCraft and the aggregate values used in Figure 3(b).

D.2 Target-Task Performance and Interaction Behavior

Target-task decomposition. Averaging equally across the three training sources, RACE improves success and progress on all three target environments (Figure 5). Absolute gains in success/progress are $+11.61\%/+17.41\%$ on BabyAI and $+6.67\%/+14.84\%$ on PDDL. TextCraft gains 5.67% on both metrics.

Interaction quality and efficiency. To characterize transferred interaction behavior, we examine action acceptance and episode length, averaged equally across the three training sources (Figure 6). RACE increases the mean accepted-action rate on all three target environments, with an absolute gain of 16.22% on PDDL. Mean episode length decreases by 2.81, 1.14, and 0.67 steps on BabyAI, PDDL, and TextCraft, respectively. These changes accompany the higher success and progress rates.

Table 7: Zero-shot transfer performance of Qwen3-8B models (%; higher is better). Panels (a) and (b) report mean success and progress, respectively. Macro is the equal-weight average over BabyAI, PDDL, and TextCraft. Shading denotes RACE; bold marks the higher mean within each matched pair, including ties.

(a) Success

		Target environments			
Training source	Method	BabyAI	PDDL	TextCraft	**Macro**
ALFWorld	GiGPO	7.14	0.00	48.50	18.55
	RACE	**29.91**	**11.67**	**51.00**	**30.86**
ScienceWorld	GiGPO	**8.48**	2.50	37.50	16.16
	RACE	6.70	**10.83**	**55.00**	**24.18**
WebShop	GiGPO	14.29	**1.67**	**56.50**	24.15
	RACE	**28.12**	**1.67**	53.50	**27.76**

(b) Progress

		Target environments			
Training source	Method	BabyAI	PDDL	TextCraft	**Macro**
ALFWorld	GiGPO	9.08	0.21	48.50	19.26
	RACE	**41.53**	**32.01**	**51.00**	**41.51**
ScienceWorld	GiGPO	**12.98**	9.65	37.50	20.04
	RACE	9.98	**23.24**	**55.00**	**29.40**
WebShop	GiGPO	19.21	**5.94**	**56.50**	27.22
	RACE	**42.00**	5.08	53.50	**33.53**

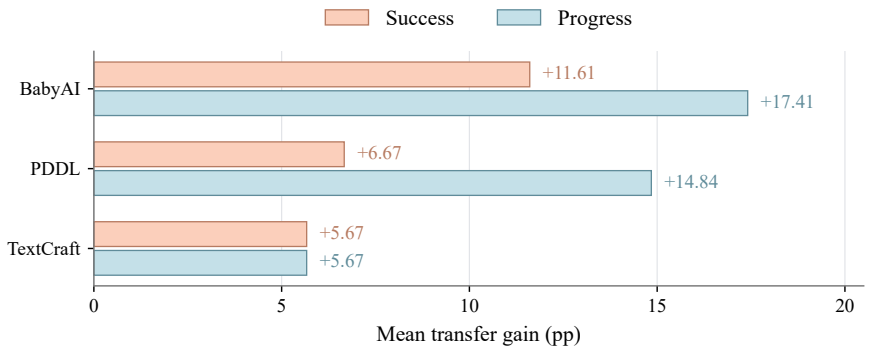


Figure 5: Qwen3-8B absolute zero-shot transfer gains (%) by target environment. Bars show RACE minus GiGPO for success and progress, averaged equally across ALFWorld-, ScienceWorld-, and WebShop-trained models. Both metrics improve on every target environment after this averaging.

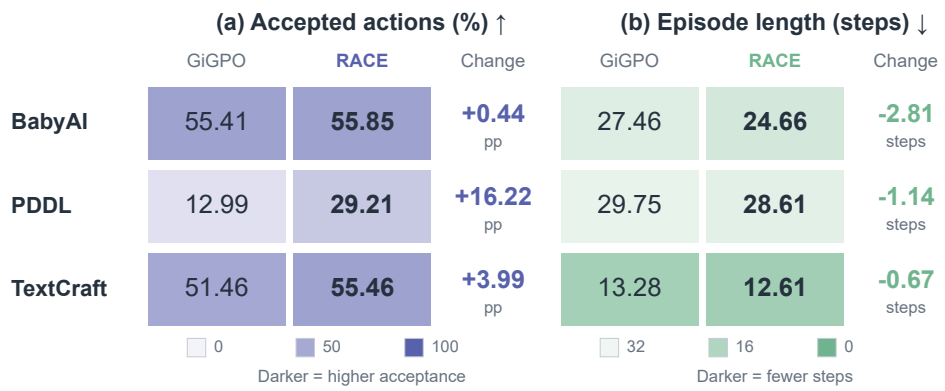


Figure 6: Qwen3-8B transferred interaction behavior averaged equally across the three training sources. (a) Accepted-action rate (%). (b) Mean environment steps per episode. Cells report method means; darker colors indicate higher acceptance or fewer steps within each metric. Annotations show absolute acceptance-rate gains (%) and changes in episode length (steps), computed as RACE minus GiGPO.

