LEARNING TO ASSIGN CREDIT IN HINDSIGHT FOR LONG-HORIZON LLM AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Sparse outcome rewards make it hard to tell which actions actually mattered, and training on them directly risks treating pivotal actions and irrelevant ones alike. Existing work has approached this issue through both prospective credit estimation and hindsight prompting. We take a complementary approach: rather than prompt for hindsight, we explicitly learn it. We ask: Given the outcome and the state, how likely was the action that was taken? Comparing this to current policy likelihood provides a fine-grained attribution of each action to the realized future, letting us amplify or suppress its share of reward accordingly. We train a lightweight parameter-efficient estimator of this hindsight conditional, avoiding repeated prompting and reliance on instruction following. We integrate the resulting action-level attribution into the GRPO policy-gradient objective to construct our method, Learned Hindsight Credit Assignment (LHCA). Through extensive experiments and ablations on ALFWorld, Sokoban and BabyAI, we show LHCA outperforms prior credit-assignment methods, including prompted hindsight and preference-trained process-reward estimators, highlighting the effectiveness of explicitly learning hindsight attribution.

1 INTRODUCTION

Long-horizon language-model agents increasingly solve tasks by interacting with external environments over many turns: searching for information, invoking tools, interpreting observations, and revising their plans before producing a final answer. Reinforcement learning (RL) provides a natural way to optimize this behavior from task outcomes, but the feedback available for training is often extremely sparse. An agent may take dozens of actions before receiving a single terminal reward indicating whether the task was completed successfully. This creates a fundamental credit-assignment problem: **which of the many actions along the trajectory actually drove the final outcome?**

Value-free methods such as Group Relative Policy Optimization (GRPO; Shao et al., 2024) offer a particularly attractive training paradigm for LLMs. Rather than learning a separate value network, GRPO compares the terminal rewards of multiple trajectories sampled for the same query and uses their relative performance as the learning signal. This avoids the memory and estimation costs of a learned critic, but provides only coarse credit assignment: the same trajectory-level advantage is broadcast to every action. Consequently, a pivotal decision that determines task success receives the same credit as an incidental action whose effect on the outcome is negligible.

A growing body of work addresses this limitation by exploiting the information available after an interaction has unfolded. Retrospective methods use environmental feedback or completed trajectories to identify critical decisions, generate intermediate supervision, or refine step-level values (Chen et al., 2025; Xu et al., 2026; Yeo et al., 2026; Tan et al., 2026). These results establish hindsight as a powerful source of information for long-horizon agents. However, obtaining this signal typically requires an LLM to reason about the completed trajectory: the model is prompted to reflect on its behavior, generate feedback, or judge which earlier decisions were responsible for the outcome. Credit assignment is therefore mediated by an auxiliary instruction-following task rather than estimated directly from the trajectory distribution.

We revisit a probabilistic alternative through Hindsight Credit Assignment (HCA; Harutyunyan et al., 2019). HCA defines a hindsight distribution $h(a_t \mid s_t, f(\tau_{>t}))$, which asks how likely an action becomes after observing a statistic of its realized future. Comparing this posterior with the

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.

action's probability under the policy, $\rho_t = \frac{h(a_t|s_t, f(\tau_{>t}))}{\pi_\theta(a_t|s_t)}$, has a precise interpretation: by Bayes' rule, ρ_t measures how much taking the action changes the likelihood of the observed future. When the future makes an action substantially more likely, $\rho_t > 1$; when the future is un-informative about the action, $\rho_t \approx 1$. Hindsight can, therefore, provide fine-grained, value-free credit without requiring a separately learned critic.

Applying this idea to LLM agents raises a basic estimation problem. The rollout policy naturally provides $\pi_\theta(a_t \mid s_t)$, since actions are generated auto-regressively from their past. The hindsight distribution instead requires evaluating the same action conditioned on both its preceding context and information that occurs after it in the original trajectory. Under the trajectory's natural ordering, however, a decoder-only policy cannot condition an action on this future information.

Our key observation is that hindsight attribution can be formulated directly as a conditional sequence-likelihood estimation problem. During training, the complete trajectory is already observed: the past s_t, action a_t, and future statistic $f(\tau_{>t})$ are available. Rather than prompting an LLM to generate a retrospective judgment about the action, we directly estimate the likelihood of a_t conditioned on both s_t and $f(\tau_{>t})$. We implement this using a lightweight causal adapter, placing the future statistic before the state and action so that a standard decoder-only model can condition on all required information when scoring the action. This turns hindsight attribution into a sequence-likelihood estimation problem. A single additional forward pass produces the posterior likelihood needed for action-level credit assignment, without auxiliary feedback generation or reliance on the policy's ability to follow a hindsight prompt.

Building on this estimator, we introduce Learned Hindsight Credit Assignment (LHCA), which brings the HCA construction into value-free group policy optimization. LHCA replaces GRPO's trajectory-level value estimate with a hindsight-weighted, turn-specific estimate before applying the same group-relative standardization and policy optimization machinery. This retains GRPO's efficient value-free training while allowing different actions within the same trajectory to receive different credit according to how strongly they are associated with the realized future. When the future is uninformative, the hindsight ratio approaches one and the resulting update recovers the GRPO signal; informative futures introduce a turn-specific correction that redistributes credit toward consequential decisions.

Our main contributions are as follows:

Main Contributions.

- We formulate turn-level credit assignment for long-horizon LLM agents directly through the probabilistic hindsight distribution of HCA and derive a value-free, hindsight-weighted estimator of state-action values compatible with group-relative policy optimization.
- We introduce Learned Hindsight Credit Assignment (LHCA), which learns the hindsight conditional directly using a lightweight causal adapter by placing the observed future statistic before the state and action, avoiding auxiliary retrospective generation or hindsight instruction following.
- We evaluate LHCA on long-horizon agentic tasks and compare it with existing outcome-level and retrospective credit-assignment approaches, together with ablations that isolate the effects of future conditioning, hindsight estimation, and optimization design.

2 METHODOLOGY

2.1 PRELIMINARIES AND NOTATIONS

Turn-level MDP. We formalize the long-horizon agentic task as a Markov Decision Process over interaction contexts. At turn t we take the state to be the full interaction context $s_t = (o_0, a_0, \ldots, a_{t-1}, o_t)$, where each o_i is an observation emitted by the environment (e.g., a webpage or tool output). The policy conditions on s_t and produces an action $a_t = (y_{t,1}, \ldots, y_{t,|a_t|})$

following $\pi_\theta(a_t \mid s_t)$. The environment emits an observation o_{t+1}, and the state updates as $s_{t+1} = (s_t, a_t, o_{t+1})$. The episode runs for T turns yielding a trajectory: $\tau = (o_0, a_0, o_1, a_1, \ldots, o_T)$. We write $\tau_{>t}$ for the future suffix of the trajectory strictly after a_t. These tasks are characterized by *sparse* rewards, i.e., the environment provides a scalar outcome reward $R(\tau)$ only at termination ($t = T$), based on the completion status. The learning objective is:

$$J(\pi_\theta) = \mathbb{E}_{\tau\sim\pi_\theta}[R(\tau)]\,. \tag{1}$$

Value-free group policy optimization. GRPO samples a group of G trajectories $\{\tau_1, \ldots, \tau_G\}$ from the starting state for each input query following the current policy and uses group statistics as a baseline to compute the advantage:

$$Q_i^{\mathrm{GRPO}} := R(\tau_i), \qquad A_i^{\mathrm{GRPO}} = \mathcal{Z}\big(Q_i^{\mathrm{GRPO}}\big), \qquad \mathcal{Z}(x_i) := \frac{x_i - \mu}{\sigma}, \tag{2}$$

where $Q_i^{\mathrm{GRPO}} = R(\tau_i)$ is GRPO's (implicit) per-trajectory value estimate, $\mathcal{Z}(\cdot)$ denotes the group z-score (standardization), and μ, σ are the mean and standard deviation of the outcome rewards $\{R(\tau_i)\}_{i=1}^G$ within the group[1]. This reduces gradient variance through intra-group comparison without any additional value-network parameters. GRPO and the method we develop optimize π_θ with the same clipped policy-gradient surrogate, differing only in the advantage $\hat{A}_t$ assigned to each turn. For a sampled trajectory τ of T turns,

$$\mathcal{J}(\theta) = \mathbb{E}_{\substack{q\sim\mathcal{D} \\ \tau\sim\pi_{\theta_{\text{old}}}(\cdot|q)}} \left[\frac{1}{|a|} \sum_{t=1}^{T} \sum_{j=1}^{|a_t|} \mathrm{MC}_\epsilon\big(r_{t,j}(\theta),\, \hat{A}_t\big) \right] - \beta\, D_{\mathrm{KL}}\big(\pi_\theta \,\|\, \pi_{\text{ref}}\big), \tag{3}$$

estimated over the G trajectories sampled per query, where $\mathrm{MC}_\epsilon(r, A) := \min\big(r\,A,\ \mathrm{clip}(r,\, 1-\epsilon,\, 1+\epsilon)\,A\big)$ is the clipped surrogate term of PPO (Schulman et al., 2017) with clipping range ϵ, $|a| = \sum_{t=1}^{T} |a_t|$ is the number of action tokens, $\pi_{\theta_{\text{old}}}$ and π_{ref} are the previous-iteration and reference policies, $\hat{A}_t$ is the advantage of turn t broadcast to its tokens, and $r_{t,j}(\theta) = \pi_\theta(y_{t,j} \mid s_t, y_{t,<j}) / \pi_{\theta_{\text{old}}}(y_{t,j} \mid s_t, y_{t,<j})$ is the token-level importance ratio. Vanilla GRPO sets $\hat{A}_t = A_i^{\mathrm{GRPO}}$, broadcasting a single trajectory-level scalar to every action token regardless of its contribution to the outcome.

