CONTEXTUAL CREDIT POLICY OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Group-based reinforcement learning enables large language model agents to learn from sparse outcome rewards in long-horizon tasks. Recent methods derive step-level credit by comparing observation-matched experiences across trajectories. However, identical observations can occur in different contexts, making some peers more informative for a given decision. Moreover, favorable terminal outcomes can obscure redundant actions or detours within successful trajectories. We propose **Contextual Credit Policy Optimization (CCPO)**, which combines contextual peer weighting with history-conditioned return credit and realized future progress. CCPO combines frozen LLM representations and accumulated trajectory statistics to weight observation-matched peers from other trajectories by contextual similarity. Historical credit compares the realized return against the weighted peer baseline, while future credit measures the change in contextual outcome value over a short realized continuation. CCPO reuses collected rollouts without an auxiliary value critic or additional environment interactions; future information is used only during training. With Qwen2.5-1.5B-Instruct and Qwen2.5-7B-Instruct, CCPO reaches up to $97.4 \pm 0.5\%$ success on ALFWorld and $79.4 \pm 1.3\%$ on WebShop; across variants, 7B gains over GRPO are 12.3–15.4 and 5.5–6.5 percentage points. Ablations support query-trajectory exclusion, combined historical and future credit, and structured context representation.

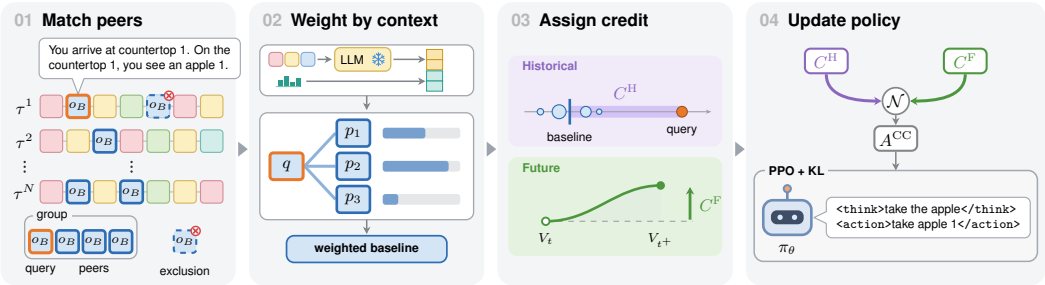

Figure 1: CCPO combines contextually weighted peer comparisons and short-horizon future progress for critic-free step-level credit assignment.

1 INTRODUCTION

Large language model (LLM) agents increasingly undertake long-horizon tasks through sustained environment interaction (Wang et al., 2024; Zhang et al., 2026). Benchmarks such as WebArena (Zhou et al., 2024a) and SWE-bench (Jimenez et al., 2024) require information gathering, coordinated actions, and adaptation to feedback. Reinforcement learning (RL) (Sutton & Barto, 2018) provides a post-training framework that optimizes policies using rewards from interaction trajectories. Recent applications span open-web research, web navigation, and multi-step software engineering (Zheng et al., 2025b; Qi et al., 2025; Luo et al., 2025).

The effectiveness of agentic learning depends on how interaction feedback is translated into policy updates. PPO-based actor–critic methods use learned value estimates to guide optimization, at the cost of training an additional value model (Schulman et al., 2017). Alternatively, group-based methods such as GRPO (Shao et al., 2024) estimate relative advantages from multiple sampled outcomes

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.

without a separately learned value critic. When applied at the trajectory level, however, a single advantage is shared across all actions in a rollout. Such comparisons differentiate outcomes across trajectories but provide limited information about the contributions of individual decisions, especially when rewards are sparse and delayed (Zhang et al., 2026). To enable fine-grained policy updates, recent methods extend this approach to construct step-level advantage estimates: GiGPO (Feng et al., 2025) compares actions encountered at shared observations, and G2PO (Wang et al., 2026d) incorporates the structure of transitions between observation groups. HGPO (He et al., 2026) accounts for historical context by organizing comparisons into hierarchical groups. Together, these advances establish the importance of constructing appropriate comparisons from existing rollouts.

Building on this foundation, we consider two challenges in step-level credit assignment. First, how strongly each comparison should influence the group advantage. Identical observations can occur after different preceding actions, reasoning, exploration depth, and accumulated contexts. Depending on historical priors, peers within a group therefore may not be equally informative for a particular query. Our premise is that, within observation-matched groups, contextual similarity should determine each peer's contribution to the query's baseline and value estimates. Second, history-based credit assesses performance relative to the preceding context. Future credit tracks contextual-value changes over a short continuation. An eventual success can include redundant actions or detours, motivating a complementary comparison of realized gains after a short continuation to distinguish between inheriting a favorable context and making progress from it.

We propose **Contextual Credit Policy Optimization (CCPO)**, which constructs step-level learning signals through contextual peer weighting. As shown in Figure 2, CCPO represents each step-level context using semantic features of an interaction window from a frozen LLM and accumulated trajectory statistics, assigning greater weight to observation-matched peers with higher contextual similarity. A shared estimator supports two temporal credit signals. Historical credit compares the realized return with its contextual peer baseline. Future credit estimates short-horizon progress as the value change between the current step and a later endpoint of the collected rollout. Values are contextually weighted averages of peers' discounted terminal rewards. Together, these signals provide step-level advantages reflecting both recent history and subsequent progress. CCPO requires neither an auxiliary value critic nor additional environment interactions. Future information is used only during training. On agentic multi-step benchmarks, ALFWorld (Shridhar et al., 2021) and WebShop (Yao et al., 2022), CCPO outperforms both prompting- and RL-based baseline methods using Qwen2.5-1.5B-Instruct and Qwen2.5-7B-Instruct backbones (Qwen et al., 2025).

Our main contributions are as follows:

- **Dual-view contextual credit estimation from history and future progress.** We assign step-level credit by combining a history-conditioned return residual with estimated value change over a future horizon, capturing outcome-relative performance and local progress. Analysis shows complementary effects of history and future credit as learning evolves.
- **Contextual similarity for step-level peer weighting.** We aggregate peer returns using continuous weights determined by the similarity between query and peer context representations, producing a context-dependent baseline. Consequently, CCPO produces learning signals that differ from standard trajectory-level advantages.
- **Semantic and structural context representations.** We combine semantic hidden states of recent interactions with accumulated trajectory statistics, enabling context-sensitive comparisons across trajectories without training an auxiliary value network.

2 Related Work

Agentic reinforcement learning for long-horizon tasks. Reinforcement learning (RL) (Sutton & Barto, 2018) enables large language model (LLM) agents to improve through extended interactions in which observations, reasoning, actions, and feedback accumulate into trajectories (Zhou et al., 2024b; Zhang et al., 2026). Applications span interactive gaming, web navigation, open-web research, and multi-step software engineering (Wang et al., 2025; Qi et al., 2025; Zheng et al., 2025b; Li et al., 2026b; Luo et al., 2025), requiring agents to gather information, coordinate sequential actions, and adapt to environmental feedback. Recent work trains agents to reason with external tools, including search engines, web browsing tools, and code interpreters (Jin et al., 2025; Tongyi

DeepResearch Team et al., 2026; Feng et al., 2026). Agentic RL can also learn from successes and failures through verbal reflection and trajectory preference learning (Shinn et al., 2023; Yao et al., 2024; Song et al., 2024; Putta et al., 2024). To collect high-quality trajectories, methods use uncertainty-guided branching and resampling (Dong et al., 2026; Wang et al., 2026a). Reward design (Wang et al., 2026e) is also crucial, spanning outcome-supervised reward models (Qi et al., 2025), process rewards from environment-aware models such as DataPRM (Qiu et al., 2026), and turn-level reward assignment (Tao et al., 2026). Despite these advances, sparse and delayed rewards, costly environment interaction, and training instability remain challenges (Tao et al., 2026; Li et al., 2026b), motivating efficient, stable policy optimization from collected trajectories.

Policy optimization for LLM agents. Existing approaches differ in their optimization objectives and advantage estimators. Preference-based methods, including DPO (Rafailov et al., 2023), ETO (Song et al., 2024), and Agent Q (Putta et al., 2024), optimize policies from response or trajectory preferences, including successful and failed agent rollouts, without explicit reward-model fitting. Actor–critic methods use learned value estimates, as in PPO-based reinforcement learning from human feedback (RLHF) with InstructGPT (Ouyang et al., 2022) and multi-step agent learning with ArCHer (Zhou et al., 2024b). To avoid a separate value critic, group-based methods estimate advantages from multiple sampled responses or trajectories. RLOO (Ahmadian et al., 2024) uses the mean reward of other samples as a baseline, while GRPO (Shao et al., 2024), used in DeepSeek-R1 (Guo et al., 2025), standardizes group rewards. Related variants include LOOP (Chen et al., 2025) with leave-one-out PPO updates and Dr. GRPO (Liu et al., 2025) with revised length and reward normalization. Further refinements target update stability and exploration efficiency through clipping, importance weighting, and sampling, including DAPO (Yu et al., 2025), GSPO (Zheng et al., 2025a), CISPO (MiniMax et al., 2025), SAPO (Gao et al., 2025), VESPO (Shen et al., 2026), SAMPO (Wang et al., 2026b), Clip-Cov (Cui et al., 2025), and T^{2}PO (Wang et al., 2026a). However, trajectory-level applications of outcome-supervised GRPO share a single advantage across all actions in a rollout, without explicitly differentiating individual contributions (Feng et al., 2025). CCPO assigns step-level credit using each decision's observation and temporal context.

Step-level credit assignment. Recent methods construct step-level learning signals for multi-step interactions (Zhang, 2026). Critic-based methods include SWEET-RL (Zhou et al., 2025) with additional training-time information and HCAPO (Tan et al., 2026) with the agent as a hindsight critic. TRACE (Tao et al., 2026) derives turn-level rewards from a frozen model's gold-answer likelihoods. OpenClaw-RL (Wang et al., 2026c) combines evaluative rewards and corrective hints extracted from next-state feedback. Other methods estimate advantages by comparing collected trajectories. GiGPO (Feng et al., 2025) combines episode-level advantages with comparisons at matching observations, while HGPO (He et al., 2026) aggregates advantages from hierarchical groups matched at different history lengths. SALT (Li et al., 2026a) constructs graphs from trajectories of the same prompt, while G2PO (Wang et al., 2026d) aggregates observations into a state-transition graph to estimate values and advantages. GEPO (Yuan et al., 2026) uses task-conditioned graph centrality to shape trajectory returns and scale step-level advantages. Peer matching faces a granularity trade-off: observation-only matching can mix different histories, whereas exact history matching reduces the available samples (He et al., 2026). CCPO weights observation-matched peers by similarity in local semantic and accumulated trajectory features, yielding query-specific return baselines and outcome values for history-conditioned credit and short-horizon progress along realized rollouts.

3 Contextual Credit Policy Optimization

3.1 Problem Formulation

We sample N trajectories per task from behavior policy π_{old}. Trajectory i for task x_i contains T_i actions $a_{i,0:T_i-1}$, observations $o_{i,0:T_i}$, and terminal outcome R_i. Prompt $p_{i,t}$ contains x_i, $o_{i,t}$, and up to K_{H} preceding observation–action pairs. Rewards come from environment feedback. GRPO (Shao et al., 2024) uses the mean reward of group rollouts as a baseline to estimate normalized advantages without a learned value critic. CCPO extends this critic-free estimation to individual steps, weighting

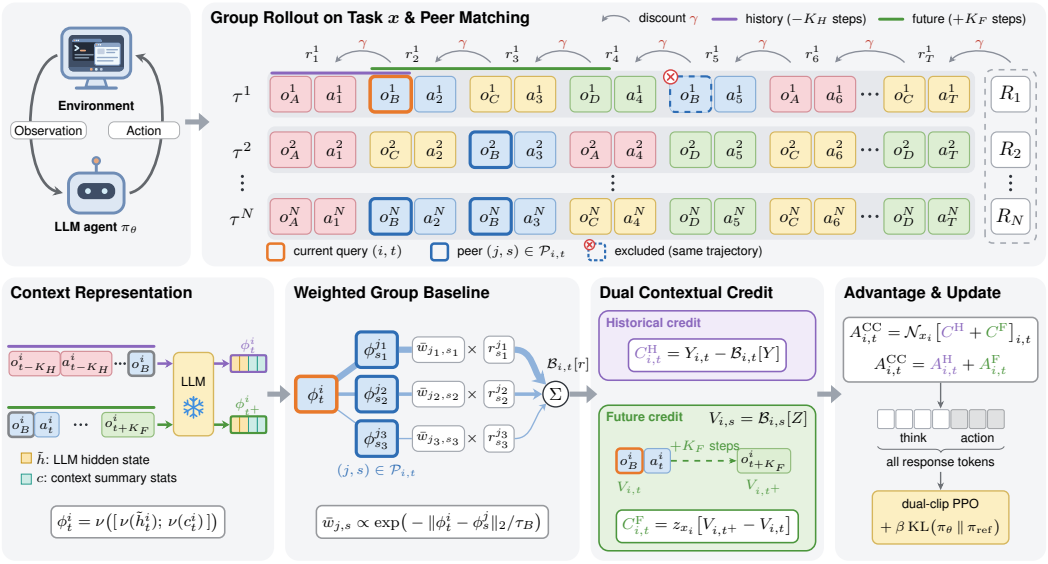

Figure 2: Overview of CCPO. Contextual similarity weights observation-matched peers. A history-conditioned return residual and estimated outcome-value change over a short future window yield step-level credit from collected rollouts without an auxiliary value critic.

peers by contextual similarity to estimate return baselines. The historical target is

$$Y_{i,t} = \sum_{u=t}^{T_i-1} \gamma^{u-t} r^{\mathrm{env}}_{i,u} - \eta\, \mathbf{1}\{a_{i,t} \text{ is invalid}\}. \tag{1}$$

Here $r^{\mathrm{env}}_{i,u}$ follows action $a_{i,u}$ and γ discounts rewards. The penalty term subtracts η for an invalid action at step t only (Appendix C.3).

