BLPO: BEYOND-LOCAL POLICY OPTIMIZATION FOR LONG-HORIZON LLM AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-horizon tasks require large language model (LLM) agents to coordinate interdependent actions, a capability central to automating complex workflows. Stepwise group-based reinforcement learning provides finer credit than trajectory-level supervision, but reliable comparisons must account for differences in interaction history. Enforcing this consistency within the limited rollouts of each task can leave insufficient support for estimating action preferences. We propose Beyond-Local Policy Optimization (BLPO), which estimates task-progress advantages from experience shared across tasks and training updates. We summarize trajectory prefixes into task-progress states and abstract actions by their function, allowing corresponding decisions from different tasks to contribute to shared estimates. To control bias and noise from heterogeneous tasks, we accumulate context-relative returns in a global memory with balanced aggregation and reliability weighting. These global advantages complement local comparisons over normalized actions and trajectory-level credit without a learned critic or additional rollouts. Experiments on ALFWorld and WebShop with 1.5B and 7B policies show improved task success over the evaluated baselines under matched training budgets. Further analyses show faster learning and shorter trajectories with little computational overhead.

1 INTRODUCTION

Large language model (LLM) agents extend automated assistance from answering individual queries to executing tasks in external environments (Yao et al., 2023; Zhou et al., 2024a). In applications such as travel planning and computational research, they offer the prospect of delegating entire workflows that otherwise require substantial human coordination (Shen et al., 2026; Li et al., 2026b). Realizing this potential requires solving long-horizon tasks, where earlier actions establish prerequisites for later steps and environmental feedback informs subsequent decisions (Zhu et al., 2026). Agents must therefore coordinate their actions over extended interactions, making the ability to learn policies that sustain progress toward task completion central to their practical utility.

Recent efforts to improve agents' performance on long-horizon tasks have increasingly explored learning from interaction, particularly through reinforcement learning (RL) (Zhou et al., 2024b; Wang et al., 2025; Jin et al., 2025; He et al., 2026b). Among these approaches, Group Relative Policy Optimization (GRPO) compares trajectory returns to optimize the policy without a separate critic (Shao et al., 2024). However, assigning one advantage to all actions can reinforce errors in successful trajectories and suppress useful steps in failed ones. When all rollouts receive the same failure reward, their relative advantages vanish entirely (Yang et al., 2026). To distinguish individual contributions, GiGPO compares actions taken from matching observed states across rollouts (Feng et al., 2025). Such comparisons can nevertheless mix different interaction histories, motivating HGPO to construct groups with varying degrees of historical consistency and adaptively combine their advantages (He et al., 2026a).

Despite these refinements, requiring greater historical consistency can leave the limited rollouts of a task insufficient for reliable comparisons (He et al., 2026a). ECPO improves the reliability of local action estimates (Li et al., 2026a), while BiPACE expands comparison groups by aggregating similar states (Wang et al., 2026); both still estimate credit within sampled rollout groups. Experience from related tasks offers additional support. Histories involving different entities may represent equivalent

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.

task progress, while different concrete actions may serve the same function. Grouping decisions on this basis allows comparisons to accumulate across tasks and training updates. Such sharing, however, must separate action effects from differences in task difficulty and prevent frequently visited contexts from dominating the estimates. Reliable global credit therefore requires correcting these biases while controlling how strongly uncertain comparisons influence policy updates.

Building on this perspective, we propose Beyond-Local Policy Optimization (BLPO), which estimates task-progress advantages by pooling experience across tasks and training updates. To construct these comparisons, predefined environment-specific rules summarize task progress from prior interactions and map concrete actions to their goal-relative functions. Applied automatically across task instances, these rules define comparison groups, while their action preferences are estimated from rollout returns. To control the heterogeneity introduced by sharing, we center returns within concrete contexts and balance their contributions to a decaying statistical memory. Preferences derived from this memory are weighted by their support and uncertainty, limiting the influence of unreliable estimates. The resulting task-progress advantages complement local comparisons over normalized actions, while trajectory advantages preserve supervision of overall task completion. All three signals enter a unified policy objective without a learned critic or additional rollouts.

Our contributions are threefold:

- We propose BLPO, a critic-free policy optimization method that reuses task-progress-aligned experience across tasks and training updates for step-level credit assignment, without additional rollouts.
- We develop a context-balanced global advantage estimator that reduces bias from context difficulty and repeated visits. Support and uncertainty scale the resulting credit to limit the influence of unreliable comparisons.
- Experiments on ALFWorld and WebShop show improved task success with both 1.5B and 7B policies. Controlled comparisons under the same abstraction support the benefits of cross-task sharing, while learning and cost analyses show faster ALFWorld learning with little additional computation.

2 Related Work

Reinforcement Learning for LLM Agents. Research on agentic RL increasingly treats sustained interaction as the basis for policy learning. ArCHer combines utterance-level value estimation with token-level optimization (Zhou et al., 2024b), while RAGEN and Search-R1 develop training around multi-turn trajectories (Wang et al., 2025; Jin et al., 2025). As training extends to longer interactions, AgentGym-RL introduces a curriculum that progressively increases interaction length (Xi et al., 2026). Complementing these algorithmic developments, Agent Lightning v1.0 addresses the systems integration of RL with existing agents while preserving the execution harness's control of the interaction loop (He et al., 2026b).

Credit Assignment in Agentic RL. Within these training frameworks, a central question is how task outcomes should supervise individual decisions. RLOO and GRPO offer relative estimates without a learned critic, but their direct trajectory-level application does not distinguish intermediate actions (Ahmadian et al., 2024; Shao et al., 2024). GiGPO makes these comparisons more granular by grouping decisions with matching observations (Feng et al., 2025). HGPO further conditions groups on interaction history, addressing inconsistencies hidden by observation matching (He et al., 2026a). Such finer comparisons also raise questions of statistical support and reliability. BiPACE increases support by grouping states in actor representation space (Wang et al., 2026), whereas ECPO calibrates comparisons over equivalent actions using support and variability (Li et al., 2026a). Alongside these group estimators, GraphGPO derives credit from goal distances in trajectory graphs (Cheng et al., 2026), while HCAPO refines action values through hindsight reasoning (Tan et al., 2026).

Experience Reuse for Agent Learning. The experience available for learning can also extend beyond current rollouts. One approach supplies earlier experience as policy context through retrieved step traces in RAPO and evolving skill documents in SkillRise (Zhang et al., 2026; Yao

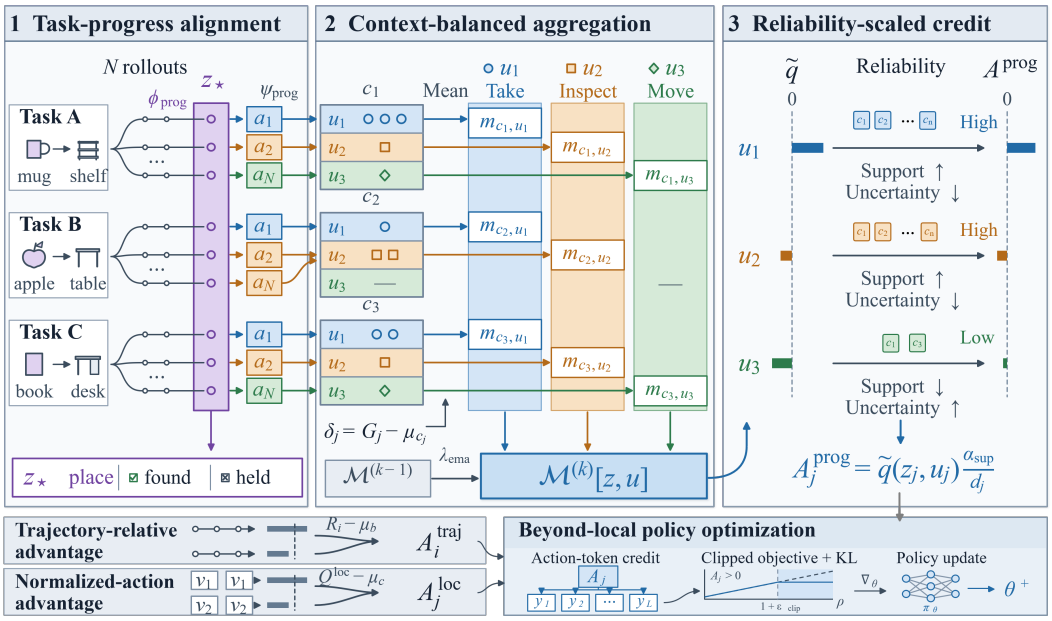

Figure 1: Overview of BLPO. (1) Trajectory histories and concrete actions from different tasks are mapped to shared task-progress states and action functions. (2) Within each concrete context, returns are centered and averaged per abstract state–action pair to update a global memory across training iterations. (3) Action contrasts estimated from this memory are scaled by effective support and uncertainty. The resulting task-progress advantages are combined with trajectory-relative advantages and local advantages over normalized actions to optimize the policy.