D.3 Transfer at the 4B Scale

For Qwen3-4B, RACE improves zero-shot transfer from ALFWorld to the same three target environments (Table 8). Success and progress increase on each target, with absolute task-macro gains of 6.70% and 9.38%, respectively.

Table 8: Zero-shot transfer performance of ALFWorld-trained Qwen3-4B models (%; higher is better). The two blocks report mean success and progress. Macro is the equal-weight average over BabyAI, PDDL, and TextCraft. Shading denotes RACE; bold marks the higher mean within each matched pair, including ties.

Metric	Method	BabyAI	PDDL	TextCraft	Macro
Success	GiGPO	21.88	5.83	39.50	22.40
	RACE	**34.82**	**7.50**	**45.00**	**29.11**
Progress	GiGPO	32.25	18.66	39.50	30.13
	RACE	**49.19**	**24.35**	**45.00**	**39.51**

E Training Dynamics

E.1 Validation and On-Policy Learning Curves

We examine learning over the full training horizon using ALFWorld validation success, ScienceWorld graded score, and WebShop task score (Figure 7). At both model scales, RACE achieves higher final validation values and larger horizon-normalized areas under the curve. Figure 3(a) summarizes the corresponding gains.

Success on trajectories sampled by the current policy is measured at each training update (Figure 8). RACE achieves higher late-training success in every benchmark–model setting, consistent with the held-out validation results.

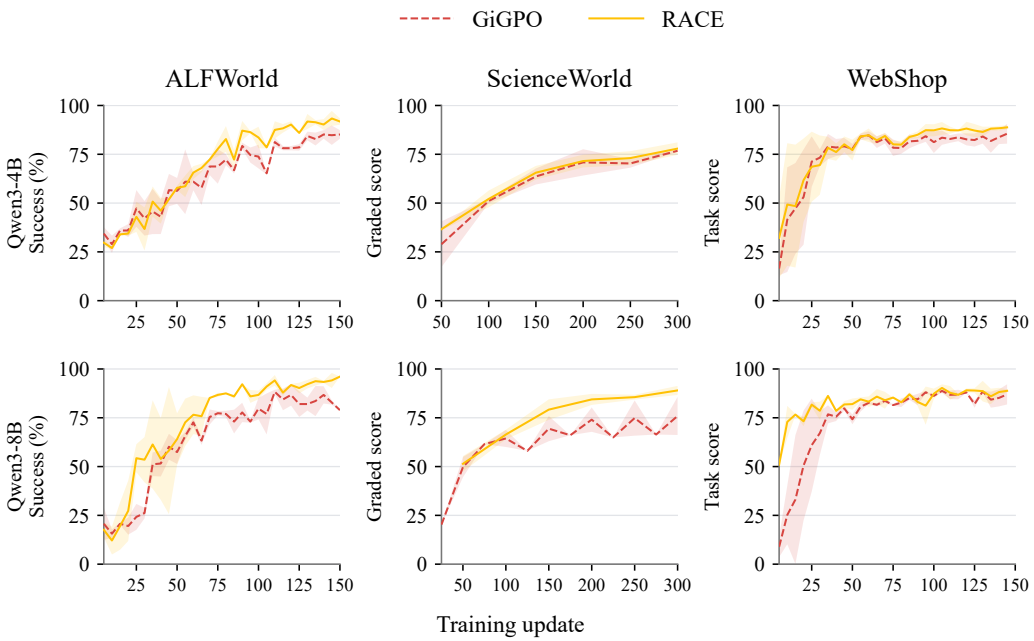


Figure 7: Validation performance during policy training. Rows correspond to Qwen3-4B and Qwen3-8B; columns correspond to ALFWorld success (%), ScienceWorld graded score, and WebShop task score. Curves show means across independent training runs, with shading indicating one sample standard deviation. Values are recorded at each benchmark's evaluation cadence.

