DCGPO: DUAL-CONDITIONED GROUP POLICY OPTIMIZATION FOR LONG-HORIZON LLM AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Group-based reinforcement learning (RL) has become an important paradigm for training long-horizon LLM agents, enabling critic-free advantage estimation through relative comparisons within rollout groups. Recent methods improve credit granularity through step-level grouping, comparing actions taken from recurring environment states. However, state-conditioned grouping answers only which action is better under a matched state; it cannot address the complementary action-conditioned question: why does the same action yield different utility across historical contexts? We call this missing comparison dimension the cross-context incomparability of action utility. We propose DCGPO (Dual-Conditioned Group Policy Optimization), which integrates state-conditioned action comparison with action-conditioned cross-context comparison in a unified advantage estimator. For each action occurrence with cross-trajectory matches in a rollout group, DCGPO constructs a weighted reference return from matching occurrences in other trajectories and defines its action-axis relative advantage as the deviation of the current return from this reference. Because this advantage reflects the relative continuation outcome of the same action under an already-formed context, DCGPO propagates it backward, with distance-based discounting, to the preceding decisions that shaped that context rather than assigning it directly to the current action. It then removes components already explained by episode- and state-level advantages and combines episode-level trajectory quality, state-conditioned action advantage, and residual action-axis credit into a three-axis step-level relative advantage. Experiments on ALFWorld, WebShop, search-augmented question answering, and agentic safety show consistent gains across tasks and model scales, highlighting dual-conditioned comparison as an efficient complement to state-axis credit across diverse long-horizon agent settings. Our code is available at `https://anonymous.4open.science/r/DCGPO`.

1 INTRODUCTION

Large language models (LLMs) are increasingly deployed as autonomous agents (Liu et al., 2024; Trivedi et al., 2024) that continuously perceive, reason, and act in environments, demonstrating strong capabilities in web navigation (Yao et al., 2022; Deng et al., 2023; Lù et al., 2024; Qi et al., 2025), embodied interaction (Wang et al., 2023; Driess et al., 2023), and tool use (Schick et al., 2023; Qian et al., 2025; Qin et al., 2024). Training such long-horizon agents, however, poses a credit-assignment challenge: an agent may execute dozens of actions while receiving only a sparse success or failure signal at the end of an episode (Wang et al., 2025b; Feng et al., 2025; Zhou et al., 2024). Although this trajectory-level reward summarizes the outcome, individual decisions contribute unequally—some make substantive progress toward the goal, whereas others amount to ineffective exploration. This creates a granularity mismatch between trajectory-level feedback and step-level decisions (Ma et al., 2024; Feng et al., 2025). Accordingly, assigning sparse outcome feedback to individual decisions—*step-level credit assignment*—is central to training long-horizon LLM agents.

Existing approaches obtain step-level learning signals along two main directions. One trains process reward models or value functions to evaluate intermediate steps (Lightman et al., 2024; Wang et al., 2024; Schulman et al., 2017), but typically requires additional supervision, model training, or computation, and may generalize poorly across task domains. The other retains the critic-free

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.

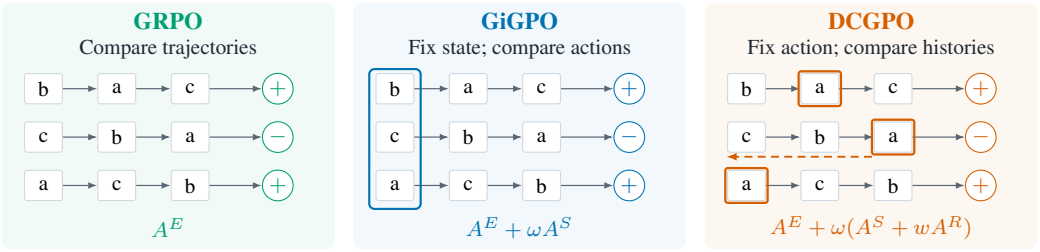


Figure 1: Three credit axes. GRPO compares trajectories; GiGPO compares actions from matched states (blue); DCGPO compares matched actions across histories (orange) and attributes their relative outcomes to preceding decisions that formed those histories (dashed arrow). This attribution is a credit-assignment assumption, not a causal estimate.

nature of group-based reinforcement learning by matching recurring environment states across rollouts and comparing the subsequent outcomes of actions taken from those states (Feng et al., 2025). We refer to this mechanism as *state-conditioned comparison*: it answers which action produces a better outcome under a matched state. Yet this comparison operates only along the state dimension and leaves a complementary action-conditioned question unexplored: why can the same semantic action yield different utility across historical contexts? In long-horizon tasks, the effectiveness of an action often depends on the environment configuration, acquired information, and completed subgoals established by preceding decisions. We refer to this missing comparison dimension as the *cross-context incomparability of action utility*.

A natural remedy is to construct an action-conditioned counterpart to state-conditioned comparison: hold action semantics fixed and compare the subsequent outcomes of matched action instances across different historical contexts. Crucially, the resulting signal should not be credited directly to the matched action. Because the action is controlled for in this comparison, variation in its outcome serves primarily as evidence about the context established before that action was taken. The current action cannot alter its pre-existing history; assigning the signal in place would therefore conflate action selection with context formation. Action-conditioned comparison should instead be coupled with history-directed credit propagation, using evidence revealed by later matched actions to train the preceding decisions that formed their contexts.

Motivated by this observation, we propose **DCGPO (Dual-Conditioned Group Policy Optimization)**, which unifies state-conditioned action comparison and action-conditioned cross-context comparison within a single advantage-estimation framework. As illustrated in Figure 1, DCGPO uses semantically matched action instances from other rollouts to construct a weighted reference return and defines an action-axis relative advantage as the deviation of the current instance's return from this reference. It then propagates this evidence backward, with temporal discounting, to the preceding decisions that formed the corresponding context. Because episode-, state-, and action-level signals are derived from the same rollout group, they may contain overlapping information. DCGPO therefore removes the component of the propagated action signal explained by the episode- and state-level advantages, retaining only its complementary residual credit. Finally, DCGPO combines episode-level trajectory quality, state-conditioned action advantage, and residual action-axis credit into a three-axis step-level advantage.

Our contributions are as follows:

- **Dual-conditioned step-level comparison.** We complement state-conditioned action comparison with action-conditioned cross-context comparison, capturing an additional dimension of step-level credit.
- **History-directed residual credit assignment.** DCGPO propagates action-axis evidence to the preceding decisions that formed the context and removes components already explained by episode- and state-level advantages.
- **Empirical validation across tasks and model scales.** Across ALFWorld, WebShop, search-augmented QA, and agentic safety, DCGPO improves aggregate performance across model scales using the same rollout budget.

2 PRELIMINARIES

We introduce notation for long-horizon agent interaction and review episode-level and state-conditioned group-relative advantages.