et al., 2026). Rather than exposing history through the policy input, TIGPO retains it in persistent transition graphs, combining past and current transitions for credit estimation (Gan, 2026). This form of temporal reuse remains confined to individual task instances. BLPO instead matches task progress and action function to pool comparisons across tasks and policy updates. To manage the resulting heterogeneity, we adjust returns within concrete contexts, balance their contributions, and attenuate uncertain comparisons.

3 Preliminaries

For a task goal g_i, a trajectory τ_i consists of T_i sequential decisions. At step t, the policy generates a textual response $a_{i,t}^{\text{raw}} \sim \pi_\theta(\cdot \mid x_{i,t})$, where $x_{i,t}$ contains the goal, current observation $s_{i,t}$, and interaction history retained within the policy's context window. The response is parsed into an environment action $a_{i,t}$, whose execution yields feedback $e_{i,t}$, including the next observation, reward $r_{i,t}$, termination status, and format and admissibility information. The complete interaction prefix available before the current action is

$$\mathcal{H}_{i,t} = \big(g_i, s_{i,1}, a_{i,1}, e_{i,1}, \ldots, a_{i,t-1}, e_{i,t-1}, s_{i,t}\big). \tag{1}$$

Unlike the policy input $x_{i,t}$, this prefix records all preceding interactions, including those outside the context window. We use $j = (i,t)$ to index an individual decision and write decision-specific quantities with either subscript.

4 BLPO

As summarized in Figure 1, we combine trajectory relative advantages (Sec. 4.1) and local advantages over normalized actions (Sec. 4.2) with task-progress advantages from global experience (Sec. 4.3). The global estimator pools context-adjusted returns across tasks and updates, accounting for support and uncertainty. Together, these signals define the policy objective in Section 4.4.

4.1 TRAJECTORY RELATIVE ADVANTAGE

For each task instance, we sample N trajectories from $\pi_{\theta_{\text{old}}}$ under the same initial state. Let b_i identify the rollout group containing τ_i. Following GRPO (Shao et al., 2024), we normalize trajectory returns within each group:

$$R_i = \sum_{t=1}^{T_i} r_{i,t}, \qquad A_i^{\text{traj}} = \frac{R_i - \mu_{b_i}}{\sigma_{b_i} + \epsilon}, \tag{2}$$

where μ_{b_i} and σ_{b_i} are the group mean and standard deviation, and $\epsilon > 0$ ensures numerical stability. Rewards follow the shared convention in Appendix A.2, including the format penalty. The resulting advantage is shared by all decisions in the trajectory.

For the subsequent step comparisons, we discount future rewards to reduce the influence of feedback separated from the current decision by many intervening actions:

$$G_j = \sum_{t'=t}^{T_i} \gamma^{t'-t} r_{i,t'}, \qquad j = (i,t), \quad \gamma \in (0,1]. \tag{3}$$

4.2 STEP ADVANTAGE OVER NORMALIZED ACTIONS

To distinguish individual decisions, we compare actions under the same task and observation. However, advantages computed from individual returns can assign opposite signs to equivalent actions when subsequent decisions differ. We therefore pool their returns before estimating local credit.

For decision $j = (i,t)$, we define

$$c_j = (b_i, s_j), \qquad v_j = \text{NormAct}(a_j^{\text{raw}}, e_j). \tag{4}$$

The context c_j matches the rollout group and current observation without conditioning on earlier history. The action mapping removes textual variations while preserving the operation and concrete arguments. Feedback is used only to distinguish malformed outputs, inadmissible actions, and executable actions; the normalization rules are specified in Appendix A.3.

Let $\mathcal{I}(c,v)$ contain all occurrences of normalized action v in context c within the current batch. We average their returns and compute the local advantage as

$$Q_{\text{loc}}(c,v) = \frac{1}{|\mathcal{I}(c,v)|} \sum_{j' \in \mathcal{I}(c,v)} G_{j'}, \qquad A_j^{\text{loc}} = \frac{Q_{\text{loc}}(c_j, v_j) - \mu_{c_j}}{\sigma_{c_j} + \epsilon}, \tag{5}$$

where μ_c and σ_c are the mean and standard deviation of all individual returns in context c, rather than of the action means. Thus, equivalent actions within a context receive the same advantage. Groups without distinct action means or return variation contribute zero local advantage.

4.3 TASK-PROGRESS ADVANTAGE OVER GLOBAL STATE–ACTION GROUPS

Matching current observations can mix histories with different task progress and bias the local baseline. Conditioning on history reduces this mismatch, but limited rollouts per task cannot reliably support detailed history groups. We therefore abstract histories and actions to pool comparable experience across tasks and training updates.

From the interaction prefix in Eq. 1, we maintain an execution state $B_{i,t}$ that records discovered entities, completed operations, and current relations:

$$B_{i,t+1} = \text{Update}(B_{i,t}, a_{i,t}, e_{i,t}). \tag{6}$$

Updates incorporate new observations and accepted action effects. This state is used for credit estimation without changing the policy input.

Using predefined rules for each environment, we map entities to goal roles, preserving task family, predicate completion, and relations governing unfinished conditions whose prerequisites are satisfied. Since concrete commands would still yield sparsely observed alternatives, we also abstract their function:

$$z_j = \phi_{\text{prog}}(g_i, B_j), \qquad u_j = \psi_{\text{prog}}(a_j; g_i, B_j). \tag{7}$$

Unlike v_j, u_j can group commands involving different entities, such as acquiring different target objects. Identical (z, u) pairs share statistics, while different u under the same z form the comparison group. Both maps are applied automatically; credit values are estimated from rollout returns. Appendix A.4 specifies these mappings.

To account for differences in context difficulty and prevent repeated visits from dominating, we compute the mean residual within each (c, z, u):

$$m_{c,z,u} = \mathrm{Mean}\{\delta_j : c_j = c,\ z_j = z,\ u_j = u\}, \qquad \delta_j = G_j - \mu_{c_j}. \tag{8}$$

Here δ_j measures the return relative to its concrete context's baseline. Each context contributes one such mean per (z, u) to the global memory in each batch. Let $\mathcal{E}_k(z, u)$ be the multiset of these observations collected at update k from contexts with multiple shareable action functions. We accumulate them in a global memory while discounting statistics from older policies:

$$\begin{aligned} W_{z,u}^{(k)} &= \lambda_{\mathrm{ema}} W_{z,u}^{(k-1)} + |\mathcal{E}_k(z,u)|, \\ S_{z,u}^{(k)} &= \lambda_{\mathrm{ema}} S_{z,u}^{(k-1)} + \sum_{m \in \mathcal{E}_k(z,u)} m, \\ M_{z,u}^{(k)} &= \lambda_{\mathrm{ema}} M_{z,u}^{(k-1)} + \sum_{m \in \mathcal{E}_k(z,u)} m^2. \end{aligned} \tag{9}$$

Here $0 < \lambda_{\mathrm{ema}} < 1$, and W, S, M denote the decayed weight, weighted sum of residuals, and weighted sum of squared residuals. After incorporating the current batch, we query $q^{(k)}(z, u) = S_{z,u}^{(k)} / W_{z,u}^{(k)}$. Distinct context contributions are tracked separately to estimate effective support. Support depends on the balance of accumulated weight across distinct contexts, not on the number of recorded decisions. Thus, repeated observations from a few contexts do not by themselves justify a strong global advantage.

