MITIGATING EARLY-TURN CREDIT FAILURE IN AGENTIC RL VIA ROLLOUT-FREE REATTRIBUTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-horizon capabilities are essential for LLM agents to solve complex tasks that require iterative tool use and multi-step interaction. Reinforcement learning, the predominant training paradigm for such agents, typically provides no intermediate rewards and reveals success only through a terminal reward. Learning from this sparse feedback requires turn-level credit assignment to estimate how each action contributes to the outcome and therefore determine its learning signal. In this work, we show theoretically and empirically that current critic-free methods suffer from *early-turn credit failure*, yielding unreliable credit in early turns. At these turns, Monte Carlo estimates are dominated by sampling noise, while final-outcome surrogate methods suffer from the weak connection between early actions and the outcome. Consequently, crucial early actions that shape the subsequent trajectory receive weak learning signals, limiting effective training on long-horizon tasks. To address this, we propose FOLIA, a rollout–free plug-in method that combines local influence with terminal hindsight for early-turn credit estimation. Specifically, FOLIA measures each action's local influence by counterfactually ablating the observed environment response, anchors this influence to the final outcome through a terminal hindsight score, and fuses the two signals with a confidence gate to reattribute credit across turns. Extensive experiments across diverse long-horizon agentic tasks show that FOLIA consistently improves existing RL algorithms through better credit assignment with negligible computational overhead.[1]

1 INTRODUCTION

Practical tasks such as multi-hop search, online shopping, and software engineering often require sequences of interdependent decisions informed by environmental feedback. To solve these long-horizon tasks, large language model (LLM) agents interleave reasoning and tool use, using feedback from each interaction to inform subsequent decisions (Zhou et al., 2024; Yao et al., 2023; Xie et al., 2024; Zheng et al., 2024; Gou et al., 2025; Li et al., 2025). Although reinforcement learning (RL) has achieved substantial success in improving LLMs on single-turn tasks like mathematics and coding (Jimenez et al., 2024; Wei et al., 2025; Guo et al., 2025; Liu et al., 2025a), extending it to long-horizon tasks presents a key challenge: agents typically receive only a terminal, trajectory-level reward after making many decisions, whereas policy optimization requires turn-level credit for each action. The outcome reward indicates only whether the trajectory succeeds, making it difficult to derive reliable turn-level credit and potentially producing weak or misleading training signals.

To mitigate the weak or misleading training signals caused by sparse outcome rewards, prior work employs dense turn-level supervision to estimate how each intermediate action contributes to the final outcome. Existing process-supervision methods, however, often rely on costly per-step targets obtained through annotation (Xiong et al., 2024; Chen et al., 2025; Abbasi et al., 2025), powerful LLM judges queried throughout training (Tan et al., 2026a; Yang et al., 2026; Liu et al., 2025b), or separately trained reward models (Liu et al., 2026; Deng et al., 2025). This limits their scalability for online credit assignment during training. Another line of critic–free work avoids external supervision by estimating turn–level credit directly from sampled trajectories, using either Monte Carlo (MC) estimation (Feng et al., 2025a; Kazemnejad et al., 2025a) or surrogate objectives anchored to the final answer (Jiang et al., 2026; Tan et al., 2026b; Wang et al., 2024). However, both approaches struggle with early-turn credit assignment. For MC methods, early-turn signals are weak relative

[1] Our code is available at Link.

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.

to sampling noise, as success depends on many subsequent decisions; reliable estimation therefore requires impractically many rollouts. Terminal-anchored methods instead rely on the association between an early action and the distant final answer, overlooking important early steps that are not directly related to the answer. Consequently, early actions often receive unreliable credit despite playing a crucial role. This failure further bottlenecks long-horizon agent training.

In this work, we theoretically characterize and empirically quantify *early-turn credit failure*, showing that sampling noise dominates early-turn credit estimates under practical rollout budgets during training. Obtaining reliable estimates requires a rollout budget that grows exponentially with the remaining horizon, making repeated generation and environment interaction prohibitively costly for long-horizon training. We therefore propose **FOLIA** (**F**using Terminal **O**utcomes and **L**ocal **I**nfluence for Credit Re**A**ttribution), a rollout-free plug-in method for more accurate turn-level credit assignment. Its core insight is to assess an action's intermediate influence and ground this assessment in outcome-based hindsight. Specifically, FOLIA uses counterfactual ablation to estimate the magnitude of each turn's influence on the subsequent action and applies terminal hindsight to determine the credit direction based on whether the turn contributes to the final outcome. A confidence gate fuses the two signals to reattribute the base method's original credit, which is then used for subsequent agent training. Extensive experiments show that FOLIA consistently improves various base RL methods with minimal computational overhead, while further analyses demonstrate its ability to provide finer-grained credit assignment.

In summary, our contributions are as follows:

- We characterize *early-turn credit failure* in long-horizon agentic RL, showing theoretically and empirically that sampling noise dominates early-turn credit estimates under practical rollout budgets and that methods using outcomes as surrogate credit signals also fail to resolve it.
- We propose FOLIA, a rollout-free, plug-in credit-reattribution method that fuses immediate counterfactual influence with an outcome-based hindsight anchor through a confidence gate.
- Extensive experiments and analyses show that FOLIA enables finer-grained credit assignment and consistently improves various RL baselines across long-horizon tasks and model scales, with negligible computational overhead.

2 PRELIMINARIES

In this section, we first formulate the agentic task under study and then introduce credit assignment and the MC methods used in agentic training.

Task formulation. We model agent–environment interaction as a finite-horizon partially observable decision process. Given a query q and its ground-truth answer $y^\star$, the LLM policy π interacts with an environment for T turns. We denote the complete interaction history after turn t by c_t, with $c_0 = q$. At turn t, the policy generates an action $r_t \sim \pi(\cdot \mid c_{t-1})$, where r_t includes the agent's reasoning tokens and issued tool call. We consider deterministic tool responses: given c_{t-1} and r_t, the environment returns a fixed response e_t, updating the history to $c_t = (q, r_1, e_1, \ldots, r_t, e_t)$. The interaction yields a trajectory $\tau = \big(q, (r_1, e_1), \ldots, (r_T, e_T), y\big)$, where y is the final answer. The environment assigns a binary terminal reward $R(\tau) = \mathbb{1}[y = y^\star] \in \{0, 1\}$. Agent training maximizes $J(\pi) = \mathbb{E}_{\tau\sim\pi}[R(\tau)]$, where the expectation is over trajectories induced by sampling from π.

Credit Assignment. In RL, credit assignment measures each action's contribution to expected return. In this setup, we study turn-level credit, formalized through the standard value functions (Sutton and Barto, 1998). The state value $V^\pi(c_t)$ is the expected terminal return from c_t, while the action value $Q^\pi(c_{t-1}, r_t)$ is the expected return after taking r_t at c_{t-1} and then following π:

$$V^\pi(c_t) := \mathbb{E}_\pi[R \mid c_t] = \mathbb{P}_\pi(R = 1 \mid c_t),\ Q^\pi(c_{t-1}, r_t) := \mathbb{E}_\pi[R \mid c_{t-1}, r_t] = \mathbb{P}_\pi(R = 1 \mid c_{t-1}, r_t).$$

The credit of turn t is its advantage, $A^\pi(c_{t-1}, r_t) = Q^\pi(c_{t-1}, r_t) - V^\pi(c_{t-1})$. We consider the setup where the environment response e_t is deterministic given the tool call issued in r_t, so taking r_t at c_{t-1} fixes the successor context c_t. Hence $Q^\pi(c_{t-1}, r_t) = V^\pi(c_t)$, and the credit reduces to the value increment across the realized transition, $\Delta_t := V^\pi(c_t) - V^\pi(c_{t-1})$. Because the reward is binary, $V^\pi(c_t) = \mathbb{P}_\pi(R = 1 \mid c_t)$ is exactly the probability of eventual success from c_t; we therefore abbreviate it as $p_t := V^\pi(c_t)$, giving $\Delta_t = p_t - p_{t-1}$. Thus Δ_t measures how much turn t changes the probability of eventual success and serves as the target signal for turn-level credit assignment.

Monte Carlo estimation. Training a reliable critic is costly and challenging; a common critic-free alternative is to estimate turn-level credit Δ_t using MC sampling (Baxter and Bartlett, 2001; Williams, 1992). From each of the two endpoint contexts, c_{t-1} and c_t, we sample K independent continuations under policy π and average their terminal rewards $R_j^{(k)}$ to obtain the sample-average estimate of p_j:

$$\hat{p}_j := \frac{1}{K}\sum_{k=1}^{K} R_j^{(k)}, \qquad j \in \{t-1, t\}.$$

Based on this, the MC-based credit estimate for turn t is $\hat{\Delta}_t := \hat{p}_t - \hat{p}_{t-1}$. This estimate is unbiased (Sutton and Barto, 1998; Thomas et al., 2015; Kazemnejad et al., 2025a): $\mathbb{E}_{\text{aux}}[\hat{\Delta}_t] = \Delta_t$, where $\mathbb{E}_{\text{aux}}$ denotes expectation over the sampled auxiliary continuations.

