CURIOPO: CURBING SPURIOUS EXPLORATION FOR STABLE MULTI-TURN AGENTIC REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent progress in multi-turn reinforcement learning (RL) has significantly advanced LLM agents on complex interactive tasks across web, embodied, and search environments. Despite improvements in credit assignment, rollout filtering, and exploration control, training instability remains pervasive and frequently leads to collapse. We identify spurious exploration as a source of this instability: agents produce valid and diverse interactions whose observations barely change a restricted next-action-token distribution, introducing weak turn-level learning signals. To address this issue, we propose `CurioPO`, a regime-aware policy optimization framework for stable multi-turn agentic RL. `CurioPO` measures each interaction through action belief shift, the observation's effect on next-action tokens, and excess surprise, a proxy for context-dependent novelty after accounting for surface token rarity. The two signals define exploration regimes, and their fused productivity score guides intervention. At the turn level, `CurioPO` evaluates candidates in sandboxed environments before committing them to rollouts and curbs spurious turns according to their regime. At the trajectory-group level, `CurioPO` reallocates rollout budget after consecutive non-productive regimes. During policy optimization, a bounded curiosity gate dampens noisy intermediate turns. Extensive experiments on WebShop, ALFWorld, and Search QA show that `CurioPO` is **(1) high-performing**, raising WebShop success by 4.46% and ALFWorld Overall by 5.61% in absolute terms over the best reported results, **(2) stable**, with lower KL divergence and smaller gradient-norm spikes, and **(3) exploration-efficient**, with fewer spurious turns.

1 INTRODUCTION

Recent advances in multi-turn agentic reinforcement learning have become central to training LLM agents that reason, act, and improve through iterative environment interaction (Liu et al., 2024; Team, 2025; Team et al., 2025; Wang et al., 2026b). Stable training depends on identifying rollouts that provide reliable learning signals for policy updates. Early actions shape later observations, and those observations influence subsequent decisions, while sparse terminal rewards assign credit only after the interaction is completed (Zhou et al., 2024; Wang et al., 2025b). Group-based optimization is therefore sensitive to rollout quality. A rollout can contain distinct actions and valid feedback while offering little useful behavioral contrast. Such surface variation can enter advantage estimates and coincide with unstable gradients, rising KL divergence, and training collapse.

Prior work addresses this instability from several directions. Fine-grained credit assignment propagates sparse outcomes across long trajectories (Feng et al., 2025). Intermediate supervision provides denser guidance (Wang et al., 2025a; Dong et al., 2025), uncertainty-guided control regulates inefficient interaction (Wang et al., 2026a), and trajectory filtering removes low-quality rollouts before optimization (Yu et al., 2025; Xue et al., 2025). These strategies improve the use of terminal feedback, rollout selection, and interaction budgets. Our focus is the feedback within a valid rollout: an observation may appear novel while providing little evidence for the agent's later action choice.

OUR KEY INSIGHT. Figure 1 shows declining success while turn validity and action diversity remain high.

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.

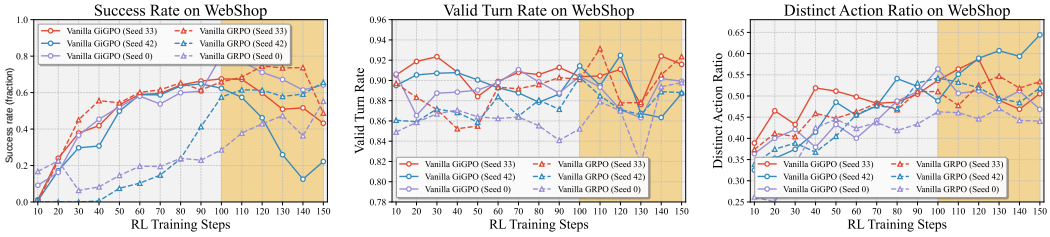


Figure 1: WebShop rollout diagnostics for GiGPO and GRPO. Success declines in the shaded stage (left), although most turns remain valid (middle) and the distinct-action ratio remains high (right). The three measurements motivate an examination of feedback within otherwise valid, diverse rollouts.

Inspection of the highlighted rollouts reveals interactions that return observations with little effect on the restricted next-action-token distribution. We call this failure mode ***spurious exploration***. In its stagnant form, agents repeatedly obtain predictable observations and continue along nearly unchanged behavioral paths. In its distracting form, agents obtain surprising observations that fail to support a better subsequent action. Both forms inflate apparent exploration while introducing weak turn-level signals into policy updates.

This diagnosis motivates measuring the effect of feedback on subsequent action selection. We characterize each observation through two complementary signals. *Action belief shift* measures its effect on the restricted distribution over next-action tokens. *Excess surprise* measures context-dependent novelty after discounting surface token rarity. The first signal represents behavioral consequence, while the second helps distinguish informative environment feedback from unusual wording. Their gated fusion scores feedback productivity and guides exploration control.

We propose `CurioPO`, a regime-aware policy optimization framework for multi-turn agentic RL. At the turn level, it evaluates candidate interactions before committing them to rollouts and applies interventions according to the observed regime. At the trajectory-group level, it reallocates rollout budget after consecutive non-productive regimes while retaining recoverable trajectories. During policy optimization, a bounded curiosity gate adjusts intermediate-turn advantages to reduce the influence of noisy feedback. Rejection-based fine-tuning (Wei et al., 2025), a memory context window, and a format penalty provide the training backbone.

Our contributions are summarized as follows:

❶ **Failure Mode Diagnosis.** We identify *spurious exploration* in interactions that appear exploratory but provide little evidence for subsequent action selection.

❷ **Regime-Aware Optimization.** We develop turn-level control, trajectory-group allocation, and curiosity-gated optimization around feedback productivity.

❸ **Empirical Validation.** Experiments on WebShop, ALFWorld, and Search QA show higher task performance than strong multi-turn RL baselines, together with steadier training and fewer spurious turns.

2 Related Work

2.1 Agentic RL Training

RAGEN (Wang et al., 2025b) provides a framework for training and evaluating agents in long-horizon environments. Subsequent work improves rollout selection and training stability. SimpleTIR (Xue et al., 2025) filters rollouts with void turns, while rStar2-Agent (Shang et al., 2025) oversamples rollout groups and retains high-quality trajectories. GiGPO (Feng et al., 2025) extends group-based advantage estimation to multi-turn interaction without an auxiliary value network. These methods curate trajectories or estimate advantages across rollout groups. `CurioPO` examines the feedback within valid trajectories and regulates non-productive interactions at the turn and trajectory-group levels.

2.2 RL WITH INTERNAL REWARDS AND SIGNALS

Model-derived signals provide another route to more stable RL training. SEED-GRPO (Chen et al., 2025) uses policy entropy to shape exploration, and DeepConf (Fu et al., 2025) filters reasoning traces through internal confidence. T^{2}PO (Wang et al., 2026a) uses uncertainty dynamics to regulate inefficient multi-turn interaction at token and turn levels. These signals describe properties of the agent's generation process. `CurioPO` studies the model-environment interface by measuring whether an observation changes subsequent action tendencies. Excess surprise and action belief shift then distinguish productive, stagnant, and distracting feedback for regime-aware control.

3 PRELIMINARIES

3.1 MULTI-TURN AGENTIC REINFORCEMENT LEARNING

We consider a multi-turn agentic reinforcement learning setting in which an LLM agent interacts with an external environment to solve a task. Each task starts from a user prompt q and an initial environment seed e_0. At turn $k \in \{1, \ldots, K\}$, the agent receives a context $\mathbf{c}^k$ that contains the task prompt and the available interaction history. The policy model π_θ generates a textual response $\mathbf{y}^k$ conditioned on $\mathbf{c}^k$, and an executable action $\mathbf{a}^k$ is parsed from this response. After executing $\mathbf{a}^k$, the environment returns an observation $\mathbf{o}^{k+1}$ and a scalar reward r^k. A trajectory is denoted as

$$\tau = \{(\mathbf{c}^k, \mathbf{a}^k, \mathbf{o}^{k+1}, r^k)\}_{k=1}^{K}. \tag{1}$$

Here τ denotes the complete interaction trajectory, K denotes the maximum number of turns, $\mathbf{c}^k$ denotes the policy input at turn k, $\mathbf{a}^k$ denotes the action committed to the environment, $\mathbf{o}^{k+1}$ denotes the environment feedback after the action, and r^k denotes the reward assigned at the same turn. In many interactive tasks, rewards are sparse and often appear only near the end of the trajectory. This property makes it difficult to decide which intermediate interactions provide reliable learning signals for policy optimization.