3.2 Contextual Peer Estimation

CCPO assigns continuous, query-specific weights to individual peers based on contextual similarity. GRPO averages rewards uniformly across same-task rollouts, but identical observations at different steps can follow different histories. Soft weighting within observation-matched groups accounts for these differences without exact history matching, which can leave few peers (He et al., 2026).

Peer selection. Exact peers share the task and observation and exclude the entire query trajectory:

$$\mathcal{E}_{i,t} = \{(j,s) : x_j = x_i,\ j \neq i,\ 0 \le s < T_j,\ o_{j,s} = o_{i,t}\}. \tag{2}$$

We use $\mathcal{P}_{i,t} = \mathcal{E}_{i,t}$ when nonempty; otherwise, we use all same-task steps with $j \neq i$. We quantify fallback rates and peer availability during training in Appendix H.

Context representation. Each step combines a frozen LLM representation $h_{i,t}$ with a trajectory summary $c_{i,t}$. The former is the final-layer hidden state of the last prompt token; the latter records elapsed steps, distinct-observation count, novelty fraction, and a revisit indicator through step t. The summary supplements the bounded prompt history. We L2-normalize each preprocessed component, concatenate them with equal weights, and L2-normalize to obtain $\phi_{i,t}$.

Contextual readout. Bucket B contains all batch steps matching the query's task and observation, or task alone for fallback. We aggregate scalar targets r over nonempty $\mathcal{P}_{i,t}$:

$$\bar{w}_{(i,t),(j,s)} \propto \exp\big(-\|\phi_{i,t} - \phi_{j,s}\|_2/\tau_B\big), \qquad \sum_{(j,s)\in\mathcal{P}_{i,t}} \bar{w}_{(i,t),(j,s)} = 1. \tag{3}$$

$$\mathcal{B}_{i,t}[r] = \sum_{(j,s)\in\mathcal{P}_{i,t}} \bar{w}_{(i,t),(j,s)} r_{j,s}. \tag{4}$$

The bandwidth τ_B is proportional to the median pairwise context distance in B, computed before trajectory exclusion. This readout estimates both return baselines and outcome values.

3.3 Temporal Credit Estimation

CCPO conditions return comparisons on a K_{H}-step history window and assesses progress over a short realized continuation of K_{F} steps. The future window compares discounted contextual outcome values at its endpoints, even without intermediate rewards.

History-conditioned return credit. We subtract the contextual peer baseline from the return:

$$C^{\mathrm{H}}_{i,t} = Y_{i,t} - \mathcal{B}_{i,t}[Y]. \tag{5}$$

Contextual future progress. We estimate contextual values from unpenalized terminal outcomes:

$$Z_{i,s} = \gamma^{T_i - s} R_i, \qquad V_{i,s} = \begin{cases} \mathcal{B}_{i,s}[Z], & s < T_i, \\ R_i, & s = T_i. \end{cases} \tag{6}$$

For future horizon K_{F}, let $t^+ = \min(t + K_{\mathrm{F}}, T_i)$. We compute

$$\Delta V_{i,t} = V_{i,t^+} - V_{i,t}, \qquad C^{\mathrm{F}}_{i,t} = z_{x_i}[\Delta V]_{i,t}, \tag{7}$$

where z_x standardizes eligible differences within task x. Each nonterminal endpoint uses its own context and peer set, and future endpoints are used only during training. The endpoint difference measures growth in discounted contextual value, including changes due to shorter remaining time to success. Compared with a discount-aligned TD residual (Schulman et al., 2016), this difference includes an extra time-dependent term that can survive task-wise centering (Appendix C.3).

3.4 Policy Optimization

We add the historical return residual and task-standardized future progress with unit coefficients:

$$A^{\mathrm{CC}}_{i,t} = \mathcal{N}_{x_i}[C^{\mathrm{H}} + C^{\mathrm{F}}]_{i,t}. \tag{8}$$

The contributions satisfy $A^{\mathrm{CC}} = A^{\mathrm{H}} + A^{\mathrm{F}}$. The default uses $A_{i,t} = A^{\mathrm{CC}}_{i,t}$. To test whether trajectory-level signals complement step-level credit, the episode-augmented variant adds a GRPO-style episode advantage (Shao et al., 2024), $A_{i,t} = A^{\mathrm{CC}}_{i,t} + A^{\mathrm{EP}}_{i,t}$, with unit weight and no further normalization. Each step's advantage weights response tokens in a dual-clipped objective (Ye et al., 2020), with a KL penalty toward the frozen reference policy. See Appendix C for details.

4 Experiments

4.1 Experimental Setup

Benchmarks and metrics. We evaluate CCPO against baseline methods on ALFWorld (Shridhar et al., 2021) and WebShop (Yao et al., 2022), covering household task execution and online shopping through multi-step interaction. ALFWorld comprises six household task categories, with 3,553 training tasks and 140 held-out tasks of the in-distribution evaluation split. For WebShop, we follow the goal partition implemented in verl-agent (Feng et al., 2025), reserving the first 500 goals as the evaluation pool and sampling training tasks from the remaining goals. Following prior work (Feng et al., 2025; Wang et al., 2026d), we report overall and per-category success rates on ALFWorld. For WebShop, we report success rate and mean task score.

Models and baselines. We use Qwen2.5-1.5B-Instruct and Qwen2.5-7B-Instruct (Qwen et al., 2025) as policy backbones. We compare CCPO with three categories of baselines: (1) **Off-the-shelf LLMs:** Gemini 3.8 Flash (Google DeepMind, 2026), DeepSeek-V4.1-Flash (DeepSeek-AI et al., 2026), and Qwen3.8-27B (Qwen Team, 2026). (2) **Prompt-based agents:** the Qwen2.5-Instruct backbones without task-specific RL fine-tuning, ReAct (Yao et al., 2023), and Reflexion (Shinn et al., 2023). (3) **RL training methods:** PPO (Schulman et al., 2017), RLOO (Ahmadian et al., 2024), GRPO (Shao et al., 2024), GiGPO (Feng et al., 2025), HGPO (He et al., 2026), and G2PO (Wang et al., 2026d). Baseline configurations are detailed in Appendix B.

Implementation details. We train CCPO using verl-agent (Feng et al., 2025), built on veRL (Sheng et al., 2025), for 150 iterations and sampling 16 tasks with eight trajectories each. We use a learning rate of 10^{-6} and discount factor 0.95. Interaction limits are 50 steps for ALFWorld and 15 for WebShop. Settings are shared across both backbones and all group-based baselines, except that HGPO retains its published limits of 4,096 prompt tokens on ALFWorld and 30 interaction steps on WebShop. CCPO uses a $K_H = 2$ history window and a $K_F = 2$ future horizon. We evaluate two CCPO variants. The default uses A_{CC} from Equation (8). The variant uses $A_{\mathrm{CC}} + A_{\mathrm{EP}}$, adding a standard episode-level advantage without further normalization. All training and evaluation are completed on 8 $\times$ H200 GPUs. We report means and standard deviations across three random seeds. Appendix B gives the shared training budgets and optimization settings.

4.2 Main Results

Comparison with baselines. Across CCPO variants, mean success gains over GRPO are 16.6–21.6 percentage points on ALFWorld and 12.3–18.5 on WebShop at 1.5B; at 7B, they are 12.3–15.4 and 5.5–6.5 points, respectively (Table 1). WebShop score gains are 13.1–14.8 at 1.5B and 1.7–6.5 at 7B. CCPO leads the listed RL baselines in mean success: A_{CC} reaches 96.9%/97.4% on ALFWorld at 1.5B/7B; on WebShop, $A_{\mathrm{CC}} + A_{\mathrm{EP}}$ at 1.5B and A_{CC} at 7B reach 75.0%/79.4%.

4.3 Ablation Studies

We ablate CCPO with Qwen2.5-1.5B-Instruct, using A_{CC} on ALFWorld and $A_{\mathrm{CC}} + A_{\mathrm{EP}}$ on WebShop as reference configurations. All ablations follow the main evaluation protocol.

Peer selection and contextual weighting. Relative to CCPO, including same-trajectory peers lowers mean success by 6.3 on ALFWorld and 19.0 percentage points on WebShop (Table 2). Uniform weighting lowers success by 3.7 and 2.6 points, and cosine weighting by 4.2 and 4.7.

Structured context summaries alter peer influence. Context summaries add accumulated trajectory statistics to semantic representations: step count, distinct-observation count, novelty fraction, and a revisit indicator. Removing them lowers mean success by 2.6 percentage points on ALFWorld and 1.3 on WebShop (Table 2). Adding summaries changes normalized weights by more than one percentage point for 33.8% of matched query–peer pairs (Table 3).

Historical and future credit. We compare the combined, history-only, and future-only configurations, retaining episode-level credit on WebShop. On ALFWorld, success falls from 96.9% to 91.4% when using only historical context and 84.6% when using only future context; on WebShop, it decreases from 75.0% to 72.7% and 74.0%, respectively (Table 2). The combined configuration achieves the highest mean success on both benchmarks. History-only exceeds future-only in mean success on ALFWorld but trails it on WebShop, suggesting task-dependent credit utility.

Temporal window allocation. We vary symmetric windows on history length and future horizon. Increasing the symmetric window from $(2, 2)$ to $(3, 3)$ reduces success from 96.9% to 91.4%. We then compare asymmetric temporal allocations at fixed context span. The default CCPO (2 step each way) outperforms all tested asymmetric windows. We restrict these variations to ALFWorld, where successful 1.5B trajectories average 8.7 steps versus 5.1 on WebShop under the reference configurations (Appendix G, Figure S3). These results favor compact context windows.

4.4 Contextual Credit Analysis

Identical observations can conceal different task progress. In Figure 3, GiGPO and G2PO repeatedly visit an empty toiletpaper hanger without acquiring the target; CCPO encounters the same observation after acquisition and completes the task (Appendix F.1). Appendix F.2 presents a complementary failure mode: G2PO repeatedly removes and replaces one spraybottle, whereas CCPO places two distinct objects (Figure S2). Across three evaluation draws, CCPO's repeated-observation rate is 22.5%, versus 26.8–41.0% for the baselines (Table 4).

Contextual credit provides unique learning signals. Figure 5a compares A_{CC} with counterfactual A_{EP} from the same A_{CC}-only ALFWorld rollouts; A_{EP} is logged for comparison and excluded

Table 1: Results on WebShop and ALFWorld. We report the success rate (%) for each sub-task and overall results for ALFWorld, and the success rate (%) and score for WebShop, averaged over 3 seeds with standard deviations. Within each section, the **best** means are highlighted and the second best are underlined; higher is better. * marks baselines reproduced or off-the-shelf models evaluated by us. Light-gray rows show results from the original papers and are excluded from ranking highlights. Δ GRPO gives the min–max gain over GRPO across both CCPO variants.

