PACT: FROM CREDIT ASSIGNMENT TO CRITIC ALIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning has become a central component of large language model (LLM) post-training, yet token-level credit lacks a generally accepted mathematical definition, leaving its relationship to commonly used training signals unclear. We formulate three regularity conditions, namely Completeness, Prefix Consistency, and Neutrality, and prove that they uniquely determine token-level credit. This characterization provides a unified basis for explaining phenomena across existing algorithms and guides the development of an improved actor-critic training procedure. Through this lens, an ideal teacher in On-Policy Distillation (OPD) acts as an implicit critic, yielding an expected policy gradient proportional to that induced by token-level credit. Response-level REINFORCE Leave-One-Out (RLOO) signals match the expected policy-gradient contribution of token-level credit despite their coarser granularity. We further establish approximate credit sparsity under bounded outcome rewards and show how intermediate critic errors in Generalized Advantage Estimation (GAE) can become comparable to the underlying credit. These motivate **P**olicy **A**ligned **C**ritic **T**raining (PACT), which adopts an Actor-then-Critic update order to apply importance sampling correction to critic training and better align the critic with the updated policy. In agentic mathematical reasoning, PACT achieves 72.87% average accuracy across four benchmarks, outperforming GRPO and PPO by 8.80 and 13.16 percentage points, respectively. On SWE-bench Verified, PACT achieves a pass rate of 67.4%, outperforming PPO, GRPO, and SAO by 2.4, 2.0, and 3.8 percentage points, respectively.

1 INTRODUCTION

"The real justification of these definitions, however, will reside in their implications."

— Claude E. Shannon, 1948

Recent advances have demonstrated that reinforcement learning can substantially improve the capabilities of large language models across general tasks (Guo et al., 2025; Ma et al., 2026; Cheng et al., 2026), general agents (Wang et al., 2025), and coding agents (Da et al., 2025; Wei et al., 2025). In these settings, a language model may generate a long sequence of tokens and repeatedly interact with an external environment. However, the training signal is often a scalar outcome reward revealed only after the completion of the trajectory, while optimization updates are applied at the level of individual generated tokens. This difference in granularity gives rise to a fundamental credit assignment problem of determining how the final outcome should be attributed to the tokens in the trajectory. Furthermore, despite its central role in reinforcement learning, credit itself lacks a generally accepted mathematical characterization, and existing methods operationalize it through different algorithm-dependent quantities (Pignatelli et al., 2023).

A common formulation models autoregressive language model reinforcement learning as a token-level Markov Decision Process, where the complete generation history is treated as the state and the next generated token as the action (Ramamurthy et al., 2022). In interactive settings, taking the complete history of the prompt, generated tokens, and environment observations as the state yields a Markov representation. However, this representation alone does not characterize how the realized outcome should be attributed across the generated tokens. In particular, it does not describe

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.

how the statistical information relevant to the final reward changes as the trajectory unfolds. Credit assignment therefore requires an additional characterization beyond the Markov formulation.

In this work, we formulate three regularity conditions for credit assignment, namely Completeness, Prefix Consistency, and Neutrality, and prove that credit is uniquely determined under these conditions. The resulting representation subsumes related forms previously derived under specific objectives or algorithmic constructions (Arjona-Medina et al., 2019; Kazemnejad et al., 2024; Setlur et al., 2025). The same uniqueness result holds for credit assignment at coarser granularities, such as the turn level, with the corresponding representation obtained by aggregating token-level credit over consecutive segments.

Our contributions are twofold:

- **A unique representation theorem and its consequences.** We establish a unique representation theorem for credit under three regularity conditions, namely Completeness, Prefix Consistency, and Neutrality. This representation provides a unified perspective on several phenomena in LLM reinforcement learning.

  Under an ideal teacher, the On-Policy Distillation (OPD) (Lu & Lab, 2025) update is equivalent in expectation, up to a scaling factor, to the policy-gradient update induced by the unique credit, with the teacher acting as an implicit critic.

  For REINFORCE Leave-One-Out (RLOO) (Ahmadian et al., 2024), the response-level signal, although not itself token-level credit, induces the same expected policy-gradient contribution as the unique credit representation.

  We further establish approximate sparsity of token-level credit under bounded outcome rewards, providing a perspective on the sensitivity of fine-grained credit estimation to critic errors and the empirical difficulty of Generalized Advantage Estimation (GAE) based Actor-Critic methods in long-horizon settings (Schulman et al., 2015).
- **Policy Aligned Critic Training.** Motivated by these findings, we propose Policy Aligned Critic Training (PACT). PACT adopts an Actor-then-Critic update order to enable importance sampling correction in critic training, better aligning the critic with the updated policy. We also use Binary Cross Entropy (BCE) instead of Mean Squared Error (MSE) to train the critic to better approximate the true values.

  On four mathematical reasoning benchmarks, PACT achieves 72.87% average accuracy, outperforming GRPO and PPO by 8.80 and 13.16 percentage points, respectively. On SWE-bench Verified, PACT achieves a pass rate of 67.4%, outperforming PPO, GRPO, and SAO by 2.4, 2.0, and 3.8 percentage points, respectively.

All proofs are deferred to the appendices, which also contain additional discussion and a detailed review of related work.

2 Unique Representation of Token-Level Credit

Credit assignment can be defined at different granularities. For example, one may assign rewards to complete responses, interaction turns, or individual tokens. In autoregressive language model reinforcement learning, however, coarser-grained credit assignments can be viewed as special cases of token-level credit by aggregating consecutive tokens into larger units. Therefore, we shall firstly focus on the finest-grained formulation and study token-level credit assignment.

2.1 Notation: Reward Information Flow

We first introduce the necessary notation to formalize the information flow of rewards during autoregressive generation. Our formulation covers both standard autoregressive generation and agentic interaction. Given a prompt q, a realized trajectory takes the form $Y = (q, T_1, O_1, T_2, O_2, \ldots, T_\tau, O_\tau)$, where T_i is the i-th token generated by the policy and O_i is the observation returned by the environment after T_i. An observation may be a tool response, an environment transition, user feedback, or any other information revealed to the model. We allow $O_i = \varnothing$; consequently, ordinary autoregressive language generation is recovered as the special case in which every observation is empty. The terminal reward is defined as a measurable function of

the complete trajectory, $R = \mathcal{R}(Y)$. In particular, R need not be assigned to any specific token or observation.

Let $\mathcal{F}_0 = \sigma(q), \mathcal{F}_i = \sigma(q, T_1, O_1, \ldots, T_i, O_i)$ denote the information available after the i-th token and its associated observation have been revealed. Here, $\sigma(\cdot)$ denotes the generated sigma-algebra. We define the conditional reward prediction at this point as $V_i = \mathbb{E}[R \mid \mathcal{F}_i]$.

2.2 Regularity Conditions of Credit Assignment

Let C_i denote the credit assigned to the i-th generated token T_i. Although environmental observations O_i are incorporated into the information filtration and affect future token generation, credit assignment aims to attribute the final outcome only to the model's generated tokens, rather than to environment-provided information. Here we also define the accumulated token credit up to step i as $S_i = \sum_{j=1}^{i} C_j, S_0 = 0$. We characterize token-level credit assignments through the following three regularity conditions.

Completeness. The assigned credits should fully explain the deviation of the final reward from its initial prediction:

$$\sum_{i=1}^{\tau} C_i = R - \mathbb{E}[R \mid \mathcal{F}_0].$$

Prefix Consistency. Credit assigned to a generated prefix should only reflect the information contained in that prefix. In particular, once a prefix has been generated, the credit accumulated by this prefix should be fixed from the perspective of the available information. Different realizations of future tokens or observations should be attributed to future decisions, rather than modifying the credit already assigned to previous tokens. Formally, for any two trajectories Y and $\widetilde{Y}$, if they share the same prefix information up to step i,

$$(q, T_1, O_1, \ldots, T_i, O_i) = (\widetilde{q}, \widetilde{T}_1, \widetilde{O}_1, \ldots, \widetilde{T}_i, \widetilde{O}_i),$$

then the accumulated credits at step i should satisfy $S_i(Y) = S_i(\widetilde{Y})$.

Neutrality. A token credit should neither systematically overestimate nor underestimate the contribution of the corresponding token from the perspective of the available information. The expected credit assigned to the next token should be zero:

$$\mathbb{E}[C_i \mid \mathcal{F}_{i-1}] = 0.$$

2.3 Unique Representation Theorem

The following theorem is the central result of our characterization and provides the foundation for the subsequent analysis. It establishes that a token-level credit assignment satisfying the three regularity conditions introduced above exists and is unique, and that the unique assignment is given by the differences of a martingale. We further establish in Appendix B.2.3 that all three conditions are necessary for uniqueness by showing that omitting any one admits alternative credit assignments.

Theorem 1 (Unique Representation of Token-Level Credit)**.** *Let L denote the maximum number of generated tokens in a trajectory, and let $\tau \leq L$ be the stopping time induced by the generation process. Assume that R is integrable and $\mathcal{F}_\tau$-measurable. Define*

$$V_i := \mathbb{E}[R \mid \mathcal{F}_i], \qquad i = 0, \ldots, \tau.$$

There exists a unique integrable token-level credit assignment, up to almost-sure equality, $\{C_i\}_{i=1}^{\tau}$ satisfying Completeness, Prefix Consistency, and Neutrality.

It admits the representation

$$C_i = V_i - V_{i-1} = \mathbb{E}[R \mid \mathcal{F}_i] - \mathbb{E}[R \mid \mathcal{F}_{i-1}] \quad \textit{a.s.}, \qquad i = 1, \ldots, \tau.$$

Moreover, $\{C_i\}_{i=1}^{\tau}$ is a martingale difference sequence with respect to $\{\mathcal{F}_i\}_{i=0}^{\tau}$.

Corollary 1 (Aggregation of the Unique Token-Level Representation)**.** *Consider any partition of the token sequence into consecutive segments $\mathcal{I}_1, \mathcal{I}_2, \ldots, \mathcal{I}_K$, where each segment $\mathcal{I}_k$ contains a set of consecutive token indices. Define the segment-level credit as $C^{(k)} = \sum_{i \in \mathcal{I}_k} C_i$. The resulting segment-level credit is*

$$C^{(k)} = \sum_{i \in \mathcal{I}_k} (V_i - V_{i-1}) = V_{b_k} - V_{a_k - 1},$$

where a_k and b_k denote the first and last token indices of segment $\mathcal{I}_k$.