3.2 GROUP-BASED POLICY OPTIMIZATION

Following group-based critic-free optimization, we collect G trajectories under the same task prompt q and the same initial environment seed e_0. This design makes trajectories comparable within a rollout group and allows relative advantage estimation without an auxiliary value model. For the ith trajectory, the discounted return from turn k is defined as

$$R_i^k = \sum_{j=k}^{K_i} \beta^{j-k} r_i^j. \tag{2}$$

Here R_i^k denotes the future return of trajectory i starting from turn k, K_i denotes the length of that trajectory, $\beta \in (0, 1]$ denotes the turn discount factor, and r_i^j denotes the reward of trajectory i at turn j. The trajectory-level advantage compares each full trajectory against other trajectories in the same group:

$$A_i^{\text{traj}} = \frac{R_i^1 - \text{mean}(\{R_j^1\}_{j=1}^{G})}{\text{std}(\{R_j^1\}_{j=1}^{G}) + \epsilon_A}. \tag{3}$$

Here A_i^{traj} denotes the relative quality of trajectory i, and ϵ_A is a small constant for numerical stability. For turn-level comparison, let $\mathcal{I}_k = \{j : K_j \geq k\}$ denote the trajectories that reach turn k:

$$A_i^{\text{turn},k} = \frac{R_i^k - \text{mean}(\{R_j^k\}_{j \in \mathcal{I}_k})}{\text{std}(\{R_j^k\}_{j \in \mathcal{I}_k}) + \epsilon_A}. \tag{4}$$

Here $A_i^{\text{turn},k}$ denotes the relative return assigned to action $\mathbf{a}_i^k$ at turn k. GRPO uses trajectory-level grouping, while GiGPO introduces step-level comparisons through repeated anchor states (Shao et al., 2024; Feng et al., 2025). Our backbone combines trajectory returns with turns aligned by index under a shared prompt and environment seed. The contribution of `CurioPO` begins with measuring feedback productivity and using it during rollout collection and updating.

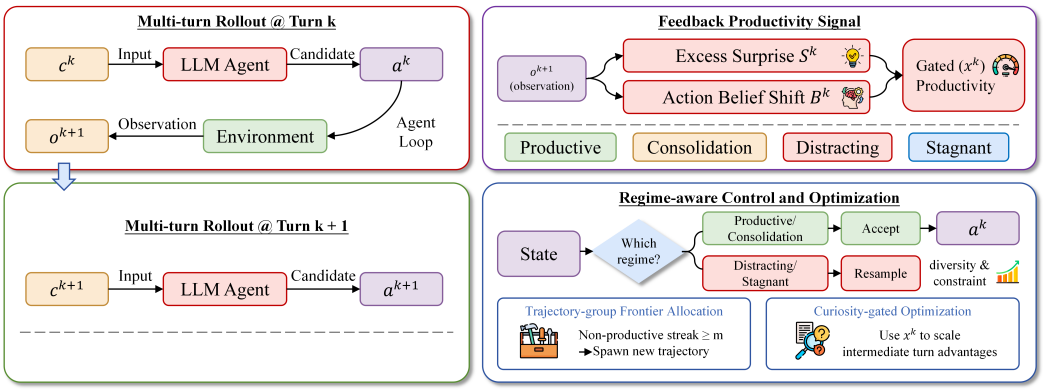


Figure 2: Overview of `CurioPO`. `CurioPO` measures excess surprise and action belief shift for each candidate, assigns a regime from the two signals, and uses their fused productivity score to guide rollout control, frontier allocation, and policy optimization. This design curbs spurious exploration at both rollout-collection and policy-update stages.

4 METHODOLOGY

As shown in Figure 2, `CurioPO` follows a feedback-first design. It first evaluates whether an observation provides useful behavioral evidence, then uses this signal to control turn generation, allocate trajectory budget, and modulate turn-level policy updates. We describe these components in order.

4.1 FEEDBACK PRODUCTIVITY SIGNAL

Our core motivation is that a valid interaction is not always a useful interaction. A turn may contain a well-formed action and a non-empty observation, yet the observation may fail to change the agent behavior in later turns. We therefore evaluate each candidate interaction before it is committed to the rollout. Given a context $\mathbf{c}^k$, a candidate action $\mathbf{a}^k$, and the sandboxed observation $\mathbf{o}^{k+1}$ returned by the environment, we first measure how unexpected this observation is beyond surface rarity. The context-conditioned negative log-likelihood is

$$\mathrm{NLL}_\theta(\mathbf{o}^{k+1} \mid \mathbf{c}^k, \mathbf{a}^k) = -\frac{1}{|\mathbf{o}^{k+1}|} \sum_{t=1}^{|\mathbf{o}^{k+1}|} \log \pi_\theta(o_t^{k+1} \mid o_{<t}^{k+1}, \mathbf{c}^k, \mathbf{a}^k). \tag{5}$$

In this expression, $|\mathbf{o}^{k+1}|$ is the number of tokens in the observation, o_t^{k+1} is the tth observation token, and $o_{<t}^{k+1}$ is its prefix. We compare this value with a reference likelihood that does not condition on the interaction context:

$$S^k = \mathrm{NLL}_\theta(\mathbf{o}^{k+1} \mid \mathbf{c}^k, \mathbf{a}^k) - \mathrm{NLL}_{\mathrm{ref}}(\mathbf{o}^{k+1}). \tag{6}$$

Here S^k denotes excess surprise. The reference model is the RFT initialization frozen at the beginning of RL and scores the observation text without interaction context. This subtraction reduces the influence of rare surface tokens such as product identifiers, object names, and long numbers. We next measure whether the observation changes the immediate action tendency. Let $\mathcal{U}^k$ contain the first tokens of admissible next-turn action strings returned by the sandboxed candidate interaction. Both distributions below use this same set. A fixed reasoning-start probe $\mathbf{u}$ is appended to each context:

$$\mathbf{p}_{\mathrm{pre}}^k = \pi_\theta(\cdot \mid \mathbf{c}^k, \mathbf{a}^k, \mathbf{u})\big|_{\mathcal{U}^k}. \tag{7}$$

$$\mathbf{p}_{\mathrm{post}}^k = \pi_\theta(\cdot \mid \mathbf{c}^k, \mathbf{a}^k, \mathbf{o}^{k+1}, \mathbf{u})\big|_{\mathcal{U}^k}. \tag{8}$$

Here $\mathbf{p}_{\mathrm{pre}}^k$ denotes the next-action distribution before observing $\mathbf{o}^{k+1}$, and $\mathbf{p}_{\mathrm{post}}^k$ denotes the distribution after observing it. The action belief shift is defined by the Jensen-Shannon divergence

$$B^k = \mathrm{JS}(\mathbf{p}_{\mathrm{pre}}^k, \mathbf{p}_{\mathrm{post}}^k). \tag{9}$$

Here B^k measures an immediate change in the next-action-token distribution. Its relation to later task success is evaluated in Section 5.3. Excess surprise and belief shift enter a gated score. We first normalize excess surprise within the current rollout group:

$$\widehat{S}_i^k = \frac{S_i^k - \min_{j,l} S_j^l}{\max_{j,l} S_j^l - \min_{j,l} S_j^l + \epsilon_S}. \tag{10}$$

Here $\widehat{S}_i^k$ is the normalized excess surprise for trajectory i at turn k, and ϵ_S prevents division by zero. The feedback productivity score is

$$\chi_i^k = B_i^k \left[\lambda + (1-\lambda)\widehat{S}_i^k\right]. \tag{11}$$

Here χ_i^k weights behavior-shaping feedback by context-dependent surprise. The floor $\lambda \in (0,1)$ retains low-surprise turns with a substantial belief shift. Appendix A gives the action-interface construction and score interpretation.