2.2 HINDSIGHT: INTRODUCTION

Process reward models (PRMs; Lightman et al., 2024; Wang et al., 2024), used both to guide generation and to supply reward during RL training, ask the *forward* question: from the current state, which action leads to a successful outcome? Such methods tend to have high-variance and difficult to train (DeepSeek-AI, 2025; Zhang et al., 2025b). In this work, we ask the *reverse* question: given that the outcome (or the full trajectory) was ultimately observed, how likely was this action? This formalizes the everyday intuition of *hindsight*: once a sequence of events has played out, we can look back and identify the actions that were pivotal to the final outcome. Concretely, we build on the hindsight distribution introduced by Harutyunyan et al. (2019),

$$h\big(a_t \mid s_t, f(\tau_{>t})\big) \;:=\; \mathbb{P}_{\tau\sim\mathcal{T}^{\pi_\theta}(s_t)}(A_t = a_t \,|\, f(\tau_{>t}))\,, \tag{4}$$

where $\mathcal{T}^{\pi_\theta}(s_t)$ is the distribution over trajectories that continue from s_t under π_θ. After unrolling a trajectory under the student policy $\pi_\theta(a_t \mid s_t)$, h re-scores each action given a chosen statistic $f(\tau_{>t})$ of that action's future.

The credit signal is the *hindsight attribution* ρ_t, or the ratio between the two distributions:

$$\rho_t = \frac{h\big(a_t \mid s_t, f(\tau_{>t})\big)}{\pi_\theta(a_t \mid s_t)} = \frac{\mathbb{P}(f(\tau_{>t}) \mid A_t = a_t)}{\mathbb{P}(f(\tau_{>t}))}, \tag{5}$$

where all probabilities are taken over $\tau \sim \mathcal{T}^{\pi_\theta}(s_t)$ and the second equality follows from a simple application of Bayes' rule. Intuitively, $\rho_t > 1$ implies that knowing the future makes the action

[1]The std. normalization can be removed as presented in Dr. GRPO (Liu et al., 2025), without affecting our comparison. We apply the same z-score to the GRPO baseline and the LHCA advantage.

more likely than the policy judged it at decision-time or, equivalently, taking the action a_t makes the observed future more likely than an average rollout. When $\rho_t \approx 1$ the action is future-independent and when $\rho_t < 1$ the future is evidence against the action. Consequently, we can use this term to amplify or suppress credit as it measures the *consistency* with the observed future (through f). Note that, ρ_t does not label an action "good" or "bad" and the outcome reward still decides whether that future should be reinforced.

Under the sparse-reward setting of Section 2.1, hindsight attribution ρ_t recovers the exact state–action Q–value Harutyunyan et al. (2019)[Theorem 2] as following.

Proposition 1 (Hindsight Q-value identity)**.** *Let $G_t = \gamma^{T-t} R(\tau)$ be the discounted return of a trajectory and a statistic f such that G_t is a deterministic function of $f(\tau_{>t})$. Then, for any state-action pair s_t, a_t with $\pi_\theta(a_t \mid s_t) > 0$, the state–action value satisfies*

$$Q^{\pi_\theta}(s_t, a_t) \;:=\; \mathbb{E}_{\tau \sim \mathcal{T}^{\pi_\theta}(s_t, a_t)}[G_t] \;=\; \mathbb{E}_{\tau \sim \mathcal{T}^{\pi_\theta}(s_t)}[\rho_t \, G_t]\,, \tag{6}$$

where, ρ_t is the hindsight attribution in equation 5.

We provide the proof of Proposition 1 in Appendix A.2. Unlike GRPO's advantage equation 2, which reuses a single scalar for every action in a trajectory, the per-action estimate $\widehat{Q}_{i,t}$ credits each action in proportion to how much the observed future supports it. We form advantages, similar to GRPO, by z-score normalization of the per-action estimates $\widehat{Q}_{i,t}$ within each group as elaborated in Section 2.3. Although this proposition motivates our core hindsight estimation procedure, practical estimation requires additional empirical considerations for training stability as we outline below.

2.3 Hindsight: Estimation and Optimization

Several prior works estimate some notion of hindsight to perform credit assignment. Chen et al. (2025); Yeo et al. (2026); Xu et al. (2026) prompt an LLM to generate feedback and identify the pivotal actions in a trajectory. To avoid autoregressive generation at training time, Tan et al. (2026) instead inject a trajectory summary into the existing policy and score the current action in a single forward pass. This is cheaper, but still leans on the instruction-following ability of the policy, since the *trajectory-feedback prompt* can be out of distribution for the model being RL-trained. Concurrent work likewise prompts a feedback-conditioned policy for distillation or credit (Hübotter et al., 2026; Li et al., 2026c; Ma et al., 2026; Li et al., 2026d).

Future Conditioning. Notice that $h(a_t \mid s_t, f(\tau_{>t}))$ is the likelihood of the action tokens a_t under a model that can attend to the future statistic $f(\tau_{>t})$ in addition to the current state s_t. Therefore, our intuition is to turn the 'hindsight estimation into a sequence-likelihood problem' rather than an instruction-following one. The state s_t and the future statistic $f(\tau_{>t})$ are both fully observed context, whereas the current action a_t is the token sequence for which we generate likelihood score. This hindsight likelihood h can be estimated with an auto-regressive approach by simply ordering $f(\tau_{>t})$ before s_t and a_t. Since $\pi_\theta(a_t \mid s_t)$ is already available from the rollout generation phase, computing the hindsight attribution $\rho_t = h/\pi_\theta$ only requires one additional forward pass per action, which is fully parallelizeable across the trajectory. We can then use this ratio to assign credit to each turn in the trajectory, which we term Learned Hindsight Credit Assignment (LHCA). The choice of $f(\tau_{>t})$ depends on the task as different environments provide observations with varying structure. We ablate on the choice of $f(\tau_{>t})$ in section 3.2

Hindsight Adapter Training. We train the adapter ϕ with an action-modeling objective on the sampled trajectories. For a trajectory τ with T number of turns,

$$\mathcal{L}^{\text{lhca}}(\phi, \tau) = -\sum_{t=1}^{T} \sum_{j=1}^{|a_t|} \log \pi_\theta^{\phi}\big(y_{t,j} \mid s_t,\, y_{t,<j},\, f(\tau_{>t})\big), \tag{7}$$

where $\pi_\theta^{\phi}(y_{t,j} \mid s_t, y_{t,<j}, f(\tau_{>t}))$ is the likelihood the causal hindsight model assigns to the action token $y_{t,j}$, using the frozen base policy π_θ together with the trainable LoRA (Hu et al., 2022) adapter ϕ. We prepend $f(\tau_{>t})$ to the causal context, followed by s_t and the preceding action tokens $y_{t,<j}$, so that the model can condition on the observed future while retaining the standard decoder-only attention structure. Since trajectories are sampled from the frozen base policy, the adapter need not

learn any new task knowledge, instead, it only teaches the model to attend to the future tokens via $f(\tau_{>t})$. Consequently, we keep it lightweight adapting only the query and key projections. We warm up ϕ on rollouts from the initial policy before training begins, and during training refresh it with the sampled rollouts per training-step to track policy drift. The loss is akin to infilling the action conditioned on the state and future statistic.

From scores to advantages. Because $h \approx \pi_\theta^\phi$ by construction, we estimate it from the adapted model as a length-normalized, sharpened sequence likelihood over the tokens of a_t,

$$\widehat{h}\big(a_t \mid s_t, f(\tau_{>t})\big) = \exp\left(\frac{1}{T_{\text{temp}}\,|a_t|}\sum_{j=1}^{|a_t|}\log \pi_\theta^\phi\big(y_j \mid y_{<j}, s_t, f(\tau_{>t})\big)\right), \tag{8}$$

where T_{temp} is a sharpening temperature that controls the estimation noise. The hindsight attribution is the clipped ratio of this estimate to the rollout policy,

$$\rho_{i,t} = \frac{\widehat{h}\big(a_t \mid s_t, f(\tau_{>t})\big)}{\pi_\theta(a_t \mid s_t)}, \tag{9}$$

where θ is the rollout policy, so $\pi_\theta(a_t \mid s_t)$ is already available from generation. Following Proposition 1, we form the single-sample hindsight Q-estimate and standardize it into an advantage with the same group z-score used for GRPO in equation 2. The only change from GRPO is the value estimate fed into this standardization:

$$\begin{aligned} &\text{GRPO:} \quad Q_i^{\text{GRPO}} = R(\tau_i), && A_i^{\text{GRPO}} = \mathcal{Z}\big(Q_i^{\text{GRPO}}\big), \\ &\text{LHCA:} \quad \widehat{Q}_{i,t}^{H} = \rho_{i,t}\,\gamma^{T-t}\,R(\tau_i), && A_{i,t}^{\text{HA-GRPO}} = \mathcal{Z}\big(\widehat{Q}_{i,t}^{H}\big). \end{aligned} \tag{10}$$

LHCA then plugs its per-turn advantage into the same objective equation 3, setting $\hat{A}_t = A_{i,t}^{\text{HA-GRPO}}$ in place of GRPO's broadcast $\hat{A}_t = A_i^{\text{GRPO}}$.