3 Failure of Early-Turn Credit Estimation

MC-based credit estimators are unbiased. With a sufficiently large rollout budget, they can therefore provide reliable credit estimates. However, such large budgets are impractical during training, where limited sampling yields noisy estimates. We show theoretically why this issue is severe for early turns, validate it empirically, and explain why existing methods do not resolve it.

3.1 Theoretical Analysis of Early-Turn Failure

We measure the reliability of MC-based credit estimation using the signal-to-noise ratio (SNR), $\textbf{SNR}_t := |\mathbb{E}_{\text{aux}}[\hat{\Delta}_t]|/\sqrt{\text{Var}_{\text{aux}}(\hat{\Delta}_t)}$. Our goal is to characterize how the SNR changes with the number of unresolved stages after a turn. Conditional on the fixed contexts c_{t-1} and c_t, the MC estimates are obtained from two independently sampled groups of K fresh continuations, one from each context; their sample means are therefore independent under this conditioning despite the shared history. We show that, for fixed K, the SNR collapses as successful continuations become increasingly rare: the credit signal shrinks faster than the standard deviation of its MC estimate. Recall the agentic task setting in Sec. 2, where the policy interacts with the environment over multiple turns and receives a positive terminal reward only for a correct final answer. When reaching a correct answer requires passing through intermediate steps in sequence, we use the following assumption to capture how their difficulty accumulates.

Assumption 3.1 (Sequential hardness)**.** For a context c, let $h(c)$ denote the number of unresolved ordered bottlenecks, *i.e.*, necessary intermediate steps that every successful continuation must pass in sequence, and, when $V^\pi(c) > 0$, let $q_j(c)$ be the conditional probability of passing bottleneck j after all preceding bottlenecks have been passed. There exist constants $\lambda > 0$ and $h_0 \geq 1$ such that every context with $V^\pi(c) > 0$ and $h(c) \geq h_0$ satisfies

$$\frac{1}{h(c)}\sum_{j=1}^{h(c)} -\log q_j(c) \geq \lambda.$$

This assumption suits real trajectories: it only requires non-vanishing average difficulty across the remaining stages and does not require the bottlenecks to be independent or equally difficult. Since every successful continuation must pass all remaining bottlenecks, it implies $V^\pi(c) \leq \prod_{j=1}^{h(c)} q_j(c) \leq e^{-\lambda h(c)}$ for $V^\pi(c) > 0$ and $h(c) \geq h_0$; the final bound also holds trivially when $V^\pi(c) = 0$.

Proposition 3.2 (Early-turn SNR collapse)**.** *Consider a turn t with $\min\{h(c_{t-1}), h(c_t)\} \geq h$. Under Assumption 3.1, if $h \geq \max\{h_0, (\log 2)/\lambda\}$, then*

$$\text{SNR}_t \leq 2\sqrt{K}\, e^{-\lambda h/2}.$$

Consequently, maintaining $\text{SNR}_t \geq \eta > 0$ requires $K \geq (\eta^2/4)e^{\lambda h}$. Maintaining a non-vanishing SNR with this MC estimator therefore requires an exponentially growing rollout budget.

Proof sketch. Let $\rho_t := p_t + p_{t-1}$. Sequential hardness gives $\rho_t \leq 2e^{-\lambda h}$. Using $|\Delta_t| \leq \rho_t$ and $\text{Var}_{\text{aux}}(\hat{\Delta}_t) \geq \rho_t/(2K)$, we obtain

$$\text{SNR}_t \leq \sqrt{2K\rho_t} \leq 2\sqrt{K}\, e^{-\lambda h/2}.$$

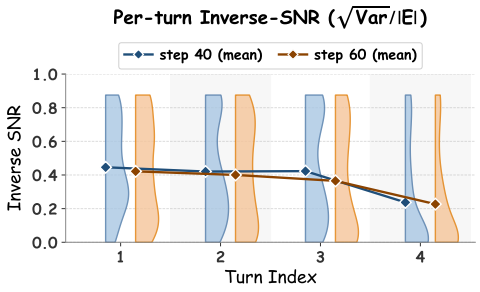
The complete proof is provided in Appendix A.1. □

The proposition identifies a fundamental signal–noise mismatch: the true credit magnitude is at most ρ_t, whereas the standard deviation of its MC estimate is of order $\sqrt{\rho_t/K}$. Moreover, $\mathbb{P}_{\text{aux}}(\hat{p}_t = \hat{p}_{t-1} = 0) \geq 1 - 2Ke^{-\lambda h}$, so both rollout groups contain only zero rewards with high probability when $2Ke^{-\lambda h} \ll 1$, as can occur at early turns with many unresolved bottlenecks (large h). In this case, $\hat{\Delta}_t = 0$ even if the true credit Δ_t is nonzero. This reliance on sampling successful continuations illustrates the unreliability of MC-based credit estimation in this regime.

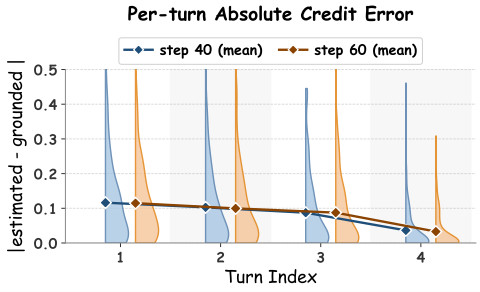
3.2 Empirical Evidence of Early-turn Failure

We further examine whether the theoretically established *early-turn failure* occurs under practical training conditions. Specifically, we train Qwen2.5-3B-Ins (Yang et al., 2024) with GRPO (Shao et al., 2024) and use checkpoints from training steps 40 and 60 (out of 100) as fixed policies to sample trajectories from the evaluation set. We then assess these trajectories using a post hoc diagnostic based on VinePPO-style prefix resampling (Kazemnejad et al., 2025a) to obtain MC-based turn-level credit estimates. For each trajectory, we truncate it at the end of each turn, sample continuations using the corresponding policy, and estimate turn-level credit as the difference between the estimates of consecutive prefixes. To mirror a common training group size for this task, we use $K = 8$ continuations per prefix for the limited-budget estimate and $K = 256$ for the high-budget reference.

With this method, we test whether the practical rollout budget yields reliable credit estimates, particularly at early turns, where our analysis predicts less reliable estimates. At each turn, we measure SNR, which captures credit signal strength relative to sampling noise, and absolute deviation from the high-budget reference, which measures realized estimation error. Together, these metrics assess reliability and accuracy across turns at a training-scale sampling budget.



(a) Distribution of inverse SNR across turns.



(b) Distribution of absolute error relative to grounded estimates.

Figure 1: Empirical unreliability of early-turn MC credit estimates. Violin width indicates density at each metric value, while dashed lines mark means. (a) Inverse SNR, where higher values indicate weaker signals relative to sampling noise. (b) Absolute deviation from high-budget estimates; higher values indicate larger errors. Early turns show higher values and greater density at large values on both metrics, indicating less reliable credit estimates under limited sampling.

We present the results in Fig. 1, using inverse SNR for clearer visualization. Both inverse SNR and absolute deviation are highest at early turns and decrease as the trajectory progresses. This pattern remains consistent at both Steps 40 and 60, indicating that the training process does not naturally alleviate the issue. These results confirm that early-turn failure is a practical training phenomenon beyond a theoretical concern. Under practical rollout budgets, early-turn MC credit has a lower SNR and higher error, yielding less reliable training signals and bottlenecking agentic policy optimization.

3.3 Can Outcome-anchored Credit Signals Resolve Early-turn Failure?

The preceding analysis traces early-turn failure in MC estimation to the impractical cost of sampling long, low-probability successful continuations. A natural remedy is to infer turn-level credit directly from the terminal outcome without additional rollouts: IGPO (Wang et al., 2026a) measures how each prefix shifts the likelihood of the ground-truth answer, and HCAPO (Tan et al., 2026b) evaluates an action's compatibility with the revealed outcome. These methods avoid auxiliary-rollout sampling noise, but they still assign credit by directly measuring an action's association with the final outcome. For early actions, this association is often weak because their influence on the outcome is mediated by many subsequent decisions, making direct outcome-based credit estimates unreliable.