Let $\mathcal{U}_z$ contain the recorded action functions for state z. Omitting the update index, we compare each function with the group mean:

$$\widetilde{q}(z, u) = q(z, u) - \frac{1}{|\mathcal{U}_z|} \sum_{u' \in \mathcal{U}_z} q(z, u'). \tag{10}$$

Omitted task details and subsequent decisions still introduce variation, while support can differ substantially between action alternatives. We therefore scale each contrast by its support and uncertainty. With $n_{\mathrm{cmp}}(z, u)$ denoting the harmonic effective support of the action and its alternatives, we define $\alpha_{\mathrm{sup}}(z, u) = n_{\mathrm{cmp}}(z, u)/(n_{\mathrm{cmp}}(z, u) + \lambda_{\mathrm{sup}})$ and obtain

$$A_j^{\mathrm{prog}} = \alpha_{\mathrm{sup}}(z_j, u_j) \frac{\widetilde{q}(z_j, u_j)}{\sqrt{\widehat{\sigma}_{z_j}^2 + \widehat{\mathrm{SE}}_{\mathrm{diff}}^2(z_j, u_j) + \eta^2}}. \tag{11}$$

Here $\widehat{\sigma}_z^2$ measures dispersion across action means, $\widehat{\mathrm{SE}}_{\mathrm{diff}}$ estimates contrast uncertainty, and $\eta > 0$ provides a numerical floor. The signal is zero unless the action is recorded and its group has at least two action functions, nonzero dispersion, and effective context support of at least two. Appendix A.5 details the support and variance estimators. For WebShop, the resulting advantages are further centered over active decisions sharing z within the current batch.

4.4 Beyond-Local Policy Optimization

We combine trajectory credit with the available local and task-progress advantages:

$$A_j = A_i^{\mathrm{traj}} + \lambda_{\mathrm{step}} A_j^{\mathrm{step}}, \qquad A_j^{\mathrm{step}} = \alpha_j A_j^{\mathrm{loc}} + (1 - \alpha_j) A_j^{\mathrm{prog}}. \tag{12}$$

Here α_j balances local and task-progress advantages according to the availability of the two estimators, while λ_{step} controls the contribution of step credit relative to trajectory credit. The weighting rule and parameter settings are specified in Appendix A.6.

We assign A_j to every token of action a_j^{raw} and maximize the clipped objective:

$$\mathcal{J}_{\mathrm{BLPO}}(\theta) = \mathbb{E}_{j,\ell}[\min(\rho_{j,\ell} A_j, \mathrm{clip}(\rho_{j,\ell}, 1 - \epsilon_{\mathrm{clip}}, 1 + \epsilon_{\mathrm{clip}}) A_j) - \beta \mathcal{D}_{\mathrm{KL},j,\ell}]\,. \tag{13}$$

For action token $y_{j,\ell}$, $\rho_{j,\ell} = \frac{\pi_\theta(y_{j,\ell}|x_j, y_{j,<\ell})}{\pi_{\theta_{\text{old}}}(y_{j,\ell}|x_j, y_{j,<\ell})}$ is the policy probability ratio. The parameter ϵ_{clip} controls clipping, while $\mathcal{D}_{\text{KL},j,\ell}$ regularizes the policy against a fixed reference with weight β. The update reuses sampled trajectories without a learned critic, process reward model, or additional environment interactions.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Benchmarks. Following prior work, we evaluate on ALFWorld (Shridhar et al., 2021) and WebShop (Yao et al., 2022) for long-horizon household and online shopping tasks. ALFWorld measures compositional manipulation success; WebShop evaluates purchases against user-specified product constraints.

Baselines. Baselines include GPT-4o, Gemini-2.5-Pro, direct Qwen2.5 prompting, ReAct (Yao et al., 2023), and Reflexion (Shinn et al., 2023). We also compare PPO (Schulman et al., 2017), RLOO (Ahmadian et al., 2024), GRPO (Shao et al., 2024), GiGPO (Feng et al., 2025), and HGPO (He et al., 2026a). Table 1 uses the baseline results reported by He et al. (2026a), with GiGPO and HGPO at history length $K = 2$ to match our policy inputs.

Implementation details. Training uses Qwen2.5-1.5B-Instruct and Qwen2.5-7B-Instruct (Qwen et al., 2025), 160 updates, 16 tasks per update, and eight trajectories per task. We measure learning curves and computational costs in our own runs. Appendix A provides evaluation and prompt details, with training and interaction parameters consolidated in Appendix A.6.

Table 1: Performance on ALFWorld and WebShop. In-Success and Out-Success measure in-distribution and out-of-distribution success. WebShop Task Score and Task Success measure partial and complete request satisfaction. RL results are mean ± standard deviation over three seeds. Bold marks the best result per backbone; dashes indicate unavailable results.

Model	Type	Method	ALFWorld		WebShop	
			In-Success	Out-Success	Task Score	Task Success
Closed	Prompting	GPT-4o	48.0	46.0	31.8	23.7
	Prompting	Gemini-2.5-Pro	60.3	50.5	42.5	35.9
Qwen2.5-1.5B-Instruct	Prompting	Qwen2.5	4.1	–	23.1	5.2
	Prompting	ReAct	12.8	–	40.1	11.3
	Prompting	Reflexion	21.8	–	55.8	21.9
	RL Training	PPO	$54.4_{\pm 3.1}$	–	$73.8_{\pm 3.0}$	$51.5_{\pm 2.9}$
	RL Training	RLOO	$69.7_{\pm 2.5}$	$68.7_{\pm 10.7}$	$73.9_{\pm 5.6}$	$52.1_{\pm 6.7}$
	RL Training	GRPO	$72.8_{\pm 3.6}$	$70.1_{\pm 2.5}$	$75.8_{\pm 3.5}$	$56.8_{\pm 3.8}$
	RL Training	GiGPO	$90.16_{\pm 0.22}$	$84.76_{\pm 2.83}$	$84.95_{\pm 2.57}$	$66.53_{\pm 1.92}$
	RL Training	HGPO	$92.77_{\pm 1.08}$	$90.16_{\pm 0.78}$	$85.56_{\pm 2.86}$	$71.54_{\pm 4.00}$
	RL Training	**BLPO**	$\mathbf{94.01}_{\pm 0.23}$	$\mathbf{91.54}_{\pm 0.60}$	$\mathbf{87.43}_{\pm 0.56}$	$\mathbf{74.02}_{\pm 1.37}$
Qwen2.5-7B-Instruct	Prompting	Qwen2.5	14.8	–	26.4	7.8
	Prompting	ReAct	31.2	–	46.2	19.5
	Prompting	Reflexion	42.7	–	58.1	28.8
	RL Training	PPO	$77.08_{\pm 1.12}$	$76.23_{\pm 1.46}$	$81.4_{\pm 3.1}$	$68.7_{\pm 5.1}$
	RL Training	RLOO	$77.86_{\pm 0.03}$	$73.95_{\pm 0.05}$	$80.3_{\pm 3.2}$	$65.7_{\pm 4.0}$
	RL Training	GRPO	$78.64_{\pm 0.73}$	$76.82_{\pm 1.47}$	$79.3_{\pm 2.8}$	$66.1_{\pm 3.7}$
	RL Training	GiGPO	$93.29_{\pm 0.40}$	$92.18_{\pm 0.39}$	$88.93_{\pm 1.49}$	$77.60_{\pm 1.68}$
	RL Training	HGPO	$95.44_{\pm 0.62}$	$92.05_{\pm 0.22}$	$\mathbf{88.96}_{\pm 1.04}$	$78.51_{\pm 1.40}$
	RL Training	**BLPO**	$\mathbf{97.14}_{\pm 1.33}$	$\mathbf{92.97}_{\pm 0.68}$	$88.90_{\pm 1.26}$	$\mathbf{79.43}_{\pm 1.57}$

5.2 Main Results

BLPO achieves the best overall performance. Table 1 reports the highest mean success for BLPO on both benchmarks at both scales. With Qwen2.5-1.5B, ALFWorld Out-Success rises from HGPO's 90.16% to 91.54%, and WebShop Task Success from 71.54% to 74.02%. Under matched rollout and update budgets, the 1.5B policy also surpasses 7B GRPO on all four metrics.