Empirical considerations. While the estimator above is principled, we found that some adjustments were helpful to stabilize training in practice. First, we normalize the raw attribution $\rho_{i,t}$ by its trajectory mean and clip the result to $[C_{\min}, C_{\max}]$,

$$\rho'_{i,t} = \text{clip}\left(\frac{\rho_{i,t}}{\frac{1}{T}\sum_k \rho_{i,k}},\, C_{\min},\, C_{\max}\right). \tag{11}$$

We hypothesize that this trajectory normalization (i) reshuffles credit within a trajectory in a zero-sum manner, leaving the trajectory's net push to GRPO while hindsight arranges credit across the turns, and/or (ii) debiases a trajectory-level offset in the adapter's likelihood estimates; the normalized $\rho'_{i,t}$ is mean-one across each trajectory and rollout group. We also apply a *do-no-harm* mask that keeps advantages non-negative on successful trajectories, so hindsight only redistributes credit among their turns rather than flipping the sign of an action that led to a correct outcome.

Algorithm 1 presents the core LHCA update. Our experiments additionally apply trajectory-wise ratio normalization and clipping, a do-no-harm mask, and an auxiliary GRPO advantage term. Additional training details and hyperparameters are provided in Appendix A.4.

Connection to vanilla GRPO. We can connect hindsight-based advantage to GRPO's by splitting the hindsight ratio into a unit term and a deviation, $\rho_{i,t} = 1 + \delta_{i,t}$.

Proposition 2 (GRPO+LHCA decomposition)**.** *Taking $\gamma = 1$ and standardizing the advantages over the same NT-token group. With $\delta_{i,t} := \rho_{i,t} - 1$, the LHCA advantage decomposes as*

$$A_{i,t}^{\textit{HA-GRPO}} = \frac{\sigma_R}{\sigma_{R+\delta R}}\,A_{i,t}^{\textit{GRPO}} \;+\; \frac{\sigma_{\delta R}}{\sigma_{R+\delta R}}\,\mathcal{Z}\left(\delta_{i,t}R(\tau_i)\right), \tag{12}$$

where σ_R, $\sigma_{\delta R}$, and $\sigma_{R+\delta R}$ are the group standard deviations of $R(\tau_i)$, $\delta_{i,t}R(\tau_i)$, and their sum. In particular, if $\delta_{i,t} \approx 0$ for every action then $A_{i,t}^{\textit{HA-GRPO}} \approx A_{i,t}^{\textit{GRPO}}$.

The hindsight advantage is thus a positive multiple of the GRPO advantage plus a z-scored hindsight correction, each weighted by the fraction of the total spread it carries; it recovers GRPO exactly

Algorithm 1 GRPO + LHCA

Require: policy π_θ; task prompts $\mathcal{D}$; LoRA adapter ϕ
+ Warm up ϕ on rollouts from π_θ via Eq. (7)
for step $= 1, \ldots, M$ **do**
  Roll out G trajectories $\{\tau_i\}_{i=1}^{G}$ per prompt with π_θ; compute rewards $R(\tau_i)$
  + Refresh ϕ on $\{\tau_i\}_{i=1}^{G}$ via Eq. (7)
  + Score hindsight ratios $\rho_{i,t}$ with π_θ^{ϕ} via Eqs. (8)–(9)
  + $A_{i,t}^{\text{HA-GRPO}} = \mathcal{Z}(\rho_{i,t}\,\gamma^{\,T-t} R(\tau_i))$ ▷ *GRPO:* $\mathcal{Z}(R(\tau_i))$
  Update π_θ by maximizing Eq. (3) with $\hat{A}_t = A_{i,t}^{\text{HA-GRPO}}$
end for
Ensure: π_θ

when the future is uninformative. Interestingly, the hindsight term enters in proportion to its own dispersion $\sigma_{\delta R}$, so a hindsight signal that barely varies and carries little information is automatically down-weighted without any tuned mixing coefficient. We defer proof to Appendix A.3. Following prior works (Feng et al., 2025; Tan et al., 2026) we consider $A_{i,t}^{\text{HA-GRPO}}$ as the *micro* signal and A_i^{GRPO} as the pure *macro* signal. We use a combination of both which protects training from noisy estimation of hindsight signal. We show an ablation without the macro term in Section 3.2.

3 Experiments and Results

3.1 Experimental Setup

Datasets and Tasks. We evaluate on three multi-turn agentic benchmarks. **ALFWorld** (Shridhar et al., 2021) is designed to assess the ability of LLM agents to perform multi-step decision-making in an embodied environment. In each episode, the agent receives a text goal and has to accomplish it through multi-turn interaction with the environment. It includes 3,287 task instances across six categories of common household activities: Pick & Place (Pick), Examine in Light (Look), Clean & Place (Clean), Heat & Place (Heat), Cool & Place (Cool), and Pick Two & Place (Pick2). **BabyAI** (Chevalier-Boisvert et al., 2019; 2023) is a 2D grid world in which the agent completes language conditioned navigation tasks. Here, each observation provides only the agent's relative position to the target locations, so the agent must track obstacles and its traversed path to solve the task effectively. We group the tasks into six categories: GoTo (plain navigation), Unlock (finding a key and unlocking a door), Open (opening a door), Boss (compositional "boss" level tasks), Pickup (locate and pickup specified object), and Other (FindRoom, AOD, SLoc). We also use **Sokoban** (Junghanns & Schaeffer, 2001) where the agent faces a spatial planning task, pushing boxes to target positions in a grid world with a time penalty. The observation consists of the entire grid state. These environments capture different challenges of multi-step alignment and planning.

Models and Baselines. We use Qwen2.5-1.5B/7B-Instruct (Yang et al., 2024) for our experiments. Our initial experiments with Qwen3 (Yang et al., 2025) showed that it produces highly verbose reasoning traces hitting the maximum response limit far more frequently while showing similar success rates. We compare with GRPO (Shao et al., 2024) as the primary baseline. iStar (Liu et al., 2026b) uses a turn-level DPO based approach to compute rewards for each action, which we use as a *foresight* baseline. Among existing hindsight literature, we use HCAPO (prompting) (Tan et al., 2026) as representative of prompting based hindsight scoring.

Training and Implementation details. We use `verl-agent` (Feng et al., 2025) codebase for ALFWorld and Sokoban environments and use `AgentGym` (Xi et al., 2025a) for BabyAI simulations. All implementations use veRL (Sheng et al., 2025) as the core engine. For ALFWorld we report results on the held out unseen tasks. For Sokoban, we report results on randomly generated 128x5 puzzles. For BabyAI, we report results on the official test set. We provide more hyperparameters and training details in Appendix A.4.

Table 1: Results on ALFWorld and Sokoban with Qwen2.5-1.5B-Instruct. *Base* and *GRPO* denote the performance of original reference and GRPO trained models respectively. *iStar* training collapses for ALFWorld and shows suboptimal performance on Sokoban data. For HCAPO (Prompting) baseline we evaluate with both $f(\tau_{>t}) = o_T$ and $f(\tau_{>t}) = \tau_{>t}$ and for LHCA we report the performance with best future conditioning. Subscripts give std. over 2 seeds. c indicates training collapse.

	ALFWorld							Sokoban
Method	Pick	Look	Clean	Heat	Cool	Pick2	All	Success Rate
Base	12.0	0.0	3.6	0.0	11.8	0.0	4.7	$17.2_{\pm 4.0}$
GRPO	$66.3_{\pm 9.6}$	$57.4_{\pm 8.5}$	$67.9_{\pm 7.1}$	$71.9_{\pm 10.7}$	$78.3_{\pm 4.7}$	$48.4_{\pm 4.9}$	$65.4_{\pm 6.0}$	$44.0_{\pm 1.0}$
iStar	c	c	c	c	c	c	c	$36.3_{\pm 4.4}$
HCAPO (o_T)	$\mathbf{88.7}_{\pm 5.7}$	$90.2_{\pm 12.4}$	$89.3_{\pm 8.1}$	$\mathbf{87.5}_{\pm 10.2}$	$91.5_{\pm 2.8}$	$93.7_{\pm 11.0}$	$89.5_{\pm 3.3}$	$48.2_{\pm 0.6}$
HCAPO ($\tau_{>t}$)	$86.3_{\pm 12.5}$	$\mathbf{94.9}_{\pm 5.6}$	$78.6_{\pm 6.4}$	$83.7_{\pm 8.8}$	$92.6_{\pm 6.4}$	$\mathbf{95.2}_{\pm 8.3}$	$87.6_{\pm 4.8}$	$43.6_{\pm 2.0}$
LHCA	$84.8_{\pm 10.2}$	$92.6_{\pm 12.8}$	$\mathbf{95.6}_{\pm 5.1}$	$81.6_{\pm 8.6}$	$\mathbf{96.2}_{\pm 3.3}$	$93.1_{\pm 8.1}$	$\mathbf{90.5}_{\pm 2.2}$	$\mathbf{51.1}_{\pm 2.5}$

Table 2: BabyAI success rate (%) with Qwen2.5-7B-Instruct; std. over 2 seeds, best in **bold**.