In particular, existing coarse-grained credit assignment schemes, such as turn-level attribution, can be recovered as special cases by aggregating the unique token-level representation over appropriate token segments.

3 Interpreting Existing RL Algorithms through the Unique Credit Representation

In this section, we use the unique credit representation to revisit several phenomena in LLM reinforcement learning. We first examine the relationship between OPD and critics, then study credit granularity in RLOO and credit estimation in GAE. All these analyses together motivate Policy Aligned Critic Training (PACT), introduced in the next section.

3.1 The Teacher in On-Policy Distillation is an Implicit Critic

On-Policy Distillation trains on student-generated trajectories using dense token-level supervision from a teacher (Lu & Lab, 2025). Prior work further showed that the OPD objective can be algebraically reformulated as dense KL-constrained RL, in which the teacher–student log-density ratio plays the role of a token-level advantage (Yang et al., 2026). This optimization-level equivalence, however, does not by itself guarantee that the induced signal is consistent with the outcome reward R, or that it constitutes a correct token-level credit assignment. We establish this missing credit-level connection by identifying an ideal teacher under which OPD induces, up to scaling, the same policy-gradient update as the unique token-level credit in Theorem 1.

Moreover, recent work suggests that effective on-policy distillation requires a teacher that improves upon the student while remaining compatible with its on-policy distribution (Li et al., 2026). We capture these requirements with an idealized teacher defined through KL-regularized reward improvement.

Fix a policy $\pi = \pi_\theta$ and a token position $t \le \tau$. Let

$$p_t(a) = \pi_\theta(a \mid \mathcal{F}_{t-1})$$

denote the current policy distribution over the vocabulary $\mathcal{V}$. Define the token-action value

$$Q_t^\pi(a) = \mathbb{E}_\pi \left[R | \mathcal{F}_{t-1}, T_t = a \right].$$

The expectation includes the subsequent environment observation and all future tokens and observations generated under π.

Definition 1 (Ideal Teacher)**.** *Let $\beta > 0$. At each prefix $\mathcal{F}_{t-1}$, the ideal teacher distribution is defined by*

$$q_t^\star \in \arg\max_{q \in \Delta(\mathcal{V})} \left\{ \mathbb{E}_{a \sim q} \left[Q_t^\pi(a) \right] - \beta D_{\mathrm{KL}}(q \| p_t) \right\}.$$

We assume that R is bounded and that p_t has full support over $\mathcal{V}$. The distribution $q_t^\star$ is defined relative to the current policy and is held fixed during the corresponding OPD update.

For a teacher distribution q_t, sampled-token OPD assigns the signal

$$A_t^{\mathrm{OPD}}(q_t) = \log \frac{q_t(T_t)}{p_t(T_t)}, \qquad T_t \sim p_t.$$

Let

$$Z_t = \nabla_\theta \log \pi_\theta(T_t \mid \mathcal{F}_{t-1})$$

denote the policy score. The on-policy update direction induced by $A_t^{\mathrm{OPD}}(q_t)$ is

$$G_t^{\mathrm{OPD}}(q_t) := \mathbb{E}_\pi \left[Z_t A_t^{\mathrm{OPD}}(q_t) \middle| \mathcal{F}_{t-1} \right] = \mathbb{E}_\pi \left[Z_t \log \frac{q_t(T_t)}{p_t(T_t)} \middle| \mathcal{F}_{t-1} \right].$$

The following theorem identifies the OPD teacher as an implicit critic under the ideal-teacher assumption. Unlike a conventional critic represented by a value head, the OPD teacher provides credit through its KL-based objective. From this perspective, the empirical success of OPD further highlights the role of critics, even when no explicit value head is used.

Theorem 2 (Teacher Credit Equivalence)**.** *Let $q_t^\star$ be the ideal teacher in Definition 1. Then*

$$G_t^{\mathrm{OPD}}(q_t^\star) = \frac{1}{\beta} \mathbb{E}_\pi \left[Z_t C_t^\pi | \mathcal{F}_{t-1} \right] \qquad a.s.,$$

where

$$C_t^\pi = V_t^\pi - V_{t-1}^\pi, \qquad V_t^\pi = \mathbb{E}_\pi[R \mid \mathcal{F}_t]$$

is the unique token-level credit from Theorem 1.

3.2 Group-Level Baselines as Gradient-Equivalent Credit Surrogates

The unique representation theorem shows that the unique token-level credit depends on the evolution of conditional reward predictions: $C_i = V_i - V_{i-1}$. However, recent group-based RL algorithms for language models, such as GRPO (Shao et al., 2024) and RLOO (Ahmadian et al., 2024), construct advantages using multiple sampled responses from the same prompt. The resulting baseline is defined at the response level, since it aggregates rewards from different sampled trajectories. This response-level baseline is then broadcast to every token position within the same response during policy optimization. This introduces a granularity gap between response-level baseline estimation and token-level credit assignment.

We now examine group-level baseline methods, taking RLOO as a representative example. We show that, although the RLOO baseline is constructed at the response level, the resulting token-wise policy gradient contribution coincides in expectation with that of the unique token-level credit.

Theorem 3 (Gradient Unbiasedness of the RLOO Estimator)**.** *Conditioned on a prompt q, let $G \geq 2$ trajectories be sampled independently from the current policy π_θ. Let R_j denote the reward of trajectory j. For trajectory i, define the leave-one-out baseline*

$$\bar{R}_{-i} = \frac{1}{G-1} \sum_{j \neq i} R_j.$$

Then its token-level policy gradient estimator is equivalent in expectation to the gradient induced by the unique token-level credit representation:

$$\mathbb{E}\left[\nabla_\theta \log \pi_\theta(T_t | \mathcal{F}_{t-1}) (R_i - \bar{R}_{-i}) \right] = \mathbb{E}\left[\nabla_\theta \log \pi_\theta(T_t | \mathcal{F}_{t-1}) C_t \right].$$

The preceding equivalence holds only in expectation and does not imply equal statistical efficiency. The conditional expectation $V_t = \mathbb{E}[R \mid \mathcal{F}_t]$ minimizes the mean squared prediction error among $\mathcal{F}_t$-measurable predictors. In contrast, the response-level RLOO baseline retains randomness from finite outcomes. In long-horizon settings, these fluctuations can be substantial relative to the local credit signal. This motivates estimating credit at a finer granularity.

3.3 Approximate Credit Sparsity and Error Sensitivity of GAE

Recent empirical observations suggest that GAE with λ close to one can be beneficial in LLM reinforcement learning. DeepSeek-R1 (Guo et al., 2025) reports that PPO (Schulman et al., 2017) with $\lambda = 1$ outperforms the commonly used setting $\lambda = 0.95$. SAO (Hou et al., 2026) adopts a length-adaptive GAE coefficient that approaches one as the response length increases. To understand these observations, we first establish an approximate sparsity property of credit and then examine how critic estimation errors enter GAE.

For any bounded reward, we can apply a positive affine transformation to normalize it into $[0, 1]$ without changing the optimal policy. Therefore, we consider $R \in [0, 1]$ without loss of generality. In this case, the reward variance satisfies $\mathrm{Var}(R) \leq \frac{1}{4}$.

Theorem 4 (Approximate Credit Sparsity)**.** *For any fixed prompt q,*

$$\mathbb{E}\left[\sum_{i=1}^{\tau} C_i^2 \mid \mathcal{F}_0\right] = \mathrm{Var}(R \mid \mathcal{F}_0) \le \frac{1}{4}.$$

Consequently, for any $\epsilon > 0$,

$$\mathbb{E}\left[\sum_{i=1}^{\tau} \mathbf{1}\{|C_i| > \epsilon\} \middle| \mathcal{F}_0\right] \le \frac{1}{4\epsilon^2}.$$

We now examine how this sparsity property relates to GAE in the outcome-only reward setting, where the reward is revealed only after the generation trajectory is complete. Let the critic estimate be $\widehat{V}_i = V_i + \varepsilon_i$, where ε_i denotes the value estimation error. With $\gamma = 1$ and the terminal prediction set to $\widehat{V}_\tau = R$, the TD residual becomes $\widehat{\delta}_i = \widehat{V}_i - \widehat{V}_{i-1} = C_i + \varepsilon_i - \varepsilon_{i-1}$. For GAE with parameter λ, the estimated advantage is computed as $\widehat{A}_t^\lambda = \sum_{i=t}^{\tau} \lambda^{i-t}\widehat{\delta}_i$. Substituting the above TD decomposition gives $\widehat{A}_t^\lambda = \sum_{i=t}^{\tau} \lambda^{i-t} C_i - \varepsilon_{t-1} + (1-\lambda)\sum_{i=t}^{\tau-1} \lambda^{i-t}\varepsilon_i$, where the terminal value error is assumed to be zero.

This decomposition separates the true credit signal from critic estimation errors. The first term corresponds to the accumulated token-level credit, while the remaining terms arise from imperfect value estimation. When $\lambda < 1$, intermediate critic errors are retained through the last term. However, Theorem 4 bounds the expected number of credit increments exceeding any fixed magnitude, independently of the maximum response length. Consequently, the critic error term can become comparable to, or even dominate, the true credit signal.

In contrast, when $\lambda = 1$, the intermediate critic error term vanishes: $\widehat{A}_t^1 = \sum_{i=t}^{\tau} C_i - \varepsilon_{t-1} = R - \widehat{V}_{t-1}$. Therefore, $\lambda = 1$ eliminates intermediate value estimation errors, leaving only the prefix value error $-\varepsilon_{t-1}$.

Following the analysis above, our proposed method, i.e., PACT, uses $\lambda = 1$.

4 Policy Aligned Critic Training

4.1 Motivation

Theorems 1 and 4 highlight the importance of accurate value estimation for fine-grained credit assignment. By Theorem 1, the unique token-level credit under policy π is

$$C_i^\pi = V_i^\pi - V_{i-1}^\pi, \qquad V_i^\pi = \mathbb{E}_\pi[R \mid \mathcal{F}_i].$$

Thus, recovering local credit requires not only accurate value predictions, but also values corresponding to the policy currently being optimized. Moreover, Theorem 4 shows that most local credits can be small in long-horizon generation. Consequently, even a modest value error or policy mismatch may dominate the underlying credit signal.