Global experience complements local credit assignment. GiGPO distinguishes decisions within trajectories; HGPO further conditions comparisons on history. BLPO pools decisions with shared task progress and action function across tasks and updates. Its gains over HGPO suggest that broader support, with context adjustment and reliability weighting, complements history conditioning within individual tasks.

Scaling and generalization. With the 7B backbone, BLPO reaches 97.14% In-Success and 92.97% Out-Success on ALFWorld, exceeding both GiGPO and HGPO. On WebShop, it attains a Task Score of 88.90, comparable to HGPO's 88.96, while increasing Task Success from 78.51% to 79.43%. Thus, BLPO improves complete task fulfillment even when partial request satisfaction remains similar.

5.3 Ablation Study

Local credit. Table 2 shows that removing action-level pooling reduces ALFWorld Out-Success by 8.40 points and WebShop Task Success by 7.81 points. Using individual returns allows different continuations to assign inconsistent local credit to the same action.

Table 2: Component ablations with Qwen2.5-1.5B. Gray values show absolute drops from full BLPO.

	ALFWorld		WebShop	
Variant	In-Suc.	Out-Suc.	Score	Success
BLPO	**94.01**	**91.54**	**87.43**	**74.02**
Local credit				
w/o Act. pool.	91.34 ▼ 2.67	83.14 ▼ 8.40	82.79 ▼ 4.64	66.21 ▼ 7.81
Task-progress groups				
w/o A^{prog}	92.25 ▼ 1.76	83.53 ▼ 8.01	82.75 ▼ 4.68	68.49 ▼ 5.53
w/o Act. abstr.	88.80 ▼ 5.21	82.94 ▼ 8.60	86.83 ▼ 0.60	71.22 ▼ 2.80
Global estimation				
w/o Residual	88.41 ▼ 5.60	81.97 ▼ 9.57	85.14 ▼ 2.29	73.24 ▼ 0.78
w/o EMA	90.89 ▼ 3.12	84.11 ▼ 7.43	86.09 ▼ 1.34	69.92 ▼ 4.10
w/o Conf.	89.97 ▼ 4.04	81.05 ▼ 10.49	85.00 ▼ 2.43	71.09 ▼ 2.93

Task-progress groups. The loss of task-progress credit lowers WebShop Task Score from 87.43 to 82.75 and Task Success from 74.02% to 68.49%. When local comparisons are uninformative, trajectory credit alone cannot distinguish useful steps from later failures. Removing action abstraction also lowers success in both environments despite retaining progress states and global memory. Concrete labels split equivalent decisions across objects and tasks, limiting comparison support.

Global estimation. Removing residualization or confidence adjustment lowers ALFWorld Out-Success below the variant without task-progress credit, supporting control of context bias and noise in shared estimates. On WebShop, removing confidence adjustment lowers Task Success by 2.93 points; removing EMA lowers it by 4.10 points. The latter uses only the current batch, supporting temporal accumulation for sparse action groups.

5.4 Further Analysis

Learning dynamics. BLPO learns faster and uses fewer actions in Figure 2. At update 80, success reaches 80.7% with 20.4 actions per task, versus HGPO's 65.0% and 24.9 actions. BLPO retains the highest final success and shortest trajectories. Success rises as the interaction count falls, indicating that shorter trajectories do not come at the expense of task completion. Later gains in success are smaller for all methods, but trajectories continue to shorten. Length averages cover all 140 tasks, including failures and timeouts, so the reduction reflects lower overall interaction cost across the complete set of evaluation tasks.

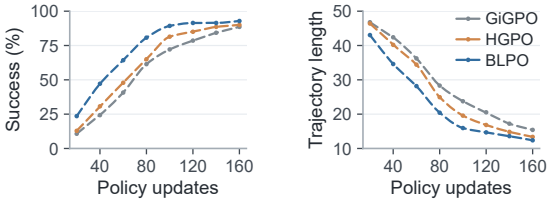

Figure 2: ALFWorld learning dynamics.

Computational cost. Global credit adds little cost in Table 3. Computing advantages takes 2.12 seconds per update, or 0.61% of update time. The increase over GiGPO and HGPO is approximately 1.2 seconds per update. This cost includes local grouping, memory maintenance, and reliability estimation. Memory retains aggregated statistics rather than replaying past trajectories, so queries reuse historical evidence without reevaluating earlier rollouts. Maintaining this broader comparison pool requires neither a separate value network nor additional environment interactions. The full run takes 15.54 hours, within the baseline range. Overall time also depends on the trajectories produced by the policy, not just the estimator's arithmetic. Appendix A.7 gives timing details.

Table 3: Training cost on ALFWorld (4 A100s, 160 updates). Adv. and Update are mean times per update; Share is their ratio. Evaluation time is excluded.

Method	Adv. (s)	Update (s)	Share	Total (h)
GiGPO	0.89	335.2	0.26%	14.90
HGPO	0.96	362.6	0.26%	16.11
BLPO	2.12	349.7	0.61%	15.54

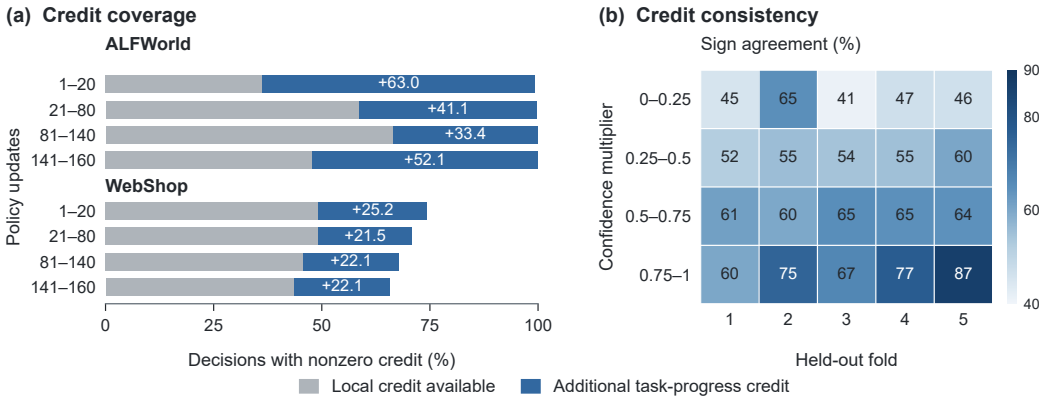

Figure 3: Coverage and consistency of global credit. (a) Fractions of decisions with nonzero local credit (gray) and additional coverage from task-progress credit (blue). (b) Sign agreement across confidence bins and held-out task folds in ALFWorld.

Coverage of task-progress credit. Figure 3a shows that task-progress credit remains available where local advantages are zero. During updates 141–160, it covers 52.1% of ALFWorld decisions and 22.1% of WebShop decisions. Local advantages can vanish when available rollouts do not distinguish alternatives within a concrete context. These decisions retain trajectory credit, but the common outcome does not distinguish their intermediate contributions. Task-progress groups supply comparisons across tasks, so preferences can remain available when the current task offers little local contrast. This additional coverage persists late in training rather than merely rescaling nonzero local signals. Appendix B.1 defines the diagnostics.