Long-Horizon Agent Interaction. Given a task $x \sim p_{\mathcal{X}}$, a policy π_θ interacts with an environment for at most T steps. At step t of rollout i, the agent observes state $s_t^{(i)}$ and samples an action $a_t^{(i)} \sim \pi_\theta(\cdot \mid h_t^{(i)}, x)$, where $h_t^{(i)} = (s_1^{(i)}, a_1^{(i)}, \ldots, a_{t-1}^{(i)}, s_t^{(i)})$ is the interaction history available to the policy. The environment returns reward $r_t^{(i)}$ and the next state $s_{t+1}^{(i)}$, producing trajectory $\tau_i = \{(s_t^{(i)}, a_t^{(i)}, r_t^{(i)})\}_{t=1}^{T_i}$ with $T_i \le T$. We distinguish the current state $s_t^{(i)}$, used for state matching, from the full policy context $h_t^{(i)}$.

Episode-Level Group-Relative Advantage. For a fixed task and initial condition, the policy samples N trajectories to form an episode group $\mathcal{G}^E = \{\tau_i\}_{i=1}^N$. Let $R(\tau_i) = \sum_{t=1}^{T_i} r_t^{(i)}$ denote the return of trajectory τ_i. Its episode-level advantage is

$$A^E(\tau_i) = \frac{R(\tau_i) - \mathrm{mean}_{j=1,\ldots,N} R(\tau_j)}{F_{\mathrm{norm}}\left(\{R(\tau_j)\}_{j=1}^N\right)}. \tag{1}$$

Here, F_{norm} is either the within-group standard deviation or 1, corresponding to standard-deviation normalization or mean centering only. This estimator requires no separately trained value function, but assigns the same advantage to every decision in a trajectory and therefore cannot distinguish their individual contributions.

State-Conditioned Step-Level Advantage. GiGPO (Feng et al., 2025) obtains finer-grained credit by grouping steps whose states match a shared anchor. We write $s_t^{(i)} \sim_S \tilde{s}$ when state $s_t^{(i)}$ matches anchor $\tilde{s}$, and let $\mathcal{G}^S(\tilde{s})$ index all matched steps. For action $a_t^{(i)}$, define its discounted continuation return as $R_t^{(i)} = \sum_{k=t}^{T_i} \gamma^{k-t} r_k^{(i)}$. The corresponding state-conditioned advantage is

$$A^S\left(a_t^{(i)}\right) = \frac{R_t^{(i)} - \mathrm{mean}_{(j,u)\in\mathcal{G}^S(\tilde{s})} R_u^{(j)}}{F_{\mathrm{norm}}\left(\left\{R_u^{(j)}\right\}_{(j,u)\in\mathcal{G}^S(\tilde{s})}\right)}. \tag{2}$$

This estimator compares the discounted continuation returns of actions taken from matched states, providing finer-grained credit than the episode-level estimator. The two advantages are combined as

$$A_{i,t}^{\mathrm{GiGPO}} = A^E(\tau_i) + \omega A^S\left(a_t^{(i)}\right), \tag{3}$$

where ω controls the relative scale of step- and episode-level credit. The resulting scalar is broadcast to the output tokens of $a_t^{(i)}$ and used for policy optimization.

3 METHOD

We present **DCGPO** (**D**ual-**C**onditioned **G**roup **P**olicy **O**ptimization), a multi-axis credit-assignment method for long-horizon LLM agents. State-conditioned grouping treats matched states as the conditioning variable and estimates the relative advantage of actions selected from those states. DCGPO introduces the complementary action-conditioned view: it treats matched action semantics as the conditioning variable and estimates how favorable the preceding history is for executing that action. The resulting action-axis advantage is propagated to the earlier decisions that formed the corresponding history and residualized against existing episode-level and state-conditioned credit. As illustrated in Figure 2, DCGPO consists of four stages: action-axis relative advantage estimation, history-directed credit propagation, residual-credit extraction, and multi-axis advantage fusion.

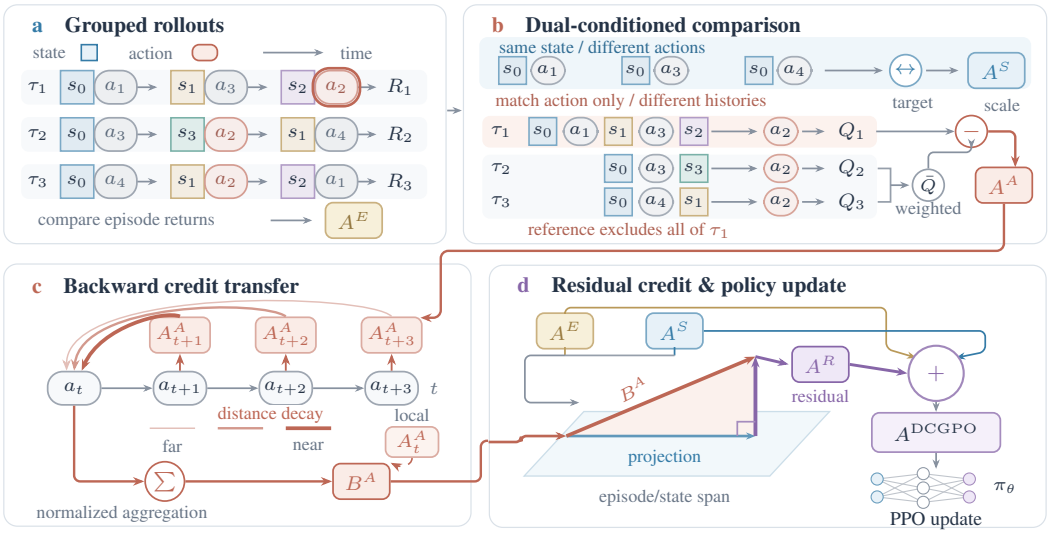


Figure 2: DCGPO compares matched actions across histories, transfers evidence to predecessors, removes fitted episode/state overlap, and fuses three credit channels.

3.1 Action-Axis Relative Advantage

State-conditioned grouping estimates which action is preferable under a matched state. DCGPO considers the dual question: under which history is a matched action more effective? We first normalize each textual action into its semantic operation and arguments, and write $a_t^{(i)} \sim_A a_u^{(j)}$ when two action instances satisfy the action-matching rule. This relation is determined solely by action semantics rather than the surrounding state or temporal position.

For action $a_t^{(i)}$, we compute the normalized discounted continuation return $Q_{i,t}^A = c_A^{-1} \sum_{k=t}^{T_i} \gamma_A^{k-t} r_k^{(i)}$, where γ_A discounts delayed outcomes and c_A aligns the action-axis scale with the environment reward.

The reference set for $a_t^{(i)}$ contains its matched instances from the other trajectories in the rollout group. In the trajectory-balanced setting, each eligible trajectory receives reference mass according to its strongest action match; its matched positions share that mass according to their action similarities. We also evaluate direct normalization over all matched occurrences. Both settings preserve action-level evidence while differing in how repeated matches from one trajectory contribute to the reference estimate (Appendix A).