4.2 Regime-Aware Turn Control

The score above allows us to separate qualitatively different interaction regimes. We use two thresholds δ_S and δ_B for normalized excess surprise and action belief shift. A turn is productive when $\widehat{S}_i^k \geq \delta_S$ and $B_i^k \geq \delta_B$. A turn is consolidation when $\widehat{S}_i^k < \delta_S$ and $B_i^k \geq \delta_B$. A turn is distracting when $\widehat{S}_i^k \geq \delta_S$ and $B_i^k < \delta_B$. A turn is stagnant when $\widehat{S}_i^k < \delta_S$ and $B_i^k < \delta_B$.

Productive and consolidation turns are accepted because both affect subsequent action selection. Distracting and stagnant turns are treated as spurious candidates and are resampled before they enter the committed rollout. At each turn k, the environment state is saved as a snapshot before execution. A candidate action is generated from the current policy and executed in a cloned copy of this snapshot. The signals S^k and B^k are computed from the resulting observation. If the candidate falls into a productive or consolidation regime, it is accepted and committed to the real trajectory. Otherwise, a regime-specific resampling strategy is applied and a new candidate is generated. This process repeats until an acceptable candidate is found or a maximum resampling budget $B_{\max}$ is reached, after which the candidate with the highest χ_i^k among all attempts is committed.

For stagnant candidates, the agent appears trapped in a narrow behavior mode, so we increase sampling diversity while preserving task anchors. If $z_t(v)$ is the logit of token v at decoding step t, the adjusted logit is

$$\widetilde{z}_t(v) = z_t(v) - \mu f_i^k(v)\mathbb{I}[v \notin \mathcal{A}_{\text{anchor}}]. \tag{12}$$

Here $\widetilde{z}_t(v)$ is the modified logit, $f_i^k(v)$ is the frequency of token v in recent action history, μ controls the penalty strength, and $\mathcal{A}_{\text{anchor}}$ contains task-anchor tokens extracted from the prompt. The sampling temperature is also increased to encourage broader exploration. For distracting candidates, the agent already reaches a novel but unhelpful region, so adding more randomness may worsen exploration. We instead suppress the spurious action signature:

$$\widetilde{z}_t(v) = z_t(v) - \gamma\mathbb{I}[v \in \mathcal{A}_{\text{spur}}^k \setminus \mathcal{A}_{\text{anchor}}]. \tag{13}$$

Here γ controls the constraint strength, and $\mathcal{A}_{\text{spur}}^k$ contains non-anchor tokens from the action that led to the distracting observation. The sampling temperature is lowered to produce more focused candidates. Stagnant turns require more diversity to escape narrow behavior, and distracting turns require more constraint to avoid irrelevant novelty. This asymmetry justifies treating the two categories differently under the same rollout budget.

4.3 Trajectory-Group Frontier Allocation

Turn-level control can correct a single candidate, but a full trajectory may still remain in non-productive regimes for several consecutive turns. Such trajectories consume rollout budget and contribute weak learning signals. We therefore track the suffix length of consecutive distracting or stagnant regimes:

$$N_i^k = \max\{m' \mid \ell_i^{k-m'+1}, \ldots, \ell_i^k \in \{\mathcal{R}_{\text{D}}, \mathcal{R}_{\text{S}}\}\}. \tag{14}$$

Here N_i^k denotes the non-productive streak of trajectory i at turn k, ℓ_i^k denotes the regime label assigned to this turn, $\mathcal{R}_{\text{D}}$ denotes the distracting regime, and $\mathcal{R}_{\text{S}}$ denotes the stagnant regime. If

$N_i^k \geq m$ and $k \geq k_{\min}$, we terminate the current partial trajectory and spawn a new trajectory from the same prompt and the same initial environment seed. The threshold m prevents premature termination after a single bad turn, and $k_{\min}$ ensures that each trajectory has enough context before allocation is applied.

A productive or consolidation turn at any point resets the streak, preserving trajectories that transition between exploration and decision-making phases. The terminated partial trajectory is used only for logging and regime statistics, and the spawned trajectory replaces it in the rollout group for policy optimization. The spawned trajectory uses an elevated temperature $T' = T + \Delta T_{\text{spawn}}$ to encourage a different exploration path. Each trajectory is allowed at most one spawn event to avoid recursive termination. This allocation rule keeps group comparison valid and avoids optimizing on trajectories that repeatedly fail to produce useful feedback.

4.4 Curiosity-Gated Policy Optimization

After rollout collection, `CurioPO` optimizes the policy with the same trajectory-group structure as standard group-based training, but it adjusts the turn contribution according to feedback productivity. Since directly concatenating an entire trajectory for optimization produces excessively long sequences, we adopt a memory context window of size P. The input for turn k contains only the states and actions from turns $k-P$ through $k-1$, rather than the full interaction history. We define a bounded gate from χ_i^k:

$$g_i^k = \phi_{\min} + (\phi_{\max} - \phi_{\min})\sigma\left(\frac{\chi_i^k - \text{med}(\chi)}{\text{std}(\chi) + \epsilon_g}\right). \tag{15}$$

Here g_i^k is the productivity gate, σ is the sigmoid function, $\text{med}(\chi)$ and $\text{std}(\chi)$ are the median and standard deviation of productivity scores in the current batch, ϵ_g prevents division by zero, and $0 < \phi_{\min} \leq \phi_{\max}$ bounds the weight. The gate is detached from gradient computation and only scales intermediate turn advantages. For terminal actions, we set $g_i^K = 1$ because the final outcome should remain governed by task reward. The final advantage is

$$A_i^{\star,k} = A_i^{\text{traj}} + \omega g_i^k A_i^{\text{turn},k}. \tag{16}$$

Here $A_i^{\star,k}$ is the advantage used for action $\mathbf{a}_i^k$, and ω balances trajectory-level and turn-level credit. Since g_i^k is positive and bounded, it changes the magnitude of the turn signal but never flips its sign. The clipped policy objective is

$$\mathcal{J}(\theta) = \mathbb{E}\left[\min\left(\rho_i^k A_i^{\star,k}, \text{clip}(\rho_i^k, 1-\epsilon, 1+\epsilon)A_i^{\star,k}\right)\right] - \beta_{\text{KL}}\mathbb{D}_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}}). \tag{17}$$

Here $\rho_i^k = \pi_\theta(\mathbf{a}_i^k \mid \mathbf{c}_i^k)/\pi_{\theta_{\text{old}}}(\mathbf{a}_i^k \mid \mathbf{c}_i^k)$ is the policy ratio on committed actions, ϵ is the clipping range, β_{KL} controls reference-policy regularization, and $\pi_{\theta_{\text{old}}}$ denotes the rollout policy. The feedback-productivity signal enters this objective through the intermediate-turn advantage.

5 Experiments

5.1 Experiment Setup

Tasks. We evaluate `CurioPO` on three categories of interactive agentic benchmarks. WebShop (Yao et al., 2022) evaluates web-based decision making, where agents search, compare, and purchase products under natural-language user instructions. ALFWorld (Shridhar et al., 2020) evaluates embodied multi-step decision making across household tasks, including pick-and-place, examine, clean, heat, cool, and pick-two. Search QA evaluates tool-augmented information seeking across single-hop and multi-hop question answering tasks, including NQ (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017), PopQA (Mallen et al., 2023), HotpotQA (Yang et al., 2018), 2Wiki (Ho et al., 2020), MuSiQue (Trivedi et al., 2022), and Bamboogle (Press et al., 2023). These tasks cover different interaction structures and therefore allow us to test whether spurious exploration appears beyond a single environment type.