Under current RL training frameworks, PPO-style training can introduce a policy mismatch between the actor and the critic. Let π_k denote the policy used to generate the trajectories $\mathcal{D}_k$ in iteration k. The critic values used to compute advantages on $\mathcal{D}_k$ are produced before the critic is updated on $\mathcal{D}_k$. Hence, these values are generated by critic parameters learned from the previous trajectories $\mathcal{D}_{k-1} \sim \pi_{k-1}$:

$$\mathcal{D}_k \sim \pi_k, \qquad \widehat{V}_{\phi_{k-1}} \approx V^{\pi_{k-1}}.$$

Although the critic is subsequently updated on $\mathcal{D}_k$, the resulting critic is not used to recompute the advantages for the current actor update. After the actor is updated from π_k to π_{k+1}, the critic is therefore again one policy update behind.

This policy–critic lag is especially consequential when token-level credit is obtained by differencing consecutive conditional values. It motivates training procedures that improve value estimation while explicitly synchronizing the critic with the updated policy. We shall develop such a procedure in this section.

4.2 PROBABILISTIC VALUE ESTIMATION

As discussed above and proved in Appendix B.1, any bounded reward can be normalized to $[0, 1]$ without changing the optimal policy. We therefore assume $R \in [0, 1]$, which also implies $V_i^\pi = \mathbb{E}_\pi[R \mid \mathcal{F}_i] \in [0, 1]$.

Instead of the conventional mean squared error, we parameterize the critic as $\widehat{V}_{\phi,i} = \sigma(z_{\phi,i})$ and optimize the soft binary cross-entropy loss

$$\mathcal{L}_{\mathrm{BCE}}(\phi) = \mathbb{E}\left[-R \log \widehat{V}_{\phi,i} - (1 - R) \log(1 - \widehat{V}_{\phi,i})\right].$$

For $R \in [0, 1]$, BCE and MSE have the same optimal prediction:

$$\arg\min_{v \in [0,1]} \mathbb{E}[-R \log v - (1 - R) \log(1 - v) \mid \mathcal{F}_i] = \arg\min_{v \in \mathbb{R}} \mathbb{E}[(v - R)^2 \mid \mathcal{F}_i] = V_i^\pi.$$

At the endpoints, BCE is defined by its limits, allowing the value $+\infty$, with $0 \log 0 = 0$. Thus, using BCE does not change the value being estimated. Classification-based objectives have shown favorable empirical performance for value estimation in deep reinforcement learning Farebrother et al. (2024). Furthermore, a controlled critic-pretraining ablation, reported in Section 5.3, shows that the probabilistic BCE critic converges substantially faster than the conventional MSE critic.

4.3 OFF-POLICY CRITIC SYNCHRONIZATION

Policy Aligned Critic Training (PACT) introduces an *Actor-then-Critic* dependency through importance sampling corrected critic targets, whereas standard PPO allows independent actor and critic updates. Within each iteration, the current critic first provides the value predictions required for actor optimization. After all actor updates have been completed, we perform one additional forward pass over the same rollout batch using the updated actor. Together with the log-probabilities recorded before the update, this produces importance ratios between the pre-update and post-update policies. We then train the critic using these ratios, so that the iteration ends with the critic better aligned to the updated actor.

Let $\mu = \pi_k$ be the policy before the actor update and $\pi = \pi_{k+1}$ the updated policy. For a prefix $\mathcal{F}_{t-1}$, define the continuation importance ratio

$$I_t = \prod_{k=t}^{\tau} \frac{\pi(T_k \mid \mathcal{F}_{k-1})}{\mu(T_k \mid \mathcal{F}_{k-1})}.$$

Assuming unchanged environment dynamics and absolute continuity of the continuation distributions, a change of measure gives $V_{t-1}^\pi = \mathbb{E}_\pi[R \mid \mathcal{F}_{t-1}] = \mathbb{E}_\mu[I_t R \mid \mathcal{F}_{t-1}]$.

We now regard $I_t R$ as a single random variable. As established in the previous subsection, binary cross-entropy recovers the conditional mean: for any integrable random variable Y whose conditional mean lies in $[0, 1]$,

$$\mathbb{E}[Y \mid \mathcal{F}] = \arg\min_{x \in [0,1]} \mathbb{E}\left[\mathcal{L}_{\mathrm{BCE}}(x, Y) \mid \mathcal{F}\right].$$

Throughout this subsection, the conditional expected BCE is extended to $x = 0$ and $x = 1$ by taking one-sided limits after the conditional expectation. For a sigmoid-parameterized critic, optima at 0 and 1 are approached as the logit tends to $-\infty$ and $+\infty$, respectively.

Applying this result with $Y = I_t R$ immediately gives

$$V_{t-1}^\pi = \arg\min_{x \in [0,1]} \mathbb{E}_\mu\left[\mathcal{L}_{\mathrm{BCE}}(x, I_t R) \mid \mathcal{F}_{t-1}\right].$$

Therefore, trajectories sampled before the actor update can be used to train a critic for the updated actor by replacing the original reward target R with the importance-corrected target $I_t R$. Although an individual realization of $I_t R$ need not lie in $[0, 1]$, this does not alter the conditional minimizer above; a logit-space gradient analysis is provided in Appendix B.6. In practice, the exact continuation ratio can have high variance for long responses. We use the detached current-token importance ratio and mask out token-level critic losses whose ratios fall outside $[\rho_{\min}, \rho_{\max}]$. These ratios require only one additional forward pass after actor training and no additional rollout generation, yielding a stable and inexpensive surrogate for exact policy synchronization.

Table 1: Mathematical reasoning performance on Avg@16 accuracy (%).

Method	AIME 2025	AIME 2026	BeyondAIME	HMMT Nov. 2025	Average
Base Model	46.67	51.25	28.63	37.50	41.01
GRPO ($\epsilon_{\text{high}} = 0.28$)	76.50	74.78	41.81	63.19	64.07
PPO ($\lambda = 0.95$)	32.33	29.38	17.86	25.67	26.31
PPO ($\lambda = 1.0$)	66.04	73.96	40.31	58.54	59.71
SAO	51.25	63.33	36.63	53.33	51.14
PACT w/o IS	76.04	82.50	51.38	61.04	67.74
PACT	**83.12**	**85.21**	**51.69**	**71.46**	**72.87**

Table 2: SWE-bench Verified pass rates (%) with Qwen3.6-35B-A3B.

Method	Base Model	GRPO	PPO ($\lambda = 1.0$)	SAO	**PACT**
Pass@1 Rate	60.8	65.4	65.0	63.6	**67.4**

5 Experiments

5.1 Experimental Setup

All RL training experiments use Dressage (Huang et al., 2026), an agentic RL framework built on slime (Zhu et al., 2025) that integrates agent execution, trajectory collection, and policy training.

Training Details. For mathematical reasoning, we train Qwen3.5-4B (Qwen Team, 2026a) on a subset of DAPO-Math-17k (Yu et al., 2026) using the OpenCode (OpenCode Contributors, 2025) agentic harness, with final-answer correctness as the outcome reward. The training subset contains 3,200 problems, constructed by prioritizing problems that the initial policy fails under pass@1 evaluation and randomly sampling additional problems from the remaining pool. For agentic coding, we train Qwen3.6-35B-A3B (Qwen Team, 2026b) on OpenSWE (Fu et al., 2026a) using a Codex (OpenAI, 2025) agent through Harbor (Harbor Framework Team, 2026), with terminal rewards provided by the task verifier. For both tasks, each rollout round produces 512 trajectories. GRPO samples 8 trajectories for each of 64 prompts. The optimization batch size is 128, yielding four minibatches per rollout round. SAO uses DIS with importance ratio ranges of $[0.7, 6.0]$ for mathematical reasoning and $[0.6, 3.0]$ for coding. PACT uses the same actor-side DIS range as SAO for mathematical reasoning and PPO clipping for coding. For critic training, PACT masks out samples whose importance ratios fall outside $[0, 6]$. For both tasks, the context window is 128k tokens, and the maximum generation length per interaction turn is 64k tokens.

Evaluation. For mathematical reasoning, we evaluate on AIME 2025, AIME 2026, HMMT Nov. 2025 (Balunović et al., 2026), and BeyondAIME (ByteDance-Seed, 2025), reporting Avg@16 accuracy. We compare PACT with GRPO using Clip-Higher, PPO with $\lambda \in \{0.95, 1.0\}$, and SAO. For agentic coding, we evaluate on SWE-bench Verified (Jimenez et al., 2024) and compare PACT with SAO, GRPO with Clip-Higher, and PPO with $\lambda = 1.0$.

5.2 Main Results

Tables 1 and 2 report the results on mathematical reasoning and agentic coding, respectively. On mathematical reasoning, PACT achieves the highest accuracy on all four benchmarks, averaging 72.87% and outperforming GRPO, PPO with $\lambda = 1.0$, and SAO by 8.80, 13.16, and 21.73 percentage points, respectively. PPO with $\lambda = 1.0$ also outperforms $\lambda = 0.95$, which undergoes policy collapse during training, consistent with our analysis of intermediate critic errors in GAE. On SWE-bench Verified, PACT achieves the highest pass rate of 67.4%, outperforming GRPO, PPO with $\lambda = 1.0$, and SAO by 2.0, 2.4, and 3.8 percentage points, respectively.