Type	Method	ALFWorld							WebShop	
		All	Pick	Look	Clean	Heat	Cool	Pick2	Succ.	Score
Off-the-Shelf Model										
Prompting	Gemini 3.8 Flash*	**96.4**	**100.0**	**84.6**	**100.0**	**100.0**	**92.0**	**95.8**	**49.2**	**54.0**
Prompting	DeepSeek-V4.1-Flash*	75.0	88.6	**84.6**	66.7	62.5	76.0	66.7	22.0	27.9
Prompting	Qwen3.8-27B*	70.7	85.7	61.5	66.7	56.3	72.0	66.7	29.0	39.5
Qwen2.5-1.5B-Instruct										
Prompting	Qwen2.5*	6.4	11.4	7.7	3.7	6.3	4.0	4.2	5.0	21.9
Prompting	ReAct	12.8	17.4	20.5	15.7	6.2	7.7	2.0	11.3	40.1
Prompting	Reflexion	21.8	35.3	22.2	21.7	13.6	19.4	3.7	21.9	55.8
RL Training	PPO (with critic)	$54.4_{\pm 3.1}$	$64.8_{\pm 3.5}$	$40.5_{\pm 6.9}$	$57.1_{\pm 4.9}$	$60.6_{\pm 6.6}$	$46.4_{\pm 4.0}$	$47.4_{\pm 1.9}$	$51.5_{\pm 2.9}$	$73.8_{\pm 3.0}$
RL Training	RLOO	$69.7_{\pm 2.5}$	$88.3_{\pm 3.0}$	$52.8_{\pm 8.6}$	$71.0_{\pm 5.9}$	$62.8_{\pm 8.7}$	$66.4_{\pm 5.5}$	$56.9_{\pm 4.7}$	$52.1_{\pm 6.7}$	$73.9_{\pm 5.6}$
RL Training	GRPO*	$75.3_{\pm 1.0}$	$81.8_{\pm 0.8}$	$66.5_{\pm 8.7}$	$77.2_{\pm 0.4}$	$73.8_{\pm 3.4}$	$78.0_{\pm 3.0}$	$65.7_{\pm 1.9}$	$56.5_{\pm 2.9}$	$70.5_{\pm 2.2}$
RL Training	GiGPO*	$83.9_{\pm 0.7}$	$84.8_{\pm 0.7}$	$77.0_{\pm 10.0}$	$92.4_{\pm 0.1}$	**$100.0_{\pm 0.0}$**	$75.0_{\pm 0.9}$	$74.6_{\pm 1.9}$	$67.4_{\pm 2.2}$	$80.1_{\pm 1.8}$
RL Training	GiGPO	$86.7_{\pm 1.7}$	$94.4_{\pm 5.9}$	$67.5_{\pm 4.6}$	$94.8_{\pm 3.8}$	$94.4_{\pm 7.8}$	$79.8_{\pm 4.7}$	$76.4_{\pm 5.4}$	$65.0_{\pm 3.2}$	$83.1_{\pm 1.6}$
RL Training	HGPO*	$89.1_{\pm 0.6}$	$93.9_{\pm 0.3}$	$80.1_{\pm 6.0}$	$92.4_{\pm 0.1}$	**$100.0_{\pm 0.0}$**	$88.3_{\pm 1.6}$	$76.2_{\pm 1.6}$	$65.4_{\pm 1.5}$	$82.1_{\pm 0.8}$
RL Training	HGPO	$92.8_{\pm 1.1}$			Unreported				$71.5_{\pm 4.0}$	$85.6_{\pm 2.9}$
RL Training	G2PO*	$94.0_{\pm 0.7}$	$93.9_{\pm 0.3}$	$90.6_{\pm 7.4}$	**$100.0_{\pm 0.0}$**	**$100.0_{\pm 0.0}$**	$92.7_{\pm 1.8}$	$86.6_{\pm 0.3}$	$71.6_{\pm 0.7}$	$83.5_{\pm 0.8}$
RL Training	G2PO	$95.0_{\pm 0.8}$	$97.2_{\pm 3.9}$	$96.3_{\pm 5.2}$	$95.1_{\pm 1.8}$	$94.9_{\pm 3.6}$	$97.1_{\pm 2.0}$	$91.4_{\pm 4.1}$	$71.2_{\pm 2.6}$	$85.1_{\pm 1.4}$
RL Training	**CCPO (ours)**									
	A_{CC} only	**$96.9_{\pm 0.8}$**	**$100.0_{\pm 0.0}$**	$75.2_{\pm 4.5}$	**$100.0_{\pm 0.0}$**	$93.1_{\pm 0.7}$	**$100.0_{\pm 0.0}$**	**$97.0_{\pm 2.6}$**	$68.8_{\pm 2.8}$	$83.6_{\pm 2.1}$
	$A_{\mathrm{CC}} + A_{\mathrm{EP}}$	$91.9_{\pm 0.5}$	$93.9_{\pm 0.2}$	**$100.0_{\pm 0.0}$**	$97.5_{\pm 2.2}$	$93.1_{\pm 0.7}$	$86.9_{\pm 0.6}$	$83.9_{\pm 2.2}$	**$75.0_{\pm 0.6}$**	**$85.3_{\pm 1.3}$**
	Δ GRPO	16.6–21.6	12.1–18.2	8.7–33.5	20.3–22.8	19.3	8.9–22.0	18.2–31.3	12.3–18.5	13.1–14.8
Qwen2.5-7B-Instruct										
Prompting	Qwen2.5*	19.3	37.1	38.5	14.8	12.5	8.0	4.2	3.0	26.6
Prompting	ReAct	31.2	48.5	35.4	34.3	13.2	18.2	17.6	19.5	46.2
Prompting	Reflexion	42.7	62.0	41.6	44.9	30.9	36.3	23.8	28.8	58.1
RL Training	PPO (with critic)	$80.4_{\pm 2.7}$	$92.3_{\pm 4.0}$	$64.0_{\pm 8.4}$	$92.5_{\pm 2.4}$	$89.5_{\pm 7.0}$	$80.3_{\pm 2.0}$	$68.8_{\pm 8.3}$	$68.7_{\pm 5.1}$	$81.4_{\pm 3.1}$
RL Training	RLOO	$75.5_{\pm 4.6}$	$87.6_{\pm 4.3}$	$78.2_{\pm 8.3}$	$87.3_{\pm 5.8}$	$81.3_{\pm 7.6}$	$71.9_{\pm 5.2}$	$48.9_{\pm 8.4}$	$65.7_{\pm 4.0}$	$80.3_{\pm 3.2}$
RL Training	GRPO*	$82.0_{\pm 0.8}$	$85.8_{\pm 2.1}$	$71.9_{\pm 10.3}$	$92.1_{\pm 0.7}$	$84.2_{\pm 2.7}$	$81.2_{\pm 1.8}$	$69.1_{\pm 0.8}$	$72.9_{\pm 3.2}$	$82.9_{\pm 0.6}$
RL Training	GiGPO*	$92.2_{\pm 0.8}$	**$100.0_{\pm 0.0}$**	$93.3_{\pm 11.5}$	$96.0_{\pm 0.3}$	$95.7_{\pm 3.7}$	$89.9_{\pm 2.1}$	$76.5_{\pm 2.3}$	$74.7_{\pm 4.0}$	$84.3_{\pm 3.5}$
RL Training	GiGPO	$90.8_{\pm 1.3}$	$97.7_{\pm 1.6}$	$82.7_{\pm 7.9}$	$98.8_{\pm 1.6}$	$83.7_{\pm 7.2}$	$89.3_{\pm 8.2}$	$79.2_{\pm 6.6}$	$72.8_{\pm 3.2}$	$84.4_{\pm 2.9}$
RL Training	HGPO*	$95.8_{\pm 0.5}$	**$100.0_{\pm 0.0}$**	$96.7_{\pm 5.8}$	**$100.0_{\pm 0.0}$**	**$100.0_{\pm 0.0}$**	$85.6_{\pm 1.9}$	$92.6_{\pm 2.6}$	$77.6_{\pm 4.3}$	$86.1_{\pm 3.0}$
RL Training	HGPO	$95.4_{\pm 0.6}$			Unreported				$78.5_{\pm 1.4}$	$89.0_{\pm 1.0}$
RL Training	G2PO*	$94.0_{\pm 0.5}$	**$100.0_{\pm 0.0}$**	**$100.0_{\pm 0.0}$**	$96.0_{\pm 0.3}$	$95.7_{\pm 3.7}$	$85.6_{\pm 1.9}$	$88.2_{\pm 2.7}$	$72.4_{\pm 3.5}$	$83.8_{\pm 2.6}$
RL Training	G2PO	$96.9_{\pm 1.3}$	$98.7_{\pm 2.3}$	$90.6_{\pm 11.5}$	$97.6_{\pm 2.1}$	$100.0_{\pm 0.0}$	$98.8_{\pm 2.1}$	$93.7_{\pm 8.4}$	$78.3_{\pm 0.6}$	$89.8_{\pm 1.7}$
RL Training	**CCPO (ours)**									
	A_{CC} only	**$97.4_{\pm 0.5}$**	**$100.0_{\pm 0.0}$**	**$100.0_{\pm 0.0}$**	**$100.0_{\pm 0.0}$**	$95.7_{\pm 3.7}$	**$94.3_{\pm 2.2}$**	**$94.1_{\pm 2.5}$**	**$79.4_{\pm 1.3}$**	**$89.4_{\pm 0.8}$**
	$A_{\mathrm{CC}} + A_{\mathrm{EP}}$	$94.3_{\pm 0.5}$	**$100.0_{\pm 0.0}$**	$85.9_{\pm 5.1}$	$96.0_{\pm 0.3}$	**$100.0_{\pm 0.0}$**	$89.9_{\pm 2.1}$	$88.2_{\pm 2.7}$	$78.4_{\pm 3.9}$	$85.4_{\pm 1.6}$
	Δ GRPO	12.3–15.4	14.2	14.0–28.1	3.9–7.9	11.5–15.8	8.7–13.1	19.1–25.0	5.5–6.5	2.5–6.5

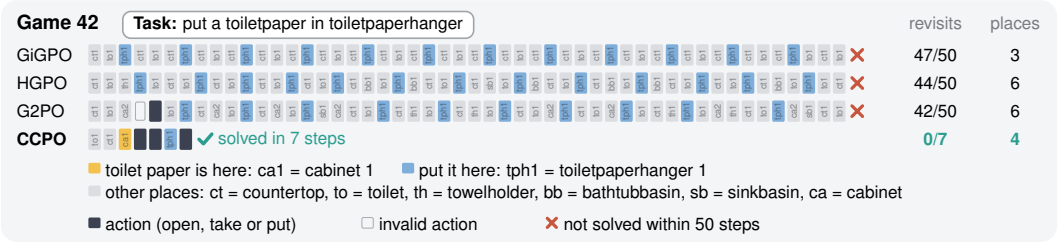

Figure 3: ALFWorld evaluation trajectories (Qwen2.5-1.5B-Instruct). Each cell denotes one step; revisit counts track navigation to previously visited locations. CCPO uses A_{CC} only.

from policy updates. Their mean absolute magnitudes diverge as validation success improves: A_{EP} diminishes while A_{CC} remains elevated, retaining effective learning signals.

The composition of contextual advantage changes during training. Figure 5b shows historical credit dominating initially, with future credit becoming more prominent despite fixed combination

Table 2: Ablations with Qwen2.5-1.5B-Instruct on ALFWorld and WebShop. **best** and second best are marked. $H = K_H$, $F = K_F$, and $W = H + F + 1$ denote history, future, and total context lengths. Markers: ⊙ current, ● history, and ◎ future.