Baselines. We compare `CurioPO` with prompting baselines, instruction tuning, and RL training methods. For prompting, we include GPT-4o (Achiam et al., 2023), Gemini-2.5-Pro (Comanici et al., 2025), Claude Sonnet 4 (Anthropic, 2025), and Qwen3-32B (Team, 2025). For RL training, we

Table 1: Main results on WebShop and ALFWorld. RL methods within each backbone block share an RFT checkpoint and are evaluated under three environment seeds. The subscript denotes standard deviation across seeds.

Method	WebShop		ALFWorld	
	Task Score	Success Rate	Overall	Heat
Prompting				
GPT-4o (Achiam et al., 2023)	31.80	23.70	48.00	56.70
Gemini-2.5-Pro (Comanici et al., 2025)	42.50	35.90	60.30	69.00
Claude Sonnet 4 (Anthropic, 2025)	45.63	39.82	63.71	70.14
Qwen3-32B (Team, 2025)	25.17	5.89	25.63	24.31
Instruction Tuning				
Qwen3-4B + SFT	70.91	26.56	64.06	59.26
RL Training based on `Qwen3-4B-RFT`				
PPO (Schulman et al., 2017)	$70.34_{\pm 8.63}$	$61.93_{\pm 5.93}$	$75.39_{\pm 3.81}$	$62.57_{\pm 1.56}$
GRPO (Shao et al., 2024)	$80.02_{\pm 7.94}$	$68.56_{\pm 4.11}$	$77.35_{\pm 0.62}$	$\mathbf{90.18_{\pm 7.15}}$
GiGPO (Feng et al., 2025)	$86.03_{\pm 4.18}$	$73.83_{\pm 3.04}$	$80.47_{\pm 2.43}$	$86.88_{\pm 4.26}$
GiGPO + DAPO (Yu et al., 2025)	$86.54_{\pm 9.81}$	$74.02_{\pm 8.18}$	$80.86_{\pm 1.37}$	$79.05_{\pm 7.45}$
T^2PO (Wang et al., 2026a)	$89.41_{\pm 2.92}$	$77.18_{\pm 2.37}$	$84.62_{\pm 2.18}$	$84.90_{\pm 5.33}$
`CurioPO`	$\mathbf{93.84_{\pm 0.22}}$	$\mathbf{81.64_{\pm 0.39}}$	$\mathbf{90.23_{\pm 1.38}}$	$85.11_{\pm 7.64}$
RL Training based on `Qwen3-8B-RFT`				
GRPO (Shao et al., 2024)	$79.56_{\pm 9.67}$	$69.47_{\pm 8.01}$	$80.67_{\pm 6.36}$	$70.58_{\pm 3.67}$
GiGPO (Feng et al., 2025)	$88.76_{\pm 5.63}$	$77.92_{\pm 4.87}$	$85.15_{\pm 4.77}$	$81.25_{\pm 7.59}$
T^2PO (Wang et al., 2026a)	$90.72_{\pm 2.38}$	$79.11_{\pm 2.04}$	$88.64_{\pm 2.76}$	$\mathbf{88.16_{\pm 4.72}}$
`CurioPO`	$\mathbf{91.65_{\pm 0.84}}$	$\mathbf{82.42_{\pm 0.61}}$	$\mathbf{92.41_{\pm 1.42}}$	$80.45_{\pm 7.79}$

Table 2: Exact-match accuracy (%) on search-augmented QA tasks. The Qwen3-4B-RFT methods are trained on NQ and HotpotQA, and other backbone blocks are listed separately. † and ⋆ mark datasets inside and outside that training set.

Method	Single-Hop QA			Multi-Hop QA			
	NQ†	TriviaQA⋆	PopQA⋆	HotpotQA†	2Wiki⋆	MuSiQue⋆	Bamboogle⋆
Prompting							
Qwen3-32B (Team, 2025)	13.56	41.32	14.28	18.24	25.77	3.98	12.32
RL Training based on `Qwen2.5-7B-Instruct`							
R1-Instruct	21.00	44.90	17.10	20.80	27.50	6.00	19.20
Search-R1 (Jin et al., 2025)	39.30	61.00	39.70	37.00	40.10	14.60	36.80
ZeroSearch (Sun et al., 2025)	43.60	61.80	**51.50**	34.60	35.20	18.40	27.80
StepSearch (Wang et al., 2025c)	–	–	–	38.60	36.60	**22.60**	40.00
RL Training based on `Qwen3-4B-RFT`							
GiGPO (Feng et al., 2025)	44.36	63.67	46.26	39.28	39.86	13.40	70.97
T^2PO (Wang et al., 2026a)	45.18	63.90	46.91	39.56	40.73	15.71	71.35
`CurioPO`	**46.13**	**64.08**	47.85	**39.80**	**42.51**	16.64	**72.58**

compare with PPO (Schulman et al., 2017), GRPO (Shao et al., 2024), GiGPO (Feng et al., 2025), GiGPO with DAPO-style dynamic sampling (Yu et al., 2025), and T^2PO (Wang et al., 2026a). For Search QA, we further compare with Search-R1 (Jin et al., 2025), ZeroSearch (Sun et al., 2025), and StepSearch (Wang et al., 2025c). Matched RL comparisons within each Qwen3-RFT backbone block use the same initialization and format penalty. Search-specific results based on Qwen2.5-7B-Instruct are shown in a separate block.

Implementation Details. We initialize RL training from `Qwen3-4B-RFT` and `Qwen3-8B-RFT`. For WebShop and ALFWorld, we use the same rollout group size, maximum turn budget, and reward definition across all RL methods. For Search QA, we follow the retrieval setting of Search-R1 (Jin et al., 2025), use E5 as the retriever, and limit the maximum number of interaction turns to four. We train all models with the verl framework (Sheng et al., 2024) and run experiments under three environment random seeds. For the matched Qwen3-RFT comparisons, task prompts, action parsing, rollout group size, and turn budgets are shared across methods. Appendix A.3 and Table 6 specify the configuration details and the supplementary comparison protocol.

5.2 Main Results (RQ1)

RQ1: Can `CurioPO` improve task performance and training stability across diverse multi-turn agentic environments?

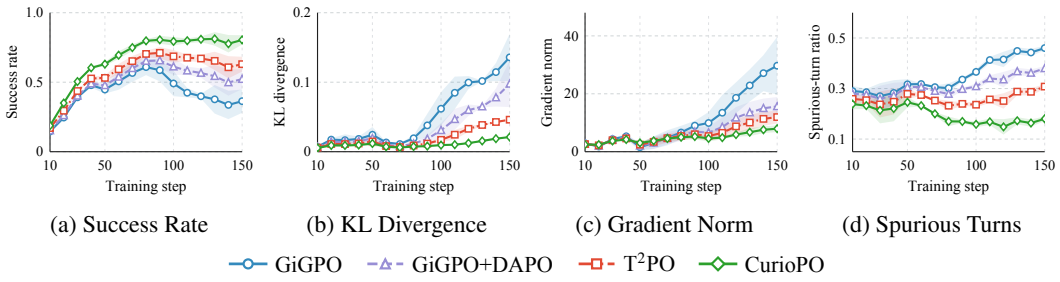


Figure 3: Training dynamics on WebShop. `CurioPO` improves success rate more steadily, reduces KL divergence and gradient-norm spikes, and prevents spurious turns from accumulating during the late training stage.