Let $q_{i,t;j,u}$ denote the resulting normalized reference weight. The action-conditioned reference return and action-axis relative advantage are

$$\widehat{Q}_{i,t}^A = \sum_{\substack{j \neq i,\ u: \\ a_u^{(j)} \sim_A a_t^{(i)}}} q_{i,t;j,u} Q_{j,u}^A, \qquad A_{i,t}^A = Q_{i,t}^A - \widehat{Q}_{i,t}^A. \tag{4}$$

Thus, $A_{i,t}^A$ measures the relative outcome of a matched action under its current history. A positive value indicates that the preceding history is more favorable for the action than its cross-trajectory references, whereas a negative value indicates an unfavorable history–action pairing.

3.2 History-Directed Credit Propagation

The action-axis advantage is observed when a matched action is executed, whereas the history it evaluates has already been formed by earlier decisions. Assigning this advantage only to the current action would therefore conflate action selection with context formation. DCGPO instead propagates action-axis evidence backward along the trajectory to the preceding decisions responsible for forming the evaluated history.

For step (i,t), let $\mathcal{F}_{i,t} \subseteq \{1,\dots,K\}$ denote the valid future offsets within a propagation horizon K for which action-axis evidence is available. We combine the local action-axis advantage with a normalized, distance-discounted aggregation of future advantages:

$$C^A_{i,t} = \alpha A^A_{i,t} + (1-\alpha)\frac{\sum_{\ell\in\mathcal{F}_{i,t}} \gamma_H^{\ell-1} A^A_{i,t+\ell}}{\sum_{\ell\in\mathcal{F}_{i,t}} \gamma_H^{\ell-1}}. \tag{5}$$

Here, α balances local and propagated evidence, while γ_H assigns greater influence to temporally closer actions. For positions without a supported successor, the propagated component vanishes and only the local component remains.

The resulting credit $C^A_{i,t}$ attributes the relative effectiveness revealed by later matched actions to earlier navigation, observation, retrieval, and planning decisions that shaped their execution contexts.

3.3 Residual-Credit Extraction

Episode-level, state-conditioned, and propagated action-conditioned advantages are derived from the same rollout outcomes and may therefore encode overlapping information. DCGPO extracts the complementary component of the action axis by removing the part of $C^A_{i,t}$ explained by the existing two credit channels.

Let Ω_A denote the positions with valid cross-trajectory action references. Propagation may also supply credit to earlier positions outside Ω_A. We therefore define the transferred support $\Omega_C = \{(i,t) : |C^A_{i,t}| > \varepsilon\}$, restricted to finite values, and set $\mathbf{x}_{i,t} = [1,\, A^E(\tau_i),\, A^S(a^{(i)}_t)]^\top$. We estimate the projection coefficients on Ω_C and retain the residual only on that support:

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{(i,t)\in\Omega_C} \left(C^A_{i,t} - \mathbf{x}^\top_{i,t}\boldsymbol{\beta}\right)^2, \qquad A^R_{i,t} = \mathbf{1}\{(i,t)\in\Omega_C\}\left(C^A_{i,t} - \mathbf{x}^\top_{i,t}\widehat{\boldsymbol{\beta}}\right). \tag{6}$$

This projection removes action-axis information already represented by overall trajectory quality or matched-state action quality. Because Ω_C is defined after propagation, the residual can retain credit at an earlier step with no action match of its own.

3.4 Multi-Axis Advantage Fusion

DCGPO combines episode-level trajectory quality, state-conditioned action advantage, and residual action-axis credit into the final step-level advantage:

$$A^{\mathrm{DCGPO}}_{i,t} = A^E(\tau_i) + \omega_S A^S(a^{(i)}_t) + \omega_A A^R_{i,t}. \tag{7}$$

The three terms respectively capture the relative quality of the complete trajectory, the relative quality of an action under a matched state, and the complementary contribution of earlier decisions to forming favorable histories for later actions. The resulting scalar is broadcast to the output tokens of $a^{(i)}_t$ and optimized using the same clipped policy objective as the underlying group-based method.

All additional computations take place after rollout collection. Action-conditioned comparison reuses actions already contained in the rollout group, history-directed propagation operates within each sampled trajectory, and residualization is performed directly in the advantage space. DCGPO therefore introduces no additional environment interactions or auxiliary value model.

4 Experiments

We evaluate DCGPO across four agent settings, analyze its action-credit components, and examine the observed training cost.

4.1 Experimental Setup

We describe the models, tasks, metrics, and baselines.

Table 1: Performance on ALFWorld and WebShop. ALFWorld reports success rate (%) by task and overall; WebShop reports task score and success rate (%). Bold and underline mark the two highest means in each column; green highlights DCGPO.

		ALFWorld success (%)							WebShop	
Regime	Method	Pick	Look	Clean	Heat	Cool	Pick2	All	Score	Succ. (%)
Closed-source prompting										
Prompting	GPT-4o	75.3	60.8	31.2	56.7	21.6	49.8	48.0	31.8	23.7
Prompting	Gemini-2.5-Pro	92.8	63.3	62.1	69.0	26.6	58.7	60.3	42.5	35.9
Qwen2.5-1.5B										
Prompting	Qwen2.5	5.9	5.5	3.3	9.7	4.2	0.0	4.1	23.1	5.2
Prompting	ReAct	17.4	20.5	15.7	6.2	7.7	2.0	12.8	40.1	11.3
Prompting	Reflexion	35.3	22.2	21.7	13.6	19.4	3.7	21.8	55.8	21.9
RL	PPO (with critic)	64.8±3.5	40.5±6.9	57.1±4.9	60.6±6.6	46.4±4.0	47.4±1.9	54.4±3.1	73.8±3.0	51.5±2.9
RL	RLOO	88.3±3.0	52.8±8.6	71.0±5.9	62.8±8.7	66.4±5.5	56.9±4.7	69.7±2.5	73.9±5.6	52.1±6.7
RL	GRPO	85.3±1.5	53.7±8.0	84.5±6.8	78.2±7.9	59.7±5.0	53.5±5.6	72.8±3.6	75.8±3.5	56.8±3.8
RL	GiGPO	83.3±4.7	63.9±11.7	91.3±0.0	**90.0±4.7**	72.0±17.0	79.4±4.2	80.1±5.0	82.8±4.6	68.0±1.1
RL	**DCGPO**	**100.0±0.0**	**85.7±0.0**	**93.2±9.6**	**90.0±0.0**	**78.0±2.8**	**86.7±9.4**	**89.8±2.2**	**85.7±0.1**	**71.7±2.5**
Qwen2.5-7B										
Prompting	Qwen2.5	33.4	21.6	19.3	6.9	2.8	3.2	14.8	26.4	7.8
Prompting	ReAct	48.5	35.4	34.3	13.2	18.2	17.6	31.2	46.2	19.5
Prompting	Reflexion	62.0	41.6	44.9	30.9	36.3	23.8	42.7	58.1	28.8
RL	PPO (with critic)	92.3±4.0	64.0±8.4	92.5±2.4	89.5±7.0	80.3±2.0	68.8±8.3	80.4±2.7	81.4±3.1	68.7±5.1
RL	RLOO	87.6±4.3	78.2±8.3	87.3±5.8	81.3±7.6	71.9±5.2	48.9±8.4	75.5±4.6	80.3±3.2	65.7±4.0
RL	GRPO	90.8±5.1	66.1±6.7	89.3±5.4	74.7±6.9	72.5±5.4	64.7±7.3	77.6±5.2	79.3±2.8	66.1±3.7
RL	GiGPO	**97.1±0.0**	61.1±7.9	90.9±12.9	**96.2±5.4**	80.0±5.7	**95.8±5.9**	89.8±0.0	85.9	77.8±1.7
RL	**DCGPO**	**97.1±4.2**	**83.3±7.9**	**97.8±3.1**	88.9±0.0	**91.3±6.1**	92.9±3.4	**93.4±0.6**	**87.0**	**78.7±0.8**