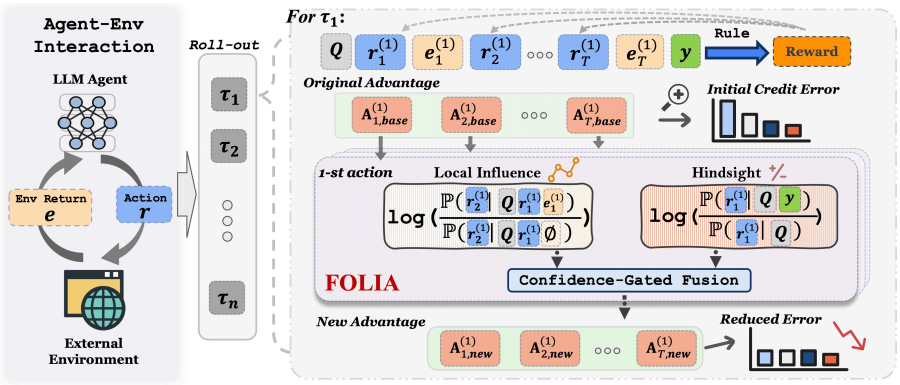


Figure 3: Pipeline for our FOLIA method. FOLIA addresses the challenge of high early-turn credit estimation error by computing the observed response's influence on the realized next turn and then grounding it with hindsight-based credit. A confidence gate then fuses the two signals to redistribute the base credit.

Fig. 2 validates this observation. Extending the analysis protocol in Sec. 3.2 to outcome-anchored methods, we compute credit using each defined surrogate, normalize the estimates to $[-1, 1]$, and measure their absolute error relative to the high-budget grounded credit. Both methods show higher error at early turns: IGPO's first-turn error is 0.49, compared with 0.33 at the final turn. HCAPO has the lowest overall error, yet its error rises from 0.25 near the answer to 0.33 at the first turn. This analysis reveals that reliable early-turn credit requires a signal that does not rely solely on the distant outcome, which we develop next.

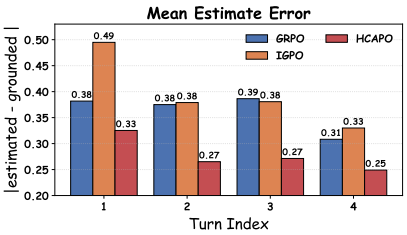


Figure 2: Per-turn average absolute error between each method's estimate and the grounded credit. Outcome-anchored credit does not resolve early-turn failure.

4 FOLIA: Fusing Terminal Outcomes and Local Influence

To mitigate early-turn credit failure caused by relying on terminal outcomes alone, we propose FOLIA. Requiring no additional rollouts, it uses counterfactual perturbation to measure how each turn immediately changes subsequent behavior and then uses hindsight to determine whether each action supports the final outcome. An uncertainty-aware fusion of the two signals reweights the base RL method's advantages, which are then used in subsequent training.

4.1 Capturing Immediate Effect through Counterfactual Influence

Directly estimating an early turn's contribution from the terminal reward is difficult. We instead use its earliest observable footprint in the trajectory. In agent–environment interaction, the action r_t elicits an observation e_t, which updates the decision context and shifts the agent's next-turn behavior. We quantify this response-mediated policy shift as an ablation-based proxy for turn t's effect on subsequent behavior, without yet determining whether that effect supports the final outcome.

Formally, let $\bar{c}_t = (c_{t-1}, r_t)$ denote the context immediately before observing e_t, such that $c_t = (\bar{c}_t, e_t)$. We construct an ablated context $c_t^{\emptyset} = (\bar{c}_t, \emptyset)$ by replacing e_t with $\emptyset$, the tool's empty-result message. The substitution is textual and therefore requires no further environment interaction. Under the frozen behavior policy $\pi_{\theta_{\text{old}}}$, we define the counterfactual influence score of turn t as

$$\delta_t = \frac{1}{|r_{t+1}|} \sum_{i=1}^{|r_{t+1}|} \log \frac{\pi_{\theta_{\text{old}}}(r_{t+1,i} \mid c_t, r_{t+1,<i})}{\pi_{\theta_{\text{old}}}(r_{t+1,i} \mid c_t^{\emptyset}, r_{t+1,<i})}, \qquad t = 1, \ldots, T-1.$$

Thus, δ_t is the average token-level log-likelihood ratio of the observed next turn under the factual and ablated contexts. Letting $\pi_t = \pi_{\theta_{\text{old}}}(\cdot \mid c_t)$ and $\pi_t^{\emptyset} = \pi_{\theta_{\text{old}}}(\cdot \mid c_t^{\emptyset})$, we get

$$|r_{t+1}|\,\delta_t = \log \frac{\pi_t(r_{t+1})}{\pi_t^{\emptyset}(r_{t+1})}, \quad \mathbb{E}_{r_{t+1}\sim\pi_t}[|r_{t+1}|\,\delta_t] = D_{\text{KL}}\left(\pi_t \,\|\, \pi_t^{\emptyset}\right),$$

where π_t and $\pi_t^{\emptyset}$ are distributions over complete next turns. Under this formulation, δ_t measures the policy shift along the realized next turn, while its length-weighted expectation quantifies the

overall divergence between the factual and ablated next-turn policies. Computing δ_t requires only teacher-forced likelihood evaluation, without additional sampling.

For an individual trajectory, δ_t reveals the direction of this shift: positive, near-zero, and negative values indicate that e_t increases, barely changes, or decreases the likelihood of the observed next turn relative to the ablated context, respectively. However, δ_t measures behavioral influence rather than reward value, as misleading feedback can strongly shift behavior away from success. Thus, δ_t captures a turn's effect on subsequent behavior without determining whether that effect contributes to the final outcome.

4.2 Grounding Immediate Influence through Terminal Hindsight

Immediate influence reveals whether a turn changes subsequent behavior, but not whether that change is beneficial. We therefore apply a rollout-free terminal anchor based on the existing trajectory's outcome, inspired by hindsight credit (Harutyunyan et al., 2019), to determine whether the change supports the final answer.

Let G denote terminal information used as a hindsight anchor. The ideal hindsight score for turn t is

$$\eta_t^\star := \log \frac{\mathbb{P}_{\pi_{\theta_{\text{old}}}}(r_t, e_t \mid c_{t-1}, G)}{\mathbb{P}_{\pi_{\theta_{\text{old}}}}(r_t, e_t \mid c_{t-1})}$$

by definition. Both terms are joint turn likelihoods, so Bayes' rule can be applied directly without separately estimating the environment likelihood. By $c_t = (c_{t-1}, r_t, e_t)$,

$$\eta_t^\star = \log \frac{\mathbb{P}_{\pi_{\theta_{\text{old}}}}(G \mid c_t)}{\mathbb{P}_{\pi_{\theta_{\text{old}}}}(G \mid c_{t-1})}.$$

Thus, hindsight measures how a completed turn changes the log-likelihood of terminal information. Computing this exactly requires marginalizing over all future policy actions and is intractable. We instead set $G = y$, using the trajectory's answer as the hindsight anchor, and approximate its likelihood by teacher-forcing y under each prefix. Normalizing by $|y|$ makes the scores comparable across answer lengths and yields a tractable, rollout-free surrogate for $\eta_t^\star/|y|$:

$$v_t := \frac{1}{|y|} \sum_{i=1}^{|y|} \log \pi_{\theta_{\text{old}}}(y_i \mid c_t, y_{<i}), \qquad \eta_t := v_t - v_{t-1}.$$

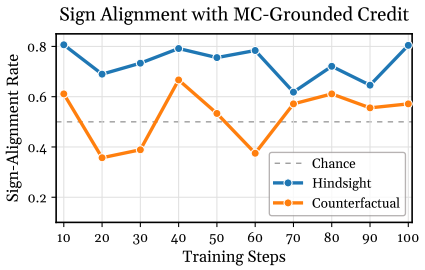


Figure 4: Sign alignment of hindsight and counterfactual signals with MC-grounded credit across training steps. Hindsight shows strong alignment, while Counterfactual approximates random guessing (Chance).

Under this approximation, η_t is a rollout-free proxy for ideal outcome lift: $\eta_t > 0$ indicates increased support for y under the frozen policy. Following Sec. 3.2, we compare hindsight and counterfactual credit directions with a high-budget MC reference on successful trajectories sampled during training, reporting the fraction of turns with matching signs. As shown in Fig. 4, hindsight aligns well with grounded credit, whereas the counterfactual signal's alignment remains near random. This supports using success-conditioned hindsight to correct counterfactual credit, consistent with prior work (Tan et al., 2026b). We exclude the final answer turn throughout to avoid trivial self-conditioning.