Method	GoTo	Unlock	Open	Boss	Pickup	Other	All
Base	54.5	0.0	0.0	0.0	30.0	40.0	22.0
GRPO	$76.4_{\pm 0.0}$	$1.8_{\pm 0.0}$	$52.0_{\pm 4.0}$	$\mathbf{16.0}_{\pm 0.0}$	$35.0_{\pm 0.0}$	$37.5_{\pm 2.5}$	$36.4_{\pm 0.2}$
iStar	63.6	1.8	0.0	16.0	40.0	40.0	28.0
HCAPO ($R(\tau)$)	$79.1_{\pm 0.9}$	$1.8_{\pm 0.0}$	$54.0_{\pm 2.0}$	$12.0_{\pm 0.0}$	$37.5_{\pm 7.5}$	$42.5_{\pm 2.5}$	$37.8_{\pm 1.2}$
LHCA ($R(\tau)$)	$\mathbf{80.9}_{\pm 0.9}$	$\mathbf{3.6}_{\pm 3.6}$	$\mathbf{56.0}_{\pm 4.0}$	$14.0_{\pm 2.0}$	$\mathbf{47.5}_{\pm 7.5}$	$\mathbf{45.0}_{\pm 5.0}$	$\mathbf{41.2}_{\pm 2.5}$

3.2 Results and Ablations

Main Results. Table 1 reports our main results on ALFWorld and Sokoban with Qwen2.5-1.5B-Instruct, and Table 2 on BabyAI with the larger Qwen2.5-7B-Instruct. Hindsight-based methods improve over vanilla GRPO on all benchmarks. Our experiments with the *foresight*-based method led to mixed results. On ALFWorld, rewards and *valid action ratio* plummet to zero while the model trails baseline GRPO on sokoban. We outline our findings in Appendix A.5. Within the hindsight family, LHCA is marginally better than HCAPO on ALFWorld (90.5 vs. 89.5) but improves more substantially on navigation tasks, Sokoban (51.1 vs. 48.2), and BabyAI (41.2 vs. 37.8). Overall, LHCA matches or outperforms the prompting-based hindsight baseline on all three benchmarks and both model scales, underscoring the value of a trainable hindsight signal over a fixed prompting based one.

Hindsight signal Calibration matters. Our trainable hindsight adapter trains to estimate h directly, whereas methods like Tan et al. (2026) use the moving actor policy to measure it directly by prompting the training policy. To test how a good estimation of hindsight signal affects downstream metrics, we design a probe where we use the initial base policy model to calculate h throughout the training. The three step progression of hindsight calculation: (a) base initial policy, (b) training policy and (c) an adapter to learn h highlight how an improved hindsight calibration helps improve RL training (Figure 1). On both ALFWorld and sokoban, a better calibrated h helps one future conditioning but hurts the other. On ALFWorld, o_T improves monotonically as calibration sharpens ($84.3 \rightarrow 89.5 \rightarrow 90.5$, a $+6.2$ point gain from base to adapter), whereas $\tau_{>t}$ rises with the training policy but falls back once the adapter is introduced ($85.8 \rightarrow 87.6 \rightarrow 85.1$). On Sokoban we see reversed roles where $\tau_{>t}$ benefits from improved calibration whereas o_T does not. In each case the conditioning better suited to the task (§3.2) is the one that benefits from improved calibration, suggesting that a better-calibrated h amplifies an already-informative future signal but cannot rescue an uninformative one.

Choosing the right $f(\tau_{>t})$ The choice of what to condition the hindsight signal on depends on the environment and the estimation method. For HCAPO, that relies on instruction following and prompting, a short final observation o_T that includes enough information about the outcome is suf-

Table 3: Removing the macro (GRPO) term tests the effect of the hindsight micro-term on downstream metrics. Sokoban success rate (std. over 2 seeds). Both rows use $f(\tau_{>t}) = o_T$

	Original	No-macro	Drop (%)
HCAPO	$48.2_{\pm 0.6}$	39.7	17.6
LHCA	$45.0_{\pm 1.13}$	41.3	8.2

Table 4: Different choices of $f(\tau_{>t})$ for hindsight ratio estimation with different methods. Average success rate on ALFWorld and Sokoban tasks (std. over 2 seeds).

	ALFWorld	Sokoban
GRPO	$65.4_{\pm 6.0}$	$44.0_{\pm 1.0}$
HCAPO $f(\tau_{>t}) = o_T$	$89.5_{\pm 3.3}$	$48.2_{\pm 0.6}$
HCAPO $f(\tau_{>t}) = \tau_{>t}$	$87.6_{\pm 4.8}$	$43.6_{\pm 2.0}$
LHCA $f(\tau_{>t}) = o_T$	$\mathbf{90.5}_{\pm 2.2}$	$45.0_{\pm 1.13}$
LHCA $f(\tau_{>t}) = \tau_{>t}$	$85.1_{\pm 2.12}$	$\mathbf{51.1}_{\pm 2.5}$

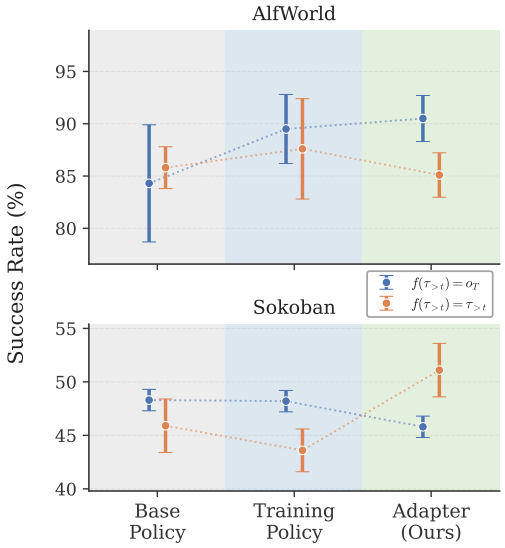


Figure 1: Downstream success rate as hindsight-signal calibration improves across the three-step progression (base policy → training policy → trained adapter), for both scoring functions $f(\tau_{>t}) = o_T$ and $f(\tau_{>t}) = \tau_{>t}$.

ficient. With full trajectory $\tau_{>t}$, the input to LLM becomes very out-of-distribution and therefore it cannot make use of the extra context. This is especially apparent in Sokoban, where the final state is not informative enough to determine the importance of actions. Consequently LHCA outperforms HCAPO on Sokoban when using $\tau_{>t}$ (51.1 vs. 43.6). Interestingly, it degrades on o_T (45.0 vs. 48.2), necessitating more careful design of the hindsight signal. On ALFWorld, final observation o_T seems to be better with both methods, with the extra context of $\tau_{>t}$ not being useful for either of them.

Removing the GRPO macro term. As we show in Prop 2, the hindsight-level advantage automatically weights the macro and micro level signals, and the additional GRPO (macro) term mainly serves as a safeguard against over-relying on a noisy hindsight estimate. Removing it isolates how much each method actually depends on that safeguard (Table 3). We observe that both methods *regress* without the macro term, but with very different margins. HCAPO drops from 48.2 to 39.7, a 17.6% relative loss, while LHCA drops from 45.0 to only 41.3, a 8.2% relative loss. Since any hindsight estimation can be noisy in practice, micro signal alone cannot act as a standalone objective (alone it under performs GRPO) and the smaller drop for LHCA indicates a less noisy estimate.

Ablating on T_{temp} We ablate the downstream effect of sharpening temperature T_{temp}, which controls the spread of hindsight distribution and attribution used for training. Proposition 2 shows that a highly dispersed hindsight ratio suppresses the contribution of GRPO's advantage to the combined signal and hindsight estimates can be noisy for individual actions. Consequently, T_{temp} ends up being an important hyperparameter in practice (Figure 2).

Hindsight ratio analysis. To test whether the learned hindsight ratio ρ corresponds to intuitively important actions, we examine its distribution over actions taken by a trained *ALFWorld* policy (Qwen2.5-1.5B-Instruct). We analyze the action corresponding to `examine` (pure inspection, which does not change the environment) and compare its ρ distribution against (a) all other actions, (b) navigation (`go`), and (c) the task-defining transformations (`heat`/`cool`/`clean`). The idea is to see whether the learned hindsight ratio $\rho = h/\pi$ assigns credit sensibly, and whether it separates actions by their relevance to the outcome. Intuitively `examine` does not change the environment and should receive lower credit than other actions, and this should be reflected in its ρ distribution. The distributions in Figure 3 separate cleanly and in the expected direction: `examine` sits below $\rho = 1$ (mean 0.974) while every comparison group sits above it, and the separation grows from $+0.087$ (vs. all other actions) to $+0.120$ (vs. task actions) as the comparison group becomes more consequential to the episode's outcome.