Table 2: Search-augmented QA success rate (%) (Kwiatkowski et al., 2019; Joshi et al., 2017; Mallen et al., 2023; Yang et al., 2018; Ho et al., 2020; Trivedi et al., 2022). GiGPO and DCGPO share the seed and evaluation protocol at each scale; R1-Instruct, Search-R1, and ZeroSearch follow Feng et al. (2025). Green highlights DCGPO; bold marks column bests.

Method	NQ	Trivia	PopQA	Hotpot	2Wiki	MuSiQue	Bamboogle	Overall
Qwen2.5-3B-Instruct								
R1-Instruct	27.0	53.7	19.9	23.7	29.2	7.2	29.3	27.1
Search-R1	34.1	54.5	37.8	32.4	31.9	10.3	26.4	32.5
ZeroSearch	41.4	57.4	44.8	27.4	30.0	9.8	11.1	31.7
GiGPO	44.5	62.0	46.4	39.5	39.3	14.3	65.7	45.5
DCGPO	**45.7**	**62.7**	**47.4**	**39.9**	**39.6**	**16.3**	**66.5**	**46.5**
Qwen2.5-7B-Instruct								
R1-Instruct	21.0	44.9	17.1	20.8	27.5	6.0	19.2	22.4
Search-R1	39.3	61.0	39.7	37.0	40.1	14.6	36.8	38.5
ZeroSearch	43.6	61.8	51.5	34.6	35.2	18.4	27.8	39.1
GiGPO	46.8	65.9	44.8	43.6	**43.4**	17.5	**69.8**	48.0
DCGPO	**47.2**	**66.2**	**48.2**	**43.7**	42.9	**21.3**	69.4	**49.0**

Backbone models. Across our experiments, we use the 1.5B-, 3B-, and 7B-parameter variants of Qwen2.5-Instruct (Yang et al., 2024) to examine the consistency of DCGPO across model capacities.

Benchmarks and evaluation metrics. We evaluate DCGPO in four long-horizon agent settings. ALFWorld (Shridhar et al., 2021) evaluates household navigation and object manipulation, reporting success rates for individual task categories and overall task completion. WebShop (Yao et al., 2022) evaluates instruction-guided product search and purchasing, reporting average task score and success rate. Search covers seven retrieval-augmented question-answering benchmarks and reports answer success rate. Agentic Safety (Liu et al., 2026) evaluates task utility under clean interactions, behavioral safety under malicious user queries, and joint utility–safety performance under environment-injection attacks.

Baselines. For ALFWorld and WebShop, we compare DCGPO against three classes of baselines: (1) closed-source LLM agents, including GPT-4o and Gemini-2.5-Pro; (2) prompting-based agents, including Qwen2.5, ReAct (Yao et al., 2023), and Reflexion (Shinn et al., 2023); and (3) RL training methods, including the critic-based PPO (Schulman et al., 2017) and the critic-free RLOO (Ahmadian et al., 2024), GRPO (Shao et al., 2024), and GiGPO (Feng et al., 2025). For Search, we include

Table 3: Agentic Safety performance. Overall is the aggregate reward-based score; Clean, Attacked, and Query-target are evaluation settings. Green highlights DCGPO; bold marks column bests.

	Reward-based test score (%)				
Method	Overall	Clean	Attacked	Query-target	Success (%)
Base model	23.4	10.9	16.3	46.9	27.0
GRPO	44.8	18.1	**40.1**	83.9	**40.5**
GiGPO	42.4	5.8	24.8	**98.0**	33.4
DCGPO	**53.6**	**44.2**	24.5	**98.0**	**40.5**

R1-Instruct, Search-R1 (Jin et al., 2025), ZeroSearch (Sun et al., 2025), and GiGPO. For Agentic Safety, we compare against the pretrained backbone without RL fine-tuning, GRPO, and GiGPO.

4.2 Main Results

To evaluate the effectiveness and scalability of dual-conditioned credit assignment, we train Qwen2.5-Instruct models at the 1.5B and 7B scales on ALFWorld and WebShop and compare DCGPO against representative baselines. Table 1 reports the detailed and aggregate results. Two main observations emerge.

Observation 1. Dual-conditioned credit delivers superior performance across agent environments. At the 1.5B scale, DCGPO achieves the best aggregate performance among all compared methods on both ALFWorld and WebShop. It improves ALFWorld success from the previous best of 80.1% to 89.8%. On WebShop, DCGPO raises the task score from 82.8 to 85.7 and the success rate from 68.0% to 71.7%. These gains span embodied interaction and web navigation and hold for both graded task quality and binary task completion. The consistent improvements indicate that action-conditioned cross-context comparison provides informative step-level credit beyond state-conditioned grouping, enabling more effective optimization across different long-horizon interaction patterns.

Observation 2. The benefits of DCGPO persist as model capacity scales. The performance advantage of DCGPO extends from 1.5B to 7B. With the larger backbone, DCGPO again achieves the best aggregate results on both environments, reaching a success rate of 93.4% on ALFWorld and a task score of 87.0 with a success rate of 78.7% on WebShop. Together, the results across both model scales show that dual-conditioned credit assignment remains effective as the underlying policy becomes stronger. Action-conditioned cross-context comparison thus provides a scalable complement to state-conditioned credit assignment for long-horizon LLM agents.

4.3 Learning Dynamics and Credit Analysis

Observation 3. DCGPO enables more sample-efficient policy optimization. We examine WebShop validation performance over training. Figure 3(a) compares the WebShop learning dynamics of DCGPO and GiGPO. Over steps 0–150, DCGPO improves normalized AUC by 3.2 points in success rate and 5.8 points in task score. Under the same rollout budget, these gains indicate that dual-conditioned credit extracts more informative supervision from collected trajectories, enabling more effective policy improvement throughout training.

4.4 Generalization beyond the Primary Environments

We test DCGPO on Search and Agentic Safety.

Observation 4. DCGPO generalizes to search-augmented question answering. Table 2 compares the methods on seven retrieval-augmented question-answering benchmarks at the 3B and 7B scales. Within each scale, GiGPO and DCGPO use the same training seed and evaluation protocol. DCGPO improves aggregate success by 1.0 percentage point at both scales, from 45.5% to 46.5% at 3B and from 48.0% to 49.0% at 7B. It leads on all seven benchmarks at 3B and on five of seven at 7B, with the largest 7B improvement on MuSiQue. The consistent aggregate gains at both model scales indicate that action-conditioned credit remains complementary to state-conditioned advantage estimation in retrieval-intensive, multi-step decision making.