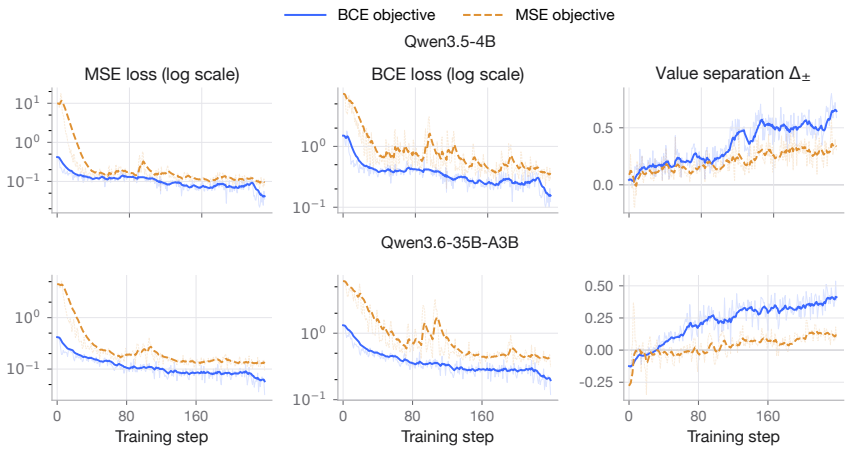


Figure 1: Fixed-policy critic pretraining with BCE and MSE objectives. Critics are trained from the same initialization on the same on-policy rollout data. We compare BCE loss, MSE loss, and value separation $\Delta_{\pm}$.

5.3 Ablation Studies

Critic Objective. We isolate the effect of the critic objective by fixing the rollout policy and training BCE and MSE critics from the same initialization on the same on-policy rollout data. We conduct this comparison in both settings described above: Qwen3.5-4B on mathematical reasoning and Qwen3.6-35B-A3B on agentic coding.

As shown in Figure 1, across both model scales, BCE-trained critics achieve lower BCE and MSE losses and greater value separation than their MSE-trained counterparts. Here, $\Delta_{\pm}$ denotes the mean predicted value on positive samples minus that on negative samples. The larger separation indicates that BCE-trained critics assign more distinct values to successful and unsuccessful trajectories.

Importance Sampling Correction. PACT w/o IS retains the Actor-then-Critic update order, BCE critic objective, and actor-side optimization settings, but removes the critic-target importance sampling correction. As shown in Table 1, adding importance correction improves average accuracy from 67.74% to 72.87%, a gain of 5.13 percentage points, with improvements on all four benchmarks. It also leads to more stable training, as illustrated by the training reward curves in Appendix D.3. This comparison supports the effectiveness of importance sampling correction in PACT.

6 Conclusion

We studied credit assignment under three regularity conditions, namely Completeness, Prefix Consistency, and Neutrality, and proved that credit is uniquely determined under these conditions. The resulting representation provides a unified basis for explaining phenomena across existing algorithms. Under an ideal teacher, OPD yields an expected policy gradient proportional to that induced by credit, identifying the teacher as an implicit critic. RLOO yields the same expected policy gradient contribution despite its coarser granularity. We further established approximate credit sparsity under bounded outcome rewards and analyzed how intermediate critic errors affect GAE in long-horizon settings.

Motivated by these findings, we proposed Policy Aligned Critic Training (PACT). Its Actor-then-Critic update order enables importance sampling correction in critic training to better align the critic with the updated policy. PACT also uses BCE instead of MSE to improve value estimation. Across four mathematical reasoning benchmarks, PACT achieves 72.87% average accuracy, outperforming GRPO and PPO by 8.80 and 13.16 percentage points, respectively. On SWE-bench Verified, PACT achieves a pass rate of 67.4%, outperforming GRPO, PPO, and SAO by 2.0, 2.4, and 3.8 percentage points, respectively. Together, these results show how a mathematical characterization of credit can explain existing algorithms and guide improvements to actor-critic training.

AI USE STATEMENT

We used generative AI tools to assist in reviewing mathematical proofs for logical gaps, checking notation and internal consistency, refining the presentation and wording of the manuscript, assisting with code implementation, and identifying potentially relevant related work. All AI-assisted code was reviewed and tested by the authors, and all cited literature was independently verified against the original sources. The central research question, the three regularity conditions, all theoretical results, and the selection and interpretation of relevant literature originated with or were determined by the authors. We made all final scientific decisions and take full responsibility for the content of this work.

REFERENCES

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting reinforce-style optimization for learning from human feedback in llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12248–12267, 2024.

Jose A Arjona-Medina, Michael Gillhofer, Michael Widrich, Thomas Unterthiner, Johannes Brandstetter, and Sepp Hochreiter. Rudder: Return decomposition for delayed rewards. *Advances in Neural Information Processing Systems*, 32, 2019.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Mislav Balunović, Jasper Dekoninck, Ivo Petrov, Nikola Jovanović, and Martin Vechev. Matharena: Evaluating llms on uncontaminated math competitions, 2026. URL https://arxiv.org/abs/2505.23281.

ByteDance-Seed. BeyondAIME: Advancing math reasoning evaluation beyond high school olympiads. https://huggingface.co/datasets/ByteDance-Seed/BeyondAIME, 2025.

Jorge Zhoujun Cheng, Shibo Hao, Tianyang Liu, Fan Zhou, Yutao Xie, Feng Yao, Yuexin Bian, Nilabjo Dey, Yonghao Zhuang, Yuheng Zha, et al. Revisiting reinforcement learning for llm reasoning from a cross-domain perspective. *Advances in Neural Information Processing Systems*, 38, 2026.

Jeff Da, Clinton Wang, Xiang Deng, Yuntao Ma, Nikhil Barhate, and Sean Hendryx. Agent-rlvr: Training software engineering agents via guidance and environment rewards. *arXiv preprint arXiv:2506.11425*, 2025.

Jesse Farebrother, Jordi Orbay, Quan Vuong, Adrien Ali Taïga, Yevgen Chebotar, Ted Xiao, Alex Irpan, Sergey Levine, Pablo Samuel Castro, Aleksandra Faust, et al. Stop regressing: Training value functions via classification for scalable deep rl. *arXiv preprint arXiv:2403.03950*, 2024.

Dayuan Fu, Shenyu Wu, Yunze Wu, Zerui Peng, Yaxing Huang, Jie Sun, Ji Zeng, Mohan Jiang, Lin Zhang, Yukun Li, Jiarui Hu, Liming Liu, Jinlong Hou, and Pengfei Liu. davinci-env: Open swe environment synthesis at scale, 2026a. URL https://arxiv.org/abs/2603.13023.

Wei Fu, Jiaxuan Gao, Xujie Shen, Chen Zhu, Zhiyu Mei, Chuyi He, Shusheng Xu, Guo Wei, Jun Mei, Jiashu Wang, et al. Areal: A large-scale asynchronous reinforcement learning system for language reasoning. *Advances in Neural Information Processing Systems*, 38:36256–36282, 2026b.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Harbor Framework Team. Harbor: A framework for evaluating and optimizing agents and models in container environments, 2026. URL https://doi.org/10.5281/zenodo.20953922.

Anna Harutyunyan, Will Dabney, Thomas Mesnard, Mohammad Gheshlaghi Azar, Bilal Piot, Nicolas Heess, Hado P van Hasselt, Gregory Wayne, Satinder Singh, Doina Precup, et al. Hindsight credit assignment. *Advances in neural information processing systems*, 32, 2019.

Zhenyu Hou, Yujiang Li, Jie Tang, and Yuxiao Dong. Single-rollout asynchronous optimization for agentic reinforcement learning. *arXiv preprint arXiv:2607.07508*, 2026.

Jian Hu, Jason Klein Liu, Haotian Xu, and Wei Shen. Reinforce++: Stabilizing critic-free policy optimization with global advantage normalization. *arXiv preprint arXiv:2501.03262*, 2025.

Liangmeng Huang, Qingchuan Li, Hongwei Xue, Shilin Yan, and Dressage Contributors. Dressage: Scalable RL for any agent and any sandbox. `https://github.com/Accio-Lab/Dressage`, 2026.

Chia-Chun Hung, Timothy Lillicrap, Josh Abramson, Yan Wu, Mehdi Mirza, Federico Carnevale, Arun Ahuja, and Greg Wayne. Optimizing agent behavior over long time scales by transporting value. *Nature communications*, 10(1):5223, 2019.

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`.

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. *arXiv preprint arXiv:2410.01679*, 2024.

Yaxuan Li, Yuxin Zuo, Bingxiang He, Jinqian Zhang, Chaojun Xiao, Cheng Qian, Tianyu Yu, Huanang Gao, Wenkai Yang, Zhiyuan Liu, et al. Rethinking on-policy distillation of large language models: Phenomenology, mechanism, and recipe. *arXiv preprint arXiv:2604.13016*, 2026.

Ziniu Li, Tian Xu, Yushun Zhang, Zhihang Lin, Yang Yu, Ruoyu Sun, and Zhi-Quan Luo. Remax: A simple, effective, and efficient reinforcement learning method for aligning large language models. *arXiv preprint arXiv:2310.10505*, 2023.

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. *arXiv preprint arXiv:2503.20783*, 2025.

Kevin Lu and Thinking Machines Lab. On-policy distillation. *Thinking Machines Lab: Connectionism*, 2025. doi: 10.64434/tml.20251026. https://thinkingmachines.ai/blog/on-policy-distillation.

Xueguang Ma, Qian Liu, Dongfu Jiang, Ge Zhang, Zejun Ma, and Wenhu Chen. General-reasoner: Advancing llm reasoning across all domains. *Advances in Neural Information Processing Systems*, 38:56596–56618, 2026.

Thomas Mesnard, Théophane Weber, Fabio Viola, Shantanu Thakoor, Alaa Saade, Anna Harutyunyan, Will Dabney, Tom Stepleton, Nicolas Heess, Arthur Guez, et al. Counterfactual credit assignment in model-free reinforcement learning. *arXiv preprint arXiv:2011.09464*, 2020.

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, 2023. URL `https://arxiv.org/abs/2306.16803`.

Marvin Minsky. Steps toward artificial intelligence. *Proceedings of the IRE*, 49(1):8–30, 1961.

OpenAI. Codex CLI. `https://github.com/openai/codex`, 2025. GitHub repository.

OpenCode Contributors. OpenCode: The open source ai coding agent. `https://github.com/anomalyco/opencode`, 2025. GitHub repository.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

Eduardo Pignatelli, Johan Ferret, Matthieu Geist, Thomas Mesnard, Hado van Hasselt, Olivier Pietquin, and Laura Toni. A survey of temporal credit assignment in deep reinforcement learning. *arXiv preprint arXiv:2312.01072*, 2023.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026a. URL https://qwen.ai/blog?id=qwen3.5.

Qwen Team. Qwen3.6-35B-A3B: Agentic coding power, now open to all, April 2026b. URL https://qwen.ai/blog?id=qwen3.6-35b-a3b.