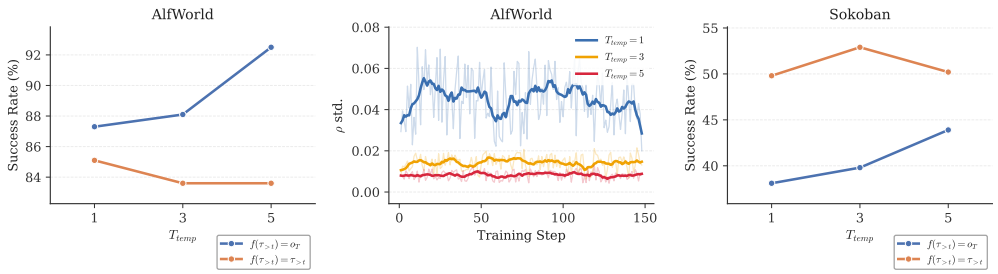


Figure 2: Effect of the sharpening temperature $T_{temp} \in \{1, 3, 5\}$. **Left, Right:** downstream success rate on ALFWorld and Sokoban for both scoring functions f. **Middle:** the raw hindsight-ratio dispersion (ρ std.) over training on ALFWorld for each T_{temp}, showing that lower temperatures yield a more dispersed hindsight signal.

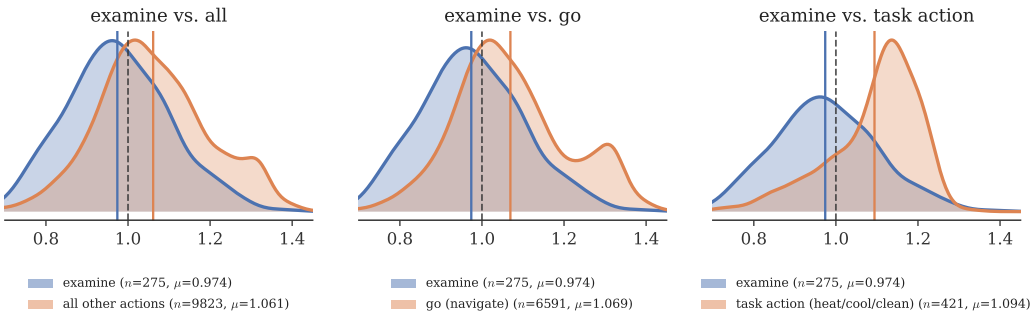


Figure 3: Hindsight ratio ρ separates actions by their relevance to the outcome. `examine` (pure inspection, ■, n=275, mean 0.974) receives lower credit than all other actions, and the separation grows as the comparison group becomes more consequential to the episode's outcome. The dashed line marks $\rho = 1$; solid lines mark each distribution's mean. Left to right, the comparison group becomes more consequential to the outcome and the separation between the two distributions grows accordingly ($+0.087, +0.095, +0.120$)

4 Conclusions

We introduced Learned Hindsight Credit Assignment (LHCA), a value-free approach for assigning fine-grained credit in long-horizon LLM agents with sparse outcome rewards. Building on the probabilistic formulation of hindsight credit assignment, LHCA asks how likely an observed action becomes after conditioning on its realized future and compares this hindsight likelihood with the action probability under the rollout policy. Rather than eliciting this signal through retrospective prompting, LHCA directly learns the hindsight conditional with a lightweight causal adapter, turning hindsight estimation into a sequence-likelihood problem. The resulting attribution provides turn-specific credit while retaining the group-relative optimization framework of GRPO.

Across ALFWorld, Sokoban, and BabyAI, LHCA consistently improves over vanilla GRPO and matches or improves upon prompting-based hindsight credit assignment, with particularly strong gains on Sokoban and BabyAI. Our ablations further show that the usefulness of hindsight depends on both how its conditional distribution is estimated and what information from the realized future is used for conditioning. Finally, analysis of the learned attribution shows that LHCA assigns systematically lower hindsight ratios to purely observational actions than to increasingly consequential actions, providing evidence that the learned signal captures meaningful differences in action relevance. Together, these results suggest that explicitly learning the probabilistic relationship between actions and their realized futures provides a practical route to finer-grained credit assignment without requiring a learned value function or retrospective generation.

AI USE STATEMENT

The authors used LLM assistants to edit prose, check consistency between tables, figures, and text, and obtain feedback on drafts of the manuscript. [If applicable: LLM assistants were also used to implement plotting functions for figures.] The authors designed the method and experiments, derived and verified all theoretical results, verified every reported number against the run artifacts, and take full responsibility for the content.

ETHICS STATEMENT

This work studies credit assignment for reinforcement-learning training of language-model agents in simulated, publicly available environments (ALFWorld, Sokoban, and BabyAI) using openly released models. It involves no human subjects, personal data, or interaction with real-world systems. More sample-efficient agent training lowers the cost of producing capable autonomous agents, which could be deployed in settings where their actions have real consequences. In addition, the learned hindsight ratio is a statistical association between an action and its observed future, not a causal explanation of agent behavior, and should not be interpreted as one. We evaluate only task success in simulated environments and make no claims about the safety-relevant behavior of the trained agents.

REPRODUCIBILITY STATEMENT

Section 2.3 specifies the hindsight adapter, its training objective, the hindsight likelihood and ratio estimates (Eqs. 9–10), trajectory normalization and clipping (Eq. 12), the do-no-harm mask, and the resulting advantage; Algorithm 1 summarizes the full training loop. Section 3 describes the models, environments, reward definitions, group sizes, sampling settings, adapter warm-up and refresh schedule, and all LHCA-specific hyperparameters. Appendices A.2 and A.3 contain complete proofs of Propositions 1 and 2. All experiments build on the publicly available verl-agent, AgentGym, and veRL codebases and Qwen2.5-Instruct models. Each configuration is trained with 2 random seeds, and we report the mean and standard deviation across seeds. For every run we retain the configuration, adapter and policy checkpoints, and raw evaluation outputs. We will release code and configurations upon publication.

REFERENCES

Ilze Amanda Auzina, Joschka Strüber, Sergio Hernández-Gutiérrez, Shashwat Goel, Ameya Prabhu, and Matthias Bethge. Intrinsic credit assignment for long horizon interaction, 2026. URL https://arxiv.org/abs/2602.12342.

Mohammad Bavarian, Heewoo Jun, Nikolas Tezak, John Schulman, Christine McLeavey, Jerry Tworek, and Mark Chen. Efficient training of language models to fill in the middle. *arXiv preprint arXiv:2207.14255*, 2022. URL https://arxiv.org/abs/2207.14255.

Wen-Tse Chen, Jiayu Chen, Fahim Tajwar, Hao Zhu, Xintong Duan, Ruslan Salakhutdinov, and Jeff G. Schneider. Retrospective in-context learning for temporal credit assignment with large language models. In *Advances in Neural Information Processing Systems*, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/680be66a383c5e7bfce021c47aedbb9e-Abstract-Conference.html.

Yanjun Chen, Yirong Sun, Hanlin Wang, Jinghan Wang, Xinming Zhang, Xiaoyu Shen, Wenjie Li, and Wei Zhang. Exact is easier: Credit assignment for cooperative llm agents, 2026. URL https://arxiv.org/abs/2603.06859.

Jie Cheng, Gang Xiong, Ruixi Qiao, Lijun Li, Chao Guo, Junle Wang, Yisheng Lv, and Fei-Yue Wang. Stop summation: Min-form credit assignment is all process reward model needs for reasoning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=3Sxby0hH1q.

Maxime Chevalier-Boisvert, Dzmitry Bahdanau, Salem Lahlou, Lucas Willems, Chitwan Saharia, Thien Huu Nguyen, and Yoshua Bengio. BabyAI: A platform to study the sample efficiency of

grounded language learning. In *7th International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=rJeXCo0cYX.

Maxime Chevalier-Boisvert, Bolun Dai, Mark Towers, Rodrigo Perez-Vicente, Lucas Willems, Salem Lahlou, Suman Pal, Pablo Samuel Castro, and Jordan K. Terry. Minigrid & miniworld: Modular & customizable reinforcement learning environments for goal-oriented tasks. In *Advances in Neural Information Processing Systems 36*, pp. 73383–73394, 2023. doi: 10.52202/075280-3209. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/e8916198466e8ef218a2185a491b49fa-Abstract-Datasets_and_Benchmarks.html.

Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Yuchen Zhang, Jiacheng Chen, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, Qixin Xu, Weize Chen, Jiarui Yuan, Huayu Chen, Kaiyan Zhang, Xingtai Lv, Shuo Wang, Yuan Yao, Xu Han, Hao Peng, Yu Cheng, Zhiyuan Liu, Maosong Sun, Bowen Zhou, and Ning Ding. Process reinforcement through implicit rewards, 2025. URL https://arxiv.org/abs/2502.01456.

DeepSeek-AI. Deepseek-r1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, 2025. doi: 10.1038/s41586-025-09422-z.

Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. Group-in-group policy optimization for LLM agent training. In *Advances in Neural Information Processing Systems 38*, pp. 51797–51830, 2025. doi: 10.52202/085713-1544. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/420c9f777c0b4f78d515e53cf74d58b2-Abstract-Conference.html.

Anna Harutyunyan, Will Dabney, Thomas Mesnard, Mohammad Gheshlaghi Azar, Bilal Piot, Nicolas Heess, Hado van Hasselt, Gregory Wayne, Satinder Singh, Doina Precup, and Rémi Munos. Hindsight credit assignment. In *Advances in Neural Information Processing Systems*, volume 32, pp. 12467–12476, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/195f15384c2a79cedf293e4a847ce85c-Abstract.html.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://mlanthology.org/iclr/2022/hu2022iclr-lora/.

Jonas Hübotter, Frederike Lübeck, Lejs Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, and Andreas Krause. Reinforcement learning via self-distillation. In *International Conference on Machine Learning (ICML)*, 2026. arXiv:2601.20802.

Zishang Jiang, tingyun li, Jinyi Han, Xinyi Wang, Sihang Jiang, Yizhou Ying, Xiaojun Meng, Jiansheng Wei, Jiaqing Liang, and Yanghua Xiao. From outcomes to actions: Leveraging hindsight for long-horizon language agent training. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=iK3yDEvQ4y.

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.

Andreas Junghanns and Jonathan Schaeffer. Sokoban: Enhancing general single-agent search methods using domain knowledge. *Artificial Intelligence*, 129(1–2):219–251, 2001. doi: 10.1016/S0004-3702(01)00109-6. URL https://doi.org/10.1016/S0004-3702(01)00109-6.

Amirhossein Kazemnejad, Milad Aghajohari, Eva Portelance, Alessandro Sordoni, Siva Reddy, Aaron Courville, and Nicolas Le Roux. VinePPO: Refining credit assignment in RL training of LLMs. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=Myx2kJFzAn.

Chenliang Li, Adel Elmahdy, Alex Boyd, Zhongruo Wang, Siliang Zeng, Alfredo Garcia, Parminder Bhatia, Taha Kass-Hout, Cao Xiao, and Mingyi Hong. Stabilizing off-policy training for long-horizon llm agent via turn-level importance sampling and clipping-triggered normalization, 2026a. URL https://arxiv.org/abs/2511.20718.

Junbo Li, Peng Zhou, Rui Meng, Meet P. Vadera, Lihong Li, and Yang Li. Turn-ppo: Turn-level advantage estimation with ppo for improved multi-turn rl in agentic llms, 2026b. URL https://arxiv.org/abs/2512.17008.

Xiaozhe Li, Tianyi Lyu, Yang Li, Yichuan Ma, Peiji Li, Linyang Li, Qipeng Guo, Dahua Lin, and Kai Chen. What and when to distill: Selective hindsight distillation for multi-turn agents. *arXiv preprint arXiv:2605.19447*, 2026c.

Yu Li, Rui Miao, Tian Lan, and Zhengling Qi. OPPO: Bayesian value recursion for token-level credit assignment in LLM reasoning. *arXiv preprint arXiv:2605.21851*, 2026d.

Zhongyi Li, Wan Tian, Jinju Chen, Huiming Zhang, Yang Liu, Yikun Ban, and Fuzhen Zhuang. Counterfactual credit policy optimization for multi-agent collaboration, 2026e. URL https://arxiv.org/abs/2603.21563.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations, ICLR 2024*, 2024. URL https://arxiv.org/abs/2305.20050.

Dongyi Liu, Yifan Niu, Qinwen Wang, Han Xiao, and Jia Li. Pica: Pivot-based credit assignment for search agentic reinforcement learning, 2026a. URL https://arxiv.org/abs/2605.09287.

Xiaoqian Liu, Ke Wang, Yuchuan Wu, Fei Huang, Yongbin Li, Jianbin Jiao, and Junge Zhang. Agentic reinforcement learning with implicit step rewards. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=ooROvpmxMV.

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. *CoRR*, abs/2503.20783, March 2025. doi: 10.48550/arXiv.2503.20783. URL https://doi.org/10.48550/arXiv.2503.20783.

Lichao Ma, Yang Sun, Shuaitao Zhao, Yangyi Fang, Cong Qin, Xiaoliang Fu, Yuhang Tian, Yuchen Wei, Junbo Zhu, Yang Wei, Lu Pan, and Jiaye Lin. How much, then where: Credit-conserving action-to-token allocation for multi-turn agent reinforcement learning. *arXiv preprint arXiv:2608.07118*, 2026.

Zibin Meng and Kani Chen. Craft: Counterfactual credit assignment from free sibling rollouts for self-distilled agentic reinforcement learning, 2026. URL https://arxiv.org/abs/2606.29476.

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

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 *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 27730–27744, 2022.

Prasanna Parthasarathi, Mathieu Reymond, Boxing Chen, Yufei Cui, and Sarath Chandar. Grpo-λ: Credit assignment improves llm reasoning, 2025. URL https://arxiv.org/abs/2510.00194.

John Schulman, Philipp Moritz, Sergey Levine, Michael I. Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. In *International Conference on Learning Representations (ICLR)*, 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/1707.06347.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *CoRR*, abs/2402.03300, 2024. doi: 10.48550/arXiv.2402.03300. URL https://doi.org/10.48550/arXiv.2402.03300.

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, 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 *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 8634–8652, 2023.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew J. Hausknecht. Alfworld: Aligning text and embodied environments for interactive learning. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=0IOX0YcCdTn.

Richard S. Sutton, David A. McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *NIPS*, pp. 1057–1063, 1999. URL http://papers.nips.cc/paper/1713-policy-gradient-methods-for-reinforcement-learning-with-function-approximatio

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. *CoRR*, abs/2603.08754, March 2026. doi: 10.48550/arXiv.2603.08754. URL https://doi.org/10.48550/arXiv.2603.08754.

Guoqing Wang, Sunhao Dai, Guangze Ye, Zeyu Gan, Wei Yao, Yong Deng, Xiaofeng Wu, and Zhenzhe Ying. Information gain-based policy optimization: A simple and effective approach for multi-turn search agents, 2026a. URL https://arxiv.org/abs/2510.14967.

Haoyu Peter Wang, Yuxin Chen, Liang Luo, Buyun Zhang, Ellie Dingqiao Wen, and Pan Li. Implicit turn-wise policy optimization for proactive user-LLM interaction. In *Forty-third International Conference on Machine Learning*, 2026b. URL https://openreview.net/forum?id=QJC5WymFK1.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-shepherd: Verify and reinforce LLMs step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9426–9439. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.510. URL https://aclanthology.org/2024.acl-long.510/.

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. *CoRR*, abs/2504.20073, April 2025. doi: 10.48550/arXiv.2504.20073. URL https://doi.org/10.48550/arXiv.2504.20073.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3–4):229–256, 1992. doi: 10.1007/BF00992696.

Zhiheng Xi, Yiwen Ding, Wenxiang Chen, Boyang Hong, Honglin Guo, Junzhe Wang, Xin Guo, Dingwen Yang, Chenyang Liao, Wei He, Songyang Gao, Lu Chen, Rui Zheng, Yicheng Zou, Tao Gui, Qi Zhang, Xipeng Qiu, Xuanjing Huang, Zuxuan Wu, and Yu-Gang Jiang. AgentGym: Evaluating and training large language model-based agents across diverse environments. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 27914–27961, 2025a. doi: 10.18653/v1/2025.acl-long.1355. URL https://aclanthology.org/2025.acl-long.1355/.

Zhiheng Xi, Chenyang Liao, Guanyu Li, Yajie Yang, Wenxiang Chen, Zhihao Zhang, Binghai Wang, Senjie Jin, Yuhao Zhou, Jian Guan, Wei Wu, Tao Ji, Tao Gui, Qi Zhang, and Xuanjing Huang. Agentprm: Process reward models for llm agents via step-wise promise and progress, 2025b. URL https://arxiv.org/abs/2511.08325.

Yutao Xie, Nathaniel Thomas, Nicklas Hansen, Yang Fu, Li Erran Li, and Xiaolong Wang. TIPS: Turn-level information-potential reward shaping for search-augmented LLMs. *CoRR*, abs/2603.22293, 2026. doi: 10.48550/arXiv.2603.22293. URL https://doi.org/10.48550/arXiv.2603.22293.

Yuanda Xu, Zhengze Zhou, Hejian Sang, Xiaomin Li, Jiaxin Zhang, Xinchen Du, Zhipeng Wang, and Alborz Geramifard. TRIAGE: Role-typed credit assignment for agentic reinforcement learning. *CoRR*, abs/2606.32017, June 2026. doi: 10.48550/arXiv.2606.32017. URL https://doi.org/10.48550/arXiv.2606.32017.

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, 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. *CoRR*, abs/2412.15115, 2024. doi: 10.48550/arXiv.2412.15115. URL https://doi.org/10.48550/arXiv.2412.15115.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Woongyeong Yeo, Yumin Choi, Taekyung Ki, and Sung Ju Hwang. HINT-SD: Targeted hindsight self-distillation for long-horizon agents. *CoRR*, abs/2605.17873, May 2026. doi: 10.48550/arXiv.2605.17873. URL https://doi.org/10.48550/arXiv.2605.17873.