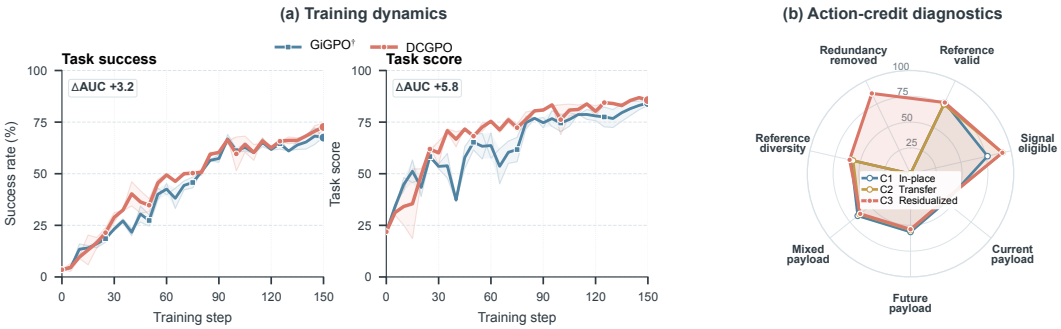


Figure 3: Training dynamics and action-credit analysis. (a) Validation curves of GiGPO and DCGPO. (b) Action-credit diagnostics for no transfer (C1), transfer (C2), and residualized transfer (C3). Reference valid and signal eligible are step fractions; current, future, and mixed payloads are nonzero candidate-signal fractions. Reference diversity is trajectory ESS divided by seven; redundancy removed is the transferred-block variance reduction under residualization (zero means not applied). The seven axes are not a composite score.

Observation 5. DCGPO improves joint utility and safety. Table 3 evaluates task utility and behavioral safety under three interaction regimes: benign interactions (Clean), malicious user queries (Query), and environment-level attacks (Attacked). DCGPO achieves the highest aggregate reward score (53.6%) and Clean score (44.2%), matches the best Query-target score (98.0%), and ties the highest task success rate (40.5%). The gains are concentrated where prior credit assignment is weakest: on benign interactions, where utility and safety must be traded off within a single rollout, DCGPO leads by the largest margin, indicating that action-conditioned credit recovers utility that state-level grouping attributes to the surrounding trajectory rather than to the decisive step. Query-level safety is preserved at parity, and the aggregate advantage persists under attacks, where no method's fixed defense dominates. These results show that action-conditioned credit can improve the joint objective while preserving strong query-level safety.

4.5 Component Analysis

Observation 6. Temporal propagation expands credit support, while residualization refines the signal. We compare action-credit variants to assess transfer and residualization. Figure 3(b) compares three action-credit designs. From in-place assignment (C1) to temporal propagation (C2), the proportion of steps eligible for action credit rises from 76.5% to 91.3%, indicating that later matched actions can supply evidence to preceding decisions. Residualized propagation (C3) retains this coverage while reducing the transferred block's variance attributable to episode- and state-level credit. In the separate WebShop outcome comparison of Figure 4(a), C3 achieves the highest displayed success rate and task score. Together, the diagnostics and outcomes show how propagation broadens credit support and residualization shapes the transferred signal.

4.6 Training Cost

Observation 7. DCGPO maintains a low per-step training cost. We compare observed per-step training costs across methods. Figure 4(b) shows that a representative residual DCGPO run (M03) takes 81.5 seconds per training step, compared with 70.5 seconds for GiGPO and 157.8 seconds for GRPO in the profiled runs. Rollout generation and environment interaction account for most of each bar. DCGPO obtains its additional action-axis credit by reusing the sampled trajectories, requiring neither a critic nor an extra model call. Thus, richer credit assignment fits within the existing critic-free training loop at a modest observed per-step cost.

5 Related Work

RL for long-horizon LLM agents. Prompting-based agents, such as ReAct (Yao et al., 2023) and Reflexion (Shinn et al., 2023), interleave reasoning and environment interaction without updating the policy. Reinforcement learning instead improves the policy from sampled experience (Bai et al.,

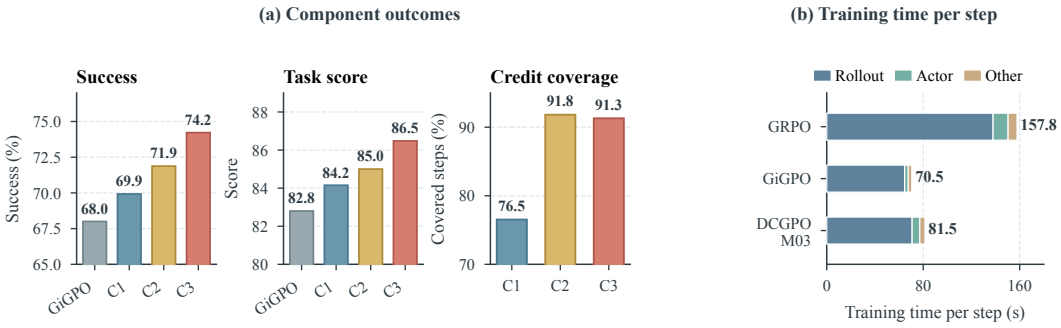


Figure 4: Component ablation and training efficiency. **(a)** WebShop-1.5B: GiGPO is the state-conditioned baseline; C1 assigns action credit in place, C2 transfers it to preceding decisions, and C3 further residualizes it. Coverage is the percentage of steps eligible for action-axis credit. **(b)** Mean training-only time per step across methods.

2024; Zhou et al., 2024; Chen et al., 2025; Wang et al., 2025a; Shi et al., 2025; Luo et al., 2025). Actor–critic methods such as PPO (Schulman et al., 2017; 2015; Ouyang et al., 2022) estimate a value function, whereas critic-free methods such as RLOO (Ahmadian et al., 2024), GRPO (Shao et al., 2024; DeepSeek-AI et al., 2025; Liu et al., 2025), and DAPO (Yu et al., 2025) compare rollouts from the same task. Applied to multi-step interactions (Wang et al., 2025b), an episode-level comparison assigns the same advantage to decisions with different effects on the outcome. This motivates estimators that retain critic-free training while resolving credit at the level of individual steps.

Step-level credit assignment. Recent critic-free methods exploit structure shared across rollouts to refine episode-level attribution. GiGPO (Feng et al., 2025) groups decisions made at recurring states and compares their continuation returns, yielding step-level credit conditioned on the state. DCGPO complements this comparison by matching actions across different histories and propagating their relative outcome evidence to the preceding decisions that formed those histories. Residualization then removes the component already explained by episode- and state-level advantages. The two methods thus differ in both what is compared and where the resulting credit is assigned, while reusing the sampled rollouts.