4.3 Confidence-Gated Fusion for Credit Redistribution

Local influence δ_t quantifies the behavioral effect of a turn but is agnostic to the terminal outcome; hindsight credit η_t provides an outcome-aligned direction, but is less reliable in magnitude for the early turns. We therefore fuse them asymmetrically to combine their complementary strengths: δ_t sets the credit weight, while η_t provides a bounded, outcome-based directional correction.

Specifically, we first transform the noisy local influence into a stable within-trajectory salience score. Given the raw signed local influence δ_t, let $v_{\delta,t}$ denote its squared standard error, which captures the variability of the per-token log-ratios. We scale $|\delta_t|$ by a reliability factor that decreases as its variability increases to obtain $\tilde{\delta}_t$, compute the length-weighted trajectory average $\bar{\delta}$, and normalize $\tilde{\delta}_t$ against this reference to obtain the final salience $\hat{\delta}_t$:

$$\tilde{\delta}_t := \frac{\delta_t^2}{\delta_t^2 + v_{\delta,t} + \epsilon_0}|\delta_t|, \qquad \bar{\delta} := \frac{\sum_j |r_j|\tilde{\delta}_j}{\sum_j |r_j|}, \qquad \hat{\delta}_t := \frac{\tilde{\delta}_t}{\tilde{\delta}_t + \bar{\delta} + \epsilon_0} \in [0, 1),$$

where $\epsilon_0 > 0$ is a stability constant. This transformation is applied to suppress small estimates that may reflect noise while preserving strong, consistent signals.

We then map the hindsight lift to a bounded confidence score, $\kappa_t = \tanh(a|\eta_t|/2)$, where a controls sensitivity. We choose $\tanh$ here because it remains sensitive to small lifts near zero while saturating for large inputs, preventing extreme hindsight values from dominating. With $s(\tau) = \mathbb{1}[R(\tau) = 1]$, we fuse the two signals as

$$u_t = \hat{\delta}_t \left[1 + s(\tau)\kappa_t \mathbb{1}[\eta_t > 0]\right].$$

The bounded multiplier lies in $[1, 2)$, so hindsight won't overwhelm the local signal.

Finally, let $p_t^0 := \frac{|r_t||A_t^{\text{base}}|}{\sum_j |r_j||A_j^{\text{base}}|}$ be turn t's share of the token-weighted base advantage, where A_t^{base} is supplied by the base algorithm and shared across all tokens in r_t. We then compute the credit redistribution weight m_t as follows:

$$m_t := \frac{\exp(u_t)}{\sum_j p_j^0 \exp(u_j)}, \qquad A_t^{\text{new}} = \text{clip}(m_t, 1 - \epsilon_{\text{clip}}, 1 + \epsilon_{\text{clip}})\, A_t^{\text{base}},$$

where $\epsilon_{\text{clip}} \in (0, 1)$ and all sums run over the reattributed turns $t = \{1, \ldots, T-1\}$. Before clipping, the normalization preserves the total token-weighted absolute advantage, since $\sum_j p_j^0 m_j = 1$. Clipping is then applied to bound changes in each advantage's magnitude relative to its initial value, after which A_t^{new} replaces A_t^{base} for training. In this way, the new credit signal contains both $\hat{\delta}_t$, which determines the magnitude of the reallocated credit, and the outcome-anchored signal η_t, which acts as a grounding gate.

5 Experiments

In this section, we evaluate the effectiveness of FOLIA. We first describe the experimental setup and then investigate the following research questions: **RQ1**: How does FOLIA improve agentic training performance? **RQ2**: How does FOLIA enhance credit assignment? **RQ3**: How do the individual credit signals and their fusion contribute to the performance improvement?

5.1 Experimental Setup

Benchmarks. We train and evaluate our method on search-based QA benchmarks and the long-horizon agentic benchmark. For search-based QA, the LLM searches the knowledge base to answer questions. We focus on multi-hop benchmarks requiring multiple search turns: HotpotQA (Yang et al., 2018), 2Wiki (Ho et al., 2020), MuSiQue (Trivedi et al., 2022), and Bamboogle (Press et al., 2023). We also include the single-hop PopQA benchmark (Mallen et al., 2023) to verify performance on shorter-horizon tasks. We further use WebShop (Yao et al., 2022), which evaluates long-horizon decision-making in a simulated e-commerce environment where an agent must search and navigate product listings, select options, and purchase an item satisfying a natural-language instruction.

Baselines and Implementation Details. We conduct training with Qwen2.5-3B-Ins and Qwen2.5-7B-Ins (Yang et al., 2024) and evaluate FOLIA as a plug-in for GRPO (Shao et al., 2024), IGPO (Wang et al., 2026a), HCAPO (Tan et al., 2026b), and GiGPO (Feng et al., 2025a). For each method, FOLIA redistributes the original advantages while keeping other settings, including tool interfaces, hyperparameters, and evaluation protocols fixed, isolating the effect of redistribution. For search-based QA, we follow established protocols (Wang et al., 2026a; Zeng et al., 2026), use E5 (Wang et al., 2022) as the retriever, and report token-level F1 against the reference. For WebShop, we follow prior setups (Feng et al., 2025a; Tan et al., 2026b) and report Task Score (mean episode reward) and Success Rate (the percentage of instructions with $r = 1$), where $r \in [0, 1]$ measures satisfaction of the requested product type, attributes, options, and price constraint. Because WebShop lacks a textual answer, we use the trajectory's final observation as the hindsight target, *i.e.*, the raw product page viewed at the agent's final action. Further setup details appear in App. A.5.

5.2 RQ1: FOLIA Improves Agentic Performance

We first evaluate the effect of FOLIA on the performance of existing RL algorithms. Tab. 1 reports token-level F1 for Search-QA benchmarks and Task Score and Success Rate for WebShop, where

Table 1: Performance of RL methods and their FOLIA-enhanced variants (shaded). We report token-level F1 on the search datasets and compute the average across four multi-hop QA benchmarks. IGPO is omitted on WebShop because it collapses beyond 10 turns. FOLIA consistently improves performance across benchmarks, demonstrating its effectiveness as a credit-reattribution method.

Method	Single-hop	Multi-hop					WebShop	
	PopQA	MuSiQue	2Wiki	HotpotQA	Bamboogle	Avg.	Score	Succ.Rate
Qwen2.5-3B-Instruct								
Prompting	0.159	0.070	0.249	0.197	0.100	0.154	0.081	0.018
GRPO	0.476	0.254	0.473	0.495	0.495	0.429	$0.774_{\pm 0.007}$	$0.510_{\pm 0.028}$
GRPO+FOLIA	**0.482**	**0.269**	**0.490**	**0.514**	**0.537**	**0.453**	$\mathbf{0.795}_{\pm 0.026}$	$\mathbf{0.567}_{\pm 0.009}$
IGPO	0.474	0.261	0.477	0.501	0.502	0.435	—	—
IGPO+FOLIA	**0.480**	**0.280**	**0.492**	**0.517**	**0.539**	**0.457**	—	—
HCAPO	0.452	0.240	0.473	0.496	0.472	0.420	$0.782_{\pm 0.017}$	$0.589_{\pm 0.020}$
HCAPO+FOLIA	**0.465**	**0.266**	**0.482**	**0.508**	**0.489**	**0.437**	$\mathbf{0.835}_{\pm 0.017}$	$\mathbf{0.606}_{\pm 0.018}$
GiGPO	0.479	0.272	0.481	0.504	0.511	0.442	$0.831_{\pm 0.011}$	$0.573_{\pm 0.033}$
GiGPO+FOLIA	**0.484**	**0.273**	**0.496**	**0.516**	**0.533**	**0.456**	$\mathbf{0.852}_{\pm 0.008}$	$\mathbf{0.648}_{\pm 0.039}$
Qwen2.5-7B-Instruct								
Prompting	0.198	0.095	0.264	0.244	0.310	0.228	0.059	0.016
GRPO	0.517	0.331	0.573	0.564	0.574	0.511	$0.831_{\pm 0.010}$	$0.670_{\pm 0.024}$
GRPO+FOLIA	**0.520**	**0.340**	**0.587**	**0.570**	**0.597**	**0.524**	$\mathbf{0.857}_{\pm 0.007}$	$\mathbf{0.697}_{\pm 0.004}$
IGPO	0.511	0.328	0.573	0.561	0.601	0.516	—	—
IGPO+FOLIA	**0.517**	**0.335**	**0.597**	**0.572**	**0.608**	**0.528**	—	—
GiGPO	0.516	0.328	0.574	0.576	0.595	0.518	$0.865_{\pm 0.016}$	$0.701_{\pm 0.019}$
GiGPO+FOLIA	0.509	**0.339**	**0.579**	**0.588**	**0.599**	**0.527**	$\mathbf{0.878}_{\pm 0.004}$	$\mathbf{0.721}_{\pm 0.017}$

FOLIA consistently improves performance across the long-horizon benchmarks. On the 3B model, FOLIA raises average F1 across multi-hop benchmarks from 0.429 to 0.453 for GRPO, from 0.435 to 0.457 for IGPO, and from 0.442 to 0.456 for GiGPO. Although designed for long-horizon tasks, it also improves performance on the shorter-horizon PopQA benchmark. The gains extend to WebShop, where human-expert trajectories average about 10 interaction turns: FOLIA raises GRPO Success Rate from 0.510 to 0.567 and improves the stronger GiGPO baseline from 0.573 to 0.648. Improvements are also observed on the larger 7B model, further illustrating FOLIA's applicability to larger models. Together, the gains across tasks, algorithms, and model scales support FOLIA as a broadly applicable plug-in method, particularly for long-horizon settings.