Method	ALFWorld							WebShop	
	All	Pick	Look	Clean	Heat	Cool	Pick2	Succ.	Score
Qwen2.5-1.5B-Instruct				A_{CC} only				$A_{\text{CC}} + A_{\text{EP}}$	
w/o self-trajectory peer exclusion	$90.6_{\pm 0.8}$	$97.0_{\pm 0.1}$	$85.9_{\pm 5.1}$	$88.1_{\pm 1.0}$	**$100.0_{\pm 0.0}$**	$86.9_{\pm 0.6}$	$83.8_{\pm 2.8}$	$56.0_{\pm 5.5}$	$75.1_{\pm 2.6}$
w/o context summary	$\underline{94.3}_{\pm 0.5}$	$97.0_{\pm 0.1}$	$81.9_{\pm 7.1}$	$92.1_{\pm 0.7}$	$\underline{95.7}_{\pm 3.7}$	$\underline{98.6}_{\pm 2.4}$	$\underline{92.6}_{\pm 2.6}$	$73.7_{\pm 3.5}$	$84.5_{\pm 3.0}$
Uniform weighting	$93.2_{\pm 0.5}$	$94.9_{\pm 1.9}$	**$92.6_{\pm 6.4}$**	$\underline{96.1}_{\pm 0.3}$	$\underline{95.7}_{\pm 3.7}$	$87.0_{\pm 0.6}$	$\underline{92.6}_{\pm 2.7}$	$72.4_{\pm 3.9}$	$83.1_{\pm 3.4}$
Cosine-based context similarity	$92.7_{\pm 0.5}$	$\underline{97.9}_{\pm 1.8}$	$\underline{89.3}_{\pm 0.6}$	$92.1_{\pm 0.7}$	$\underline{95.7}_{\pm 3.7}$	$89.9_{\pm 2.1}$	$88.2_{\pm 2.7}$	$70.3_{\pm 2.7}$	**$86.6_{\pm 1.4}$**
Historical context only									
H: 2, F: 0, W: 3 ●●⊙	$91.4_{\pm 0.8}$	$93.9_{\pm 0.2}$	$85.9_{\pm 5.1}$	**$100.0_{\pm 0.0}$**	$93.1_{\pm 0.7}$	$86.9_{\pm 0.6}$	$83.8_{\pm 2.8}$	$72.7_{\pm 2.8}$	$85.1_{\pm 0.4}$
Future context only									
H: 0, F: 2, W: 3 ⊙◎◎	$84.6_{\pm 1.6}$	$91.9_{\pm 1.9}$	$71.9_{\pm 10.3}$	$88.1_{\pm 1.0}$	$82.0_{\pm 2.4}$	$91.3_{\pm 0.4}$	$70.6_{\pm 3.0}$	$\underline{74.0}_{\pm 2.4}$	$85.1_{\pm 2.4}$
Symmetric context variation									
H: 1, F: 1, W: 3 ●⊙◎	$86.7_{\pm 0.0}$	$91.9_{\pm 1.9}$	$78.5_{\pm 1.3}$	$92.1_{\pm 0.7}$	$91.0_{\pm 3.1}$	$85.6_{\pm 1.9}$	$75.0_{\pm 2.9}$	—	—
H: 3, F: 3, W: 7 ●●●⊙◎◎◎	$91.4_{\pm 0.8}$	$97.0_{\pm 0.1}$	$85.9_{\pm 5.1}$	$96.0_{\pm 0.3}$	$93.1_{\pm 0.7}$	$91.3_{\pm 0.4}$	$79.4_{\pm 2.8}$	—	—
Asymmetric context variation									
H: 4, F: 0, W: 5 ●●●●⊙	$\underline{94.3}_{\pm 0.9}$	$97.0_{\pm 0.1}$	$\underline{89.3}_{\pm 0.6}$	**$100.0_{\pm 0.0}$**	$\underline{95.7}_{\pm 3.7}$	$91.3_{\pm 0.4}$	$88.2_{\pm 2.7}$	—	—
H: 3, F: 1, W: 5 ●●●⊙◎	$92.7_{\pm 0.9}$	**$100.0_{\pm 0.0}$**	$\underline{89.3}_{\pm 0.6}$	$96.0_{\pm 0.3}$	$91.0_{\pm 3.1}$	**$100.0_{\pm 0.0}$**	$73.6_{\pm 3.9}$	—	—
H: 1, F: 3, W: 5 ●⊙◎◎◎	$89.3_{\pm 1.6}$	$93.9_{\pm 0.2}$	$82.6_{\pm 10.9}$	$89.6_{\pm 1.5}$	$82.0_{\pm 2.4}$	$91.3_{\pm 0.4}$	$88.2_{\pm 2.7}$	—	—
H: 0, F: 4, W: 5 ⊙◎◎◎◎	$62.0_{\pm 1.8}$	$71.7_{\pm 2.6}$	$82.6_{\pm 10.9}$	$58.2_{\pm 6.1}$	$66.1_{\pm 3.4}$	$59.4_{\pm 1.3}$	$44.1_{\pm 1.1}$	—	—
CCPO (ours)									
H: 2, F: 2, W: 5 ●●⊙◎◎	**$96.9_{\pm 0.8}$**	**$100.0_{\pm 0.0}$**	$75.2_{\pm 4.5}$	**$100.0_{\pm 0.0}$**	$93.1_{\pm 0.7}$	**$100.0_{\pm 0.0}$**	**$97.0_{\pm 2.6}$**	**$75.0_{\pm 0.8}$**	$\underline{85.4}_{\pm 1.6}$

Table 3: Absolute changes in matched peer weights from context summaries.

Change (pp)	> 0.25	> 0.5	> 1	> 5	> 10
Pairs (%)	62.1	48.7	33.8	8.6	3.6

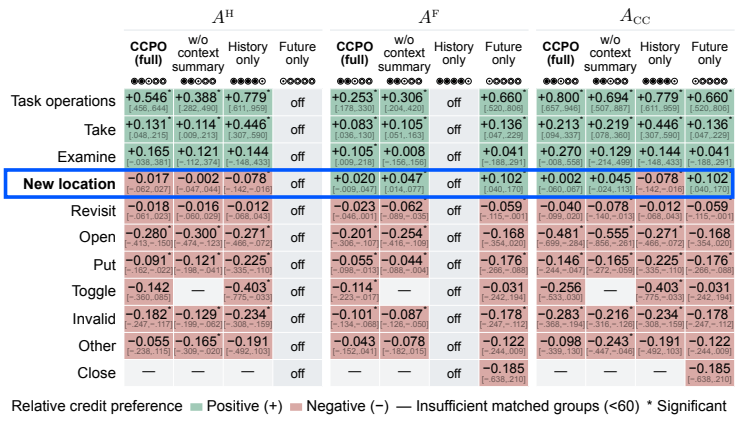

Figure 4: Credit preferences on ALFWorld (Qwen2.5-1.5B-Instruct). Mean category-relative credit within matched task–observation groups and iterations, from successful rollouts at iterations 101–150. Asterisks: Bonferroni-adjusted CIs excluding zero. Dashes: fewer than 60 groups; `off`: inactive component. Context markers denote ⊙ current, ● history, and ◎ future.

weights. Between iterations 1–30 and 121–150, the future-to-history mean absolute magnitude ratio rises from 0.25 to 0.60 at 1.5B and from 0.32 to 1.01 at 7B. Correlations between historical and future magnitudes within 10-iteration windows become less consistently positive later, especially at 7B. The two components thus evolve differently across training, supporting complementary roles.

What does future credit add beyond historical credit? Using successful 1.5B ALFWorld rollouts from iterations 101–150, we compare category-relative credit within task–observation-matched groups per iteration, with bootstrap CIs and Bonferroni correction across ten categories (Figure 4). Relative to other actions, A^{F} favors new-location transitions more than A^{H} (difference: $+0.037$;

Table 4: Revisit rates on ALFWorld evaluation rollouts.

Method	GiGPO	G2PO	HGPO	CCPO (A_{CC} only)
Revisit rate (%)	$41.0_{\pm 0.9}$	$26.8_{\pm 0.7}$	$36.4_{\pm 0.7}$	$\mathbf{22.5}_{\pm 1.8}$

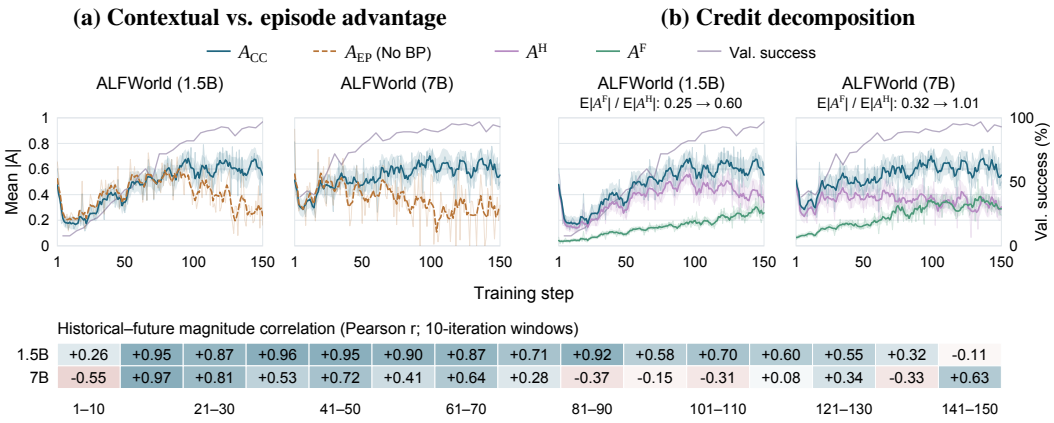

Figure 5: Contextual credit dynamics on ALFWorld (Qwen2.5-1.5B and 7B). (a) A_{CC} versus A_{EP} from the same rollouts; only A_{CC} is backpropagated. (b) Historical, future, and combined credit. Future-to-history ratios compare iterations 1–30 and 121–150. The heatmap correlates unsmoothed historical and future magnitudes by non-overlapping 10-iteration windows. Curves: mean absolute credit, centered 5-iteration means, and faint raw traces. Bands: 95% trajectory-bootstrap CIs.

adjusted CI: $[0.001, 0.073]$), while both favor cleaning, heating, and cooling. The $(H, F) = (4, 0)$, $(0, 4)$, and w/o-context-summary variants show similar patterns. Paired action timing also differs (Appendix E.2, Figure S1). The failure case in Figure 3 shows that identical observations can accompany different task progress; the credit preferences in Figure 4 complement this behavioral evidence by showing how historical and future credit evaluate exploratory transitions differently.

5 Conclusion

We introduced Contextual Credit Policy Optimization (CCPO), which combines contextual peer weighting with historical and realized-future views of step-level credit. By comparing returns against contextually relevant peers and measuring changes in contextual outcome value, CCPO uses collected rollouts without training an auxiliary value critic. Experiments on ALFWorld and WebShop with 1.5B and 7B backbones demonstrate improvements over prompting and RL baselines. Ablations support combining the dual contextual credit. Our analysis shows that their relative contributions evolve during training while remaining complementary.

Limitations. Despite its success on ALFWorld and WebShop, CCPO has several limitations. First, our evaluation covers two multi-step benchmarks and one model family; broader generalization and applicability to short-horizon or single-step tasks remain untested. Second, exact observation matching can limit peer availability, while fixed temporal windows and hand-designed trajectory statistics may require adaptation across environments. Finally, future credit vanishes when all rollouts in a task group receive zero terminal reward, limiting the future channel's contribution under sparse rewards, particularly early in training when successful trajectories are scarce.

Future work. The changing balance between historical and future credit motivates adaptive horizons and stage-dependent weighting. Further directions include streaming credit estimation from incomplete rollouts, multimodal contexts combining visual observations and textual histories, semantic peer retrieval beyond exact observation matching, and uncertainty estimates for correlated visits. Broader evaluation should examine generalization across model families and domains, computational scaling, and the effects of realized-future conditioning on policy updates.

AI USE STATEMENT

In this work, we used generative AI tools to assist with method implementation, experimental design, and interpretation of experimental results, including investigating potential causes of observed behavior. We also used these tools to suggest, implement, and execute statistical significance analyses. Additionally, we used generative AI tools for literature organization and citation management, brainstorming, creating graphical components for the pipeline figure, polishing manuscript language and grammar, and adjusting visual presentation and layout.

We reviewed AI-assisted work through the following procedures. AI assisted citation organization, with links checked manually. The authors approved AI-proposed statistical analysis plans. Generated code followed an author-defined structure for targeted manual inspection. Authors checked experimental configurations; AI-assisted runtime checks tested whether configuration switches activated the intended paths. The authors assembled the pipeline figure in PowerPoint using AI-assisted graphical components and refinements to colors and spacing.

We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Appendix B specifies training settings and ablation configurations, while Appendix C details context representation, peer selection and weighting, credit construction and normalization, and policy updates. Appendix E describes data filtering, matched comparisons, and statistical procedures for our analyses. We will provide the code and environment setup instructions with the submission. These details support independent reimplementation of CCPO.

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 Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12248–12267, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.662. URL https://aclanthology.org/2024.acl-long.662/.

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. URL https://arxiv.org/abs/2502.01600v3.

Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, Zhiyuan Liu, Hao Peng, Lei Bai, Wanli Ouyang, Yu Cheng, Bowen Zhou, and Ning Ding. The entropy mechanism of reinforcement learning for reasoning language models. arXiv preprint arXiv:2505.22617, 2025. URL https://arxiv.org/abs/2505.22617v1.

DeepSeek-AI et al. DeepSeek-V4.1-Flash: Pushing the limits of KV cache compression. arXiv preprint arXiv:2609.19969, 2026. URL https://arxiv.org/abs/2609.19969v1.

Guanting Dong, Hangyu Mao, Kai Ma, Licheng Bao, Yifei Chen, Zhongyuan Wang, Zhongxia Chen, Jiazhen Du, Huiyang Wang, Fuzheng Zhang, Guorui Zhou, Yutao Zhu, Ji-Rong Wen, and Zhicheng Dou. Agentic reinforced policy optimization. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 16981–17017, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/1c58e53bdf1fb045440256fd567531ae-Paper-Conference.pdf.

Jiazhan Feng, Shijue Huang, Xingwei Qu, Ge Zhang, Yujia Qin, Baoquan Zhong, Chengquan Jiang, Jinxin Chi, and Wanjun Zhong. ReTool: Reinforcement learning for strategic tool

use in LLMs. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 37909–37926, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/4038c9208dfc22644c60ad39c24e5c53-Paper-Conference.pdf.

Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. Group-in-group policy optimization for LLM agent training. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, pp. 46375–46408. Curran Associates, Inc., 2025. doi: 10.52202/085713-1544. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/420c9f777c0b4f78d515e53cf74d58b2-Paper-Conference.pdf.