Lifan Yuan, Wendi Li, Huayu Chen, Ganqu Cui, Ning Ding, Kaiyan Zhang, Bowen Zhou, Zhiyuan Liu, and Hao Peng. Free process rewards without process labels. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 73511–73525. PMLR, 2025. URL https://proceedings.mlr.press/v267/yuan25c.html.

Yaocheng Zhang, Haohuan Huang, Zijun Song, Yuanheng Zhu, Qichao Zhang, Zijie Zhao, and Dongbin Zhao. Criticsearch: Fine-grained credit assignment for search agents via a retrospective critic. *arXiv preprint arXiv:2511.12159*, 2025a.

Zhenru Zhang, Chujie Zheng, Yangzhen Wu, Beichen Zhang, Runji Lin, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. The lessons of developing process reward models in mathematical reasoning. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 10495–10516, Vienna, Austria, July 2025b. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.547. URL https://aclanthology.org/2025.findings-acl.547/.

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, 2025. URL https://arxiv.org/abs/2503.15478.

A Appendix

A.1 Related Work

Reinforcement learning with LLMs has been in wide use to train agents capable of solving complex multi-turn interactive tasks (Feng et al., 2025; Jin et al., 2025). As the horizon of the task increases, credit assignment becomes a bottleneck for effective and efficient learning. Preliminary works focused on redistributing a single outcome reward to individual tokens or reasoning steps. At the token level, apart from critic-based PPO (Schulman et al., 2017), VinePPO (Kazemnejad et al., 2025) and OPPO (Li et al., 2026d) discard the learned critic and estimate token values from Monte-Carlo rollouts or value recursion, while GRPO-λ (Parthasarathi et al., 2025) spreads the group-level advantage across tokens through λ-return / eligibility-trace weighting. At the step level, PURE (Cheng et al., 2025) aggregates a step-level process reward model with a min-form rule to curb reward hacking, and iStar and PRIME (Liu et al., 2026b; Cui et al., 2025) extract implicit per-step rewards from training rollouts. These methods evaluate each action's likelihood of reaching a successful outcome, therefore we use Liu et al. (2026b) as a foresight baseline in our experiments.

Turn-level. For multi-turn agents, several families of methods assign credit at the granularity of a whole turn rather than a token. *Value-based turn PRMs* learn a turn-level value function, e.g. AgentPRM (Xi et al., 2025b) (TD + GAE turn values) and SWEET-RL (Zhou et al., 2025) (an asymmetric critic given privileged training-time information). *Turn-level policy-gradient* methods recast the interaction as a turn-level MDP and use turn-level importance ratios, as in Turn-PPO (Li et al., 2026b) and SORL (Li et al., 2026a). *Information-theoretic* methods such as IGPO (Wang et al., 2026a) credit a turn by how much it increases the probability of eventual task success. *Implicit / log-prob* methods, e.g. ITPO (Wang et al., 2026b) and ΔBelief-RL (Auzina et al., 2026), read off dense per-turn rewards from shifts in the model's own log-probabilities. Finally, *potential-based reward shaping* methods (Liu et al., 2026a; Xie et al., 2026) add a dense per-turn signal derived from a state potential while preserving the optimal policy.

Hindsight-based. Several works utilize hindsight in different forms to assign credit to each turn. HCAPO (Tan et al., 2026) uses generative scoring to prompt the training actor policy to score each action directly during training. C3 (Chen et al., 2026) estimates counterfactual credit through leave-one-out ablations of actions, extending the construction to the multi-agent setting. CCPO (Li et al., 2026e) formalizes the trajectory as a structural causal model and assigns credit via the average treatment effect of each action on the final return. CriticSearch (Zhang et al., 2025a) specializes the retrospective asymmetric critic to search agents, producing turn-level credit from privileged hindsight information. CRAFT (Meng & Chen, 2026) derives signed, per-token counterfactual credit from free sibling rollouts branched off the same trajectory. HINT-SD (Yeo et al., 2026) targets long-horizon agents by self-distilling hindsight-relabeled trajectories, using autoregressively generated feedback, teaching the policy to imitate the pivotal actions that hindsight identifies as responsible for success. Similarly, HPO (Jiang et al., 2026) converts final outcomes into per-action supervision by projecting both the current policy distribution and the hindsight distribution into an intent space and extracting low-variance learning signals from the Wasserstein distance between them. Finally, RICL (Chen et al., 2025) prompts a judge LLM for retrospective feedback on intermediate actions. Common to these methods is that the hindsight signal is obtained either by external annotation, through additional counterfactual rollouts or causal modeling, through autoregressive generation, or through handcrafted distance metrics to positive outcomes. The primary goal of our work is to show an efficient way to compute a task-agnostic hindsight signal from the existing rollout trajectories, without requiring any additional computation or external supervision.

A.2 Proof of Proposition 1

Proof. Throughout we write Q for Q^{π_θ}, and all probabilities and expectations are taken over trajectories $\tau \sim \mathcal{T}^{\pi_\theta}(s_t)$ that continue from s_t under the policy π_θ. By construction the policy is the marginal action distribution at s_t, $\pi_\theta(a_t \mid s_t) = \mathbb{P}(A_t = a_t)$, and the hindsight distribution is the posterior $h(a_t \mid s_t, f(\tau_{>t})) = \mathbb{P}(A_t = a_t \mid f(T_{>t}) = f(\tau_{>t}))$, where $T_{>t}$ is the random future suffix. Write $F = f(T_{>t})$ for the random future statistic, and let f' range over its support. By assumption, $G_t = g(F)$ for some deterministic function g.

By definition, $Q(s_t, a_t) = \mathbb{E}_{\tau \sim \mathcal{T}^{\pi_\theta}(s_t,a_t)}[G_t] = \mathbb{E}[G_t \mid A_t = a_t]$, since conditioning a policy rollout on $A_t = a_t$ yields exactly the action-conditioned distribution $\mathcal{T}^{\pi_\theta}(s_t, a_t)$. Expanding the hindsight-weighted expectation and substituting $\rho_t = h/\pi_\theta$ together with $G_t = g(F)$,

$$\begin{aligned}\mathbb{E}[\rho_t\, G_t] &= \sum_{f'} \mathbb{P}(F = f')\, \frac{\mathbb{P}(A_t = a_t \mid F = f')}{\mathbb{P}(A_t = a_t)}\, g(f') \\ &= \sum_{f'} \mathbb{P}(F = f' \mid A_t = a_t)\, g(f') \\ &= \mathbb{E}[g(F) \mid A_t = a_t] = \mathbb{E}[G_t \mid A_t = a_t] = Q(s_t, a_t),\end{aligned}$$

where the second line uses Bayes' rule and the sums are replaced by integrals in the continuous case. □

A.3 Proof of Proposition 2

Proof. With $\gamma = 1$ we have $\widehat{Q}^{H}_{i,t} = \rho_{i,t}\, R(\tau_i)$, so the z-scored advantage expands as

$$\begin{aligned}A^{\text{HA-GRPO}}_{i,t} &= \frac{\rho_{i,t}\, R(\tau_i) - \frac{1}{NT}\sum \rho_{i,t}\, R(\tau_i)}{\sigma(\rho_{i,t}\, R(\tau_i))} \\ &= \frac{R(\tau_i) + \delta_{i,t} R(\tau_i) - \frac{1}{NT}\sum \big(R(\tau_i) + \delta_{i,t} R(\tau_i)\big)}{\sigma(R(\tau_i) + \delta_{i,t} R(\tau_i))} \\ &= \frac{R(\tau_i) - \frac{1}{NT}\sum R(\tau_i)}{\sigma_{R+\delta R}} + \frac{\delta_{i,t} R(\tau_i) - \frac{1}{NT}\sum \delta_{i,t} R(\tau_i)}{\sigma_{R+\delta R}},\end{aligned}$$

where the sums run over all NT action tokens in the group. The second line substitutes $\rho_{i,t} = 1 + \delta_{i,t}$, and the third splits the mean-subtracted numerator (mean subtraction is linear) over the shared denominator $\sigma_{R+\delta R} := \sigma(R(\tau_i) + \delta_{i,t} R(\tau_i))$. Recognizing each centered term as a z-score, and using $A^{\text{GRPO}}_{i,t} = \big(R(\tau_i) - \frac{1}{NT}\sum R(\tau_i)\big)/\sigma_R$ standardized over the same group,

$$A^{\text{HA-GRPO}}_{i,t} = \frac{\sigma_R}{\sigma_{R+\delta R}}\, A^{\text{GRPO}}_{i,t} + \frac{\sigma_{\delta R}}{\sigma_{R+\delta R}}\, \mathcal{Z}\big(\delta_{i,t} R(\tau_i)\big).$$

If $\delta_{i,t} = 0$ for all (i, t), the second term vanishes and $\sigma_{R+\delta R} = \sigma_R$, so $A^{\text{HA-GRPO}}_{i,t} = A^{\text{GRPO}}_{i,t}$. □

A.4 Training Hyperparameters

Hyperparameters for ALFWorld. All methods share identical base hyperparameters. The maximum prompt length is 2048 tokens and the maximum response length is 512 tokens. Each episode allows up to 50 environment steps. The actor learning rate is 1e-6 (all methods are critic-free, GRPO-style). The reward is rule-based: +10 on task success plus a fractional goal-condition-success shaping term in [0,1]; invalid actions are penalized with a policy-loss coefficient of 0.1. For all group-based methods (GRPO, HCAPO, LHCA, and the frozen-adapter variant) we use a group size of G = 8 and sample 16 groups per rollout, giving 16 x 8 = 128 parallel environments. The rollout temperature is 1.0 and the validation temperature is 0.4 (sampling enabled), the PPO mini-batch size is 256, the clip ratio is 0.2, and the KL-loss coefficient β_{KL} is 0.01 (low_var_kl applied as a loss). We use in distribution eval for validation and unseen tasks as final test set.