Beyond performance, our analyses in the appendix show that FOLIA's rollout-free design introduces only negligible computational overhead (App. A.2) and enables base methods to break through their late-training performance plateaus and continue improving (App. A.3).

5.3 RQ2: FOLIA Enables Fine-Grained Credit Assignment

We further examine whether FOLIA improves turn-level credit assignment during training. Following Sec. 3.2, we evaluate 200 examples using fixed intermediate checkpoints obtained during training and derive turn-level reference credits through large-scale MC sampling, providing a grounded and accurate reference. We then measure how much the original and redistributed credits deviate from this reference, focusing on the first two turns of the full trajectory.

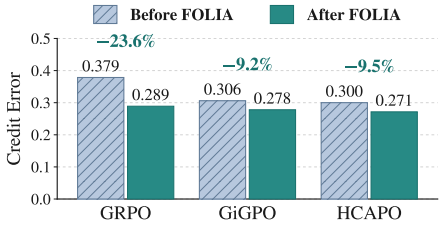


Figure 5: Average early-turn credit-estimation error across methods. FOLIA consistently improves estimation accuracy across all methods.

As shown in Fig. 5, the redistributed advantages show lower deviation from the high-budget MC reference at early turns across different methods. For GRPO, it reduces early-turn credit error by over 20%, indicating that the redistribution significantly corrects noise in the original estimates. For HCAPO and GiGPO, FOLIA still reduces error by nearly 10%, showing that its benefits generalize across base estimators. These improvements occur in the early turns identified as least reliable in our preceding analysis, while remaining stable in later-turn estimates. These observations establish FOLIA as a practical solution to early-turn credit estimation and show that, beyond improving final task performance, it provides more accurate and fine-grained training signals. We further provide a case study in App. A.4 to show how it functions.

5.4 RQ3: Ablation Study of Credit Signals in FOLIA

Finally, we isolate the effect of each FOLIA component through an ablation study on the Qwen2.5-3B-Ins model. Using GRPO and GiGPO as base algorithms with all other settings fixed, we evaluate a hindsight-only variant that retains the outcome anchor but removes the local influence signal. As shown in Table 2, hindsight alone improves GRPO from 0.510 to 0.546 and remains close to the baseline under GiGPO. This indicates that anchoring the direction of credit to the final outcome can be beneficial in some settings, although hindsight alone cannot estimate the magnitude of each step's credit. Adding local influence yields Success Rates of 0.567 and 0.648, respectively. These comparisons show that local influence improves upon the hindsight-only variant for both base algorithms. The results support the complementary roles of the two signals in the proposed redistribution rationale: terminal hindsight anchors credit redistribution to the final outcome, while local counterfactual influence provides fine-grained estimates of turn-level credit magnitude. Their confidence-gated fusion is therefore critical to achieving the full performance gains.

Table 2: Ablation study of FOLIA on WebShop using Success Rate.

Variant	GRPO	GiGPO
Vanilla	0.510	0.573
hindsight-only	0.546	0.563
w/ FOLIA	**0.567**	**0.648**

6 Related Work

Reinforcement learning for LLM agents. RL has become central to training LLM-based agents for complex tasks, including search (Jin et al., 2025; Zheng et al., 2025a), coding (Da et al., 2025; Wei et al., 2025), and other agentic tasks (Feng et al., 2025b; Wang et al., 2025). Schulman et al. (2017) uses a learned critic under a clipped surrogate objective, whereas Shao et al. (2024) removes the critic and estimates advantages from the relative rewards of grouped samples, proving effective for eliciting reasoning (Guo et al., 2025). Recent variants improve stability and scalability: Yu et al. (2025) decouple clipping and introduce dynamic sampling, Liu et al. (2025a) remove length and difficulty biases from advantage estimation, and Zheng et al. (2025b) use sequence-level clipping. However, these methods largely assign uniform trajectory-level credit, overlooking differences in step-level contributions and the noise inherent in long-horizon tasks.

Credit Assignment for Agentic Training. Credit assignment is a key challenge in RL training. MC-based methods estimate an action's credit from sampled future returns (Sutton and Barto, 1998), temporal-difference learning from immediate rewards and bootstrapped values (Sutton, 1988), and hindsight credit assignment from how much more likely an action was given the realized outcome than a priori (Harutyunyan et al., 2019). Recent work extends these ideas into LLM-agent training (Tao et al., 2026; Wang et al., 2026b). Kazemnejad et al. (2025b) apply an MC method to refine step-level credit estimates. Feng et al. (2025a) recover step-level advantages by grouping actions taken at repeated environment states across a trajectory group; Wang et al. (2026a) define turn-level credit as the marginal increase in the policy's probability of producing the ground-truth answer, while Kong et al. (2026) derive credit from information gain. In contrast, Tan et al. (2026b) and Jiang et al. (2026) use the LLM itself as a post-hoc critic, leveraging the completed trajectory to refine step-level value estimates in hindsight. Although these methods offer various solutions to credit assignment, none explicitly identifies or addresses early-turn credit estimation failure. Our method is therefore complementary to these approaches by recovering the missing effects of early turns.

7 Conclusion

In this work, we first identify and illustrate *early-turn credit failure* as a bottleneck in long-horizon agentic RL. Theoretically and empirically, we show that MC estimates of early-turn credit are affected by sampling noise under practical rollout budgets and are thus unreliable, and that existing outcome-anchored alternatives also suffer from the weak connection between early actions and the final answer, which hinders accurate early-turn credit estimation. Motivated by these findings, we propose FOLIA, a rollout-free plug-in method that reattributes credit within existing trajectories. Specifically, FOLIA fuses each turn's local counterfactual influence with a terminal anchor through a confidence gate, using local influence to measure each turn's effect on the next response and hindsight to ground the redistribution direction. Across diverse long-horizon benchmarks, FOLIA consistently improves various RL baselines on models of different scales while introducing negligible computational and memory overhead. Further analyses show that FOLIA enables finer-grained turn-level credit assignment, allowing it to overcome the base method's training plateau.

AI USE STATEMENT

During the research process, we used LLMs as search tools to investigate related work and as coding assistants during the initial development of our ideas. The authors reviewed the resulting code, verified its correctness with a broad range of test cases, and carefully examined the training logs. During manuscript preparation, we used LLMs only to improve grammar, clarity, and readability. All key conceptual contributions, experiments, analyses, and interpretations were developed by the authors.

REPRODUCIBILITY STATEMENT

We provide the methodological and experimental details needed to reproduce our findings. Sec. 4 describes the setup for our analysis, while Sec. 5 details the models, datasets, evaluation protocols, and training configurations. App. A.5 provides further setup details. All models and datasets used in our experiments are publicly available, and we have released our code via the anonymous repository.

REFERENCES

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 *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=oKn9c6ytLx.

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

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh Jing Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, Yitao Liu, Yiheng Xu, Shuyan Zhou, Silvio Savarese, Caiming Xiong, Victor Zhong, and Tao Yu. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang, editors, *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/5d413e48f84dc61244b6be550f1cd8f5-Abstract-Datasets_and_Benchmarks_Track.html.

Boyuan Zheng, Boyu Gou, Jihyung Kil, Huan Sun, and Yu Su. Gpt-4v(ision) is a generalist web agent, if grounded. In Ruslan Salakhutdinov, Zico Kolter, Katherine A. Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pages 61349–61385. PMLR / OpenReview.net, 2024. URL https://proceedings.mlr.press/v235/zheng24e.html.

Boyu Gou, Ruohan Wang, Boyuan Zheng, Yanan Xie, Cheng Chang, Yiheng Shu, Huan Sun, and Yu Su. Navigating the digital world as humans do: Universal visual grounding for GUI agents. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=kxnoqaisCT.

Xiaoxi Li, Guanting Dong, Jiajie Jin, Yuyao Zhang, Yujia Zhou, Yutao Zhu, Peitian Zhang, and Zhicheng Dou. Search-o1: Agentic search-enhanced large reasoning models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 5420–5438. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.EMNLP-MAIN.276. URL https://doi.org/10.18653/v1/2025.emnlp-main.276.