Chang Gao, Chujie Zheng, Xiong-Hui Chen, Kai Dang, Shixuan Liu, Bowen Yu, An Yang, Shuai Bai, Jingren Zhou, and Junyang Lin. Soft adaptive policy optimization. arXiv preprint arXiv:2511.20347, 2025. URL https://arxiv.org/abs/2511.20347v2.

Google DeepMind. Gemini 3.8 Flash. Model card, September 2026. URL https://deepmind.google/models/model-cards/gemini-3-8-flash/. Published September 2, 2026. Accessed September 23, 2026.

Daya Guo, Dejian Yang, Haowei Zhang, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, 2025. doi: 10.1038/s41586-025-09422-z. URL https://doi.org/10.1038/s41586-025-09422-z.

Shuo He, Lang Feng, qi wei, Xin Cheng, Lei Feng, and Bo An. Hierarchy-of-groups policy optimization for long-horizon agentic tasks. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 27572–27593, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/2e7eb932f93d5db3b61c7a001e72fd0f-Paper-Conference.pdf.

Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R Narasimhan. SWE-bench: Can language models resolve real-world GitHub issues? In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=VTF8yNQM66.

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 *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=Rwhi91ideu.

Jiazheng Li, Yawei Wang, Qiaojing Yan, Yijun Tian, Zhichao Xu, Huan Song, Panpan Xu, and Lin Lee Cheong. SALT: Step-level advantage assignment for long-horizon agents via trajectory graph. In Vera Demberg, Kentaro Inui, and Lluís Marquez (eds.), *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 4709–4725, Rabat, Morocco, March 2026a. Association for Computational Linguistics. ISBN 979-8-89176-386-9. doi: 10.18653/v1/2026.findings-eacl.247. URL https://aclanthology.org/2026.findings-eacl.247/.

Wanli Li, Bince Qu, Bo Pan, Jianyu Zhang, Zheng Liu, Pan Zhang, Wei Chen, and Bo Zhang. LiteResearcher: A scalable agentic RL training framework for deep research agent. In *Third Conference on Language Modeling*, 2026b. URL https://openreview.net/forum?id=6qDucSpMvz.

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 *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=5PAF7PAY2Y.

Michael Luo, Naman Jain, Jaskirat Singh, Sijun Tan, Ameen Patel, Qingyang Wu, Alpay Ariyak, Colin Cai, Tarun Venkat, Shang Zhu, Ben Athiwaratkun, Manan Roongta, Ce Zhang, Li Erran Li, Raluca Ada Popa, Koushik Sen, and Ion Stoica. DeepSWE: Training a fully open-sourced, state-of-the-art coding agent by scaling RL. Together AI and Agentica research blog, 2025. URL https://www.together.ai/blog/deepswe.

MiniMax et al. MiniMax-M1: Scaling test-time compute efficiently with lightning attention. arXiv preprint arXiv:2506.13585, 2025. URL https://arxiv.org/abs/2506.13585v1.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744. Curran Associates, Inc., 2022. doi: 10.52202/068431-2011. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.

Pranav Putta, Edmund Mills, Naman Garg, Sumeet Motwani, Chelsea Finn, Divyansh Garg, and Rafael Rafailov. Agent Q: Advanced reasoning and learning for autonomous AI agents. arXiv preprint arXiv:2408.07199, 2024. URL https://arxiv.org/abs/2408.07199v1.

Zehan Qi, Xiao Liu, Iat Long Iong, Hanyu Lai, Xueqiao Sun, Jiadai Sun, Xinyue Yang, Yu Yang, Shuntian Yao, Wei Xu, Jie Tang, and Yuxiao Dong. WebRL: Training LLM web agents via self-evolving online curriculum reinforcement learning. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 79791–79821, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/c66e1fcc9691aae706250638f36f681b-Paper-Conference.pdf.

Zhisong Qiu, Shuofei Qiao, Kewei Xu, Yuqi Zhu, Lun Du, Ningyu Zhang, and Huajun Chen. Rewarding the scientific process: Process-level reward modeling for agentic data analysis. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, pp. 11866–11877. ACM, 2026. doi: 10.1145/3770855.3819049. URL https://doi.org/10.1145/3770855.3819049.

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. arXiv preprint arXiv:2412.15115, 2025. URL https://arxiv.org/abs/2412.15115v2.

Qwen Team. Qwen3.8-Max: A new bar for coding and cowork, August 2026. URL https://qwen.ai/blog?id=qwen3.8.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 53728–53741. Curran Associates, Inc., 2023. doi: 10.52202/075280-2338. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/a85b405ed65c6477a4fe8302b5e06ce7-Paper-Conference.pdf.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. In *International Conference on Learning Representations*, 2016. URL https://arxiv.org/abs/1506.02438.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347, 2017. URL https://arxiv.org/abs/1707.06347v2.

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. URL https://arxiv.org/abs/2402.03300v3.

Guobin Shen, Chenxiao Zhao, Xiang Cheng, Lei Huang, and Xing Yu. VESPO: Variational sequence-level soft policy optimization for stable off-policy LLM training. arXiv preprint arXiv:2602.10693, 2026. URL https://arxiv.org/abs/2602.10693v3.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. HybridFlow: A flexible and efficient RLHF framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, pp. 1279–1297. ACM, 2025. doi: 10.1145/3689031.3696075. URL https://doi.org/10.1145/3689031.3696075.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: language agents with verbal reinforcement learning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 8634–8652. Curran Associates, Inc., 2023. doi: 10.52202/075280-0377. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/1b44b878bb782e6954cd888628510e90-Paper-Conference.pdf.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. ALFWorld: Aligning text and embodied environments for interactive learning. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2010.03768v2.

Yifan Song, Da Yin, Xiang Yue, Jie Huang, Sujian Li, and Bill Yuchen Lin. Trial and error: Exploration-based trajectory optimization of LLM agents. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7584–7600, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.409. URL https://aclanthology.org/2024.acl-long.409/.

Richard S Sutton and Andrew G Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2nd edition, 2018. ISBN 9780262039246. URL https://mitpress.mit.edu/9780262039246/reinforcement-learning/.

Hui-Ze Tan, Xiao-Wen Yang, Hao Chen, Jie-Jing Shao, Yi Wen, Yuteng Shen, Weihong Luo, Xiku Du, Lan-Zhe Guo, and Yu-Feng Li. Hindsight credit assignment for long-horizon LLM agents. arXiv preprint arXiv:2603.08754, 2026. URL https://arxiv.org/abs/2603.08754v1.

Leitian Tao, Baolin Peng, Wenlin Yao, Tao Ge, Hao Cheng, Mike Hang Wang, Jianfeng Gao, and Sharon Li. TRACE: Turn-level reward assignment via credit estimation for long-horizon agents. arXiv preprint arXiv:2607.13988, 2026. URL https://arxiv.org/abs/2607.13988v1.

Tongyi DeepResearch Team et al. Tongyi DeepResearch technical report. arXiv preprint arXiv:2510.24701, 2026. URL https://arxiv.org/abs/2510.24701v3.

Haixin Wang, Hejie Cui, Chenwei Zhang, Xin Liu, Shuowei Jin, Shijie Geng, Xinyang Zhang, Nasser Zalmout, Zhenyu Shi, and Yizhou Sun. T^2PO: Uncertainty-guided exploration control for stable multi-turn agentic reinforcement learning. In *Forty-third International Conference on Machine Learning*, 2026a. URL https://openreview.net/forum?id=aD1zjvdJN4.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Jirong Wen. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6): 186345, 2024. doi: 10.1007/s11704-024-40231-1. URL https://doi.org/10.1007/s11704-024-40231-1.

Xiaoxuan Wang, Han Zhang, Haixin Wang, Yidan Shi, Ruoyan Li, Kaiqiao Han, Chenyi Tong, Haoran Deng, Alexander K Taylor, Renliang Sun, Yanqiao Zhu, Jason Cong, Yizhou Sun, and Wei Wang. ARLArena: A unified framework for stable agentic reinforcement learning. In *Forty-third International Conference on Machine Learning*, 2026b. URL https://openreview.net/forum?id=90kxFi9VGP.

Yinjie Wang, Xuyang Chen, Xiaolong Jin, Mengdi Wang, and Ling Yang. OpenClaw-RL: Train any agent simply by talking. arXiv preprint arXiv:2603.10165, 2026c. URL https://arxiv.org/abs/2603.10165v2.

Yunan Wang, Minghui Song, Zihan Zhang, Shaohan Huang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun, and Qi Zhang. Group-graph policy optimization for long-horizon agentic reinforcement learning. arXiv preprint arXiv:2606.22995, 2026d. URL https://arxiv.org/abs/2606.22995v1.

Zihan Wang, Kangrui Wang, Qineng Wang, Pingyue Zhang, Linjie Li, Zhengyuan Yang, Xing Jin, Kefan Yu, Minh Nhat Nguyen, Licheng Liu, Eli Gottlieb, Yiping Lu, Kyunghyun Cho, Jiajun Wu, Li Fei-Fei, Lijuan Wang, Yejin Choi, and Manling Li. RAGEN: Understanding self-evolution in LLM agents via multi-turn reinforcement learning. arXiv preprint arXiv:2504.20073, 2025. URL https://arxiv.org/abs/2504.20073v2.

Zihan Wang, Chi Gui, Xing Jin, Qineng Wang, Licheng Liu, Kangrui Wang, Shiqi Chen, Linjie Li, Zhengyuan Yang, Pingyue Zhang, Yiping Lu, Jiajun Wu, Li Fei-Fei, Lijuan Wang, Yejin Choi, and Manling Li. RAGEN-2: Reasoning collapse in agentic RL. In *Forty-third International Conference on Machine Learning*, 2026e. URL https://openreview.net/forum?id=01caH9oj7C.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. WebShop: Towards scalable real-world web interaction with grounded language agents. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 20744–20757. Curran Associates, Inc., 2022. doi: 10.52202/068431-1508. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/82ad13ec01f9fe44c01cb91814fd7b8c-Paper-Conference.pdf.

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

Weiran Yao, Shelby Heinecke, Juan Carlos Niebles, Zhiwei Liu, Yihao Feng, Le Xue, Rithesh Ramapura Narasimha Murthy, Zeyuan Chen, Jianguo Zhang, Devansh Arpit, Ran Xu, Phil Mui, Huan Wang, Caiming Xiong, and Silvio Savarese. Retroformer: Retrospective large language agents with policy gradient optimization. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 10091–10111, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/29f421fbdcc82aeb349d784d3aaccdb3-Paper-Conference.pdf.

Deheng Ye, Zhao Liu, Mingfei Sun, Bei Shi, Peilin Zhao, Hao Wu, Hongsheng Yu, Shaojie Yang, Xipeng Wu, Qingwei Guo, Qiaobo Chen, Yinyuting Yin, Hao Zhang, Tengfei Shi, Liang Wang, Qiang Fu, Wei Yang, and Lanxiao Huang. Mastering complex control in MOBA games with deep reinforcement learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34 (04):6672–6679, 2020. doi: 10.1609/aaai.v34i04.6144. URL https://doi.org/10.1609/aaai.v34i04.6144.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Juncai Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Ru Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Yonghui Wu, and Mingxuan Wang. DAPO: An open-source LLM reinforcement learning system at scale. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, pp. 113222–113244. Curran Associates, Inc., 2025. doi: 10.52202/085713-3775. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/a4277440d50f1f15d2cb4c14f7e0c0d2-Paper-Conference.pdf.

Jiazhen Yuan, Zhike Gong, Jinquan Hang, Zhengbiao Bai, and Wei Zhao. Graph-enhanced policy optimization in LLM agent training. arXiv preprint arXiv:2510.26270, 2026. URL https://arxiv.org/abs/2510.26270v2.

Chenchen Zhang. From reasoning to agentic: Credit assignment in reinforcement learning for large language models. arXiv preprint arXiv:2604.09459, 2026. URL https://arxiv.org/abs/2604.09459v3.

Guibin Zhang, Hejia Geng, Xiaohang Yu, Zhenfei Yin, Zaibin Zhang, Zelin Tan, Heng Zhou, Zhong-Zhi Li, Xiangyuan Xue, Yijiang Li, Yifan Zhou, Yang Chen, Chen Zhang, Yutao Fan, Zihu Wang, Songtao Huang, Francisco Piedrahita Velez, Yue Liao, Hongru Wang, Mengyue Yang, Heng Ji, Jun Wang, Shuicheng Yan, Philip Torr, and Lei Bai. The landscape of agentic reinforcement learning for LLMs: A survey. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=RY19y2RI1O. Survey Certification.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. Group sequence policy optimization. arXiv preprint arXiv:2507.18071, 2025a. URL https://arxiv.org/abs/2507.18071v2.

Yuxiang Zheng, Dayuan Fu, Xiangkun Hu, Xiaojie Cai, Lyumanshan Ye, Pengrui Lu, and Pengfei Liu. DeepResearcher: Scaling deep research via reinforcement learning in real-world environments. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 414–431, Suzhou, China, November 2025b. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.22. URL https://aclanthology.org/2025.emnlp-main.22/.