Rajkumar Ramamurthy, Prithviraj Ammanabrolu, Kianté Brantley, Jack Hessel, Rafet Sifa, Christian Bauckhage, Hannaneh Hajishirzi, and Yejin Choi. Is reinforcement learning (not) for natural language processing: Benchmarks, baselines, and building blocks for natural language policy optimization. *arXiv preprint arXiv:2210.01241*, 2022.

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

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

Amrith Setlur, Chirag Nagpal, Adam Fisch, Xinyang Geng, Jacob Eisenstein, Rishabh Agarwal, Alekh Agarwal, Jonathan Berant, and Aviral Kumar. Rewarding progress: Scaling automated process verifiers for llm reasoning. In *International Conference on Learning Representations*, volume 2025, pp. 60808–60838, 2025.

Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in neural information processing systems*, 33:3008–3021, 2020.

Richard S Sutton. Learning to predict by the methods of temporal differences. *Machine learning*, 3 (1):9–44, 1988.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 1998.

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999.

Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. Kimi k1.5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*, 2025.

Zihan Wang, Kangrui Wang, Qineng Wang, Pingyue Zhang, Linjie Li, Zhengyuan Yang, Xing Jin, Kefan Yu, Minh Nhat Nguyen, Licheng Liu, et al. Ragen: Understanding self-evolution in llm agents via multi-turn reinforcement learning. *arXiv preprint arXiv:2504.20073*, 2025.

Yuxiang Wei, Zhiqing Sun, Emily McMilin, Jonas Gehring, David Zhang, Gabriel Synnaeve, Daniel Fried, Lingming Zhang, and Sida Wang. Toward training superintelligent software agents through self-play swe-rl. *arXiv preprint arXiv:2512.18552*, 2025.

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

Wenkai Yang, Weijie Liu, Ruobing Xie, Kai Yang, Saiyong Yang, and Yankai Lin. Learning beyond teacher: Generalized on-policy distillation with reward extrapolation. *arXiv preprint arXiv:2602.12125*, 2026.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *Advances in Neural Information Processing Systems*, 38:113222–113244, 2026.

Yufeng Yuan, Yu Yue, Ruofei Zhu, Tiantian Fan, and Lin Yan. What's behind ppo's collapse in long-cot? value optimization holds the secret. *arXiv preprint arXiv:2503.01491*, 2025.

Yu Yue, Yufeng Yuan, Qiying Yu, Xiaochen Zuo, Ruofei Zhu, Wenyuan Xu, Jiaze Chen, Chengyi Wang, TianTian Fan, Zhengyin Du, et al. Vapo: Efficient and reliable reinforcement learning for advanced reasoning tasks. *arXiv preprint arXiv:2504.05118*, 2025.

Kaiyan Zhang, Yuxin Zuo, Bingxiang He, Youbang Sun, Runze Liu, Che Jiang, Yuchen Fan, Kai Tian, Guoli Jia, Pengfei Li, et al. A survey of reinforcement learning for large reasoning models. *arXiv preprint arXiv:2509.08827*, 2025.

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, 2025. URL https://arxiv.org/abs/2507.18071.

Zilin Zhu, Chengxing Xie, Xin Lv, and slime Contributors. slime: An llm post-training framework for rl scaling. https://github.com/THUDM/slime, 2025. GitHub repository. Corresponding author: Xin Lv.

A Additional Related Work

A.1 Reinforcement Learning for Large Language Models

Reinforcement learning has become an important component of large language model post-training, supporting both preference alignment and the development of complex reasoning capabilities (Ouyang et al., 2022; Guo et al., 2025). Its early applications focused primarily on preference alignment (Zhang et al., 2025). Representative RLHF methods train scalar reward models from human preferences over model responses and use PPO to optimize language model policies (Stiennon et al., 2020; Ouyang et al., 2022; Bai et al., 2022). Their PPO implementations use learned value functions to construct token-level advantage estimates (Stiennon et al., 2020; Ouyang et al., 2022). More recently, reinforcement learning has increasingly been used to improve complex reasoning through outcome-level or verifiable rewards (Shao et al., 2024; Guo et al., 2025; Team et al., 2025; Zhang et al., 2025).

To avoid the computational cost and optimization difficulty of training a critic, ReMax (Li et al., 2023) uses the reward of a greedily decoded response as its baseline, whereas RLOO (Ahmadian et al., 2024) constructs a leave-one-out baseline from the rewards of other responses sampled for the same prompt. GRPO (Shao et al., 2024) instead replaces the learned value model with group-relative advantage estimates. REINFORCE++ (Hu et al., 2025) combines critic-free optimization with global advantage normalization. Subsequent work refines different components of critic-free optimization, including clipping, sampling, and loss aggregation in DAPO (Yu et al., 2026), normalization-bias correction in Dr. GRPO (Liu et al., 2025), and sequence-level importance weighting and clipping in GSPO (Zheng et al., 2025).

Despite the prevalence of critic-free optimization, recent work has continued to develop value-based methods for long and heterogeneous reasoning trajectories. VC-PPO (Yuan et al., 2025) attributes observed failures of PPO in long-chain-of-thought training to value-initialization bias and the attenuation of terminal reward signals in GAE, and accordingly introduces value pretraining and decoupled GAE. VAPO (Yue et al., 2025) builds on these techniques and introduces length-adaptive GAE for responses of varying lengths. The high and variable cost of generating long trajectories has also motivated asynchronous training systems (Fu et al., 2026b; Hou et al., 2026). AReaL (Fu et al., 2026b) decouples rollout generation from policy optimization and explicitly accounts for stale training samples, whereas SAO (Hou et al., 2026) combines single-rollout asynchronous training with a learned value model and a token-level GAE estimator. In contrast to these studies of optimization algorithms and training systems, our work develops a general characterization of token-level credit in long-sequence LLM reinforcement learning and studies its implications for response-level baselines and learned critics.

A.2 Credit Assignment in Reinforcement Learning

Credit assignment concerns how earlier decisions contribute to subsequent outcomes and has been recognized as a fundamental difficulty in learning systems since early work on artificial intelligence (Minsky, 1961). Many classical reinforcement learning methods connect decisions to outcomes through sampled returns and temporal structure. Monte Carlo methods use sampled returns as targets for value estimation (Sutton & Barto, 1998), whereas REINFORCE-style policy gradient methods use observed rewards or returns to construct gradient estimates (Williams, 1992). Temporal-difference learning bootstraps from temporally successive predictions (Sutton, 1988), while eligibility traces distribute subsequent TD errors to previously visited states and actions (Sutton & Barto, 1998). Actor-critic methods use learned value functions to estimate policy gradients (Sutton et al., 1999). Generalized advantage estimation combines TD residuals to control the bias-variance trade-off in advantage estimation (Schulman et al., 2015).

When consequential decisions are separated from their outcomes by long temporal gaps, subsequent work has developed more targeted mechanisms. RUDDER learns return-equivalent reward redistributions through contribution analysis of return predictions (Arjona-Medina et al., 2019), while Temporal Value Transport uses attentional memory retrieval to transport value estimates from later events to relevant events in the distant past (Hung et al., 2019). Hindsight Credit Assignment assigns credit to past decisions according to the likelihood that they led to an observed outcome (Harutyunyan et al., 2019). Counterfactual Credit Assignment constructs future-conditioned baselines to dis-

entangle an action's influence from external factors and subsequent actions (Mesnard et al., 2020), whereas COCOA estimates an action's contribution by asking whether a subsequent reward would still have been obtained under an alternative action (Meulemans et al., 2023).

Pignatelli et al. (2023) organize these developments into several methodological families, including approaches based on temporal contiguity, return decomposition, and future conditioning. Collectively, these methods address the propagation of delayed rewards, the decomposition of trajectory returns, or the estimation of action influence, but they do so through different target quantities and estimation procedures. In long-horizon reinforcement learning for large language models, this problem becomes especially salient, as many outcome-supervised methods use a single response-level or terminal reward to train a large number of token-level decisions. Our work studies the mathematical characterization of credit under three regularity conditions and uses the resulting unique representation to analyze existing algorithms and guide critic training.

B Detailed Proofs

B.1 Reward Normalization and Variance Bound

For a bounded reward R satisfying $R_{\min} \le R \le R_{\max}$ with $R_{\min} < R_{\max}$, define the normalized reward by

$$\widetilde{R} = \frac{R - R_{\min}}{R_{\max} - R_{\min}}.$$

Since this is a positive affine transformation, i.e.,

$$\widetilde{R} = aR + b, \qquad a > 0,$$

maximizing the expected reward is equivalent to maximizing the expected normalized reward.

It remains to show the variance bound for $R \in [0, 1]$. Let

$$\mu = \mathbb{E}[R].$$

Since $0 \le R \le 1$, we have

$$R^2 \le R.$$

Therefore,

$$\mathrm{Var}(R) = \mathbb{E}[R^2] - \mu^2 \le \mu - \mu^2 = \mu(1 - \mu) \le \frac{1}{4}.$$

B.2 Unique Representation Theorem

B.2.1 Prefix Consistency Implies Adaptedness

Lemma 1. *Suppose that the prompt and all environmental observations are represented as finite token sequences over a finite vocabulary. If a token-level credit assignment satisfies Prefix Consistency, then its accumulated-credit process $\{S_i\}_{i=0}^{L}$ is adapted to the prefix filtration $\{\mathcal{F}_i\}_{i=0}^{L}$; that is, S_i is $\mathcal{F}_i$-measurable for every $i = 0, \dots, L$.*

Proof. Let $(\Omega, \mathcal{F}, \mathbb{P})$ denote the underlying probability space. We assume that all information presented to the language model, including the prompt q and each environmental observation O_j, is represented as a finite token sequence over the finite vocabulary $\mathcal{V}$. Let

$$\mathcal{V}^* = \bigcup_{m=0}^{\infty} \mathcal{V}^m$$

denote the set of all finite token sequences. Since $\mathcal{V}$ is finite, $\mathcal{V}^*$ is countable. For any fixed step i, define the observed prefix

$$H_i = (q, T_1, O_1, \dots, T_i, O_i).$$

Then H_i takes values in the countable prefix space

$$\mathcal{H}_i = \mathcal{V}^* \times (\mathcal{V} \times \mathcal{V}^*)^i,$$

which we equip with the discrete sigma-algebra $2^{\mathcal{H}_i}$. By construction,

$$\mathcal{F}_i = \sigma(H_i).$$

Prefix Consistency requires that, for any $\omega, \widetilde{\omega} \in \Omega$,

$$H_i(\omega) = H_i(\widetilde{\omega}) \quad \Longrightarrow \quad S_i(\omega) = S_i(\widetilde{\omega}).$$

Hence, for each $h \in H_i(\Omega)$, the random variable S_i takes a common value on the fiber

$$H_i^{-1}(\{h\}).$$

Define $g_i : \mathcal{H}_i \to \mathbb{R}$ by assigning this common value to $g_i(h)$ for $h \in H_i(\Omega)$, and set $g_i(h) = 0$ for $h \notin H_i(\Omega)$. Prefix Consistency ensures that g_i is well defined. Because $\mathcal{H}_i$ is equipped with the discrete sigma-algebra, g_i is measurable. Moreover, by construction,

$$S_i = g_i \circ H_i.$$

Since H_i is $\mathcal{F}_i$-measurable, it follows that S_i is also $\mathcal{F}_i$-measurable. As this argument applies to every i, the accumulated-credit process is adapted to the prefix filtration.