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.

Yuxiang Wei, Olivier Duchenne, Jade Copet, Quentin Carbonneaux, Lingming Zhang, Daniel Fried, Gabriel Synnaeve, Rishabh Singh, and Sida I. Wang. SWE-RL: advancing LLM reasoning via reinforcement learning on open software evolution. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/7107d4d2e837bde2171c6b71b5bde954-Abstract-Conference.html.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, Hao Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, Tao Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nat.*, 645(8081):633–638, 2025. doi: 10.1038/S41586-025-09422-Z. URL https://doi.org/10.1038/s41586-025-09422-z.

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, 2025a. doi: 10.48550/ARXIV.2503.20783. URL https://doi.org/10.48550/arXiv.2503.20783.

Weimin Xiong, Yifan Song, Xiutian Zhao, Wenhao Wu, Xun Wang, Ke Wang, Cheng Li, Wei Peng, and Sujian Li. Watch every step! LLM agent learning via iterative step-level process refinement. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1556–1572, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.93. URL https://aclanthology.org/2024.emnlp-main.93/.

Cong Chen, Kaixiang Ji, Hao Zhong, Muzhi Zhu, Anzhou Li, Guo Gan, Ziyuan Huang, Cheng Zou, Jiajia Liu, Jingdong Chen, Hao Chen, and Chunhua Shen. Gui-shepherd: Reliable process reward and verification for long-sequence GUI tasks. *CoRR*, abs/2509.23738, 2025. doi: 10.48550/ARXIV.2509.23738. URL https://doi.org/10.48550/arXiv.2509.23738.

Sina Abbasi, Mohammad Reza Modarres, and Mohammad Taher Pilehvar. NormXLogit: The head-on-top never lies. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 34926–34947, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1769. URL https://aclanthology.org/2025.emnlp-main.1769/.

Weiting Tan, Xinghua Qu, Ming Tu, Meng Ge, Andy T. Liu, Philipp Koehn, and Lu Lu. Process-supervised reinforcement learning for interactive multimodal tool-use agents, 2026a. URL https://openreview.net/forum?id=UtEY8lF8q2.

Rui Yang, Qianhui Wu, Yuxi Chen, Hao Bai, Wenlin Yao, Hao Cheng, Baolin Peng, Huan Zhang, Tong Zhang, and Jianfeng Gao. Openwebrl: Demystifying online multi-turn reinforcement learning for visual web agents, 2026. URL https://arxiv.org/abs/2606.02031.

Xiaoqian Liu, Ke Wang, Yongbin Li, Yuchuan Wu, Wentao Ma, Aobo Kong, Fei Huang, Jianbin Jiao, and Junge Zhang. EPO: Explicit policy optimization for strategic reasoning in LLMs via reinforcement learning. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15371–15396, Vienna, Austria, July 2025b. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.747. URL https://aclanthology.org/2025.acl-long.747/.

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*, 2026. URL https://openreview.net/forum?id=ooROvpmxMV.

Yuchen Deng, Shichen Fan, Naibo Wang, Xinkui Zhao, and See-Kiong Ng. AgentPro: Enhancing LLM agents with automated process supervision. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 9981–10006, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.506. URL https://aclanthology.org/2025.emnlp-main.506/.

Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. Group-in-group policy optimization for LLM agent training. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025a. URL http://papers.nips.cc/paper_files/paper/2025/hash/420c9f777c0b4f78d515e53cf74d58b2-Abstract-Conference.html.

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*, 2025a. URL https://openreview.net/forum?id=Myx2kJFzAn.

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.

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, 2026b. doi: 10.48550/ARXIV.2603.08754. URL https://doi.org/10.48550/arXiv.2603.08754.

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 Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok,*

Thailand, August 11-16, 2024, pages 9426–9439. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-LONG.510. URL https://doi.org/10.18653/v1/2024.acl-long.510.

Richard S. Sutton and Andrew G. Barto. *Reinforcement learning - an introduction*. Adaptive computation and machine learning. MIT Press, 1998. ISBN 978-0-262-19398-6. URL http://www.incompleteideas.net/book/first/the-book.html.

Jonathan Baxter and Peter L. Bartlett. Infinite-horizon policy-gradient estimation. *J. Artif. Intell. Res.*, 15:319–350, 2001. doi: 10.1613/JAIR.806. URL https://doi.org/10.1613/jair.806.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Mach. Learn.*, 8:229–256, 1992. doi: 10.1007/BF00992696. URL https://doi.org/10.1007/BF00992696.

Philip S. Thomas, Scott Niekum, Georgios Theocharous, and George Dimitri Konidaris. Policy evaluation using the Ω-return. In Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 334–342, 2015. URL https://proceedings.neurips.cc/paper/2015/hash/0e65972dce68dad4d52d063967f0a705-Abstract.html.

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.

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.

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. In *The Fourteenth International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=qkWP6phrvZ.

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 Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 12467–12476, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/195f15384c2a79cedf293e4a847ce85c-Abstract.html.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2369–2380. Association for Computational Linguistics, 2018. doi: 10.18653/V1/D18-1259. URL https://doi.org/10.18653/v1/d18-1259.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing A multi-hop QA dataset for comprehensive evaluation of reasoning steps. In Donia Scott, Núria Bel, and Chengqing Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 6609–6625. International Committee on Computational Linguistics, 2020. doi: 10.18653/V1/2020.COLING-MAIN.580. URL https://doi.org/10.18653/v1/2020.coling-main.580.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022. doi: 10.1162/tacl_a_00475. URL https://aclanthology.org/2022.tacl-1.31/.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, volume EMNLP 2023 of *Findings of ACL*, pages 5687–5711. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.FINDINGS-EMNLP.378. URL https://doi.org/10.18653/v1/2023.findings-emnlp.378.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 9802–9822. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.ACL-LONG.546. URL https://doi.org/10.18653/v1/2023.acl-long.546.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. Webshop: Towards scalable real-world web interaction with grounded language agents. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/82ad13ec01f9fe44c01cb91814fd7b8c-Abstract-Conference.html.

Hansi Zeng, Liam Collins, Bhuvesh Kumar, Neil Shah, and Hamed Zamani. Cosearch: Joint training of reasoning and document ranking via reinforcement learning for agentic search. *CoRR*, abs/2604.17555, 2026. doi: 10.48550/ARXIV.2604.17555. URL https://doi.org/10.48550/arXiv.2604.17555.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training. *CoRR*, abs/2212.03533, 2022. doi: 10.48550/ARXIV.2212.03533. URL https://doi.org/10.48550/arXiv.2212.03533.

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.

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, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 414–431. Association for Computational Linguistics, 2025a. doi: 10.18653/V1/2025.EMNLP-MAIN.22. URL https://doi.org/10.18653/v1/2025.emnlp-main.22.

Jeff Da, Clinton Wang, Xiang Deng, Yuntao Ma, Nikhil Barhate, and Sean Hendryx. Agent-rlvr: Training software engineering agents via guidance and environment rewards. *CoRR*, abs/2506.11425, 2025. doi: 10.48550/ARXIV.2506.11425. URL https://doi.org/10.48550/arXiv.2506.11425.

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. *CoRR*, abs/2504.11536, 2025b. doi: 10.48550/ARXIV.2504.11536. URL https://doi.org/10.48550/arXiv.2504.11536.

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, 2025. doi: 10.48550/ARXIV.2504.20073. URL https://doi.org/10.48550/arXiv.2504.20073.

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

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 Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/a4277440d50f1f15d2cb4c14f7e0c0d2-Abstract-Conference.html.

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. *CoRR*, abs/2507.18071, 2025b. doi: 10.48550/ARXIV.2507.18071. URL https://doi.org/10.48550/arXiv.2507.18071.

Richard S. Sutton. Learning to predict by the methods of temporal differences. *Mach. Learn.*, 3:9–44, 1988. doi: 10.1007/BF00115009. URL https://doi.org/10.1007/BF00115009.

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. *CoRR*, abs/2607.13988, 2026. doi: 10.48550/ARXIV.2607.13988. URL https://doi.org/10.48550/arXiv.2607.13988.

Zhitong Wang, Songze Li, Hao Peng, Shuzheng Si, Yi Wang, Maosong Sun, and Juanzi Li. Envrl: Learn from environment dynamics in agentic reinforcement learning. *CoRR*, abs/2606.17680, 2026b. doi: 10.48550/ARXIV.2606.17680. URL https://doi.org/10.48550/arXiv.2606.17680.