Reliability of global credit. Figure 3b compares preferences estimated from four task subsets with residual contrasts in the held-out subset. Held-out tasks do not contribute to the estimates, so agreement tests preferences across task instances rather than their fit to the observations used for estimation. This probes whether progress-aligned groups preserve action preferences despite differences in the concrete tasks. The highest confidence bin achieves greater sign agreement than the lowest in all five folds. Agreement is not monotonic in every fold, but the overall pattern supports reliability weighting of global preferences. Appendix B.2 gives the protocol.

Benefit of cross-task experience. Same-task-only restricts memory to each task instance, retaining the abstraction, cross-update accumulation, and all other estimation and fusion rules. Table 4 shows lower success than both full BLPO and removing task-progress credit. On WebShop, its Task Score stays near full BLPO, but Task Success falls by 7.03 points, separating partial satisfaction from completion. With the abstraction fixed, these findings support cross-task experience beyond the shared domain prior. Removing the branch also performs better than restricting it to one task, so retaining task-progress credit alone does not explain the gains. Its benefit depends on

Table 4: Qwen2.5-1.5B results under global sharing, same-task-only sharing, and removal of the task-progress credit branch.

	ALFWorld		WebShop	
Setting	In-Suc.	Out-Suc.	Score	Success
Full BLPO	**94.01**	**91.54**	**87.43**	**74.02**
Same-task-only	87.63	81.97	87.19	66.99
w/o A^{prog}	92.25	83.53	82.75	68.49

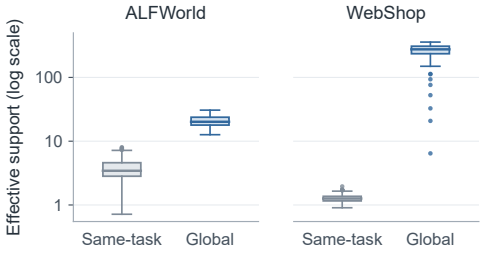

Figure 4: Effective context support from same-task and global experience. Boxes summarize per-update means in full BLPO.

the comparison evidence supplied to the estimator; limited support can yield noisy estimates that interfere with local credit.

Effective context support. Figure 4 shows greater effective context support with cross-task sharing, especially in WebShop. Effective support accounts for the concentration of contribution weights across contexts, rather than simply counting recorded decisions. Repeated visits to one context add weight, not distinct contexts. Memory can therefore persist across updates yet provide narrow support, as the same-task WebShop estimates illustrate. The reliability correction uses support on both sides of each action comparison. A broader pool can strengthen weakly supported action comparisons, helping explain the value of global credit under the same abstraction.

Case study. Figure 5 shows two actions receiving negative local credit despite advancing necessary subgoals: acquiring the target object in ALFWorld and selecting the requested size in WebShop. In the first case, taking a remote control shares its function with acquiring other target objects, including CDs and toilet paper, in placement tasks. In the second, selecting a blanket size is compared with satisfying required options for other products, including shirt sizes and shoe colors. Although the WebShop trajectory ultimately fails, selecting the requested size still resolves a necessary constraint, separating this useful step from the final outcome. Evidence from these groups turns the mixed step advantage positive in both cases, aligning credit with the actions' contribution to task progress. The global preferences remain positive after excluding the current task's contributions, confirming support from other tasks.

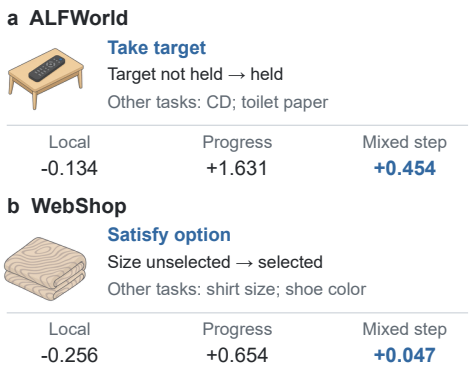

Figure 5: Examples of cross-task credit in ALFWorld and WebShop.

6 Discussion and Limitations

Our results support global credit estimation that controls context difficulty and uncertainty, rather than increasing comparison support alone. The current instantiation uses environment-specific rules to align task progress and action functions, while sharing the statistical estimator across environments. Applying BLPO to new domains calls for suitable comparison groups; learning transferable abstractions is a promising direction for broadening its applicability.

7 Conclusion

We introduced BLPO to estimate task-progress advantages from global experience while controlling context bias and uncertainty. Combined with trajectory and local credit, these advantages improve task success on ALFWorld and WebShop at both model scales. Faster learning and shorter ALFWorld trajectories further support global experience as a useful source of decision-level supervision.

AI USE STATEMENT

We used generative AI tools to assist with coding during this work, supporting the development and refinement of research software. The authors retain full responsibility for the implementation, its correctness, and the results reported in this paper.

REPRODUCIBILITY STATEMENT

Section 4 presents the advantage estimators and optimization objective. Appendix A details the reward convention, state and action construction, global memory, hyperparameters, and evaluation protocol. Appendix B describes the data and statistical procedures used in the additional analyses. Code is available for review at https://anonymous.4open.science/r/BLPO-1FDD/.

REFERENCES

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

Xin Cheng, Shuo He, Lang Feng, HaiYang Xu, Ming Yan, Lei Feng, and Bo An. Beyond trajectory-level attribution: Graph-based credit assignment for agentic reinforcement learning. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2605.26684.

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

Jinwei Gan. TIGPO: Temporal instance-graph policy optimization for long-horizon LLM agents, 2026. URL https://arxiv.org/abs/2609.03383.

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

Zhiyuan He, Siwei Zhang, Zhiwen Zhou, Yuqing Yang, Yu Kang, Yuge Zhang, Luna K. Qiu, Tin Yan Tsui, Jiahang Xu, and Chong Luo. Agent Lightning v1.0: Towards harnessed agentic RL, 2026b. URL https://arxiv.org/abs/2608.17528.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan O. Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-R1: Training LLMs to reason and leverage search engines with reinforcement learning. In *Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=Rwhi91ideu.

Yuanfan Li, Qi Zhou, Wenjing Duan, and Lu Chen. When denser credit is not enough: Evidence-calibrated policy optimization for long-horizon LLM agent training, 2026a. URL https://arxiv.org/abs/2606.05885.

Zongxia Li, Zhongzhi Li, Yucheng Shi, Ruhan Wang, Junyao Yang, Zhichao Liu, Xiyang Wu, Anhao Li, Yue Yu, Ninghao Liu, Lichao Sun, Haotao Mi, and Leowei Liang. Long-Horizon-Terminal-Bench: Testing the limits of agents on long-horizon terminal tasks with dense reward-based grading, 2026b. URL https://arxiv.org/abs/2607.08964.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang,

Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115v2.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

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, 2024. URL https://arxiv.org/abs/2402.03300.

Yuanzhe Shen, Zisu Huang, Zhengyuan Wang, Muzhao Tian, Zhengkang Guo, Chenyang Zhang, Shuaiyu Zhou, Zengjie Hu, Dailin Li, Jingwen Xu, Kaimin Wang, Wenhao Liu, Tianlong Li, Fengpeng Yue, Feng Hong, Cao Liu, and Ke Zeng. TRIP-Bench: A benchmark for long-horizon interactive agents in real-world scenarios, 2026. URL https://arxiv.org/abs/2602.01675.

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

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. ALFWorld: Aligning text and embodied environments for interactive learning. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021. URL https://arxiv.org/abs/2010.03768.

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

Hanyang Wang, Weijieying Ren, Yuxiang Zhang, Ding Cao, Zhizhao Zeng, Ke Zeng, and Tianxiang Zhao. BiPACE: Bisimulation-guided policy optimization with action counterfactual estimation for LLM agents, 2026. URL https://arxiv.org/abs/2606.25556.