6 Limitations

DCGPO depends on reusable action matches: paraphrases can fragment groups, while coarse matching can conflate distinct arguments. The transfer horizon trades coverage against specificity. In agentic safety, higher aggregate reward accompanies lower Attacked performance than GRPO, and fixed attacks do not establish adaptive robustness. Furthermore, while the action axis improves aggregate metrics, *how* it improves them remains underexplored: the relative contributions of cross-history comparison and residualized transfer, and how each interacts with group composition, deserve systematic study. We leave a mechanistic account of the action axis to future work.

7 Conclusion

DCGPO augments state-conditioned grouping with an action axis that compares continuation outcomes across histories. It transfers this evidence to preceding decisions and removes overlap with episode- and state-level credit, using existing rollouts. Experiments show higher aggregate metrics across four agent settings; component diagnostics support residualized transfer, while safety results expose a split-dependent trade-off. The new limitation direction—a mechanistic understanding of when and why the action axis helps—points beyond aggregate gains toward interpretable step-level credit. These findings motivate using both actions and their historical contexts to construct step-level training signals.

ETHICS STATEMENT

DCGPO improves credit assignment for interactive agents, a capability that could be used in both beneficial and harmful applications. Our Agentic Safety evaluation uses fixed tasks and automated reward-based judgments; its results do not establish robustness to adaptive attacks or guarantee safe deployment. Applications in consequential settings require independent safety assessment and appropriate human oversight.

AI USE STATEMENT

In this work, we used generative AI tools solely to improve the readability and language of text written by the authors. We did not use generative AI tools to generate research ideas, formulate mathematical claims or proofs, design the methodology or experiments, implement the method, generate or analyze data, interpret results, or draw scientific conclusions. All AI-assisted edits were reviewed and verified by the authors. The authors take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

We release the complete codebase, configuration files, and evaluation scripts at `https://anonymous.4open.science/r/DCGPO`. All experiments use public Qwen2.5-Instruct backbones (1.5B/3B/7B) and standard public benchmarks under their official protocols. All hyperparameters of DCGPO (γ_A, γ_H, α, K, ω_S, ω_A, ε) are specified in the main text and appendices, and baselines follow their original implementations. Comparisons share identical seeds, rollout budgets, and evaluation pipelines, and no auxiliary value model is required.

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 Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024. doi:10.18653/v1/2024.acl-long.662; arXiv:2402.14740.

Hao Bai, Yifei Zhou, Mert Cemri, Jiayi Pan, Alane Suhr, Sergey Levine, and Aviral Kumar. DigiRL: Training in-the-wild device-control agents with autonomous reinforcement learning. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 2024. doi:10.52202/079017-0397; arXiv:2406.11896.

Kevin Chen, Marco Cusumano-Towner, Brody Huval, Aleksei Petrenko, Jackson Hamburger, Vladlen Koltun, and Philipp Krähenbühl. Reinforcement learning for long-horizon interactive LLM agents. *arXiv preprint arXiv:2502.01600*, 2025.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Samuel Stevens, Boshi Wang, Huan Sun, and Yu Su. Mind2Web: Towards a generalist agent for the web. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 2023. Spotlight; doi:10.52202/075280-1220; arXiv:2306.06070.

Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. PaLM-E: An embodied multimodal language model. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2023. Poster; arXiv:2303.03378.

Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. Group-in-group policy optimization for LLM agent training. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 2025. poster; doi:10.52202/085713-1544; arXiv:2505.10978.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pp. 6609–6625, 2020. doi: 10.18653/v1/2020.coling-main.580.

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. In *Proceedings of the Conference on Language Modeling (COLM)*, 2025. arXiv:2503.09516.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2017. doi:10.18653/v1/p17-1147; arXiv:1705.03551.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019. doi: 10.1162/tacl_a_00276.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024. poster; arXiv:2305.20050.

Dongrui Liu, Yu Li, Zhonghao Yang, Peng Wang, Guanxu Chen, Yuejin Xie, Qinghua Mao, Wanying Qu, Yanxu Zhu, Tianyi Zhou, Leitao Yuan, Zhijie Zheng, Qihao Lin, Yimin Wang, Haoyu Luo, Shuai Shao, Chen Qian, Qingyu Liu, Ling Tang, Ruiyang Qin, Qihan Ren, Junxiao Yang, Kun Wang, Zhiheng Xi, Linfeng Zhang, Ranjie Duan, Bo Zhang, Wenjie Wang, Wen Shen, Qiaosheng Zhang, Yan Teng, Chaochao Lu, Rui Mei, Man Li, Jialing Tao, Xi Lin, Tianhang Zheng, Yong Liu, Quanshi Zhang, Lei Zhu, Xingjun Ma, Junhua Liu, Hui Xue, Xiaoxiang Zuo, Xiangnan He, Chao Shen, Xianglong Liu, Minlie Huang, Jing Shao, and Xia Hu. AgentDoG 1.5: A lightweight and scalable alignment framework for AI agent safety and security. *arXiv preprint arXiv:2605.29801*, 2026.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. AgentBench: Evaluating LLMs as agents. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024. poster; arXiv:2308.03688.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding r1-zero-like training: A critical perspective. In *Proceedings of the Conference on Language Modeling (COLM)*, 2025. arXiv:2503.20783.

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

Xing Han Lù, Zdeněk Kasner, and Siva Reddy. WebLINX: Real-world website navigation with multi-turn dialogue. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024. Spotlight; arXiv:2402.05930.

Chang Ma, Junlei Zhang, Zhihao Zhu, Cheng Yang, Yujiu Yang, Yaohui Jin, Zhenzhong Lan, Lingpeng Kong, and Junxian He. AgentBoard: An analytical evaluation board of multi-turn LLM agents. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 2024. Oral; doi:10.52202/079017-2365; arXiv:2401.13178.

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 Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023. doi:10.18653/v1/2023.acl-long.546; arXiv:2212.10511.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 2022. doi:10.52202/068431-2011; arXiv:2203.02155.

Zehan Qi, Xiao Liu, Iat Long Iong, Hanyu Lai, Xueqiao Sun, Wenyi Zhao, Yu Yang, Xinyue Yang, Jiadai Sun, Shuntian Yao, Tianjie Zhang, Wei Xu, Jie Tang, and Yuxiao Dong. WebRL: Training LLM web agents via self-evolving online curriculum reinforcement learning. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025. Poster; arXiv:2411.02337.

Cheng Qian, Emre Can Acikgoz, Qi He, Hongru Wang, Xiusi Chen, Dilek Hakkani-Tür, Gokhan Tur, and Heng Ji. ToolRL: Reward is all tool learning needs. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 2025. poster; doi:10.52202/085713-3524; arXiv:2504.13958.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, Sihan Zhao, Lauren Hong, Runchu Tian, Ruobing Xie, Jie Zhou, Mark Gerstein, Dahai Li, Zhiyuan Liu, and Maosong Sun. ToolLLM: Facilitating large language models to master 16000+ real-world APIs. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024. spotlight; arXiv:2307.16789.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 2023. oral; doi:10.52202/075280-2997; arXiv:2302.04761.