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 B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 15585–15606, 2024a. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/4410c0711e9154a7a2d26f9b3816d1ef-Paper-Conference.pdf.

Yifei Zhou, Andrea Zanette, Jiayi Pan, Sergey Levine, and Aviral Kumar. ArCHer: Training language model agents via hierarchical multi-turn RL. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 62178–62209. PMLR, 21–27 Jul 2024b. URL https://proceedings.mlr.press/v235/zhou24t.html.

Yifei Zhou, Song Jiang, Yuandong Tian, Jason Weston, Sergey Levine, Sainbayar Sukhbaatar, and Xian Li. SWEET-RL: Training multi-turn LLM agents on collaborative reasoning tasks. arXiv preprint arXiv:2503.15478, 2025. URL https://arxiv.org/abs/2503.15478v1.

A NOTATION

Table S1: Notation for CCPO. Subscripts index trajectories and steps; H and F denote historical and future channels. Implementation details appear in Appendix C.

Symbol	Definition
Tasks, trajectories, and temporal indexing	
$x,\ x_i$	Task instance; x_i is the task associated with trajectory i.
N	Number of trajectories sampled per task in a rollout batch.
τ_i	Trajectory i, containing observations $o_{i,0:T_i}$, actions $a_{i,0:T_i-1}$, and their outcomes.
$i,\ j$	Query-trajectory and peer-trajectory indices.
$t,\ s$	Zero-based query-step and peer-step indices: $0 \le t < T_i$, $0 \le s < T_j$.
T_i	Number of actions in trajectory i; T_i also indexes its terminal endpoint.

Continued on next page

Table S1 continued

Symbol	Definition
$o_{i,t}$	Observation before action $a_{i,t}$; matching uses exact string equality.
$a_{i,t}$	Generated response, including reasoning and the executable action.
$p_{i,t}$	Task, current observation, and up to K_{H} preceding observation–action pairs; excludes the current response.
$r^{\mathrm{env}}_{i,t},\ R_i$	Environment reward after action $a_{i,t}$, and terminal trajectory outcome.
$\gamma,\ \eta$	Return discount factor and current invalid-action penalty coefficient.
$K_{\mathrm{H}},\ K_{\mathrm{F}}$	Prompt-history length and realized future horizon, both 2 by default.
t^+	Future endpoint on the collected trajectory: $t^+ = \min(t + K_{\mathrm{F}}, T_i)$.
$H,\ F,\ W$	Ablation aliases: $H = K_{\mathrm{H}}$, $F = K_{\mathrm{F}}$, $W = H + F + 1$.
Context representation	
$h_{i,t}$	Frozen reference model's final-layer hidden state at the last prompt token.
$\bar{h},\ P_{\perp,3}$	Batch mean hidden state and projection removing its three leading principal directions.
$n_{i,t}$	Distinct-observation count: $\lvert\{o_{i,0}, \ldots, o_{i,t}\}\rvert$.
$n_{i,t}/(t+1)$	Observation-novelty ratio over the observed prefix.
$\mathbf{1}\{\mathrm{revisit}_{i,t}\}$	Whether the current observation appeared earlier in the same trajectory.
$c_{i,t}$	37-dimensional encoding of elapsed steps, distinct observations, novelty ratio, and revisit indicator.
$\nu(v)$	L2 normalization: $v/\max(\lVert v\rVert_2, 10^{-9})$.
$\phi_{i,t}$	Fused context feature; see Equation 9.
$[u; v]$	Vector concatenation.
Peer matching and contextual weighting	
$\mathcal{E}_{i,t}$	Exact task–observation peers outside the query trajectory; Equation 2.
$\mathcal{P}_{i,t}$	Peers used for readout: $\mathcal{E}_{i,t}$ when nonempty, otherwise all same-task nonterminal steps with $j \neq i$.
$u = (i,t),\ v = (j,s)$	Shorthand for query and peer step identifiers.
$w_{u,v}$	Kernel weight: $\exp(-\lVert\phi_u - \phi_v\rVert_2/\tau_B)$.
$\bar{w}_{u,v}$	Query-normalized weight: $w_{u,v}/\sum_{v' \in \mathcal{P}_{i,t}} w_{u,v'}$.
B	Observation or task bucket used to determine the bandwidth.
$m_B,\ \widetilde{m}_B$	Median pairwise distance in B and its safeguarded value; Appendix C.2.
τ_B	Kernel bandwidth $0.15\widetilde{m}_B$, distinct from trajectory τ_i.
$r_{j,s}$	Readout target: $Y_{j,s}$ for return baselines or $Z_{j,s}$ for outcome values.
$\mathcal{B}_{i,t}[r]$	Contextual weighted readout; Equation 4.
Temporal credit signals	
$Y_{i,t}$	Historical return target including the current invalid-action penalty; Equation 1.
$\mathcal{B}_{i,t}[Y]$	Contextual peer baseline for the realized return.
$C^{\mathrm{H}}_{i,t}$	Historical credit: $Y_{i,t} - \mathcal{B}_{i,t}[Y]$.
$Z_{i,s}$	Unpenalized outcome target: $\gamma^{T_i - s} R_i$.
$V_{i,s}$	Contextual outcome value: $\mathcal{B}_{i,s}[Z]$ for $s < T_i$, and R_i at $s = T_i$.
$\Delta V_{i,t}$	Outcome-value change: $V_{i,t^+} - V_{i,t}$.
$C^{\mathrm{F}}_{i,t}$	Task-standardized future credit: $z_{x_i}[\Delta V]_{i,t}$.
Support, normalization, and advantages	
$\mathcal{S}^{\mathrm{H}}_x$	Unique steps of task x with a defined historical baseline.
$\mathcal{S}^{\mathrm{F}}_x$	Unique task-x steps with both outcome-value endpoints defined.
$\mathcal{S}_x,\ \widehat{\mathcal{S}}_x$	$\mathcal{S}_x = \mathcal{S}^{\mathrm{H}}_x \cup \mathcal{S}^{\mathrm{F}}_x$; $\widehat{\mathcal{S}}_x$ is the corresponding multiset of restored training rows.
$\mathcal{R}_x$	All restored input step rows of task x, used for episode statistics.
$\bar{u}_S,\ s_S(u)$	Mean and sample standard deviation over set or row multiset S.
$z_S[u]$	Standardization: $(u - \bar{u}_S)/(s_S(u) + 10^{-6})$ for $\lvert S\rvert > 1$.
$z_x[\Delta V]$	Future-credit standardization over $\mathcal{S}^{\mathrm{F}}_x$.
$\mathcal{N}_x[\cdot]$	Final contextual transformation.
$A_{\mathrm{CC},i,t}$	Contextual advantage $\mathcal{N}_{x_i}[C^{\mathrm{H}} + C^{\mathrm{F}}]_{i,t}$, also denoted $A^{\mathrm{CC}}_{i,t}$.

Continued on next page

Table S1 continued

Symbol	Definition
$A^{\mathrm{H}}_{i,t}$	Historical contribution after the final transformation; Equation 11.
$A^{\mathrm{F}}_{i,t}$	Future contribution after the same transformation; $A_{\mathrm{CC}} = A^{\mathrm{H}} + A^{\mathrm{F}}$.
$\mu^d_x,\ \sigma_x$	Component mean and common scale for the additive decomposition; Appendix C.3.
$e_{i,t}$	Episode-channel score: $R_i - \eta \mathbf{1}\{a_{i,t}$ invalid$\}$.
$A_{\mathrm{EP},i,t}$	Episode-score advantage over $\mathcal{R}_x$.
α_{EP}	Episode-channel weight: 0 by default, 1 for the augmented variant.
$A_{i,t}$	Optimizer advantage: $A_{\mathrm{CC},i,t} + \alpha_{\mathrm{EP}} A_{\mathrm{EP},i,t}$.
Policy optimization	
$\theta,\ \pi_\theta$	Trainable policy parameters and the current LLM policy.
π_{old}	Behavior policy used to collect rollouts.
π_{ref}	Frozen reference policy for feature extraction and KL regularization.

B TRAINING AND ABLATION CONFIGURATIONS

Table S2: Training configurations shared by 1.5B and 7B models on ALFWorld and WebShop.

Setting	GRPO	GiGPO	HGPO	G2PO	CCPO
Both benchmarks					
Training iterations	150	150	150	150	150
Tasks × rollouts per iteration	16×8	16×8	16×8	16×8	16×8
Total sampled trajectories	19,200	19,200	19,200	19,200	19,200
NVIDIA H200 GPUs per run	8	8	8	8	8
Learning rate	10^{-6}	10^{-6}	10^{-6}	10^{-6}	10^{-6}
Discount factor	0.95	0.95	0.95	0.95	0.95
Actor KL coefficient	0.01	0.01	0.01	0.01	0.01
KL estimator	`low_var_kl`	`low_var_kl`	`low_var_kl`	`low_var_kl`	`low_var_kl`
KL included in reward	No	No	No	No	No
Invalid-action penalty coeff.	0.1	0.1	0.1	0.1	0.1
Environment training seed	0	0	0	0	0
Prompt history (steps)	2	2	2	2	2
Response token limit / step	512	512	512	512	512
Prompt truncation	`error`	`error`	`left`	`error`	`error`
Validation decoding	Sampling	Sampling	Sampling	Sampling	Sampling
Validation temperature	0.4	0.4	0.4	0.4	0.4
Validation interval (iterations)	5	5	5	5	5
ALFWorld					
PPO mini-batch size	256	256	256	256	256
Maximum interaction steps	50	50	50	50	50
Prompt token limit	2,048	2,048	4,096	2,048	2,048
WebShop					
PPO mini-batch size	64	64	64	64	64
Maximum interaction steps	15	15	30	15	15
Prompt token limit	4,096	4,096	4,096	4,096	4,096

Table S3: CCPO variant and ablation configurations. Each ablation changes only the listed settings relative to its corresponding reference configuration.

Variant	Configuration change
$A_{\mathrm{CC}} + A_{\mathrm{EP}}$	Add episode advantage with unit weight after contextual normalization, without renormalizing the sum.
w/o self-trajectory peer exclusion	Allow matching steps from the query trajectory, including the query itself, as peers.
w/o context summary	Set the accumulated-statistics weight from 1 to 0; retain frozen hidden-state features and prompt history.
Cosine-based context similarity	Replace the exponential kernel with cosine-based weighting.
Historical credit only	Set the future-credit weight from 1 to 0; retain historical credit with unit weight.
Future credit only	Set the historical-credit weight from 1 to 0; retain future credit with unit weight and horizon 2.
Temporal window variation	Vary the prompt-history length and future horizon from the default $(K_{\mathrm{H}}, K_{\mathrm{F}}) = (2, 2)$. A zero window disables that credit channel.

C CCPO IMPLEMENTATION DETAILS

Using complete, globally gathered rollouts, we remove identical step copies before feature construction, peer readouts, and future-credit standardization. Original rows are restored for final advantage computation before PPO minibatching. Training and ablation configurations are in Appendix B.

C.1 CONTEXT REPRESENTATION

The frozen reference model supplies the final-layer hidden state at the last prompt token. Across unique steps in the batch, we center these states, remove their three leading principal directions, and L2-normalize. The summary encodes elapsed steps t, distinct-observation count $n_{i,t}$, and novelty ratio $n_{i,t}/(t+1)$ in three 12-dimensional thermometer blocks with upper bounds 30, 25, and 1. For value v and bound b, the first $\lfloor 12\,\mathrm{clip}(v/b, 0, 1)\rfloor$ coordinates are one; the rest are zero. Appending the binary revisit indicator gives $c_{i,t} \in \mathbb{R}^{37}$. With $\nu(v) = v/\max(\|v\|_2, 10^{-9})$, the fused feature is

$$\phi_{i,t} = \nu\big([\nu(P_{\perp,3}(h_{i,t} - \bar{h})); \nu(c_{i,t})]\big)\,. \tag{9}$$

The projection is recomputed per batch without reward labels. With two or three unique steps, we skip principal-direction removal; with one, we use the raw hidden block. Nonterminal endpoints use their own prompt features and accumulated statistics.

C.2 PEER READOUTS

Readouts use $\mathcal{P}_{i,t}$ from Section 3.2; an empty set is unsupported. For the full observation bucket B, let m_B be the median distance over distinct unordered pairs before trajectory exclusion. Equation 3 uses $\tau_B = 0.15\widetilde{m}_B$, with $\widetilde{m}_B = m_B$ when $m_B > 10^{-9}$ and 1 otherwise. Task-level fallback recomputes this bandwidth in the full task bucket.

Repeated visits retain individual weights. Each peer trajectory contributes its weighted mean, weighted again by its total kernel mass. Both denominators are floored at 10^{-12}; all-zero weights yield zero while retaining support. Away from these floors, this implements Equation 4.