Zihan Wang, Kangrui Wang, Qineng Wang, Pingyue Zhang, Linjie Li, Zhengyuan Yang, Xing Jin, Kefan Yu, Minh Nhat Nguyen, Licheng Liu, Eli Gottlieb, Yiping Lu, Kyunghyun Cho, Jiajun Wu, Li Fei-Fei, Lijuan Wang, Yejin Choi, and Manling Li. RAGEN: Understanding self-evolution in LLM agents via multi-turn reinforcement learning, 2025. URL https://arxiv.org/abs/2504.20073.

Zhiheng Xi, Jixuan Huang, Chenyang Liao, Baodai Huang, Jiaqi Liu, Honglin Guo, Yajie Yang, Rui Zheng, Junjie Ye, Jiazheng Zhang, Wenxiang Chen, Wei He, Yiwen Ding, Guanyu Li, Zehui Chen, Zhengyin Du, Xuesong Yao, Yufei Xu, Jiecao Chen, Tao Gui, Zuxuan Wu, Qi Zhang, Xuanjing Huang, and Yu-Gang Jiang. AgentGym-RL: An open-source framework to train LLM agents for long-horizon decision making via multi-turn RL. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://iclr.cc/virtual/2026/oral/10008786.

Kaibing Yang, Guangfeng Cai, Shengtian Yang, Shuo He, Yu Li, Mengyi Liu, Pengwei Chen, Jun Xu, and Lei Feng. Progress-conditioned group policy optimization for long-horizon agentic tasks, 2026. URL https://arxiv.org/abs/2607.22724.

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

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

Zhiyuan Yao, Yuxin Chen, Zhengxi Lu, Zishan Xu, Yueqing Sun, Yifu Guo, Yuquan Lu, Zhengzhou Cai, Kangning Zhang, Zhuowen Han, Zi-Han Wang, Ziang Ye, Qi Gu, Xunliang Cai, Weiwen Liu, and Yongliang Shen. SkillRise: Agentic reinforcement learning for cross-task skill evolution, 2026. URL https://arxiv.org/abs/2607.26784.

Siwei Zhang, Yun Xiong, Xi Chen, Zi'an Jia, Renhong Huang, Jiarong Xu, and Jiawei Zhang. RAPO: Expanding exploration for LLM agents via retrieval-augmented policy optimization. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2026. URL https://arxiv.org/abs/2603.03078.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. WebArena: A realistic web environment for building autonomous agents. In *International Conference on Learning Representations*, 2024a. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/4410c0711e9154a7a2d26f9b3816d1ef-Abstract-Conference.html.

Yifei Zhou, Andrea Zanette, Jiayi Pan, Sergey Levine, and Aviral Kumar. ArCHer: Training language model agents via hierarchical multi-turn RL. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 62178–62209. PMLR, 2024b. URL https://proceedings.mlr.press/v235/zhou24t.html.

Zilin Zhu, Longteng Guo, Yanghong Mei, Bowen Pang, Zongxun Zhang, Xingjian He, Ruyi Ji, and Jing Liu. When robots do the chores: A benchmark and agent for long-horizon household task execution, 2026. URL https://arxiv.org/abs/2605.14504.

A Implementation Details

A.1 Evaluation protocol

Our runs share the same environment interface, prompts, action parser, and interaction limits. Following GiGPO (Feng et al., 2025), we report success rates on ALFWorld and both task scores and success rates on WebShop. We evaluate each final checkpoint with three evaluation seeds. For each seed, ALFWorld uses 512 episodes on each of the In and Out splits, and WebShop uses 512 tasks. We compute each metric separately for each seed and report the mean and sample standard deviation across seeds. For the learning curves in Figure 2, all methods and checkpoints are evaluated on the same 140 ALFWorld In tasks.

A.2 Rollout reward convention

Our training runs use the same rule-based rollout reward. Following GiGPO (Feng et al., 2025), successful task completion receives a reward of 10 and failure receives 0; a format-invalid model output incurs an additional penalty:

$$r_{i,t} = r^{\text{env}}_{i,t} - \lambda_{\text{fmt}} \mathbb{I}\left[a^{\text{raw}}_{i,t} \text{ is malformed}\right], \qquad \lambda_{\text{fmt}} = 0.1. \tag{14}$$

The penalty is applied before both the episodic return in Equation 2 and the discounted future return in Equation 3 are computed. It applies only when the raw output violates the prescribed response format or cannot be parsed into an action. A well formed but inadmissible action is recorded separately and is not treated as a format violation.

A.3 EXACT CONTEXT AND LOCAL ACTION POOLING

The concrete context in Equation 4 combines the task rollout-group identifier b_i with the action's pre-execution observation. Two decisions enter the same local group only when both components match. Observation matching preserves the complete content, including concrete entities and arguments, without entity replacement or semantic matching. Thus, identical observations from different task groups remain separate, as do observations that differ in their concrete objects. Earlier interaction history and WebShop's additional public-state fields are not used for local grouping.

For action normalization, we apply Unicode NFKC normalization to the extracted action text, trim its ends, convert letters to lowercase, and collapse consecutive whitespace. This removes differences in writing without deleting the operation, entity identifiers, or arguments. The normalized text is paired with one of three validity classes: malformed output, a well formed but inadmissible action, or an executable action. Validity is determined from the parser and environment feedback after execution; it does not change the pre-action observation used for grouping. For example, case and whitespace variants of `take apple 1 from table 1` represent the same normalized action. Changing the object identifier or validity class instead produces a distinct normalized action.

Once the rollout is complete, discounted returns are averaged within each (c, v) group. The baseline μ_c and sample standard deviation σ_c are instead computed over all individual decision occurrences in c, not over equally weighted action means. The pooled advantage is assigned to every occurrence of the same action. A group is inactive when its return variation is negligible or all pooled action means coincide with the baseline, including the single-action case. All local groups and statistics are rebuilt for each rollout batch. The w/o action pooling variant replaces $Q_{\mathrm{loc}}(c_j, v_j)$ with G_j in Equation 5, retaining the context baseline and scale. Its activity depends on individual-return variation rather than differences between action means. For the global residual in Equation 8, we subtract μ_c from each occurrence's own return G_j, not from its pooled action value.

A.4 HISTORY AND TASK-PROGRESS ABSTRACTION

Construction principle. The rules follow a common construction principle across the two environments. We begin with task completion conditions and identify observable facts that track whether their prerequisites have been met. Task descriptions define goal-relative roles, while interaction history and environment feedback update the relevant facts before each decision. States summarize task progress, and actions are mapped to their functions relative to the goal and current progress. This retains distinctions such as acquiring an unplaced target versus retaking a placed one, while allowing object identities to differ. The environment-specific rules are reused across task instances and executed automatically, without annotating individual trajectories. They define comparison groups rather than numerical advantages or additional progress rewards; action preferences are estimated from rollout returns. State construction uses only pre-action information and does not extend the policy input.

ALFWorld history and progress states. Task metadata identifies the task family, target object, destination, and required tool. Placement tasks define a destination relation; Heat, Cool, and Clean add a required property; Pick2 requires two placements; Look requires inspection under light. We reconstruct the relevant facts by replaying accepted actions, tracking possession, object locations, completed transformations, placement counts, encountered entities, and container or tool states. Rejected actions leave these facts unchanged, and removing a previously placed instance decreases the placement count.

The task-progress state combines the task family, the types of goal conditions, their progress representations, and their joint completion pattern. Each condition retains goal-relative relations, a coarse progress stage, and relevant facts such as whether the target is held or the required tool is accessible. Transformation stages distinguish an undiscovered target, discovery or possession, possession with a seen or accessible tool, placement in the tool, and completed transformation. Placement conditions separately retain whether delivery remains unfinished; Pick2 additionally records placement count, while Look retains target and lighting conditions. These stages are categorical grouping features and do not provide additional rewards.

ALFWorld action functions. Commands are parsed into operations and arguments, with entities mapped to task roles such as target, destination, and tool. Functions distinguish navigation, acquisition, transformation, placement, inspection, and exploration. For Pick2, taking a target from the destination is classified as a retake rather than acquisition, separating actions that can undo completed placement.