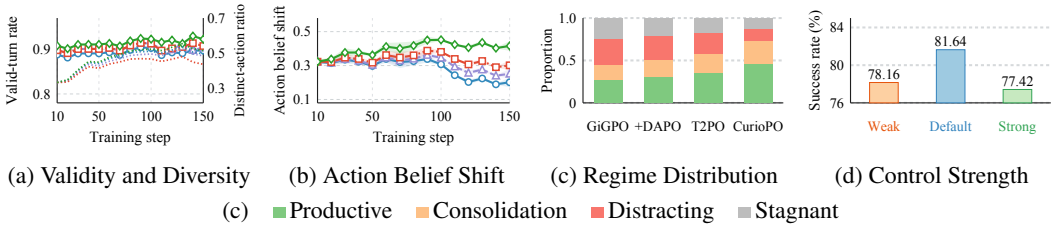


Figure 4: Rollout analysis on WebShop. The four panels compare validity and diversity, action belief shift, regime composition, and combined control strength. Panels (a,b) use the method colors and markers in Figure 3. In (a), solid and dotted lines show valid-turn rate and distinct-action ratio, respectively.

Table 1 and Table 2 show higher aggregate WebShop and ALFWorld results in both backbone blocks and higher Search QA accuracy against the Qwen3-4B-RFT comparators. On WebShop, success rises from GiGPO's 73.83% to 81.64% with `Qwen3-4B-RFT`, and from 77.92% to 82.42% with `Qwen3-8B-RFT`. Heat is the ALFWorld exception, and Appendix B.5 gives the full subtask breakdown and its analysis. Beyond final performance, Figure 3 shows that `CurioPO` produces steadier training dynamics, with lower KL divergence, smaller gradient-norm spikes, and fewer accumulated spurious turns. We attribute the gains over trajectory filtering and uncertainty-guided control to directly measuring whether feedback changes later behavior, which provides a more reliable signal for stabilizing multi-turn policy optimization.

Finding 1. Across WebShop, ALFWorld, and Search QA, `CurioPO` increases overall task success and reduces the training instability shown in Figure 3.

5.3 Model Analysis (RQ2)

RQ2: Does `CurioPO` reduce spurious exploration while preserving valid and diverse interactions?

Figure 4 shows that valid turns and distinct actions remain common across methods. `CurioPO` increases action belief shift and the share of productive and consolidation regimes while reducing stagnant and distracting turns. These measurements connect the training dynamics in Figure 3 to the quality of feedback used for later decisions. Figure 5 compares excess-surprise, belief-shift, fused-score, and random rankings under the same single-observation masking protocol. Over all turns, masking one observation lowers downstream success by 23.7%, 11.4%, 4.8%, and 2.1% from the highest to the lowest belief-shift quartile. Appendix B.2 specifies the masking protocol and separately discusses oracle-usefulness labels, rare-token dependence, and confirmatory feedback.

Finding 2. `CurioPO` retains valid, diverse interactions, and the masking check links higher belief shift to larger downstream success changes.

5.4 Hyperparameter Analysis (RQ3)

RQ3: How do the control settings affect WebShop success and rollout composition?

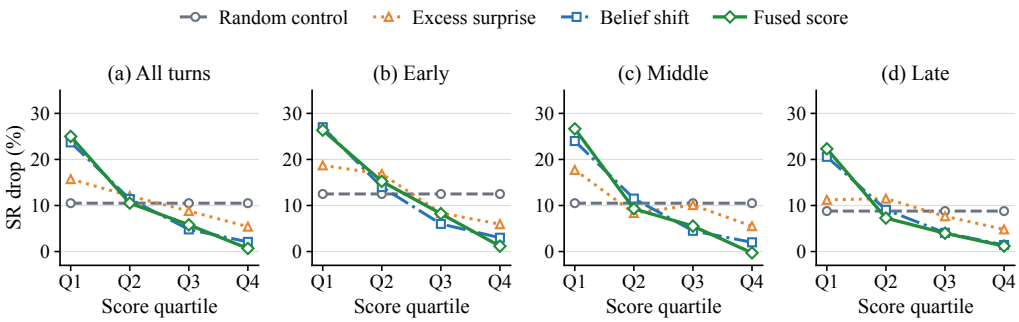


Figure 5: ALFWorld single-observation masking. The vertical axis shows the decrease in downstream success after replacing one observation with a neutral sequence of equal length and rerolling the trajectory suffix. Panels (a–d) show all, early, middle, and late turns, and each compares excess surprise, action belief shift, fused score, and random ranking across score quartiles (Q1 highest). A steeper Q1-to-Q4 decline places more influential observations near the top of the ranking.

Figure 4d compares joint changes to the belief-shift threshold, surprise threshold, resampling budget, and stagnant-streak threshold. The default uses $\delta_B = 0.15$, $\delta_S = 0.35$, $B_{\max} = 4$, and $m = 3$. Appendix B.1 gives the settings and rollout diagnostics. Strong control reduces distinct actions, while weak control retains more spurious interactions. A separate streak sweep gives WebShop success rates of 80.44%, 81.64%, and 80.97% for $m \in \{2, 3, 5\}$. The default $m = 3$ is highest in this one-variable check, which isolates streak length from the joint-control comparison. Table 7 quantifies the joint-control trade-off. Weak control retains a distinct-action ratio of 0.61 but has a spurious-turn ratio of 0.284, whereas strong control reduces these ratios to 0.41 and 0.149. At the default, distinct actions remain frequent (0.58), while the spurious-turn ratio stays near the strong setting (0.168).

Finding 3. Default control achieves the highest WebShop success among the tested settings.

5.5 Ablation Study (RQ4)

RQ4: Which components contribute to stable policy improvement?

Removing action belief shift lowers WebShop success from 81.64% to 70.24% and raises the spurious-turn ratio from 0.168 to 0.341. Table 3 also tests candidate selection, trajectory allocation, and policy updating. Under matched environment calls, `CurioPO` reaches 81.64% versus 74.91%–77.04% for the controls in Appendix B.3, and Appendix B.4 gives compute costs. Action-belief-shift removal gives the largest drop, while sandbox evaluation also changes success and spurious-turn frequency. The remaining rows isolate regime control, frontier allocation, and curiosity gating.

Table 3: WebShop leave-one-out ablations (`Qwen3-4B-RFT`). Shading marks the largest signal-removal drop and full method.

Method	Score	SR	Spur.	KL	Turns
w/o Excess Surprise	88.72	75.91	0.278	0.081	5.46
w/o Action Belief Shift	84.35	70.24	0.341	0.113	5.91
Linear Fusion	87.16	73.82	0.306	0.096	5.68
w/o Regime Control	86.03	72.47	0.322	0.104	5.84
Uniform Resampling	88.04	74.63	0.291	0.088	5.57
w/o Frontier Allocation	89.20	76.19	0.265	0.077	6.08
w/o Curiosity Gate	90.11	77.35	0.241	0.073	5.31
w/o Sandbox Evaluation	85.69	71.86	0.335	0.108	5.96
`CurioPO`	**93.84**	**81.64**	**0.168**	**0.041**	**4.72**

Finding 4. Signal-guided rollout control and policy updating both contribute.

6 Conclusion

`CurioPO` curbs *spurious exploration* in valid, diverse interactions by combining excess surprise with restricted action belief shift. The signals guide rollout selection, trajectory allocation, and policy updates. Results on WebShop, ALFWorld, and Search QA show higher task performance and steadier training.

AI Usage Statement. AI assisted only with language polishing.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. arXiv preprint arXiv:2303.08774, 2023.

Anthropic. Claude sonnet 4, 2025. URL https://www.anthropic.com. Large language model.

Minghan Chen, Guikun Chen, Wenguan Wang, and Yi Yang. Seed-grpo: Semantic entropy enhanced grpo for uncertainty-aware policy optimization. arXiv preprint arXiv:2505.12346, 2025.

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.

Guanting Dong, Licheng Bao, Zhongyuan Wang, Kangzhi Zhao, Xiaoxi Li, Jiajie Jin, Jinghan Yang, Hangyu Mao, Fuzheng Zhang, Kun Gai, et al. Agentic entropy-balanced policy optimization. arXiv preprint arXiv:2510.14545, 2025.

Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. Group-in-group policy optimization for llm agent training. arXiv preprint arXiv:2505.10978, 2025.