C.3 CREDIT CONSTRUCTION AND NORMALIZATION

With terminal outcomes $R_i \in \{0, 10\}$, the supplied historical target is $Y_{i,t} = \gamma^{T_i-1-t}R_i - \eta\mathbf{1}\{a_{i,t} \text{ invalid}\}$, whereas future values use $Z_{i,t} = \gamma^{T_i-t}R_i$ without penalties. The current-step penalty enters Y once. At $t^+ = \min(t + K_{\mathrm{F}}, T_i)$, terminal values equal R_i, including zero for unsuccessful rollouts. Both value endpoints must be defined; their difference has no additional discount factor.

Let $h = t^+ - t$. Equation 7 can be written as

$$\Delta V_{i,t} = \big(\gamma^h V_{i,t^+} - V_{i,t}\big) + (1 - \gamma^h)V_{i,t^+}. \tag{10}$$

Our value targets are nonnegative, so the second term is nonnegative and can vary across steps. This time-dependent contribution can persist after task-wise centering.

Standardization uses sample standard deviation plus 10^{-6}, with equal weight per included row. Future differences are standardized over supported unique steps within each task in both benchmarks; fewer than two eligible steps give zero future credit. Unavailable components are zero-filled before combination. Unsupported contextual advantages remain zero. All-zero outcomes give zero future credit; historical targets still include local penalties.

For the plotted ALFWorld contributions, let μ_x^d be the mean of C^d over the restored union support and $\sigma_x = s_x(C^{\mathrm{H}} + C^{\mathrm{F}}) + 10^{-6}$. For multiple supported rows,

$$A_{i,t}^d = \frac{C_{i,t}^d - \mu_{x_i}^d}{\sigma_{x_i}}, \qquad d \in \{\mathrm{H}, \mathrm{F}\}, \qquad A_{\mathrm{CC}} = A^{\mathrm{H}} + A^{\mathrm{F}}. \tag{11}$$

Both contributions are zero outside the union support. These additive contributions share a normalization scale and sum to A_{CC}.

The optional episode channel uses $e_{i,t} = R_i - \eta \mathbf{1}\{a_{i,t} \text{ invalid}\}$ over all restored task rows, including contextually unsupported steps. Hence longer trajectories contribute more rows, and local penalties can produce different episode advantages within a trajectory. The channel is added with unit weight after contextual normalization, without renormalizing.

C.4 Policy-Update Integration

The adapter broadcasts each detached step advantage to all generated response tokens, including reasoning and executable actions; prompt and padding tokens are excluded. The loss mask retains valid response tokens from steps with zero contextual credit. The dual-clipped policy objective (Ye et al., 2020) follows Section 3.4. Features and credit estimates receive no gradients. Estimation uses float64 NumPy arithmetic, while the adapter extracts and returns detached float32 features and token advantages.

For a response token y with preceding context c, `low_var_kl` uses $\ell = \mathrm{clip}(\log \pi_{\mathrm{ref}}(y \mid c) - \log \pi_\theta(y \mid c), -20, 20)$ and the penalty $\mathrm{clip}(e^\ell - \ell - 1, -10, 10)$. The penalty is aggregated with the response-token mask and added to the policy loss with coefficient 0.01.

Algorithm 1 CCPO training iteration

1 Collect complete rollouts with π_{old}; gather metadata and reward targets.
2 Extract frozen prompt features; retain unique steps and the input-row mapping.
3 Encode prefix statistics; construct ϕ by Equation 9.
4 Select exact peers or task fallback; compute $\mathcal{B}[Y]$ and $V = \mathcal{B}[Z]$.
5 Set $t^+ \leftarrow \min(t + K_{\mathrm{F}}, T_i)$ and terminal values $V_{i,T_i} \leftarrow R_i$.
6 Set $C^{\mathrm{H}}_{i,t} \leftarrow Y_{i,t} - \mathcal{B}_{i,t}[Y]$ and $C^{\mathrm{F}}_{i,t} \leftarrow z_{x_i}[\Delta V]_{i,t}$, where $\Delta V_{i,t} = V_{i,t^+} - V_{i,t}$, on eligible steps.
7 Zero-fill unavailable terms; restore rows; set $A_{\mathrm{CC}} \leftarrow \mathcal{N}_x[C^{\mathrm{H}} + C^{\mathrm{F}}]$.
8 Set $A \leftarrow A_{\mathrm{CC}} + \alpha_{\mathrm{EP}} A_{\mathrm{EP}}$, $\alpha_{\mathrm{EP}} \in \{0, 1\}$; pass this sum directly to the token adapter.
9 Broadcast detached A to response tokens; update the policy with PPO and KL.

D VALIDATION ROLLOUT EXAMPLES

D.1 ALFWORLD: PICKING UP A SOAPBAR

Table S4: ALFWorld prompt and response excerpts at step 6 (Qwen2.5-1.5B-Instruct).

Component	Excerpt
Prompt	
Role	You are an expert agent operating in the ALFRED Embodied Environment.
Task	Your task is to: put two soapbar in garbagecan.
History	Prior to this step, you have already taken 5 step(s). Below are the most recent 2 observations and the corresponding actions you took:
History excerpt	[Observation 4: 'You arrive at cabinet 1. The cabinet 1 is closed.', Action 4: 'open cabinet 1']
History excerpt	[Observation 5: 'You open the cabinet 1. The cabinet 1 is open. In it, you see nothing.', Action 5: 'go to sinkbasin 1']
Current step	You are now at step 6 and your current observation is: You arrive at sinkbasin 1. On the sinkbasin 1, you see a soapbar 3, and a soapbar 2.
Actions	Your admissible actions of the current situation are: ['examine sinkbasin 1', [...] 'take soapbar 2 from sinkbasin 1', 'take soapbar 3 from sinkbasin 1'].
Instructions	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.
Model response	

Response	`<think>`
	The agent has arrived at the sinkbasin 1 and has two soapbars available: soapbar 2 and soapbar 3. The agent needs to put one of the soapbars in the garbagecan 1.
	`</think>`
	`<action>take soapbar 2 from sinkbasin 1</action>`

E Statistical Analysis

E.1 Relative Credit Preferences

Data and filtering. For Figure 4, we analyze ALFWorld training rollouts from iterations 101–150 of four Qwen2.5-1.5B-Instruct runs: full CCPO and CCPO without the accumulated context summary, both with $(K_{\mathrm{H}}, K_{\mathrm{F}}) = (2, 2)$, together with the asymmetric $(4, 0)$ and $(0, 4)$ window variants of Table 2. All use the A_{CC}-only configuration. Table S5 specifies the temporal windows and sample counts. We retain successful trajectories, identified by a recorded episode return greater than 0.5, and interior steps satisfying $t \geq 2$ and $t_{\mathrm{last}} - t \geq 2$, where t is the zero-based logged step index and t_{last} is the last logged step in the trajectory. We apply this interior-step filter only to the analysis, consistently across all four runs. Padding duplicates are removed upstream; trajectories with missing episode returns are excluded. Each run is analyzed using its own rollouts, so comparisons across runs also reflect differences in the sampled behavior.

Table S5: Filtered samples for the credit-preference analysis. Groups match training iteration, task instance, and observation. M counts categories with at least 60 usable groups; each contrast requires both the category and other actions in a group.

Run	$(K_{\mathrm{H}}, K_{\mathrm{F}})$	Steps	Groups	M
CCPO (full)	$(2, 2)$	39,503	11,003	10
w/o context summary	$(2, 2)$	38,680	9,778	9
Asymmetric (history)	$(4, 0)$	40,972	9,914	10
Asymmetric (future)	$(0, 4)$	49,736	13,235	11

Action categories. The action at step t is classified from the environment response recorded at step $t + 1$, without using credit values. Fixed response-prefix rules identify *task operations* (clean, heat, or cool), *take*, *examine*, *new location*, *revisit*, *open*, *put*, *toggle*, *invalid*, *close*, and *other*. Arrivals are classified as new locations or revisits using the locations previously visited anywhere in the trajectory, including steps subsequently removed by filtering. Invalid actions are identified by the environment's "Nothing happens." response. Task completion may require revisiting a location.

Matched-group contrast. A comparison group B contains steps from the same training iteration and task instance with identical observation text. Only exact task-observation groups enter this analysis. For action category c, let $I_B(c)$ contain steps of that category and $O_B(c)$ contain the other steps in B. We retain groups with both sets nonempty and, for a step-level quantity m_u, define

$$
\begin{aligned}
d_B(c; m) &= \frac{1}{|I_B(c)|} \sum_{u \in I_B(c)} m_u - \frac{1}{|O_B(c)|} \sum_{u \in O_B(c)} m_u, \\
\Delta(c; m) &= \frac{1}{|\mathcal{U}(c)|} \sum_{B \in \mathcal{U}(c)} d_B(c; m),
\end{aligned}
\tag{12}
$$

where $\mathcal{U}(c)$ is the set of usable groups. We apply this definition to A^{H}, A^{F}, and A_{CC}, using the additive decomposition in Equation 11. Groups receive equal weight regardless of their size. A positive contrast indicates higher average group-relative credit for the category. Individual advantages may have either sign, and the contrast describes an association between action category and credit.

Bootstrap intervals and channel comparison. For each run, categories with fewer than 60 usable groups are excluded. We draw 6,000 bootstrap samples of usable groups with replacement, using random seed 0. Two-sided percentile intervals use quantile levels $0.05/(2M)$ and $1 - 0.05/(2M)$, where M is the number of eligible categories in that run. Asterisks indicate intervals excluding zero. Each combination of run, analysis window, and measured quantity defines a separate Bonferroni correction family.

The channel comparison is paired within groups: we bootstrap $d_B(c; A^{\mathrm{F}}) - d_B(c; A^{\mathrm{H}})$ using the same sampled groups for both terms. For new-location transitions in full CCPO, the mean paired

difference is 0.0372 with adjusted interval $[0.0011, 0.0727]$, based on 2,804 groups. The paired-contrast interval excludes zero; both channel-specific intervals include zero.

The analysis is conditional on successful trajectories and observations at which multiple action categories were sampled.

E.2 Paired Action Timing During Evaluation

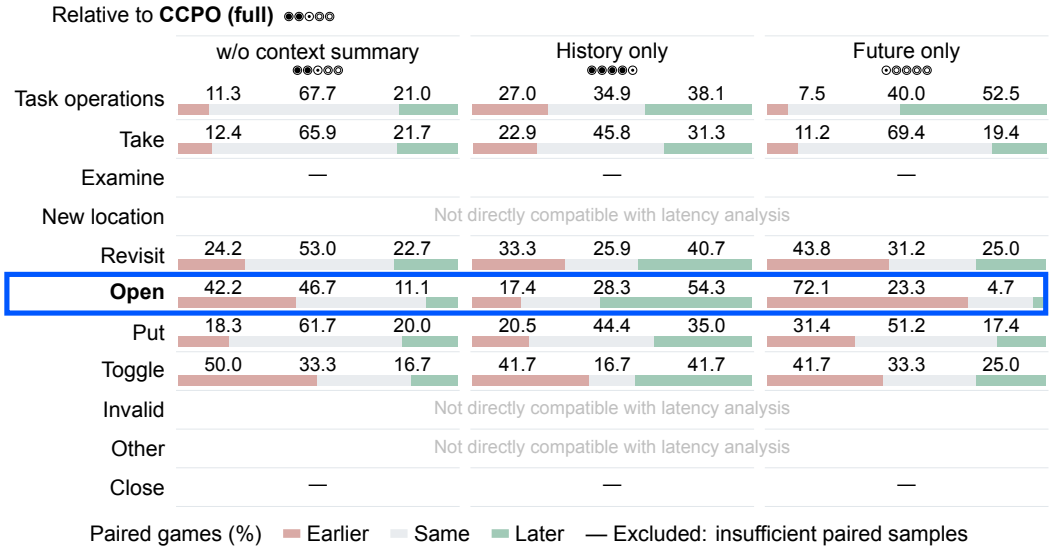

Figure S1: Action timing on ALFWorld (Qwen2.5-1.5B-Instruct). First occurrence at checkpoint 150, paired across the same 140 evaluation games against full CCPO. Percentages condition on both policies performing the action. Dashes: fewer than 12 pairs or unavailable results; gray rows are excluded. Context markers denote ⊙ current, ◉ history, and ◎ future.

Evaluation population. For Figure S1, we use evaluation rollouts from checkpoints at training iteration 150, collected following the main evaluation protocol. Each ablation is paired with full CCPO on the same 140 ALFWorld tasks using their evaluation-game identifiers. This analysis includes successful and failed episodes.