□

B.2.2 Proof of Unique Representation Theorem

Theorem 1 (Unique Representation of Token-Level Credit)**.** *Let L denote the maximum number of generated tokens in a trajectory, and let $\tau \le L$ be the stopping time induced by the generation process. Assume that R is integrable and $\mathcal{F}_\tau$-measurable. Define*

$$V_i := \mathbb{E}[R \mid \mathcal{F}_i], \qquad i = 0, \dots, \tau.$$

There exists a unique integrable token-level credit assignment, up to almost-sure equality, $\{C_i\}_{i=1}^{\tau}$ satisfying Completeness, Prefix Consistency, and Neutrality.

It admits the representation

$$C_i = V_i - V_{i-1} = \mathbb{E}[R \mid \mathcal{F}_i] - \mathbb{E}[R \mid \mathcal{F}_{i-1}] \quad \textit{a.s.}, \qquad i = 1, \dots, \tau.$$

Moreover, $\{C_i\}_{i=1}^{\tau}$ is a martingale difference sequence with respect to $\{\mathcal{F}_i\}_{i=0}^{\tau}$.

Proof. Since $\tau \le L$ almost surely, we adopt the convention that $C_i = 0$ on $\{i > \tau\}$ and regard every credit sequence as being defined on the deterministic horizon $\{1, \dots, L\}$. We similarly define

$$V_i := \mathbb{E}[R \mid \mathcal{F}_i], \qquad i = 0, \dots, L.$$

This extension is innocuous. Indeed, since R is $\mathcal{F}_\tau$-measurable, we have

$$V_i = R \qquad \text{on } \{\tau \le i\} \quad \text{a.s.}$$

Consequently, $V_i - V_{i-1} = 0$ on $\{i > \tau\}$ almost surely.

Existence. Define

$$C_i^\star := V_i - V_{i-1}, \qquad i = 1, \dots, L.$$

We verify that this assignment satisfies the three regularity conditions.

First, the sum telescopes:

$$\sum_{i=1}^{\tau} C_i^\star = V_\tau - V_0.$$

Because R is $\mathcal{F}_\tau$-measurable,

$$V_\tau = \mathbb{E}[R \mid \mathcal{F}_\tau] = R \quad \text{a.s.}$$

Therefore,

$$\sum_{i=1}^{\tau} C_i^\star = R - \mathbb{E}[R \mid \mathcal{F}_0] \quad \text{a.s.},$$

which establishes Completeness.

Next, let

$$S_i^\star := \sum_{j=1}^{i} C_j^\star = V_i - V_0.$$

Recall that $\mathcal{F}_i = \sigma(H_i)$ and that the prefix space $\mathcal{H}_i$ is countable. Hence, we may choose a version of the conditional expectation for which

$$V_i = v_i(H_i)$$

for some measurable function $v_i : \mathcal{H}_i \to \mathbb{R}$. Likewise, $V_0 = v_0(q)$ for some measurable function v_0. It follows that

$$S_i^\star = v_i(H_i) - v_0(q).$$

Thus, if two trajectories have the same prefix up to step i, they have the same value of $S_i^\star$. This proves Prefix Consistency.

Finally, the tower property of conditional expectation gives

$$\begin{aligned}\mathbb{E}[C_i^\star \mid \mathcal{F}_{i-1}] &= \mathbb{E}[V_i - V_{i-1} \mid \mathcal{F}_{i-1}] \\ &= \mathbb{E}[\mathbb{E}[R \mid \mathcal{F}_i] \mid \mathcal{F}_{i-1}] - V_{i-1} \\ &= \mathbb{E}[R \mid \mathcal{F}_{i-1}] - V_{i-1} \\ &= 0 \quad \text{a.s.}\end{aligned}$$

Therefore, $C_i^\star$ also satisfies Neutrality, proving existence.

Uniqueness. Let $\{\widetilde{C}_i\}_{i=1}^{\tau}$ be any other credit assignment satisfying Completeness, Prefix Consistency, and Neutrality. Extend it by setting $\widetilde{C}_i = 0$ on $\{i > \tau\}$, and define its accumulated credit process by

$$\widetilde{S}_i := \sum_{j=1}^{i} \widetilde{C}_j, \qquad i = 0, \dots, L,$$

with $\widetilde{S}_0 = 0$.

By Prefix Consistency, $\widetilde{S}_i$ depends only on the observed prefix H_i. Equivalently, as established in Lemma 1, $\widetilde{S}_i$ is $\mathcal{F}_i$-measurable. Moreover, Neutrality yields

$$\begin{aligned}\mathbb{E}[\widetilde{S}_i \mid \mathcal{F}_{i-1}] &= \mathbb{E}[\widetilde{S}_{i-1} + \widetilde{C}_i \mid \mathcal{F}_{i-1}] \\ &= \widetilde{S}_{i-1} + \mathbb{E}[\widetilde{C}_i \mid \mathcal{F}_{i-1}] \\ &= \widetilde{S}_{i-1}.\end{aligned}$$

Hence, $\{\widetilde{S}_i\}_{i=0}^{L}$ is a martingale with respect to $\{\mathcal{F}_i\}_{i=0}^{L}$.

By Completeness and the zero extension after τ, its terminal value is

$$\widetilde{S}_L = \widetilde{S}_\tau = R - \mathbb{E}[R \mid \mathcal{F}_0] = R - V_0 \quad \text{a.s.}$$

A finite-horizon martingale is determined by its terminal value. Therefore, for every $i = 0, \dots, L$,

$$\begin{aligned}\widetilde{S}_i &= \mathbb{E}[\widetilde{S}_L \mid \mathcal{F}_i] \\ &= \mathbb{E}[R - V_0 \mid \mathcal{F}_i] \\ &= \mathbb{E}[R \mid \mathcal{F}_i] - V_0 \\ &= V_i - V_0 \quad \text{a.s.}\end{aligned}$$

Taking successive differences gives

$$\widetilde{C}_i = \widetilde{S}_i - \widetilde{S}_{i-1} = V_i - V_{i-1} = C_i^\star \quad \text{a.s.}$$

for every $i \le \tau$. Thus, the credit assignment is unique up to almost-sure equality.

The representation also immediately implies

$$\mathbb{E}[C_i^\star \mid \mathcal{F}_{i-1}] = 0,$$

so $\{C_i^\star\}_{i=1}^{\tau}$ is a martingale difference sequence with respect to $\{\mathcal{F}_i\}_{i=0}^{\tau}$. □

B.2.3 NECESSITY OF THE REGULARITY CONDITIONS

The three regularity conditions in Theorem 1 are logically independent. The following construction shows that removing any one of them admits credit assignments that differ from the unique conditional-reward increments.

Proposition 1 (Necessity of the three regularity conditions)**.** *None of Completeness, Prefix Consistency, and Neutrality is implied by the other two. A single bounded two-step process suffices to show that none of the three conditions is implied by the other two.*

Proof. Let X and Y be independent Rademacher random variables:

$$\mathbb{P}(X=1)=\mathbb{P}(X=-1)=\mathbb{P}(Y=1)=\mathbb{P}(Y=-1)=\frac{1}{2}.$$

Consider a stochastic process with a deterministic two-step horizon with

$$\tau=2, \qquad \mathcal{F}_0=\{\varnothing,\Omega\}, \qquad \mathcal{F}_1=\sigma(X), \qquad \mathcal{F}_2=\sigma(X,Y),$$

and define the terminal reward by

$$R=\frac{1+X}{2}.$$

Thus $R\in\{0,1\}$, and

$$V_0=\mathbb{E}[R]=\frac{1}{2}, \qquad V_1=\mathbb{E}[R\mid\mathcal{F}_1]=R, \qquad V_2=\mathbb{E}[R\mid\mathcal{F}_2]=R.$$

The unique representation from Theorem 1 is therefore

$$C_1^\star=\frac{X}{2}, \qquad C_2^\star=0.$$

Completeness Consider the assignment

$$C_1=0, \qquad C_2=0.$$

Its accumulated credits are

$$S_1=0, \qquad S_2=0,$$

so Prefix Consistency holds. Neutrality also holds because

$$\mathbb{E}[C_1\mid\mathcal{F}_0]=0, \qquad \mathbb{E}[C_2\mid\mathcal{F}_1]=0.$$

However,

$$C_1+C_2=0\neq R-V_0=\frac{X}{2}.$$

Thus an assignment may be prefix-consistent and neutral while explaining none of the realized reward deviation.

Prefix Consistency Consider the assignment

$$C_1=\frac{X}{2}+Y, \qquad C_2=-Y.$$

It satisfies Completeness because

$$C_1+C_2=\frac{X}{2}=R-V_0.$$

It also satisfies Neutrality:

$$\mathbb{E}[C_1\mid\mathcal{F}_0]=\mathbb{E}\left[\frac{X}{2}+Y\right]=0$$

and, by the independence of X and Y,

$$\mathbb{E}[C_2\mid\mathcal{F}_1]=-\mathbb{E}[Y\mid X]=0.$$

Nevertheless,

$$S_1=C_1=\frac{X}{2}+Y$$

is not $\mathcal{F}_1$-measurable. Indeed,

$$\mathrm{Var}(S_1\mid\mathcal{F}_1)=\mathrm{Var}(Y\mid X)=1.$$

Consequently, two trajectories with the same first-step information X but different future realizations of Y assign different credit to the first token. The assignment therefore uses future information to revise the credit of an already generated prefix.