Yichao Fu, Xuewei Wang, Yuandong Tian, and Jiawei Zhao. Deep think with confidence. arXiv preprint arXiv:2508.15260, 2025.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. arXiv preprint arXiv:2011.01060, 2020.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, 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.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. arXiv preprint arXiv:1705.03551, 2017.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: a benchmark for question answering research. Transactions of the Association for Computational Linguistics, 7:453–466, 2019.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. arXiv preprint arXiv:2412.19437, 2024.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 9802–9822, 2023.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models. In Findings of the Association for Computational Linguistics: EMNLP 2023, pp. 5687–5711, 2023.

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

Ning Shang, Yifei Liu, Yi Zhu, Li Lyna Zhang, Weijiang Xu, Xinyu Guan, Buze Zhang, Bingcheng Dong, Xudong Zhou, Bowen Zhang, et al. rstar2-agent: Agentic reasoning technical report. arXiv preprint arXiv:2508.20722, 2025.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300, 2024.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. arXiv preprint arXiv: 2409.19256, 2024.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. Alfworld: Aligning text and embodied environments for interactive learning. arXiv preprint arXiv:2010.03768, 2020.

Hao Sun, Zile Qiao, Jiayan Guo, Xuanbo Fan, Yingyan Hou, Yong Jiang, Pengjun Xie, Yan Zhang, Fei Huang, and Jingren Zhou. Zerosearch: Incentivize the search capability of llms without searching. arXiv preprint arXiv:2505.04588, 2025.

Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. Kimi k1. 5: Scaling reinforcement learning with llms. arXiv preprint arXiv:2501.12599, 2025.

Qwen Team. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multihop questions via single-hop question composition. Transactions of the Association for Computational Linguistics, 10:539–554, 2022.

Haixin Wang, Hejie Cui, Chenwei Zhang, Xin Liu, Shuowei Jin, Shijie Geng, Xinyang Zhang, Nasser Zalmout, Zhenyu Shi, and Yizhou Sun. T^2po: Uncertainty-guided exploration control for stable multi-turn agentic reinforcement learning, 2026a. URL https://arxiv.org/abs/2605.02178.

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, 2025a. URL https://arxiv.org/abs/2509.09265.

Xiaoxuan Wang, Han Zhang, Haixin Wang, Yidan Shi, Ruoyan Li, Kaiqiao Han, Chenyi Tong, Haoran Deng, Renliang Sun, Alexander Taylor, et al. Arlarena: A unified framework for stable agentic reinforcement learning. arXiv preprint arXiv:2602.21534, 2026b.

Zihan Wang, Kangrui Wang, Qineng Wang, Pingyue Zhang, Linjie Li, Zhengyuan Yang, Kefan Yu, Minh Nhat Nguyen, Licheng Liu, Eli Gottlieb, Monica Lam, 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, 2025b. URL https://arxiv.org/abs/2504.20073.

Ziliang Wang, Xuhui Zheng, Kang An, Cijun Ouyang, Jialu Cai, Yuhang Wang, and Yichao Wu. Stepsearch: Igniting llms search ability via step-wise proximal policy optimization. arXiv preprint arXiv:2505.15107, 2025c.

Zhepei Wei, Wenlin Yao, Yao Liu, Weizhi Zhang, Qin Lu, Liang Qiu, Changlong Yu, Puyang Xu, Chao Zhang, Bing Yin, et al. Webagent-r1: Training web agents via end-to-end multi-turn reinforcement learning. arXiv preprint arXiv:2505.16421, 2025.

Zhenghai Xue, Longtao Zheng, Qian Liu, Yingru Li, Xiaosen Zheng, Zejun Ma, and Bo An. Simpletir: End-to-end reinforcement learning for multi-turn tool-integrated reasoning. arXiv preprint arXiv:2509.02479, 2025.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In Proceedings of the 2018 conference on empirical methods in natural language processing, pp. 2369–2380, 2018.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. Webshop: Towards scalable real-world web interaction with grounded language agents. Advances in Neural Information Processing Systems, 35:20744–20757, 2022.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. arXiv preprint arXiv:2503.14476, 2025.

Yifei Zhou, Andrea Zanette, Jiayi Pan, Sergey Levine, and Aviral Kumar. Archer: Training language model agents via hierarchical multi-turn rl. arXiv preprint arXiv:2402.19446, 2024.

APPENDIX

A Supplementary Method Details **14**

A.1 Notation and scope of the productivity signal 14

A.2 Action interfaces and regime assignment 14

A.3 Documented task protocol 15

B Configuration and Additional Analyses **16**

B.1 Configuration and control strength 16

B.2 Checks on the usefulness proxy 16

B.3 Matched-budget controls and ablations 17

B.4 Cost and deployment trade-offs 17

B.5 ALFWorld task breakdown 17

B.6 Failures and transfer checks 18

C Limitations and Reporting Details **18**

A Supplementary Method Details

A.1 Notation and scope of the productivity signal

The task prompt and initial environment seed define a rollout group of G trajectories. At turn k of trajectory i, the policy receives context $\mathbf{c}_i^k$, emits a response from which action $\mathbf{a}_i^k$ is parsed, and receives observation $\mathbf{o}_i^{k+1}$ and reward r_i^k. A candidate action can be evaluated on a copy of the environment state before it is committed. The two measurements made on the candidate observation are excess surprise S_i^k and action belief shift B_i^k.

Table 4: Notation used in the method and this appendix. A regime label describes the feedback from one candidate interaction rather than a ground-truth label of the action's long-term value.

Symbol	Meaning
q, e_0	task prompt and initial environment seed
$\mathbf{c}^k, \mathbf{a}^k, \mathbf{o}^{k+1}$	context, parsed action, and resulting observation
$\mathcal{U}^k$	distinct first tokens of admissible next-action strings
$S^k, \widehat{S}^k$	excess surprise and its within-group normalization
B^k, χ^k	action belief shift and fused productivity score
ℓ^k, N^k	regime label and non-productive suffix length
$B_{\max}, m, k_{\min}$	candidate-attempt budget, streak threshold, minimum turn
P, g_i^k	optimization context window and bounded advantage gate

For an observation of L tokens, the method uses the mean observation-token negative log-likelihood (NLL),

$$\mathrm{NLL}_\theta(\mathbf{o} \mid \mathbf{c}, \mathbf{a}) = -L^{-1} \sum_{t=1}^{L} \log \pi_\theta(o_t \mid o_{<t}, \mathbf{c}, \mathbf{a}), \qquad S = \mathrm{NLL}_\theta(\mathbf{o} \mid \mathbf{c}, \mathbf{a}) - \mathrm{NLL}_{\mathrm{ref}}(\mathbf{o}). \quad (18)$$

The reference likelihood conditions on the observation text alone. For the Qwen3-4B-RFT setting, the reference is a frozen copy of that backbone at the start of RL. The subtraction reduces the effect of intrinsically rare observation tokens. Usefulness is assessed separately through action belief shift.

Let $\mathbf{u}$ be the same fixed reasoning-start probe on both sides of the comparison. The method restricts the policy's next-token logits to $\mathcal{U}^k$ and renormalizes with a softmax at temperature one:

$$\mathbf{p}_{\mathrm{pre}}^k = \mathrm{softmax}\big(z_\theta(\cdot \mid \mathbf{c}^k, \mathbf{a}^k, \mathbf{u})\big|_{\mathcal{U}^k}\big), \quad \mathbf{p}_{\mathrm{post}}^k = \mathrm{softmax}\big(z_\theta(\cdot \mid \mathbf{c}^k, \mathbf{a}^k, \mathbf{o}^{k+1}, \mathbf{u})\big|_{\mathcal{U}^k}\big). \quad (19)$$

With $\mathbf{m}^k = (\mathbf{p}_{\mathrm{pre}}^k + \mathbf{p}_{\mathrm{post}}^k)/2$, action belief shift is

$$B^k = \tfrac{1}{2} D_{\mathrm{KL}}(\mathbf{p}_{\mathrm{pre}}^k \| \mathbf{m}^k) + \tfrac{1}{2} D_{\mathrm{KL}}(\mathbf{p}_{\mathrm{post}}^k \| \mathbf{m}^k). \quad (20)$$

Jensen-Shannon divergence is symmetric and bounded. The fixed probe and restricted vocabulary make the measurement a local test of whether the observation changes the model's immediate next-action tendency. It is not an estimate of long-horizon information value.