Hyperparameters for Sokoban. We use the 5x5 single-box text (ASCII) variant with search depth 30. The maximum prompt and response lengths are 2048 and 512 tokens. Each episode is limited to 15 environment steps. The actor learning rate is 1e-6. We use the standard gym-Sokoban shaped reward (step penalty -0.1, box-on-target +1, box-off-target -1, puzzle solved +10, invalid action -1) and keep the additional policy-loss invalid-action penalty coefficient as 0.1. These intermediate rewards are summed to create a trajectory-level sparse reward for our setting. All group-based methods use G = 16 and sample 32 groups per rollout, i.e. 32 x 16 = 512 parallel environments. The rollout temperature is 1.0 and the validation temperature is 0.4. The PPO mini-batch size is 256 for the HCA-family methods and 128 for the GRPO baseline; the clip ratio is 0.2 and β_{KL} = 0.001.

Hyperparameters for BabyAI. The maximum prompt length is 5120 tokens and the maximum response length is 512 tokens (max model length 5632); an observation history window of 5 turns is used. Each episode allows up to 20 environment steps. The actor learning rate is 1e-6. The reward is the native BabyAI success reward (higher for faster completion), with success defined as reward > 0; the invalid-action penalty coefficient is 0.1. All group-based methods use G = 8 and sample 16 groups per rollout (16 x 8 = 128 parallel environments). Both the rollout and validation temperatures are 1.0. The PPO mini-batch size is 64, the clip ratio is 0.2, and β_{KL} = 0.01.

Computing Details. Qwen2.5-1.5B-Instruct experiments were conducted on 4xH100 GPUs (TP=2 for Alfworld and TP=1 for Sokoban) for 150 iterations. Qwen2.5-7B-Instruct experiments were done on 8xH100 GPUs for 200 iterations. All methods use the bypass-mode rollout log-prob reuse, in which the rollout log-probs are taken directly as the behavior policy's old log-probs.

LHCA implementation details. Algorithm 1 shows the core update; in experiments, we first normalize the hindsight ratios within each trajectory and clip them as in equation 11. Using the resulting $\rho'_{i,t}$, we compute the hindsight advantage, apply the do-no-harm mask, and then add the GRPO macro term:

$$
\begin{aligned}
A^H_{i,t} &= \mathcal{Z}\left(\rho'_{i,t}\,\gamma^{T_i-t} R(\tau_i)\right), \\
\widetilde{A}^H_{i,t} &= \begin{cases} \max\{A^H_{i,t}, 0\}, & \text{if trajectory } i \text{ succeeds,} \\ A^H_{i,t}, & \text{otherwise,} \end{cases} \\
\hat{A}_{i,t} &= A^{\mathrm{GRPO}}_i + \omega\, \widetilde{A}^H_{i,t},
\end{aligned}
$$

where T_i is the number of turns in trajectory i. The final advantage $\hat{A}_{i,t}$ is broadcast to the action tokens in equation 3. Thus, the mask acts on the hindsight term before the GRPO term is added.

LHCA Specific Hyperparameters The hindsight importance ratio $\rho_{i,t}$ is clipped to [C_{min}, C_{max}] = [0.6, 1.4] to prevent training instability from extreme posterior estimates. The micro-advantage weight omega is 1.0, the discount factor gamma (used in the γ^{T-t} term) is 0.95. The online LoRA credit adapter (LHCA only) is warmed up for 10 steps before its micro-advantage is applied, and is refreshed online during training with an adapter learning rate of 2e-5 for the 1.5B models (ALFWorld, Sokoban) and 1e-5 for the 7B model (BabyAI).

A.5 iSTAR TRAINING INSTABILITY

We implement iStar (Liu et al., 2026b) as a foresight baseline and are able to reproduce stable training using the GRPO+KL setting on the tasks from Liu et al. (2026b).

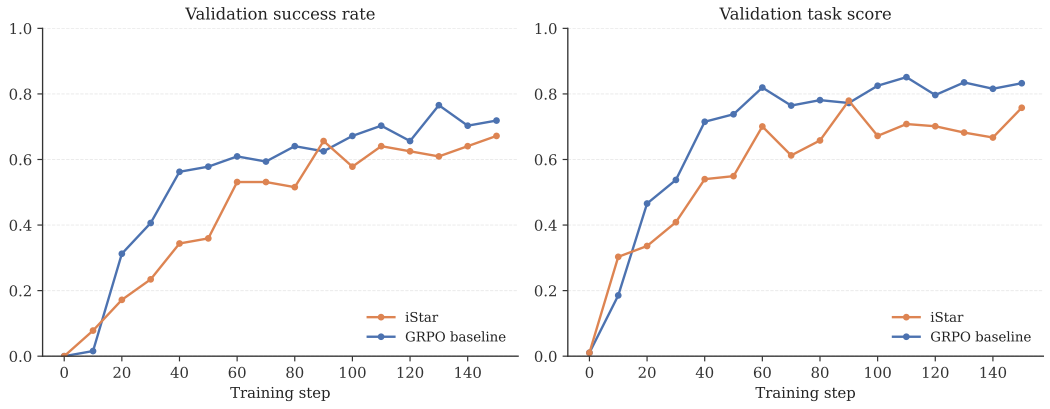


Figure 4: iStar training on a task from Liu et al. (2026b) showing stable training.

However when we switch to our tasks, Alfworld, Sokoban and BabyAI, we find that iStar is extremely unstable to train. We find that the valid-action ratio collapses from near 1.0 early in training

down to a small fraction of the trajectory before partially and noisily recovering, with episode reward tracking the same collapse. BabyAI is comparatively closer to stable, with iStar's smoothed train success rate following a similar noisy trend to the GRPO baseline rather than collapsing outright. The reason for this selective instability is unclear and we leave a thorough investigation of this discrepancy to future work.

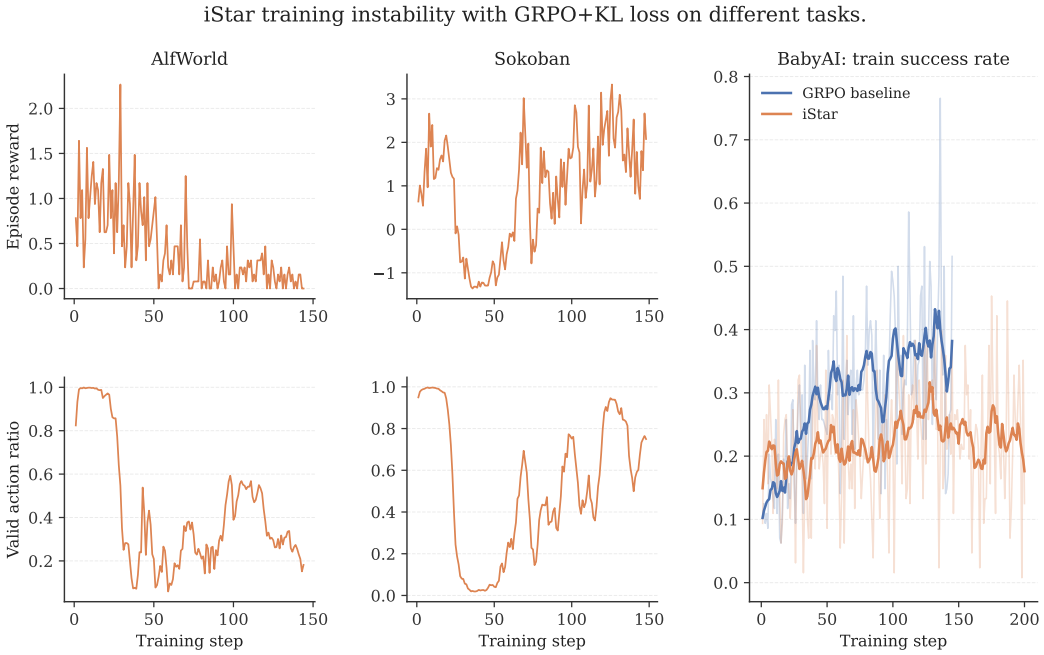


Figure 5: iStar training instability across AlfWorld, Sokoban, and BabyAI. **Left, middle columns:** raw (unsmoothed) episode reward and valid-action ratio for iStar on AlfWorld and Sokoban. **Right:** BabyAI train success rate for iStar vs. the GRPO baseline; faint lines are the raw per-step curves, bold lines are a moving average.