Neutrality Consider the assignment

$$C_1 = X, \qquad C_2 = -\frac{X}{2}.$$

It satisfies Completeness because

$$C_1 + C_2 = \frac{X}{2} = R - V_0.$$

Both $S_1 = X$ and $S_2 = X/2$ are measurable with respect to their corresponding prefix sigma-algebras. Thus, Prefix Consistency is satisfied. However,

$$\mathbb{E}[C_2 \mid \mathcal{F}_1] = -\frac{X}{2} \neq 0 \qquad \text{a.s.}$$

The second-token credit is already completely predictable before the second token is generated. Thus credit can be moved arbitrarily between token positions through predictable compensating terms while preserving both prefix measurability and the terminal sum.

The three constructions show that each condition excludes a distinct pathology. Completeness prevents the assignment from ignoring part or all of the realized outcome. Prefix Consistency prevents future information from modifying credit assigned to an earlier prefix. Neutrality prevents predictable zero-sum transfers of credit across token positions. Hence all three conditions are necessary for the unique representation in Theorem 1. □

B.2.4 Aggregation of the Unique Token-Level Representation

Corollary 1 (Aggregation of the Unique Token-Level Representation)**.** *Consider any partition of the token sequence into consecutive segments $\mathcal{I}_1, \mathcal{I}_2, \ldots, \mathcal{I}_K$, where each segment $\mathcal{I}_k$ contains a set of consecutive token indices. Define the segment-level credit as $C^{(k)} = \sum_{i \in \mathcal{I}_k} C_i$. The resulting segment-level credit is*

$$C^{(k)} = \sum_{i \in \mathcal{I}_k} (V_i - V_{i-1}) = V_{b_k} - V_{a_k - 1},$$

where a_k and b_k denote the first and last token indices of segment $\mathcal{I}_k$.

Proof. The result follows directly from Theorem 1 by summing the unique token-level credit increments within each segment. □

B.3 Teacher Credit Equivalence Theorem

Theorem 2 (Teacher Credit Equivalence)**.** *Let $q_t^\star$ be the ideal teacher in Definition 1. Then*

$$G_t^{\mathrm{OPD}}(q_t^\star) = \frac{1}{\beta}\mathbb{E}_\pi\left[Z_t C_t^\pi | \mathcal{F}_{t-1}\right] \qquad \textit{a.s.},$$

where

$$C_t^\pi = V_t^\pi - V_{t-1}^\pi, \qquad V_t^\pi = \mathbb{E}_\pi[R \mid \mathcal{F}_t]$$

is the unique token-level credit from Theorem 1.

Proof. Condition on $\mathcal{F}_{t-1}$. The optimization problem in Definition 1 can be written as

$$\max_{q \in \Delta(\mathcal{V})} \left\{ \sum_{a \in \mathcal{V}} q(a) Q_t^\pi(a) - \beta \sum_{a \in \mathcal{V}} q(a) \log \frac{q(a)}{p_t(a)} \right\}.$$

Its Lagrangian is

$$\mathcal{L}(q, \lambda) = \sum_{a \in \mathcal{V}} q(a) Q_t^\pi(a) - \beta \sum_{a \in \mathcal{V}} q(a) \log \frac{q(a)}{p_t(a)} + \lambda \left(\sum_{a \in \mathcal{V}} q(a) - 1 \right).$$

The first-order condition for each $a \in \mathcal{V}$ is

$$Q_t^\pi(a) - \beta \left(\log \frac{q(a)}{p_t(a)} + 1 \right) + \lambda = 0.$$

Consequently, the unique optimizer is

$$q_t^\star(a) = \frac{p_t(a)\exp(Q_t^\pi(a)/\beta)}{\sum_{b\in\mathcal{V}} p_t(b)\exp(Q_t^\pi(b)/\beta)}.$$

Define the prefix-dependent normalizing factor

$$\mathcal{Z}_t = \sum_{b\in\mathcal{V}} p_t(b)\exp(Q_t^\pi(b)/\beta)\,.$$

It follows that

$$\log\frac{q_t^\star(a)}{p_t(a)} = \frac{1}{\beta}Q_t^\pi(a) - \log\mathcal{Z}_t.$$

Substituting this identity into the OPD gradient yields

$$G_t^{\mathrm{OPD}}(q_t^\star) = \frac{1}{\beta}\mathbb{E}_\pi\left[Z_t Q_t^\pi(T_t)|\mathcal{F}_{t-1}\right] - \log\mathcal{Z}_t\,\mathbb{E}_\pi\left[Z_t|\mathcal{F}_{t-1}\right].$$

The conditional score-function identity gives

$$\mathbb{E}_\pi\left[Z_t|\mathcal{F}_{t-1}\right] = 0.$$

Therefore,

$$G_t^{\mathrm{OPD}}(q_t^\star) = \frac{1}{\beta}\mathbb{E}_\pi\left[Z_t Q_t^\pi(T_t)|\mathcal{F}_{t-1}\right].$$

It remains to relate the token-action value to the unique credit. Since

$$C_t^\pi = V_t^\pi - V_{t-1}^\pi,$$

the tower property gives

$$\begin{aligned}\mathbb{E}_\pi\left[C_t^\pi|\mathcal{F}_{t-1},T_t\right] &= \mathbb{E}_\pi\left[V_t^\pi|\mathcal{F}_{t-1},T_t\right] - V_{t-1}^\pi \\ &= Q_t^\pi(T_t) - V_{t-1}^\pi.\end{aligned}$$

Because Z_t is measurable with respect to $\sigma(\mathcal{F}_{t-1},T_t)$,

$$\begin{aligned}\mathbb{E}_\pi\left[Z_t C_t^\pi|\mathcal{F}_{t-1}\right] &= \mathbb{E}_\pi\left[Z_t\mathbb{E}_\pi\left[C_t^\pi|\mathcal{F}_{t-1},T_t\right]|\mathcal{F}_{t-1}\right] \\ &= \mathbb{E}_\pi\left[Z_t\left(Q_t^\pi(T_t) - V_{t-1}^\pi\right)\middle|\mathcal{F}_{t-1}\right] \\ &= \mathbb{E}_\pi\left[Z_t Q_t^\pi(T_t)|\mathcal{F}_{t-1}\right].\end{aligned}$$

The last equality again follows from the conditional score-function identity. Combining the two expressions proves

$$G_t^{\mathrm{OPD}}(q_t^\star) = \frac{1}{\beta}\mathbb{E}_\pi\left[Z_t C_t^\pi|\mathcal{F}_{t-1}\right].$$

□

B.4 Gradient Unbiasedness of the RLOO Estimator

Throughout the proof, all expectations are conditioned on the prompt q. Following the zero-extension convention, set both C_t and Z_t to zero after termination. Assume that all expectations involving policy scores below are finite.

Proof. Let

$$Z_t = \nabla_\theta \log\pi_\theta(T_t \mid \mathcal{F}_{t-1}).$$

By Theorem 1, the unique token-level credit satisfies $C_t = V_t - V_{t-1}$ and $R_i - V_0 = \sum_{k=1}^{\tau} C_k$. We first show that

$$\mathbb{E}[Z_t(R_i - C_t)] = 0.$$

For $k < t$, C_k is $\mathcal{F}_{t-1}$-measurable, and therefore

$$\mathbb{E}[Z_t C_k] = \mathbb{E}[C_k\mathbb{E}[Z_t|\mathcal{F}_{t-1}]] = 0.$$

For $k > t$, Z_t is $\mathcal{F}_{k-1}$-measurable, and the martingale difference property of C_k gives

$$\mathbb{E}[Z_t C_k] = \mathbb{E}[Z_t \mathbb{E}[C_k | \mathcal{F}_{k-1}]] = 0.$$

The same argument applies to V_0, since V_0 is $\mathcal{F}_{t-1}$-measurable. Hence

$$\mathbb{E}[Z_t(R_i - C_t)] = 0.$$

For the RLOO baseline, $\bar{R}_{-i}$ is computed from independent responses and is independent of the current trajectory.

Thus $\mathbb{E}[Z_t \bar{R}_{-i}] = 0$. Combining the above results gives

$$\mathbb{E}[Z_t(R_i - \bar{R}_{-i})] = \mathbb{E}[Z_t C_t].$$

□

B.5 Approximate Credit Sparsity Theorem

Theorem 4 (Approximate Credit Sparsity)**.** *For any fixed prompt q,*

$$\mathbb{E}\left[\sum_{i=1}^{\tau} C_i^2 \mid \mathcal{F}_0\right] = \mathrm{Var}(R \mid \mathcal{F}_0) \le \frac{1}{4}.$$

Consequently, for any $\epsilon > 0$,

$$\mathbb{E}\left[\sum_{i=1}^{\tau} \mathbf{1}\{|C_i| > \epsilon\} \middle| \mathcal{F}_0\right] \le \frac{1}{4\epsilon^2}.$$

Proof. Since $\tau \le L$ almost surely, extend the credit sequence to the deterministic horizon $\{1, \dots, L\}$ by setting $C_i = 0$ on $\{i > \tau\}$. By the unique representation, $C_i = V_i - V_{i-1}$, where $V_i = \mathbb{E}[R \mid \mathcal{F}_i]$. Because $R \in [0, 1]$, these increments are square-integrable.