Amirhossein Kazemnejad, Milad Aghajohari, Eva Portelance, Alessandro Sordoni, Siva Reddy, Aaron C. Courville, and Nicolas Le Roux. Vineppo: Refining credit assignment in RL training of llms. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu, editors, *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025b. URL https://proceedings.mlr.press/v267/kazemnejad25a.html.

Fanqi Kong, Jiayi Zhang, Mingyi Deng, Chenglin Wu, Yuyu Luo, and Bang Liu. Infopo: Information-driven policy optimization for user-centric agents. *CoRR*, abs/2603.00656, 2026. doi: 10.48550/ARXIV.2603.00656. URL https://doi.org/10.48550/arXiv.2603.00656.

GLM. GLM-5: from vibe coding to agentic engineering. *CoRR*, abs/2602.15763, 2026. doi: 10.48550/ARXIV.2602.15763. URL https://doi.org/10.48550/arXiv.2602.15763.

Zhenyu Hou, Yujiang Li, Jie Tang, and Yuxiao Dong. Single-rollout asynchronous optimization for agentic reinforcement learning. *CoRR*, abs/2607.07508, 2026. doi: 10.48550/ARXIV.2607.07508. URL https://doi.org/10.48550/arXiv.2607.07508.

A Appendix

A.1 Proof of Proposition 3.2

Proof. For any context c with $V^\pi(c) > 0$ and $h(c) \geq h_0$, every successful continuation must pass all remaining bottlenecks. Let $B_j(c)$ denote the event of passing bottleneck j. By the probability chain rule and Assumption 3.1,

$$V^\pi(c) \leq \mathbb{P}_\pi\left(\bigcap_{j=1}^{h(c)} B_j(c) \,\middle|\, c\right)$$
$$= \prod_{j=1}^{h(c)} q_j(c) = \exp\left(-\sum_{j=1}^{h(c)} -\log q_j(c)\right) \leq e^{-\lambda h(c)}.$$

The same bound holds trivially when $V^\pi(c) = 0$.

Let $p_j := V^\pi(c_j)$ and $\rho_t := p_t + p_{t-1}$. Since $\min\{h(c_{t-1}), h(c_t)\} \geq h$, the preceding result gives $p_t, p_{t-1} \leq e^{-\lambda h}$ and hence $\rho_t \leq 2e^{-\lambda h}$.

Because the two groups of K continuation rollouts are independent,

$$\mathrm{Var}_{\mathrm{aux}}(\hat{\Delta}_t) = \frac{p_t(1-p_t) + p_{t-1}(1-p_{t-1})}{K}.$$

If $\rho_t = 0$, both values and the estimated credit are identically zero, so the conclusion holds under the convention $\mathrm{SNR}_t = 0$. Assume henceforth that $\rho_t > 0$.

When $h \geq (\log 2)/\lambda$, both probabilities are at most $1/2$. Since $p/2 \leq p(1-p) \leq p$ for $p \in [0, 1/2]$, we have $\rho_t/(2K) \leq \mathrm{Var}_{\mathrm{aux}}(\hat{\Delta}_t) \leq \rho_t/K$. Moreover, $|\Delta_t| = |p_t - p_{t-1}| \leq \rho_t$. Therefore,

$$\mathrm{SNR}_t \leq \frac{\rho_t}{\sqrt{\rho_t/(2K)}} = \sqrt{2K\rho_t} \leq 2\sqrt{K}\, e^{-\lambda h/2}.$$

Thus, $\mathrm{SNR}_t \geq \eta > 0$ necessarily requires $K \geq \frac{\eta^2}{4} e^{\lambda h}$.

Consequently, with K, λ, h_0 fixed, $\mathrm{SNR}_t \to 0$ as $h \to \infty$ across problems with increasing numbers of unresolved bottlenecks. Moreover, a union bound gives

$$\mathbb{P}_{\mathrm{aux}}\left(\hat{p}_t = \hat{p}_{t-1} = 0\right) \geq 1 - K\rho_t \geq 1 - 2Ke^{-\lambda h},$$

showing that both rollout groups contain only zero rewards with high probability for fixed K and sufficiently large h. □

A.2 Efficiency Study

We further show that FOLIA can be directly integrated into base RL algorithms with modest additional overhead. We analyze its computational and memory costs and validate them using practical training measurements. Both our analysis and training logs demonstrate that FOLIA introduces only limited additional cost.

Additional computation and memory. FOLIA is applied after the base algorithm computes its advantages and therefore leaves rollout generation, environment interaction, and reward computation unchanged. Thus, its only additional computation is extracting the two credit signals from trajectories already collected for training. Computing terminal hindsight requires one teacher-forced forward pass over the turn prefixes. For local influence, the factual log-probabilities are reused from the base algorithm, so only the counterfactual contexts require an additional forward pass. The full method therefore only adds two forward-only passes. The subsequent fusion and redistribution involve only scalar operations whose cost grows linearly with the number of turns. Moreover, these forward passes are processed in micro-batches and remain below the memory peak of the actor update, while the CPU stores only a small set of per-turn scalars.

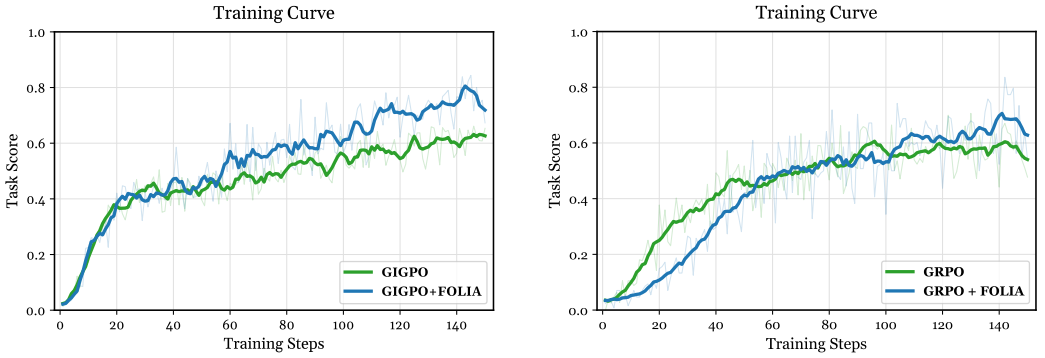
Empirical measurements. To evaluate efficiency in practical training, we measure FOLIA's computational overhead on WebShop using Qwen2.5-3B-Instruct with GRPO and GiGPO on a cluster of eight A100-80GB GPUs. Each batch contains 16 prompts with 8 trajectories per prompt. We report the average per-step runtime after discarding the first training step to exclude warm-up and CUDA-graph capture. The overhead is computed relative to the same step with this additive block removed, while the end-to-end step times are included for reference.

Table 3: Efficiency of FOLIA on WebShop with Qwen2.5-3B-Instruct. Time denotes the average wall-clock time per training iteration, while GPU memory denotes the peak allocated memory during training. FOLIA introduces negligible time and memory overhead.

	Time per training iteration (s)		GPU Memory	
Method	Base → FOLIA	Change	Base → FOLIA	Change
GRPO	216.5 → 224.1 s	+3.5%	64.1 → 64.3 GB	+0.3%
GiGPO	206.8 → 217.6 s	+5.2%	62.6 → 63.2 GB	+1.0%

As shown in Table 3, FOLIA increases the cost of each training step by only 3.5–5.2%. Meanwhile, peak allocated GPU memory increases by at most 0.6 GB. These results confirm that FOLIA can be integrated into existing RL algorithms with low computational overhead and negligible additional memory usage.

A.3 Training Efficacy



(a) Qwen2.5-3B-Instruct with GiGPO. (b) Qwen2.5-7B-Instruct with GRPO.

Figure 6: Task Score during WebShop training for (a) Qwen2.5-3B-Instruct with GiGPO and (b) Qwen2.5-7B-Instruct with GRPO.

We further examine the efficacy of FOLIA through its training dynamics. As shown in Fig. 6, the base methods like GiGPO and GRPO gradually plateau around step 100, after which their task scores show only little further improvement. In contrast, FOLIA continues to improve beyond this late-training plateau. By assigning more fine-grained credit to individual turns, FOLIA provides useful training signals even after the base methods have largely converged, allowing the policy to reach a higher final task score.

A.4 Case Study: FOLIA improves early-turn credit estimation

Extending the analysis in Sec. 5.3, we present an example from 2Wiki (Ho et al., 2020) to illustrate how FOLIA improves early-turn credit estimation in real-world data. Following the setup described before, the MC reference is built by sampling 256 continuations from each prefix and normalizing each turn's signed credit by the sum of credits across all turns.