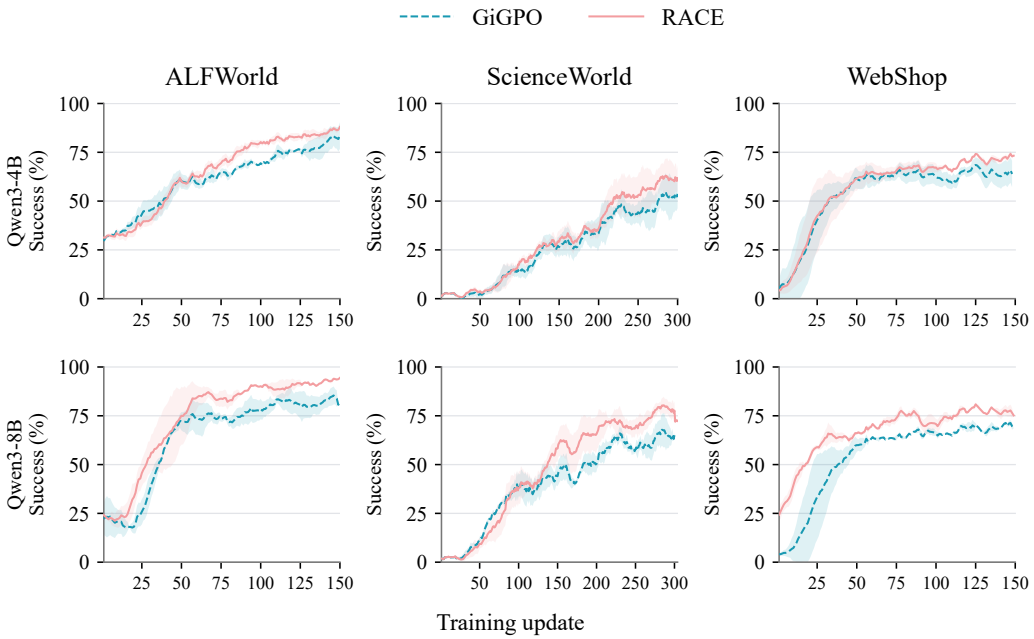


Figure 8: On-policy trajectory success (%) during training. Rows correspond to Qwen3-4B and Qwen3-8B; columns correspond to ALFWorld, ScienceWorld, and WebShop. Curves show means across independent training runs after a centered 15-update moving average; shading indicates one sample standard deviation. Success is measured on trajectories sampled by the current policy. Figure 7 reports the held-out validation metrics.

E.2 Adaptive Intervention Dynamics

Corrections are active across a large fraction of action steps on ScienceWorld, while the weak-signal gates on ALFWorld and WebShop select subsets of steps. Figure 9 tracks this activation together with changes in estimated recoverability, whose magnitude varies across environments and over training.