Writing $\mathcal{G}$ for the comparison pool specified by the group-normalization rule in the main text, the score is

$$\widehat{S}_i^k = \frac{S_i^k - \min_{(j,l)\in\mathcal{G}} S_j^l}{\max_{(j,l)\in\mathcal{G}} S_j^l - \min_{(j,l)\in\mathcal{G}} S_j^l + \epsilon_S}, \qquad \chi_i^k = B_i^k \big[\lambda + (1-\lambda)\widehat{S}_i^k\big]. \quad (21)$$

The multiplicative form makes $\chi = 0$ when the measured action belief shift is zero, regardless of surprise. The λ floor allows low-surprise observations that still alter the action tendency to retain weight. The score ranks feedback within its comparison pool rather than giving a calibrated probability of success.

A.2 Action interfaces and regime assignment

Table 5 summarizes the action-string construction. The same token set $\mathcal{U}^k$ is used to restrict both distributions in Eq. (19). This is a first-token projection: action strings with the same first token collapse to one token and are not individually distinguished by B^k.

Table 5: Action strings used to construct $\mathcal{U}^k$ in each environment.

Environment	Candidate action strings used to form $\mathcal{U}^k$
WebShop	A search action when a search bar is available, plus `click[...]` for each clickable item reported by the environment.
ALFWorld	Environment-provided admissible commands, excluding `help`.
Search QA	The two tool tags `<search>` and `<answer>`.

The regime label ℓ_i^k is the quadrant formed by $\widehat{S}_i^k$ and B_i^k:

$$\ell_i^k = \begin{cases} \text{productive}, & \widehat{S}_i^k \geq \delta_S,\ B_i^k \geq \delta_B, \\ \text{consolidation}, & \widehat{S}_i^k < \delta_S,\ B_i^k \geq \delta_B, \\ \text{distracting}, & \widehat{S}_i^k \geq \delta_S,\ B_i^k < \delta_B, \\ \text{stagnant}, & \widehat{S}_i^k < \delta_S,\ B_i^k < \delta_B. \end{cases} \tag{22}$$

Productive and consolidation candidates are accepted by the turn-control rule. A distracting or stagnant candidate triggers another candidate attempt until an acceptable candidate is found or $B_{\max}$ is exhausted. If no attempt is acceptable, the highest-χ attempt is committed. Stagnant control penalizes recently repeated non-anchor action tokens and increases decoding temperature, whereas distracting control penalizes non-anchor tokens from the candidate's action signature and lowers temperature. Task-anchor tokens are extracted from the task instruction and exempted from both penalties.

For a committed trajectory, N_i^k counts the consecutive distracting or stagnant labels at its suffix. When $N_i^k \geq m$ and $k \geq k_{\min}$, the frontier rule replaces that partial trajectory with one restarted from the same prompt and initial seed. A productive or consolidation label resets N_i^k. At most one such restart is permitted per trajectory slot. The optimization input uses a window of P previous turns, and the full initial task prompt remains part of the context definition.

The update gate rescales the turn-level relative return while retaining the trajectory-level term:

$$g_i^k = \phi_{\min} + (\phi_{\max} - \phi_{\min})\sigma\left(\frac{\chi_i^k - \operatorname{med}(\chi)}{\operatorname{std}(\chi) + \epsilon_g}\right), \qquad A_i^{\star,k} = A_i^{\text{traj}} + \omega g_i^k A_i^{\text{turn},k}. \tag{23}$$

The gate is detached from gradient computation and set to one for terminal actions. Its positive bounds change the magnitude of the intermediate turn term without reversing its sign. The clipped policy-ratio objective and reference-policy KL penalty are otherwise the same as in the main method. Setting $g = 1$ removes curiosity weighting from this episode-plus-turn estimator. The full training procedure still differs from GiGPO, whose turn groups use repeated anchor states.

Algorithmic sequence. For each (q, e_0) group, CurioPO samples candidate actions and scores their cloned observations using Eqs. (18)–(22). It resamples distracting or stagnant candidates up to $B_{\max}$ attempts, commits an accepted candidate or the highest-χ attempt, and updates the non-productive streak. A streak meeting the frontier conditions triggers one restart. The collected group supplies return-based trajectory and turn advantages, which Eq. (23) combines for the clipped policy update.

A.3 Documented task protocol

The evaluation covers WebShop product search and purchase, ALFWorld household interaction, and search-augmented question answering on NQ, TriviaQA, PopQA, HotpotQA, 2Wiki, MuSiQue, and Bamboogle. Search QA follows the Search-R1 retrieval setting with E5 and a four-turn limit. Qwen3-4B-RFT and Qwen3-8B-RFT are the RL initializations, and training uses verl and three environment random seeds. Trainable baselines share the task prompt, action parser, format penalty, initialization, group size, and turn budget within each comparison. Table 6 gives the configuration values.

For WebShop and ALFWorld, environment snapshots use state dictionaries, while Search QA retrieval is stateless. The "without sandbox evaluation" variant computes both signals after action commitment and uses them for frontier allocation and the advantage gate, with no pre-commitment resampling.

B Configuration and Additional Analyses

B.1 Configuration and control strength

Table 6: Configuration values for the control analysis. γ is the distracting-action penalty (key `gamma_sup`), and the return discount is the separate config field named `gamma`.

Parameter	Configuration key	Value
Return discount β	`gamma`	0.95
Belief-shift threshold δ_B	`delta_B`	0.15
Normalized-surprise threshold δ_S	`delta_S`	0.35
Fusion floor λ	`lam`	0.4
Maximum candidate attempts $B_{\max}$	`B_max`	4
Non-productive streak m	`m`	3
Minimum frontier turn $k_{\min}$	`k_min`	2
Stagnant-token frequency penalty μ	`mu`	1.0
Distracting-action penalty γ	`gamma_sup`	1.5
Respawn temperature increment	`delta_tau`	0.3
Gate interval $[\phi_{\min}, \phi_{\max}]$	`phi_min`, `phi_max`	$[0.5, 1.5]$
Optimization context window P	`P`	4 (Search QA: 2)
Rollout group size	`rollout.n`	8
Environment maximum steps	`env.max_steps`	15

Weak and strong control jointly scale the two thresholds and change the attempt and streak budgets. Table 7 gives the resulting parameters and WebShop diagnostics. The weaker setting permits more distinct actions and more spurious turns, whereas the stronger setting records fewer spurious turns and fewer distinct actions, consistent with excessive suppression of useful exploration.

Table 7: Control-strength comparison on WebShop. SR is shown in %, and the remaining diagnostics are proportions.

Setting	δ_B	δ_S	$B_{\max}$	m	SR	Valid	Distinct	Spurious
Weak	0.075	0.175	2	5	78.16	0.94	0.61	0.284
Default	0.150	0.350	4	3	81.64	0.95	0.58	0.168
Strong	0.225	0.525	8	2	77.42	0.96	0.41	0.149

The one-variable checks give WebShop success rates of 80.44, 81.64, and 80.97 as m varies over $2, 3, 5$, and 80.91, 81.64, and 81.48 as P varies over $2, 4, 8$. Across three seeds, the productive-regime proportion is higher for CurioPO than for GiGPO (Figure 4c).

B.2 Checks on the usefulness proxy

Across 5,000 turns, the correlations with rare-token count are 0.71 for raw observation NLL and 0.09 for excess surprise. The comparison separates surface token rarity from the observation score used for candidate selection.