The example in Fig. 7 illustrates how early-turn credit can be overestimated, leading to unreliable credit assignment. The question asks the agent to compare the death dates of `Giovanni dei Conti di Segni` and `Sakari Simelius`. In Turn 1, the agent instead retrieves general information about

Query: *Who died later, Giovanni Dei Conti Di Segni or Sakari Simelius?*
Gold: Sakari Simelius

Turn 1
Action: `search("Giovanni Dei Conti Di Segni")`
Obs.: [1] *Conti di Segni*: "...an important noble family of medieval and early modern Italy..."
[2] "**Giovanni dei Conti di Segni**, ...Giovanni Conti (d. 1493) and Francesco Conti (d. 1521)...."

Turn 2
Action: `search("Sakari Simelius death date")`
Obs.: [1] *Teuku Umar*: "...Umar was killed in an ambush on February 10, 1899 in Meulaboh."

Turn 3
Reason: "Search for the death date of Giovanni Dei Conti Di Segni to determine who died later..."
Action: `search("Giovanni Dei Conti Di Segni death date")`

Turn 4: Answer: Sakari Simelius ✓

Figure 7: Example illustrating how FOLIA improves early-turn credit estimation.

the `Conti di Segni` family and the death dates of other individuals. Although the passages mention `Giovanni`, they do not establish his date of death. After searching for `Sakari Simelius` in Turn 2, the agent revisits its original query about `Giovanni` in Turn 3 with a more specific target. In this case, the first retrieval provides relevant background but does not establish the requested date of death. Although the trajectory ends with the correct answer, the initial query's ambiguity limits this turn's contribution: the agent must search for Giovanni's death date again in Turn 3. Consistent with this observation, the MC reference assigns Turn 1 a normalized credit of only 0.032. However, GRPO and GiGPO assign credits of 0.250 and 0.299, respectively, substantially overestimating this retrieval's utility. Such credit could reinforce the first underspecified query despite its failure to obtain the necessary fact. FOLIA adjusts this allocation, reducing Turn 1's credit to 0.169 under GRPO and 0.213 under GiGPO and bringing both estimates closer to the MC reference. This case illustrates how FOLIA reduces the credit assigned to an early retrieval that provides background information but does not establish the fact needed to answer the question.

A.5 Training Setup

In Table 4, we summarize the training configurations for GRPO, GiGPO, and IGPO on `Qwen2.5-3B-Instruct`. We use the same shared hyperparameters across methods rather than tuning each separately, allowing us to compare their credit assignment mechanisms under a consistent setup. For experiments with FOLIA, we redistribute each base method's advantage estimates without changing its other training settings. The table also lists method-specific settings where applicable. Further implementation details are available in the codebase linked in the footnote on the first page.

Table 4: Training configurations for GRPO, GiGPO, and IGPO augmented with FOLIA on Search-based QA using `Qwen2.5-3B-Instruct`.

(a) GRPO

Configuration	Value
Learning rate	10^{-6}
Samples per prompt	8
PPO mini-batch	128
KL coefficient	0.001
Epochs	1
FOLIA Clip Rate	0.5
Reliability a	1.0

(b) GiGPO

Configuration	Value
Learning rate	10^{-6}
Samples per prompt	8
PPO mini-batch	128
KL coefficient	0.001
Epochs	1
GiGPO gamma	1.0
Step Advantage	0.5
FOLIA Clip Rate	0.5
Reliability a	1.0

(c) IGPO

Configuration	Value
Learning rate	10^{-6}
Samples per prompt	8
PPO mini-batch	128
KL coefficient	0.001
Epochs	1
IGPO gamma	1.0
Info gain type	`prob_diff`
FOLIA Clip Rate	0.5
Reliability a	1.0

A.6 Limitation and Broader Impact

In this work, we characterize and empirically examine the challenge of early-turn credit estimation in training long-horizon agents. We then introduce FOLIA, a rollout-free method that improves credit estimation by combining intermediate local influence with a terminal hindsight signal. As a plug-in framework, FOLIA can be integrated into different RL algorithms and potentially extended beyond search and shopping to broader agentic tasks, such as coding.

Despite its broad applicability, FOLIA has limitations. First, it redistributes the original trajectory-level learning signal rather than creating supervision when that signal is absent. Its effectiveness may therefore be limited in all-negative or all-positive settings, where terminal outcomes provide little discriminative credit. Second, FOLIA is designed to mitigate unreliable early-turn credit in critic-free methods. Integrating it with value-function-based advantage estimates, such as those used in PPO (Schulman et al., 2017) and its variants (GLM, 2026; Hou et al., 2026), remains underexplored.

A.7 Prompt Template

We follow the prompt templates used in prior work (Zeng et al., 2026; Wang et al., 2026b;a) and provide the templates for each task below:

WebShop Prompt Template Without History

```
You are an expert autonomous agent operating in the WebShop
e-commerce environment.
Your task is to: {task_description}.
Your current observation is: {current_observation}.
Your admissible actions of the current situation are:
[
{available_actions}
].
{strategy_hint}

Now it's your turn to take one action for the current step.
You should first reason step-by-step about the current situation,
then think carefully which admissible action best advances the
shopping goal. This reasoning process MUST be enclosed within
<think> </think> tags.

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

Figure 8: WebShop prompt template used at the initial step or when `history_length` $= 0$.

WebShop Prompt Template With History

```
You are an expert autonomous agent operating in the WebShop
e-commerce environment.
Your task is to: {task_description}.

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

You are now at step {current_step} and your current observation
is: {current_observation}.
Your admissible actions of the current situation are:
[
{available_actions}
].
{strategy_hint}

Now it's your turn to take one action for the current step.
You should first reason step-by-step about the current situation,
then think carefully which admissible action best advances the
shopping goal. This reasoning process MUST be enclosed within
<think> </think> tags.

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

Figure 9: WebShop prompt template used at later steps when `history_length` > 0.

Prompt Template for Wiki-Based Factoid Question Answering

```
You are a tool-augmented research agent for wiki-based factoid question answering. Your task is to answer
questions drawn from Wikipedia-style datasets. The final answer is evaluated using exact match (EM) or
token-level F1, so it must be short and precise.

You have ONE tool available:
- search(query: string) -> returns a list of Wikipedia passages
============================================================
CRITICAL OUTPUT FORMAT (MUST FOLLOW EXACTLY)
============================================================
For EVERY assistant turn, you MUST output EXACTLY TWO TAG BLOCKS in this order:
1) <reason> ... </reason>
2) EITHER:
   (A) <tool_call> ... </tool_call>
   OR
   (B) <answer> ... </answer>
No other text is allowed outside these tags. Do NOT output <tool_response>. The environment will provide
tool results separately.
Allowed patterns:
- <reason> ... </reason>
  <tool_call> ... </tool_call>
- <reason> ... </reason>
  <answer> ... </answer>
If you violate the format, your output is invalid.
============================================================
TOOL CALL JSON SCHEMA (STRICT)
============================================================
When calling the tool, the <tool_call> block MUST contain ONLY a valid JSON object:
<tool_call>
{
  "name": "search",
  "arguments": {
    "query": "<string>"
  }
}
</tool_call>
Rules:
- "name" MUST be exactly "search"
- "arguments" MUST be an object
- "query" MUST be a single string
- Do NOT add extra keys
- Do NOT wrap JSON in Markdown
- Do NOT include comments, trailing commas, or natural language
============================================================
GENERAL TOOL USAGE
============================================================
Use the search tool whenever additional evidence would help you determine the correct answer. If you
believe you already have sufficient information to answer correctly, answer directly.
You may use multiple search calls across turns.
============================================================
SEARCH GUIDELINES
============================================================
- Write search queries that are clear and specific to what you want to confirm or find.
- After receiving evidence, reassess whether you can answer; if not, search again with a refined query.
============================================================
REASONING CONTENT REQUIREMENTS
============================================================
- Do NOT include tool JSON inside <reason>.
- Do NOT include <tool_call> or <answer> tags inside <reason>.
============================================================
ANSWER REQUIREMENTS (STRICT: SHORT ANSWER)
============================================================
Inside <answer>, you MUST:
- Output ONLY the final answer string
- Do NOT include explanations, reasoning, or extra text
- Do NOT include citations, sources, or formatting
- Use a concise canonical form (Wikipedia-style when possible)
Examples of valid answers:
- Paris
- 1997
- George Washington
- The Lord of the Rings
If the expected answer type is a person, place, organization, title, or date, output only that span. If
multiple surface forms are possible, output the most standard form.
============================================================
INTEGRITY
============================================================
- Do not fabricate facts.
- If you are uncertain, use search to verify.
- If evidence is conflicting, search again with a query that
  resolves the conflict.
============================================================
BEGIN
============================================================
Question: What is the date of death of Jana Cilliers's husband?
```

Figure 10: Prompt template used for tool-augmented Wikipedia-style factoid question answering.