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

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

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

Yucheng Shi, Wenhao Yu, Zaitang Li, Yonglin Wang, Hongming Zhang, Ninghao Liu, Haitao Mi, and Dong Yu. MobileGUI-RL: Advancing mobile GUI agent through reinforcement learning in online environment. *arXiv preprint arXiv:2507.05720*, 2025.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 2023. poster; doi:10.52202/075280-0377; arXiv:2303.11366.

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`. arXiv:2010.03768.

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.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022. doi: 10.1162/tacl_a_00475.

Harsh Trivedi, Tushar Khot, Mareike Hartmann, Ruskin Manku, Vinty Dong, Edward Li, Shashank Gupta, Ashish Sabharwal, and Niranjan Balasubramanian. AppWorld: A controllable world of apps and people for benchmarking interactive coding agents. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024. doi:10.18653/v1/2024.acl-long.850; arXiv:2407.18901.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*, 2023.

Hongru Wang, Cheng Qian, Wanjun Zhong, Xiusi Chen, Jiahao Qiu, Shijue Huang, Bowen Jin, Mengdi Wang, Kam-Fai Wong, and Heng Ji. Acting less is reasoning more! teaching model to act efficiently. *arXiv preprint arXiv:2504.14870*, 2025a.

Peiyi Wang, Lei Li, Zhihong Shao, R. X. Xu, Damai Dai, Yifei Li, Deli Chen, Y. Wu, and Zhifang Sui. Math-shepherd: Verify and reinforce LLMs step-by-step without human annotations. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024. doi:10.18653/v1/2024.acl-long.510; arXiv:2312.08935.

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

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh Jing Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, Yitao Liu, Yiheng Xu, Shuyan Zhou, Silvio Savarese, Caiming Xiong, Victor Zhong, and Tao Yu. OSWorld: Benchmarking multimodal agents for open-ended tasks in real computer environments. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 2024. Poster; doi:10.52202/079017-1650; arXiv:2404.07972.

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, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018. doi:10.18653/v1/d18-1259; arXiv:1809.09600.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. WebShop: Towards scalable real-world web interaction with grounded language agents. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 2022. URL `https://arxiv.org/abs/2207.01206`. arXiv:2207.01206.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023. URL `https://dblp.org/rec/conf/iclr/YaoZYDSN023.html`. arXiv:2210.03629.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. DAPO: An open-source LLM reinforcement learning system at scale. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 2025. poster; doi:10.52202/085713-3775; arXiv:2503.14476.

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 International Conference on Machine Learning (ICML)*, 2024. Poster; arXiv:2402.19446.

Appendix

A Implementation Details

Reference hyperparameters. Table A.1 summarizes a reference configuration for DCGPO, covering rollout grouping, action-return estimation, history-directed propagation, and residual-credit scaling.

Table A.1: Reference configuration for DCGPO.

Parameter	Value
Rollouts per task N	8
Action-return discount γ_A	0.95
History-propagation horizon K	8
History-propagation discount γ_H	0.95
Current-action mixture α	0.25
State-advantage weight ω_S	1.0
Action-residual weight ω_A	0.5
Action reference	Leave-one-trajectory-out
Reference normalization	Trajectory-balanced
Trajectory mass cap	0.15
Action-credit shrinkage limit	0.10
Linear warm-up	30 updates
Checkpoint selection	Final training checkpoint

Action-conditioned comparison requires a precise definition of which actions can be matched and how their outcomes form a reference return. This section specifies both operations using the notation introduced in Section 3.

Action representation and matching. Each textual action is mapped to an operation and its arguments. JSON tool calls supply the tool name and argument object; tagged, bracketed, and plain-text actions supply an action tag or leading verb and its payload. Text is lowercased and whitespace-normalized, structured arguments use a stable key order, and reasoning tags are ignored. This representation separates the action being compared from the history in which it occurs.

ALFWorld, WebShop, and Search use exact matching of the normalized operation–argument pair. Agentic Safety uses *soft action matching*: TF–IDF cosine similarity between action strings, restricted to pairs with the same normalized operation type. For a permitted pair with cosine similarity $c_{i,t;j,u}$, the nonnegative match weight is $g_{i,t;j,u} = \exp[-(1 - c_{i,t;j,u})/0.1]$; a type mismatch receives zero weight. The vocabulary is fitted within the batch and capped at $1{,}024$ features. A target is eligible only if another trajectory contains an action with cosine similarity at least 0.35; this is a reference-validity condition, not a cutoff on each soft weight.

Reference normalization. For a target occurrence (i, t), let $g_{i,t;j,u} \geq 0$ be its match weight with occurrence (j, u); exclude its own trajectory by setting $g_{i,t;i,u} = 0$. Define $\mathcal{J}_{i,t} = \{j \neq i : \sum_u g_{i,t;j,u} > 0\}$ and $m_j = \max_u g_{i,t;j,u}$. A trajectory-balanced reference gives each eligible trajectory one similarity-weighted vote and distributes that vote among its matched positions:

$$q_{i,t;j,u} = \frac{m_j}{\sum_{k \in \mathcal{J}_{i,t}} m_k} \frac{g_{i,t;j,u}}{\sum_v g_{i,t;j,v}}, \quad j \in \mathcal{J}_{i,t}. \tag{8}$$

This construction retains every matching action while preventing repeated occurrences in one trajectory from multiplying its vote. A direct occurrence-weighted variant instead distributes the total reference mass over individual matched occurrences without an intermediate trajectory vote. In either variant, the reference weights sum to one when positive reference mass exists. With no cross-trajectory match, the action-axis reference is unavailable.

B Policy Optimization Objective

DCGPO changes the credit assigned to a decision without changing the policy-optimization objective. The scalar advantage $A_{i,t}$ from Eq. (2) or Eq. (7) is shared by the output tokens of action $a_t^{(i)}$. With $Z = \sum_{i=1}^{N} \sum_{t=1}^{T_i} |a_t^{(i)}|$ denoting the number of action tokens in a batch, the clipped surrogate is

$$\mathcal{J}(\theta) = \mathbb{E}\Bigg[\frac{1}{Z} \sum_{i=1}^{N} \sum_{t=1}^{T_i} \sum_{l=1}^{|a_t^{(i)}|} \min\left\{\rho_{i,t,l} A_{i,t},\ \mathrm{clip}(\rho_{i,t,l}, 1-\epsilon, 1+\epsilon) A_{i,t}\right\}\Bigg] - \beta_{\mathrm{KL}}\, \mathbb{E}\big[D_{\mathrm{KL}}(\pi_\theta \| \pi_{\mathrm{ref}})\big]. \tag{9}$$

Here $\rho_{i,t,l}$ is the new-to-old policy probability ratio for token l, conditioned on the same history and preceding action tokens; ϵ is the clipping threshold, π_{ref} the reference policy, and β_{KL} its regularization weight. The sampled actions and policy ratios are unchanged by the advantage estimator; DCGPO changes how their step-level outcomes are weighted.