WebShop decision frontiers. Fixed patterns extract options and price limits; normalized tokens represent the requested product. These constraints are compared with the observation and public page state, including query, selected options, clickable elements, and prior interaction counts. The task-progress state retains only the page type and active decision frontier, omitting concrete query and product identities.

Frontier rules are evaluated in order. On a results page, an affordable candidate covering at least two-thirds of the target product tokens marks readiness to inspect a candidate. Otherwise, query coverage below two-thirds indicates a need for query refinement. If coverage is sufficient and another results page exists, the frontier instead indicates further result exploration. On an item page, excessive price or zero product-token overlap first marks the item as unsuitable. Missing or conflicting options then indicate a need for option selection when a requested choice is available; otherwise, the item is unsuitable. Once required options are satisfied, sufficient token coverage and an acceptable price mark readiness to purchase. Remaining cases indicate a need for information only if product details have not yet been inspected; otherwise, the frontier remains unresolved. An absent price limit imposes no price restriction. These checks use current public information, not a hidden purchase-success label.

WebShop action functions. Action functions distinguish opening a strongly matching, affordable candidate from selecting a requested option that remains unmet. Conflicting or repeated selections are separated, as are inspection, purchase, backtracking, and result exploration. Table 5 illustrates these functional distinctions; sharing requires both the abstract state and action function to match.

Table 5: Examples of task-relative distinctions preserved by action abstraction.

Environment	Task-relative decision	Action function
ALFWorld	In Pick2, take a target from outside the destination.	Acquire an unplaced target
ALFWorld	In Pick2, take a target back out of the destination.	Retake a placed target
WebShop	Select a requested option that is not yet satisfied.	Satisfy an option
WebShop	Select a conflicting option or repeat an already satisfied choice.	Violate or repeat an option

Sharing boundary. WebShop excludes unresolved frontiers and action functions from global sharing, including open-ended search queries and results whose product titles and prices cannot be aligned reliably. Item subpages are also excluded because they generally offer only a return action. Parser and execution-validity checks exclude malformed or inadmissible commands without altering the pre-action state; excluded decisions retain access to trajectory or local credit. These gates determine whether sharing is allowed; Appendix A.5 describes how support and uncertainty scale the resulting estimate. The grouping interface and estimator are shared across environments, while task parsing, tracked facts, and functional mappings remain environment-specific.

A.5 Global memory and reliability

This section specifies the memory maintenance and reliability estimators used in Section 4.3. For each observed (z, u) pair, the memory maintains the moments in Equation 9, the sum of squared observation weights, and the last update. It also tracks decayed contributions from contexts recognized across updates, separately from the batch-local grouping c. For this support calculation, ALFWorld matches the task family and observation, whereas WebShop matches the task instance

and observation. This tracks repeated support across updates without changing local comparison groups.

Before a record is read or updated after Δ optimization steps, its moments and context weights are multiplied by $\lambda_{\mathrm{ema}}^{\Delta}$; the sum of squared weights is multiplied by $\lambda_{\mathrm{ema}}^{2\Delta}$. We set $\lambda_{\mathrm{ema}} = 2^{-1/H}$ for a half-life of H updates. A local context supplies new residual observations only when it contains at least two shareable occurrences with different abstract action functions. Repeated occurrences within each (c, z, u) are then averaged as in Equation 8, and each resulting mean adds one unit of weight. All eligible current-batch observations are written before the advantages for that update are queried.

Let $C_{z,u}(\chi)$ be the decayed contribution weight of context χ under this matching rule. The effective context support is

$$n_{\mathrm{eff}}(z,u) = \frac{\left(\sum_{\chi} C_{z,u}(\chi)\right)^2}{\sum_{\chi} C_{z,u}(\chi)^2}. \tag{15}$$

We first take the harmonic mean of the effective supports of all alternatives to u, then take the harmonic mean of that value and $n_{\mathrm{eff}}(z,u)$ to obtain $n_{\mathrm{cmp}}(z,u)$. This makes the support correction in Equation 11 sensitive to the less-supported side of the comparison.

To compute uncertainty, let $D_{z,u}$ denote the stored sum of squared observation weights. The weighted residual variance and the estimated variance of the action mean are

$$\widehat{v}(z,u) = \frac{\max\{M_{z,u} - S_{z,u}^2/W_{z,u}, 0\}}{W_{z,u} - D_{z,u}/W_{z,u}}, \qquad h(z,u) = \frac{\widehat{v}(z,u)}{\max\{n_{\mathrm{eff}}(z,u), 1\}}. \tag{16}$$

We use zero variance when its denominator is numerically negligible. For $K = |\mathcal{U}_z|$ recorded action alternatives, the squared standard error of the centered contrast is

$$\widehat{\mathrm{SE}}_{\mathrm{diff}}^2(z,u) = \left(\frac{K-1}{K}\right)^2 h(z,u) + \frac{1}{K^2} \sum_{u' \neq u} h(z,u'). \tag{17}$$

The dispersion $\widehat{\sigma}_z^2$ is the sample variance of the recorded action means, and η^2 supplies the additive floor in Equation 11.

A global group is active only when it has at least two action alternatives with non-negligible dispersion and effective context support of at least two. For this group-level support, a context's weight is the maximum of its weights across action alternatives, avoiding repeated counting within the group. A queried action must also have a memory record. After reliability correction, WebShop subtracts the current-batch mean over active decisions in each progress-state group; no further variance normalization is applied. The reported diagnostics and case values include this centering.

A.6 Training and interaction settings

We train for 160 policy updates, sampling 16 tasks and eight trajectories per task at each update, for a total of 128 trajectories. The policy learning rate is 10^{-6}, with clipping threshold $\epsilon_{\mathrm{clip}} = 0.2$ and KL coefficient $\beta = 0.01$. We use a discount factor of $\gamma = 0.95$ and format penalty $\lambda_{\mathrm{fmt}} = 0.1$.

The global memory uses an EMA half-life of eight updates, with support shrinkage $\lambda_{\mathrm{sup}} = 4$ and numerical floor $\eta = 0.05$. For advantage fusion, $\alpha_j = 2/3$ when both step estimators are active, 1 for local credit only, and 0 for task-progress credit only. If neither estimator is active, the update uses trajectory credit alone; we set $\lambda_{\mathrm{step}} = 1$ throughout.

Interactions are limited to 50 steps in ALFWorld and 30 in WebShop. Training and evaluation sample responses at temperatures 1.0 and 0.4, respectively. The policy retains up to two recent observation–action pairs and generates at most 512 tokens per response. After chat formatting, prompts are left-truncated to 2,048 tokens for ALFWorld and 4,096 for WebShop.

A.7 Computational profile

The computational measurements in Table 3 use four NVIDIA A100 40GB GPUs and cover all 160 ALFWorld policy updates. We average the logged advantage-computation and full-update times

across updates, excluding validation and checkpoint evaluation. Advantage computation includes local grouping, memory maintenance, and reliability estimation. Full-update time also includes rollout generation and policy optimization, so its variation reflects both computational overhead and the trajectory lengths produced by each policy.

A.8 PROMPT TEMPLATES

Both environments use the same prompt templates during training and evaluation. The environment prompt is passed as a single user message through the Qwen chat template, which adds the default system message and the assistant generation prefix. Each subsequent input includes the task, current observation, admissible actions, and recent history, subject to the limits in Appendix A.6. History contains actions submitted to the environment, not the model's reasoning text. This policy input is separate from the compact history state maintained by the credit estimator.

ALFWorld. The template below is used after the initial decision. At the initial decision, its opening sentence is "You are an expert agent operating in the ALFRED Embodied Environment." The task is already included in the initial observation, and the history and step counters are omitted. The admissible action list excludes `help`.