The counterfactual check covers 2,000 ALFWorld turns. One observation at a time is replaced by a neutral token sequence of equal length, and the remaining trajectory is rerolled eight times. Turns are grouped by B quartile before comparing unmasked and masked success rates. This protocol probes downstream consequences of an observation rather than only its immediate distribution shift. It never masks an entire quartile of a trajectory at once.

An ALFWorld oracle label marks an observation as useful when it reveals the location or state of a goal-relevant entity. AUCs for predicting that label are 0.58 for excess surprise, 0.79 for action belief shift, and 0.83 for the fused score χ, against 0.50 for random ranking. Separately, three annotators labeled 300 WebShop turns, yielding a Fleiss κ of 0.68 and an AUC of 0.81 for χ against the majority label. The oracle and human comparisons measure different notions of usefulness, so their AUCs are not interchangeable.

In a confirmatory-feedback check, 8.3% of turns labeled stagnant are oracle-useful. A separate analysis variant exempts turns whose top next-action probability exceeds 0.9 and reduces this fraction to 3.1%. This exemption is outside the default algorithm in the main results table. It addresses feedback that preserves a well-founded plan, while delayed information whose value appears several turns later remains unaddressed.

B.3 Matched-budget controls and ablations

Table 8 compares filtering methods under a matched environment-call budget. The random-resampling row estimates the benefit of extra candidate attempts alone. The raw-NLL row tests whether a generic observation-surprise criterion can explain the full improvement.

Table 8: WebShop environment-call-matched comparison. Score and SR are shown in %.

Method	Score	SR
GiGPO, original budget	86.03	73.83
GiGPO + random resampling	87.02	74.91
GiGPO + outcome-only filtering	87.55	75.63
GiGPO + policy-entropy filtering	88.14	76.28
GiGPO + raw-NLL filtering	88.79	77.04
CurioPO	93.84	81.64

The leave-one-out ablation includes a uniform-resampling variant. It uses a different protocol from the environment-call-matched random-resampling control in Table 8, so the two rows are not interchangeable. The build-up experiment below complements the leave-one-out analysis by adding components in sequence.

The build-up experiment starts from the GiGPO base: success rate progresses from 73.83 to 76.02 after trajectory rejection, 78.55 after regime-aware turn control, 80.09 after frontier allocation, and 81.64 after adding the curiosity gate. Later stages operate on the trajectories retained by earlier ones, so these increments are not independent causal effects. The build-up and leave-one-out designs answer different questions.

B.4 Cost and deployment trade-offs

The cost comparison gives 5.9 environment calls per trajectory, one policy forward pass per turn, 38.2 GPU-hours over 150 steps, and 9.7 wall-clock hours for GiGPO. CurioPO uses 7.1 attempted environment calls, 4.72 committed turns per trajectory, 2.3 policy forward passes per turn, 49.5 GPU-hours, and 12.6 wall-clock hours over the same number of steps. The ratio of GPU-hour totals is approximately 1.30. These run-level totals do not isolate the marginal cost of each signal.

The post-commitment variant removes sandbox evaluation and pre-commitment candidate resampling. Table 3 reports 85.69 WebShop score and 71.86 success rate for this variant. It therefore provides a lower-infrastructure option, but it does not test the same rollout-selection mechanism as full CurioPO.

B.5 ALFWorld task breakdown

The main comparison summarizes ALFWorld Overall and Heat. Table 9 gives all six task categories so that the Heat exception can be read alongside the other household tasks. Under Qwen3-4B-RFT, CurioPO reaches 85.11% on Heat versus 90.18% for GRPO, while it is higher on the other five categories. Under Qwen3-8B-RFT, CurioPO reaches 80.45% on Heat versus 88.16% for T^2PO. Overall is computed across task instances, rather than as an unweighted mean of the six category scores.

Table 9: ALFWorld success by task category (%). Entries are mean ± standard deviation for RL rows and single values elsewhere. Overall appears in the main table and is not repeated here.

Method	Pick	Look	Clean	Heat	Cool	Pick2
Prompting						
GPT-4o	75.30	60.80	31.20	56.70	21.60	49.80
Gemini-2.5-Pro	92.80	63.30	62.10	69.00	26.60	58.70
Claude Sonnet 4	90.13	65.34	66.77	70.14	29.80	61.36
Qwen3-32B	63.53	18.33	18.70	24.31	10.08	10.11
Instruction tuning						
Qwen3-4B + SFT	89.29	66.67	64.12	59.26	35.71	54.54
RL from Qwen3-4B-RFT						
PPO	83.34 ± 9.47	75.09 ± 6.25	74.50 ± 7.90	62.57 ± 1.56	84.21 ± 0.00	58.33 ± 7.38
GRPO	85.32 ± 6.77	64.59 ± 4.34	91.16 ± 0.79	90.18 ± 7.15	73.87 ± 9.64	60.20 ± 5.31
GiGPO	87.94 ± 8.91	77.31 ± 8.36	87.95 ± 6.87	86.88 ± 4.26	79.09 ± 4.68	71.41 ± 7.08
GiGPO + DAPO	89.94 ± 8.06	72.08 ± 0.08	93.05 ± 0.43	79.05 ± 7.45	83.08 ± 7.75	65.55 ± 9.12
T^2PO	92.73 ± 6.42	80.54 ± 5.12	93.48 ± 3.91	84.90 ± 5.33	82.19 ± 4.76	74.03 ± 5.41
CurioPO	97.36 ± 6.94	87.77 ± 4.89	98.33 ± 2.77	85.11 ± 7.64	85.84 ± 2.57	80.35 ± 2.86
RL from Qwen3-8B-RFT						
GRPO	90.59 ± 4.27	72.12 ± 7.37	83.33 ± 6.12	70.58 ± 3.67	88.91 ± 5.38	62.39 ± 4.72
GiGPO	92.10 ± 9.36	84.65 ± 2.84	89.47 ± 8.36	81.25 ± 7.59	80.76 ± 4.02	75.03 ± 6.94
T^2PO	96.42 ± 3.91	86.30 ± 3.84	94.36 ± 4.15	88.16 ± 4.72	86.67 ± 3.51	80.26 ± 3.12
CurioPO	99.15 ± 2.05	90.91 ± 4.37	96.67 ± 3.77	80.45 ± 7.79	90.91 ± 4.15	85.71 ± 1.46

B.6 Failures and transfer checks

In a WebShop task with a 40-dollar budget, a page revisit has $(\widehat{S}, B) = (0.11, 0.02)$ and is classified as stagnant, whereas a subsequent 89-dollar product page has $(0.71, 0.03)$ and is classified as distracting. The contrast shows how visually novel feedback can leave the immediate next decision unchanged.

On ALFWorld Heat, CurioPO reaches 85.11% with Qwen3-4B-RFT versus 90.18% for GRPO. Heat can require a fixed action sequence whose confirming observations barely change the next-action distribution. A separate top-1-probability exemption variant reaches 91.24%, suggesting that confirmatory observations benefit from a less aggressive intervention rule.

On MuSiQue, CurioPO scores 16.64 versus 13.40 for GiGPO and 15.71 for T^2PO. Among turns labeled non-productive, 14.2% contain a bridge entity needed for the eventual answer, illustrating delayed utility under a four-turn search budget. In a 200-task, one-seed τ-bench pilot, CurioPO scores 48.5 versus 43.2 for GiGPO on retail and airline tool use. This pilot provides a targeted transfer check.

C Limitations and Reporting Details

Action belief shift measures an immediate change in a restricted first-token distribution. Useful feedback can confirm an existing plan, reveal a bridge fact needed later, or affect later action tokens, and a large shift can also reflect distraction. The four regimes are operational rollout-control categories. Sandboxed resampling assumes reversible candidate actions, while non-enumerable or compound actions require an alternative action representation.

The evaluation uses three environment seeds and reports standard deviations for several comparisons. The Search QA comparison is reported separately across seven datasets to expose variation across retrieval settings.