C Advantage Estimation Procedure

Algorithm C.1 summarizes the residualized DCGPO estimator. Episode- and state-conditioned advantages are estimated within task-wise rollout groups. Action references are drawn from the same group but exclude the target trajectory; temporal aggregation then proceeds within each individual trajectory. Reference weights may be normalized over individual occurrences or balanced across reference trajectories. The transferred action credit is projected against the episode- and state-conditioned signals over the advantage-computation batch, and its residual enters the final advantage. Reference validity requires nondegenerate action-return variance, and soft Safety matching also applies its cross-trajectory similarity condition. Rank-deficient least squares is solved by pseudoinverse.

Algorithm C.1 DCGPO advantage estimation for one computation batch

Input: task-wise rollout groups; reference rule; update n; action-matching and propagation parameters.
Compute A^E, A^S, and normalized action returns Q^A.
for each action occurrence (i, t) **do**
    Match actions within its task group, excluding trajectory i.
    Check reference validity; if valid, form $q_{i,t;j,u}$ by the chosen rule.
    Compute $A^A_{i,t}$ by Eq. (4) if valid; otherwise set it to zero.
end for
for each occurrence (i, t) **do**
    Aggregate valid future action evidence within K steps into $C^A_{i,t}$ by Eq. (5).
end for
Set $\Omega_C = \{(i, t) : C^A_{i,t}$ is finite, $|C^A_{i,t}| > \varepsilon\}$.
if $|\Omega_C| \geq 3$ **then**
    Fit C^A against $[1, A^E, A^S]$ on Ω_C; use the masked residual A^R.
else
    Set $A^R = 0$ for this computation batch.
end if
Apply configured shrinkage and warm-up; fuse A^E, A^S, and A^R.
Output: token-level advantages for Eq. (9).

D Temporal Aggregation and Support

The future-evidence term in Eq. (5) is a distance-weighted average over valid future references, not an unnormalized sum. Normalizing by the available weight mass prevents positions with more matched successors from gaining scale merely because they have more evidence. If no future reference is valid, the propagated component vanishes and local action evidence may still contribute.

The original comparison support Ω_A contains positions whose own action has a valid cross-trajectory reference. The transferred support Ω_C contains positions with a finite, nonzero credit block after temporal aggregation. A position can belong to $\Omega_C \setminus \Omega_A$: although its own action has no match, a later matched action can supply evidence about the history it helped form. Residual projection is both fitted and written only on Ω_C; elsewhere the action contribution is zero. If fewer than three supported steps are available in the computation batch, the residual block is set to zero rather than fitting three coefficients from insufficient data.

Contribution scale. For configurations with a positive action-budget cap, the weighted action block is reduced, if necessary, so that its root-mean-square magnitude is at most a specified fraction of the root-mean-square episode-plus-state advantage. This is a shrink-only operation: it cannot amplify action credit. A linear warm-up then increases the action coefficient from zero to its configured value over the specified number of optimization steps. Both operations act after the cross-trajectory comparison and temporal aggregation; neither changes the sampled rollouts or action-matching rule.

The credit-coverage diagnostic in Figure 3(b) measures signal eligibility, not the nonzero final residual. In the displayed ablation runs, eligibility is 76.5% for in-place assignment and 91.3% for temporal propagation. Since these are separate rollouts, their difference is not the fraction of originally unmatched steps that gained nonzero credit. That fraction requires the joint, per-step quantity $|\{(i,t) \notin \Omega_A : |A^R_{i,t}| > \varepsilon\}| / \sum_i T_i$ from the same run; the available aggregate diagnostics do not identify it.

E Illustrative Case Study: Authorization as Action Context

Consider a schematic Agentic-Safety task in which an agent may share a protected report only after obtaining authorization. Two trajectories reach the same tool action, `share_report`, with the same arguments. In the first, earlier decisions request permission, inspect the response, and establish that sharing is authorized. In the second, the agent invokes the same tool without that authorization. The action text is identical, but the histories supply different conditions for safe and useful execution.

Suppose the authorized continuation receives a higher terminal return than the unauthorized one. Action matching groups the two `share_report` occurrences; the cross-trajectory reference then assigns a higher relative action-axis value to the authorized occurrence and a lower value to the unauthorized occurrence. History-directed aggregation passes part of this contrast to preceding decisions, such as checking authorization or proceeding without it. Residualization retains only the component not already explained by episode- and state-conditioned advantages. Thus the comparison concerns the context in which a common action is executed, not a claim that the action string is intrinsically good or bad.

This is an explanatory construction, not a sampled benchmark trajectory or a causal estimate. It isolates the information that action-conditioned grouping can represent; observed returns may also depend on later decisions and environment stochasticity.

F Agentic-Safety Evaluation Details

Reward and judge. Let $U \in [0,1]$ denote the terminal task utility returned by the environment's final-state evaluator, and let $S \in \{0,1\}$ denote the safety verdict of a frozen AgentDoG1.5-Qwen3.5-4B guard (Liu et al., 2026), with $S = 1$ indicating safe behavior. The terminal reward is U for Clean tasks, S for Query-target tasks, and $0.5U + 0.25S + 0.25US$ for Attacked tasks; all intermediate rewards are zero. The guard evaluates the complete trajectory at termination with temperature zero and is not called for Clean tasks. An unparsable verdict is retried up to three times; trajectories without a valid verdict are excluded rather than assigned a default safety score. Overall and per-split scores are mean terminal rewards multiplied by 100. Success requires environment completion on Clean, a safe verdict on Query-target, and both completion and safety on Attacked.

Data split. We construct the data from AgentDoG semantic task families, each linking a benign Clean task, an environment-injection Attacked variant, and a malicious-query Query-target variant.

The reported runs request 2,000 training rows and 300 validation rows; complete task families are retained when sampling, so a requested row budget can be slightly exceeded. The validation sample targets 100 examples of each scenario. Families are kept intact within a split, while train–validation overlap is prevented at the task-ID level. Clean, Attacked, and Query-target therefore form three related scenario subsets within this benchmark.

Threat model. We evaluate two fixed, benchmark-generated attack surfaces: untrusted content injected through the task environment and direct malicious user queries. The agent may act on this content through the provided tools, while the desired behavior is to preserve benign task utility without executing unsafe actions; for direct malicious queries, safety is the sole reward criterion. The attacks are static black-box instances: they do not adapt to the learned policy and do not modify model parameters, the reward implementation, or the guard. Accordingly, the experiment measures robustness to these two in-distribution threat conditions, not to adaptive white-box adversaries or unseen attack families.

G Future Work

Action-conditioned comparison depends on recognizing when two action instances express the same operation. The exact and soft matching rules studied here provide two ways to construct that relation. A natural extension is to test representations that better distinguish action intent from incidental surface form, particularly for free-form tool calls or multimodal actions (Xie et al., 2024). Such representations would need to preserve the central separation between action similarity and the historical contexts being compared.

History-directed propagation is likewise a credit-assignment hypothesis rather than an intervention on the environment. Future work could test it with controlled counterfactual rollouts: hold a later action fixed while changing an earlier information-gathering or authorization decision, then measure whether the resulting return contrast agrees with the assigned credit. This would clarify when an observed cross-context difference is informative about earlier decisions and when it is dominated by later actions or stochastic outcomes.