ALFWorld prompt template

You are an expert agent operating in the ALFRED Embodied Environment. Your task is to: `{task_description}`

Prior to this step, you have already taken `{step_count}` step(s). Below are the most recent `{history_length}` observations and the corresponding actions you took: `{action_history}`

You are now at step `{current_step}` and your current observation is: `{current_observation}`

Your admissible actions of the current situation are: [`{admissible_actions}`].

Now it's your turn to take an action.

You should first reason step-by-step about the current situation. This reasoning process MUST be enclosed within `<think> </think>` tags.

Once you've finished your reasoning, you should choose an admissible action for current step and present it within `<action> </action>` tags.

WebShop. The task is extracted from the instruction field, and available actions are constructed from the search interface and clickable page elements. The initial decision uses the same instructions below but omits the history sentence and replaces "You are now at step {current_step} and your current observation is:" with "Your current observation is:". History entries in both environments take the form `[Observation k: '...', Action k: '...']`.

WebShop prompt template

You are an expert autonomous agent operating in the WebShop e-commerce environment.

Your task is to: `{task_description}`.

Prior to this step, you have already taken `{step_count}` step(s). Below are the most recent `{history_length}` observations and the corresponding actions you took: `{action_history}`

You are now at step `{current_step}` and your current observation is: `{current_observation}`.

Your admissible actions of the current situation are: [`{available_actions}`].

Now it's your turn to take one action for the current step.

You should first reason step-by-step about the current situation, then think carefully which admissible action best advances the shopping goal. This reasoning process MUST be enclosed within `<think> </think>` tags.

Once you've finished your reasoning, you should choose an admissible action for current step and present it within `<action> </action>` tags.

B Additional Experimental Diagnostics

B.1 Data and statistical procedures

The credit analyses use 529,352 ALFWorld decisions and 162,121 WebShop decisions collected over 160 policy updates in each environment. Each decision is associated with its task, concrete context, task-progress state, action function, discounted return, and both credit estimates. Coverage is the proportion of decisions with a nonzero advantage from the corresponding estimator.

B.2 Cross-task consistency test

The test in Figure 3b uses 3,473 supported action contrasts. Tasks are partitioned into five disjoint folds, keeping all records from each task together. Four folds estimate context-balanced residual action means, and the held-out fold provides comparison estimates for the same progress groups. Contributions retain their EMA weights at update 160. Both estimates are centered over common action alternatives, requiring at least two alternatives and effective context support of at least two per action on each side. We compare the signs of nonzero contrasts and bin their agreement by the confidence multiplier. Away from the numerical floor, this multiplier is

$$\kappa(z,u) = \alpha_{\mathrm{sup}}(z,u)\frac{\widehat{\sigma}_z}{\sqrt{\widehat{\sigma}_z^2 + \widehat{\mathrm{SE}}_{\mathrm{diff}}^2(z,u) + \eta^2}}. \tag{18}$$

It measures shrinkage relative to the unscaled standardized contrast $\widetilde{q}(z,u)/\widehat{\sigma}_z$, not a probability of correctness. Numerically, we divide the magnitude of the reliability-scaled contrast by that of the unscaled contrast, flooring the latter at 10^{-6} and capping the ratio at one. All quantities are recomputed from the estimation folds; agreement is summarized in four bins of width 0.25. Repeating the procedure for each held-out fold produces the five columns in the figure.

B.3 Complete decision-credit decomposition

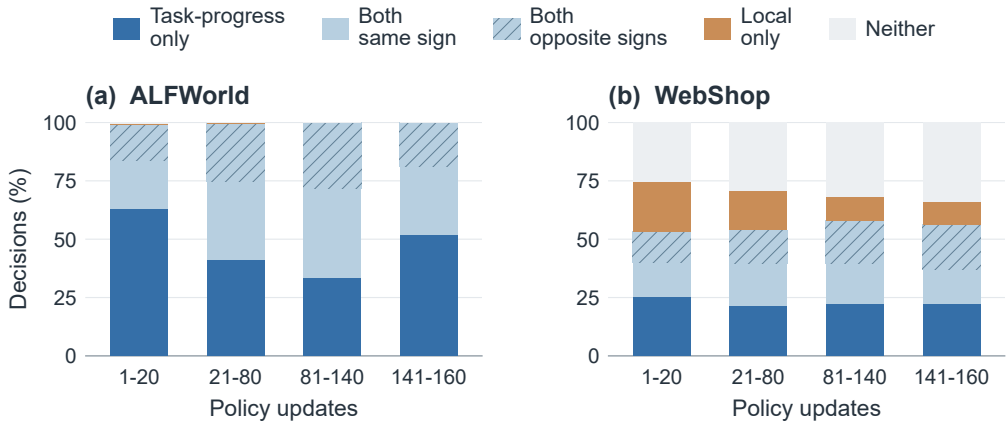

Figure 6: Local and task-progress credit across training phases. Bars partition all decisions by nonzero credit availability; hatching marks shared coverage with opposing advantage signs.

Figure 6 shows that broader coverage is a persistent contribution of task-progress credit, rather than an effect confined to early exploration. Its exclusive coverage remains substantial in every phase, while local-only credit accounts for less than 0.1% of ALFWorld decisions. In WebShop, task-progress-only coverage stays near 22% after the initial phase, even as local-only coverage decreases. Where both estimators are active, their signs can differ because they compare actions against different reference groups. The global branch therefore changes both the availability and the direction of decision-level supervision; the sign decomposition measures these differences, not which estimator is correct.

B.4 Performance across ALFWorld task types



Figure 7 shows BLPO's largest gain on Pick2, which requires retaining completed placement progress while handling a second object. Task-progress groups can credit useful intermediate steps through comparisons with similar stages in other tasks, even when subsequent execution fails. Look instead requires target possession and lighting conditions to hold jointly. This greater dependence on the immediate configuration may offer less scope for progress-based sharing, providing a possible explanation for BLPO's lower success than HGPO on Look.

Figure 7: Success by ALFWorld task type with Qwen2.5-1.5B at update 160, evaluated on the same 140 tasks as Figure 2. Axes scale linearly from 0% at the center to 100% at the boundary.

B.5 Training dynamics

Figure 8 reports training diagnostics over all 160 policy updates. Return is the mean undiscounted environment reward of the sampled rollout batch, before the format penalty. Entropy is averaged over output tokens, rather than executable action categories. BLPO increases return more rapidly during early and middle training, while the gap to HGPO narrows toward the end. Its entropy decreases earlier and approaches GiGPO's level without a sustained decline in return. Meanwhile, the logged KL estimate relative to the reference policy levels off, with late values between those of GiGPO and HGPO.

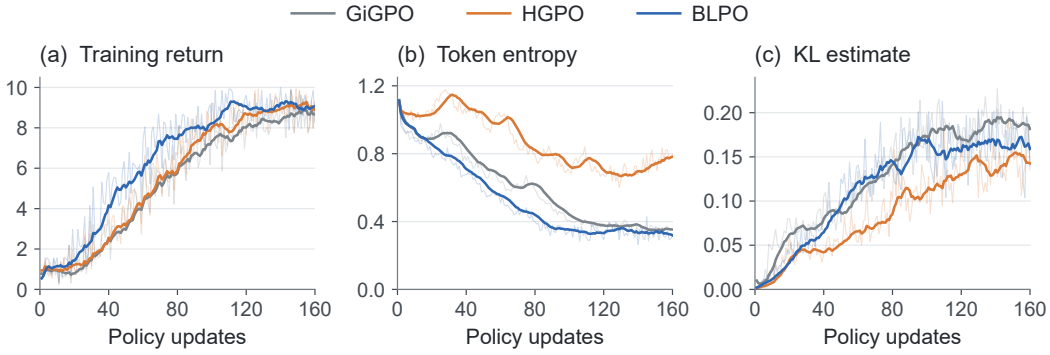

Figure 8: Training diagnostics on ALFWorld with Qwen2.5-1.5B, using one training run per method. Pale lines show per-update measurements; solid lines show 10-update trailing means.