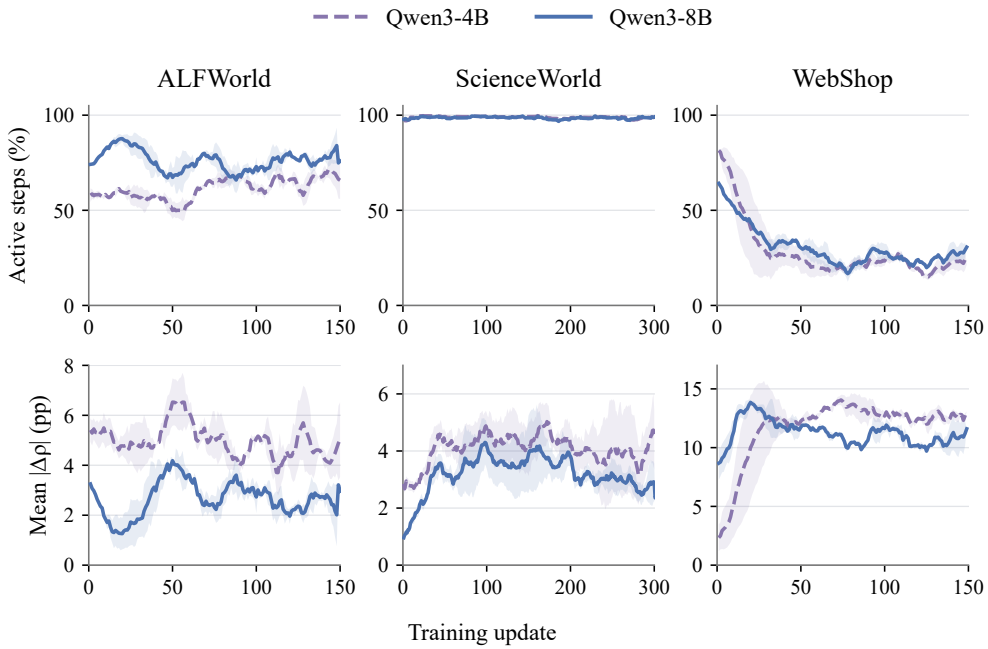


Figure 9: RACE intervention dynamics for Qwen3-4B and Qwen3-8B. Columns correspond to ALFWorld, ScienceWorld, and WebShop. The top row reports the fraction of generated action steps with an active correction (%); the bottom row reports the mean absolute recoverability change (%). Curves show means across independent RACE training runs after a centered 15-update moving average; shading indicates one sample standard deviation.

F Mechanism Analysis and Ablations

F.1 Recoverability Diagnostics

On ScienceWorld, leave-one-trajectory-out recoverability estimates are positively correlated with realized outcomes (Figure 10), with time-averaged trajectory-level correlations of 0.60 at 4B and 0.53 at 8B. Eligible episodic neighbors support 98.9% and 98.6% of action steps, respectively. The mean absolute correction is 5.7% and 5.5% of the mean absolute base advantage.

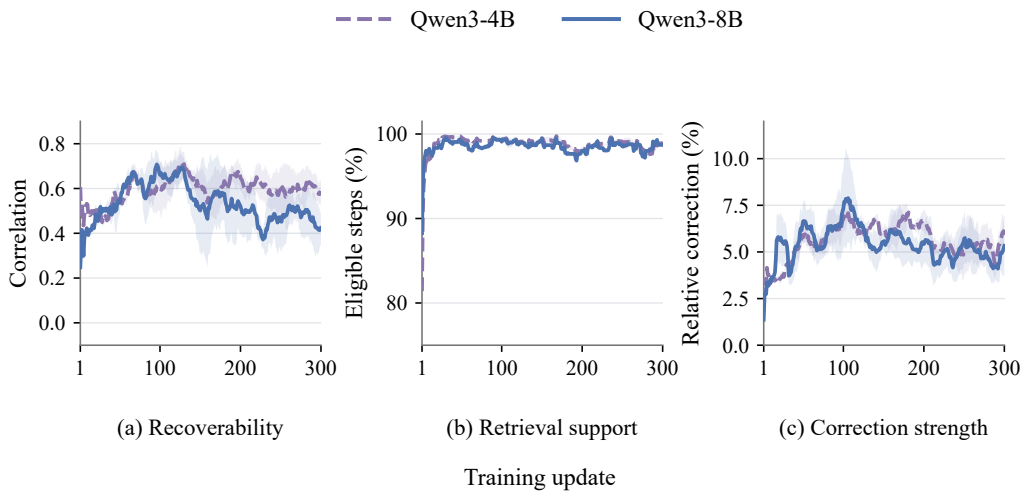


Figure 10: Recoverability diagnostics during ScienceWorld training with Qwen3-4B and Qwen3-8B. (a) Trajectory-level correlation between leave-one-trajectory-out recoverability estimates and realized outcomes. (b) Fraction of action steps with sufficient eligible neighbors (%). (c) Mean absolute RACE correction divided by mean absolute base advantage (%). Curves show means across independent training runs after a trailing 15-update moving average; shading indicates one sample standard deviation.

F.2 Task-Level Ablation Results

Table 9 reports the OOD task-family success rates shown in Figure 4(a), complementing the aggregate ablation results in Table 2.

Table 9: ALFWorld OOD success rates (%) by task family for Qwen3-4B after 100 training updates. Bold marks the highest value in each column, including ties; shading denotes full RACE. Table 2 reports aggregate results.

Configuration	Examine	Pick	Clean	Cool	Heat	Pick2
GiGPO	50.00	76.92	51.52	60.87	52.17	58.82
RACE	**100.00**	80.77	75.76	**82.61**	**78.26**	**82.35**
w/o task-type grouping	83.33	**84.62**	**78.79**	**82.61**	69.57	70.59
within-task retrieval only	83.33	69.23	72.73	**82.61**	73.91	52.94
w/o shrinkage	83.33	76.92	54.55	69.57	**78.26**	70.59
w/o weak-signal gate	58.33	76.92	60.61	47.83	56.52	52.94