First occurrence. For game g and category c, let $t_{g,c}$ denote the one-based step of its first occurrence, identified by the response-based categories above. The paired timing difference is

$$\Delta t_{g,c} = t_{g,c}^{\text{ablation}} - t_{g,c}^{\text{full}}. \tag{13}$$

Only games where both policies perform the action are eligible. The plotted *earlier*, *same*, and *later* percentages count pairs with $\Delta t_{g,c} < 0$, $= 0$, and > 0, respectively. Percentages are computed over eligible game pairs within each category. Games where only one policy performs the action enter separate coverage counts and are excluded from these percentages.

Displayed categories. Bars report descriptive proportions; dashes mark fewer than 12 eligible pairs or unavailable results. New location is excluded: its first occurrence is fixed at step 1 under this labeling rule. Invalid and other are also excluded.

E.3 Changes in Peer Contributions

Paired representation comparison. For Table 3, we use full CCPO's Qwen2.5-1.5B-Instruct ALFWorld training rollouts from iterations 101–150 to recompute normalized peer weights with and without the accumulated context summary. Rollouts, queries, recorded peer sets, trajectory

exclusions, and return targets are held fixed. The full representation has 1,573 dimensions; the comparison retains its 1,536-dimensional hidden-state block and renormalizes that block to unit length. The bandwidth rule is unchanged, but its numerical value is recomputed from distances in each representation. This comparison isolates the effect on weights for fixed rollouts; it is separate from comparing independently trained ablation policies.

Population and statistic. We retain the recorded peer-pool selection, including task-only fallback pools where used. This diagnostic covers successful and failed rollouts and retains boundary steps. It covers 99,424 queries and 1,678,986 query–peer pairs. For each pair, define the absolute change in normalized weight, measured in percentage points, as

$$D_{uv} = 100 \left| \bar{w}_{u,v}^{\text{full}} - \bar{w}_{u,v}^{\text{no-summary}} \right|. \tag{14}$$

For a threshold d, the reported percentage is

$$P(d) = \frac{100}{|\mathcal{Q}|} \sum_{(u,v)\in\mathcal{Q}} \mathbf{1}\{D_{uv} > d\}, \tag{15}$$

where $\mathcal{Q}$ is the set of query–peer step pairs (u, v). Pairs receive equal weight, so queries with more peers contribute more terms. Thresholds of 0.25, 0.5, 1, 5, and 10 percentage points yield 62.1%, 48.7%, 33.8%, 8.6%, and 3.6% of pairs, respectively. These are descriptive proportions of the recorded pairs. These proportions quantify weight redistribution over fixed query-peer pairs.

F Failure Mode Comparisons

We compare ALFWorld evaluation trajectories from Qwen2.5-1.5B-Instruct checkpoints at training step 150, with one sampled trajectory per method and task, temperature 0.4, and a 50-step limit. Tables show selected actions and observation excerpts; italicized rows summarize repeated actions.

F.1 Repeated Visits Without Acquiring the Target

Table S6: Repeated navigation versus target acquisition. Task: "put a toiletpaper in toiletpaperhanger." Revisit counts refer to navigation to previously visited locations.

Step	Action	Observation excerpt / summary
GiGPO: failed at 50 steps; 47 revisits		
3	go to toiletpaperhanger 1	On the toiletpaperhanger 1, you see nothing.
4–49	*46 navigation actions among countertop 1, toilet 1, and toiletpaperhanger 1; every returned observation repeats an earlier one.*	
G2PO: failed at 50 steps; 42 revisits		
5	open cabinet 2	In it, you see a soapbottle 1.
8–15	*Two navigation cycles: countertop 1 → cabinet 2 → toilet 1 → toiletpaperhanger 1.*	
HGPO: failed at 50 steps; 44 revisits		
4	go to toiletpaperhanger 1	On the toiletpaperhanger 1, you see nothing.
5–18	*14 navigation actions among toilet 1, countertop 1, and toiletpaperhanger 1; every returned observation repeats an earlier one.*	
CCPO (A_{CC} only): success at step 7; 0 revisits		
3	go to cabinet 1	The cabinet 1 is closed.
4	open cabinet 1	In it, you see [. . .] a toiletpaper 1.
5	take toiletpaper 1 from cabinet 1	You pick up the toiletpaper 1 from the cabinet 1.
6	go to toiletpaperhanger 1	On the toiletpaperhanger 1, you see nothing.
7	move toiletpaper 1 to toiletpaperhanger 1	You move the toiletpaper 1 to the toiletpaperhanger 1.

F.2 Repeating a Completed Subgoal

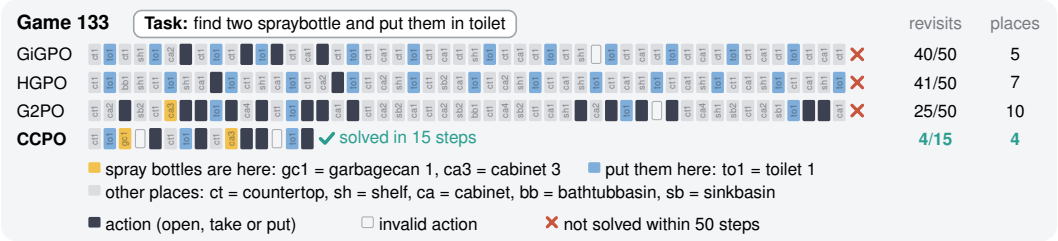

Figure S2: Evaluation trajectories for the two-object task in Table S7. Each cell denotes one step; revisit counts track navigation to previously visited locations. CCPO uses A_{CC} only.

Table S7: Repeated manipulation of one object versus placing two distinct objects. Task: "find two spraybottle and put them in toilet."

Step	Action	Observation excerpt / summary
G2PO: failed at 50 steps		
10	move spraybottle 2 to toilet 1	You move the spraybottle 2 to the toilet 1.
14	go to toilet 1	On the toilet 1, you see a cloth 3, and a spraybottle 2.
15	take spraybottle 2 from toilet 1	You pick up the spraybottle 2 from the toilet 1.
16	move spraybottle 2 to toilet 1	You move the spraybottle 2 to the toilet 1.
37–39, 48–49	*Two further take–replace cycles with spraybottle 2; step 38 is an invalid action.*	
CCPO (A_{CC} only): success at step 15		
4	"1, and you have already taken "	Nothing happens.
5	take spraybottle 1 from garbagecan 1	You pick up the spraybottle 1 from the garbagecan 1.
8	move spraybottle 1 to toilet 1	You move the spraybottle 1 to the toilet 1.
11	open cabinet 3	In it, you see a spraybottle 2.
12	take spraybottle 2 from cabinet 3	You pick up the spraybottle 2 from the cabinet 3.
13	". now, you are at step 13, and"	Nothing happens.
15	move spraybottle 2 to toilet 1	You move the spraybottle 2 to the toilet 1.
GiGPO: failed at 50 steps		
11	take soapbottle 2 from countertop 1	You pick up the soapbottle 2 from the countertop 1.
13	move soapbottle 2 to toilet 1	You move the soapbottle 2 to the toilet 1.
HGPO: failed at 50 steps		
1–50	*Navigation and cabinet opening; no object is picked up.*	

G TRAJECTORY LENGTHS

Figure S3 reports average steps among successful trajectories. At 1.5B, the ALFWorld reference configuration (A_{CC}) averages 8.65 ± 0.08 steps, compared with 5.13 ± 0.11 for the WebShop reference ($A_{\text{CC}} + A_{\text{EP}}$). WebShop's shorter successful trajectories motivate restricting the context-window variations to ALFWorld.

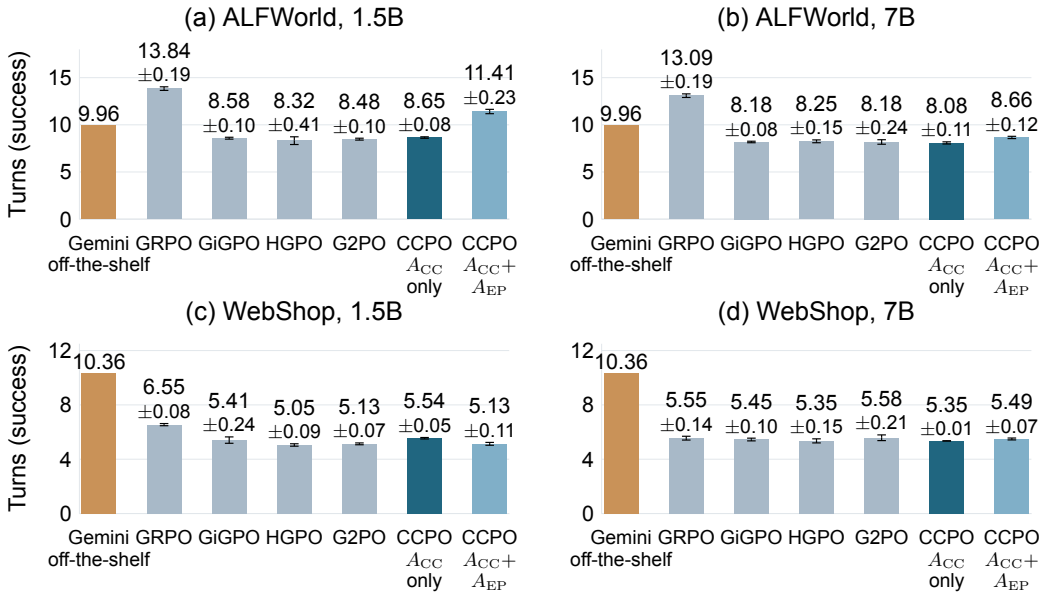

Figure S3: Average steps among successful trajectories on ALFWorld and WebShop. Panels show Qwen2.5-1.5B/7B-Instruct agents and a shared off-the-shelf Gemini reference. Error bars show sample standard deviations across three evaluation subsets; no deviation is reported for Gemini.

H EXACT-MATCH FALLBACK AND PEER AVAILABILITY

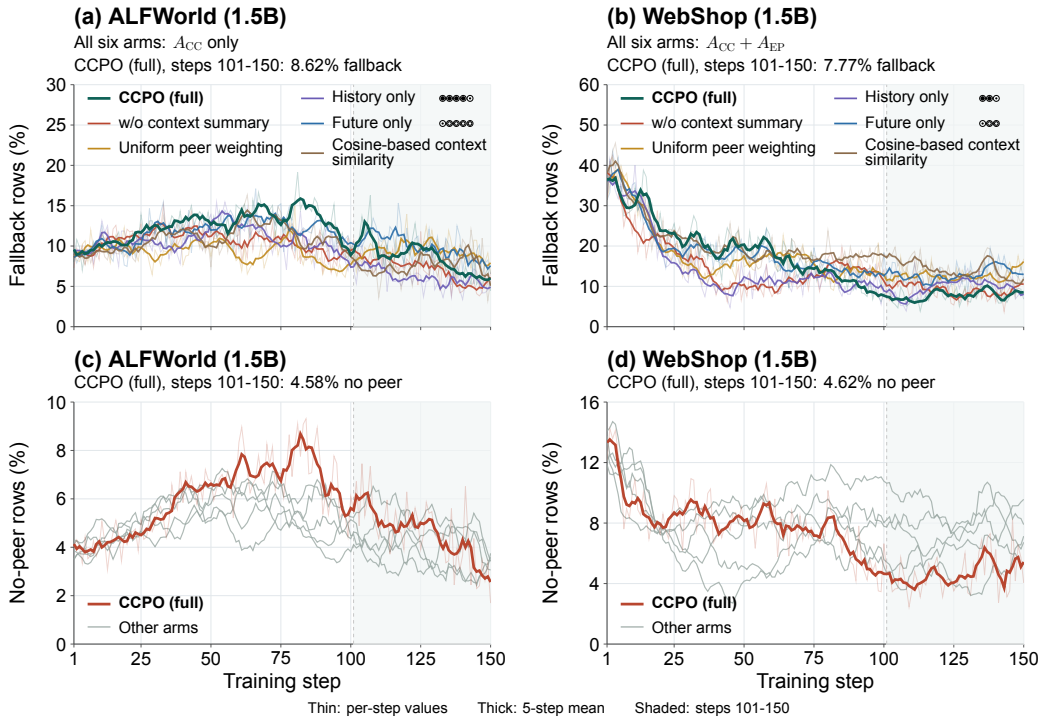

Figure S4: **Exact-match fallback and peer availability during training.** Results use Qwen2.5-1.5B-Instruct on ALFWorld and WebShop. (a,b) Percentage of rows falling back from the exact anchor-observation key to a task-only key, comparing CCPO (full) with five ablations. (c,d) No-peer row percentages, computed as the singleton-bucket fraction divided by mean bucket size; CCPO (full) is highlighted and ablations are gray. Faint traces show raw per-step values, while smoothed curves use centered five-step moving averages. Shading marks steps 101–150; annotations report unweighted means of raw per-step values over this window. All ALFWorld arms use A_{CC} only, and all WebShop arms use $A_{\text{CC}} + A_{\text{EP}}$. Each arm represents one recorded training run. Vertical-axis ranges differ between benchmarks.