For $1 \le i < j \le L$, C_i is $\mathcal{F}_{j-1}$-measurable. The tower property and the martingale difference property therefore give

$$\mathbb{E}[C_i C_j \mid \mathcal{F}_0] = \mathbb{E}[C_i \, \mathbb{E}[C_j \mid \mathcal{F}_{j-1}] | \mathcal{F}_0] = 0.$$

By Completeness and the zero extension,

$$\sum_{i=1}^{L} C_i = \sum_{i=1}^{\tau} C_i = R - V_0, \qquad V_0 = \mathbb{E}[R \mid \mathcal{F}_0].$$

Consequently,

$$\begin{aligned}
\mathbb{E}\left[\sum_{i=1}^{\tau} C_i^2 \middle| \mathcal{F}_0\right] &= \sum_{i=1}^{L} \mathbb{E}[C_i^2 \mid \mathcal{F}_0] \\
&= \mathbb{E}\left[\left(\sum_{i=1}^{L} C_i\right)^2 \middle| \mathcal{F}_0\right] \\
&= \mathbb{E}[(R - V_0)^2 \mid \mathcal{F}_0] \\
&= \mathrm{Var}(R \mid \mathcal{F}_0).
\end{aligned}$$

Since $R^2 \le R$ and $V_0 \in [0, 1]$,

$$\mathrm{Var}(R \mid \mathcal{F}_0) = \mathbb{E}[R^2 \mid \mathcal{F}_0] - V_0^2 \le V_0(1 - V_0) \le \frac{1}{4}.$$

Finally, for any $\epsilon > 0$, the pointwise inequality

$$\epsilon^2 \sum_{i=1}^{\tau} \mathbf{1}\{|C_i| > \epsilon\} \le \sum_{i=1}^{\tau} C_i^2$$

implies, upon taking conditional expectations,

$$\mathbb{E}\left[\sum_{i=1}^{\tau} \mathbf{1}\{|C_i| > \epsilon\} \middle| \mathcal{F}_0\right] \le \frac{\mathrm{Var}(R \mid \mathcal{F}_0)}{\epsilon^2} \le \frac{1}{4\epsilon^2}.$$

□

B.6 Gradient of BCE with Importance-Weighted Targets

For the prefix $\mathcal{F}_{t-1}$, let the critic prediction be

$$\widehat{V}_{\phi,t-1} = \sigma(z_{\phi,t-1}),$$

where $z_{\phi,t-1}$ is the critic logit. Let I_t be the exact continuation importance ratio satisfying

$$\mathbb{E}_\mu[I_t R \mid \mathcal{F}_{t-1}] = V^\pi_{t-1}.$$

The importance-weighted target $I_t R$ is held fixed when differentiating with respect to the critic logit. Its BCE is

$$\begin{aligned}\mathrm{BCE}\left(\sigma(z_{\phi,t-1}), I_t R\right) &= -I_t R \log \sigma(z_{\phi,t-1}) \\ &\quad - (1 - I_t R) \log\left(1 - \sigma(z_{\phi,t-1})\right) \\ &= \log\left(1 + \exp(z_{\phi,t-1})\right) - I_t R z_{\phi,t-1}.\end{aligned}$$

Therefore,

$$\frac{\partial}{\partial z_{\phi,t-1}} \mathrm{BCE}\left(\sigma(z_{\phi,t-1}), I_t R\right) = \widehat{V}_{\phi,t-1} - I_t R.$$

Taking the conditional expectation under the rollout policy gives

$$\begin{aligned}\mathbb{E}_\mu\left[\frac{\partial}{\partial z_{\phi,t-1}} \mathrm{BCE}\left(\sigma(z_{\phi,t-1}), I_t R\right) \middle| \mathcal{F}_{t-1}\right] &= \widehat{V}_{\phi,t-1} - \mathbb{E}_\mu[I_t R \mid \mathcal{F}_{t-1}] \\ &= \widehat{V}_{\phi,t-1} - V^\pi_{t-1}.\end{aligned}$$

For $0 < V^\pi_{t-1} < 1$, the expected logit gradient vanishes exactly when

$$\widehat{V}_{\phi,t-1} = V^\pi_{t-1}.$$

When $V^\pi_{t-1} = 0$ or 1, equality is approached as the logit tends to $-\infty$ or $+\infty$, respectively. In addition,

$$\frac{\partial^2}{\partial z^2_{\phi,t-1}} \mathrm{BCE}\left(\sigma(z_{\phi,t-1}), I_t R\right) = \widehat{V}_{\phi,t-1}\left(1 - \widehat{V}_{\phi,t-1}\right) \le \frac{1}{4}.$$

Thus, although individual importance-weighted targets may fall outside $[0, 1]$, the expected logit gradient has the sign of the value prediction error, and the curvature with respect to the logit remains uniformly bounded.

We further show that approaching the minimum expected BCE implies approaching the target conditional mean in mean square. Let Y be an integrable target and write

$$m = \mathbb{E}[Y \mid \mathcal{F}] \in [0, 1].$$

For an $\mathcal{F}$-measurable prediction $v \in (0, 1)$, define

$$\ell(v) = \mathbb{E}[\mathrm{BCE}(v, Y) \mid \mathcal{F}] = -m \log v - (1 - m) \log(1 - v).$$

All logarithms are natural, and the expectations below are assumed to exist and be finite. Differentiating gives

$$\ell'(v) = \frac{v - m}{v(1 - v)}.$$

Since $u(1 - u) \le 1/4$, for $v \ge m$,

$$\ell(v) - \ell(m) = \int_m^v \frac{u - m}{u(1 - u)}\, du \ge 4 \int_m^v (u - m)\, du = 2(v - m)^2.$$

Similarly, for $v < m$,

$$\ell(v) - \ell(m) = \int_v^m \frac{m - u}{u(1 - u)}\, du \ge 4 \int_v^m (m - u)\, du = 2(v - m)^2.$$

Here,

$$\ell(m) = -m \log m - (1 - m) \log(1 - m),$$

with $0 \log 0 = 0$. When $m = 0$ or $m = 1$, this expression denotes the infimum over $v \in (0, 1)$, and the inequalities follow by taking limits. Taking expectations gives

$$\mathbb{E}[(v - m)^2] \leq \frac{1}{2} \mathbb{E}[\ell(v) - \ell(m)]. \tag{1}$$

Consequently, for any sequence of predictions v_n,

$$\mathbb{E}[\ell(v_n) - \ell(m)] \longrightarrow 0 \quad \Longrightarrow \quad \mathbb{E}[(v_n - m)^2] \longrightarrow 0.$$

This applies both to the on-policy target $Y = R$ and to the exact importance-weighted target $Y = IR$, provided that $\mathbb{E}_\mu[IR \mid \mathcal{F}] = V^\pi$.

We can further relate value estimation error to credit estimation error. Fix a policy π and a trajectory distribution under which all the following expectations are taken. For $i \in \{t-1, t\}$, suppose that

$$\mathbb{E}[Y_i \mid \mathcal{F}_i] = V_i^\pi,$$

and let ℓ_i denote the corresponding conditional expected BCE. Define

$$\varepsilon_i = \widehat{V}_{\phi,i} - V_i^\pi, \qquad \widehat{C}_{\phi,t} = \widehat{V}_{\phi,t} - \widehat{V}_{\phi,t-1}.$$

Since $C_t^\pi = V_t^\pi - V_{t-1}^\pi$,

$$\widehat{C}_{\phi,t} - C_t^\pi = \varepsilon_t - \varepsilon_{t-1}.$$

Using $(a - b)^2 \leq 2a^2 + 2b^2$ and Equation 1, we obtain

$$\begin{aligned} \mathbb{E}[(\widehat{C}_{\phi,t} - C_t^\pi)^2] &\leq 2\mathbb{E}[\varepsilon_t^2] + 2\mathbb{E}[\varepsilon_{t-1}^2] \\ &\leq \mathbb{E}[\ell_t(\widehat{V}_{\phi,t}) - \ell_t(V_t^\pi)] \\ &\quad + \mathbb{E}[\ell_{t-1}(\widehat{V}_{\phi,t-1}) - \ell_{t-1}(V_{t-1}^\pi)]. \end{aligned} \tag{2}$$

Thus, if the expected BCE at both adjacent prefixes approaches its minimum, the estimated credit converges to C_t^π in mean square.

B.7 Current-Token Importance Weighting as a One-Step Policy-Tracking Surrogate

Exact synchronization with an updated policy requires correcting the entire continuation distribution. Such a correction involves a product of token-level importance ratios and can be unstable for long trajectories. We characterize here the current-token importance target used in our method and its relation to the exact updated-policy value.

Let μ denote the rollout policy and π the policy obtained after actor optimization. For the token generated after prefix $\mathcal{F}_{t-1}$, define

$$\rho_t = \frac{\pi(T_t \mid \mathcal{F}_{t-1})}{\mu(T_t \mid \mathcal{F}_{t-1})}.$$

We assume that $\pi(\cdot \mid \mathcal{F}_{t-1})$ is absolutely continuous with respect to $\mu(\cdot \mid \mathcal{F}_{t-1})$. The exact continuation importance ratio is

$$W_t = \prod_{k=t}^{\tau} \rho_k,$$

and change of measure gives

$$V_{t-1}^\pi = \mathbb{E}_\pi[R \mid \mathcal{F}_{t-1}] = \mathbb{E}_\mu[W_t R \mid \mathcal{F}_{t-1}].$$

Our practical critic target instead uses only the current-token ratio,

$$Y_t^{\text{one}} = \rho_t R, \qquad \widetilde{V}_{t-1}^{\pi,\mu} = \mathbb{E}_\mu[\rho_t R \mid \mathcal{F}_{t-1}].$$

Proposition 2 (One-step policy tracking)**.** *The surrogate value $\widetilde{V}_{t-1}^{\pi,\mu}$ is exactly the value of the hybrid policy that samples T_t from π and follows μ thereafter:*

$$\widetilde{V}_{t-1}^{\pi,\mu} = \mathbb{E}_{\pi \triangleright_t \mu}[R \mid \mathcal{F}_{t-1}].$$

Moreover, suppose that $R \in [0,1]$, $\tau \le L$, and

$$|\rho_k - 1| \le \varepsilon, \qquad k = t, \dots, \tau.$$

Writing $\delta_k = \rho_k - 1$, we have

$$V_{t-1}^{\pi} = \widetilde{V}_{t-1}^{\pi,\mu} + \mathbb{E}_{\mu}\left[R \sum_{k=t+1}^{\tau} \delta_k \middle| \mathcal{F}_{t-1} \right] + \mathcal{R}_t,$$

where

$$|\mathcal{R}_t| \le (1+\varepsilon)^{L-t+1} - 1 - (L-t+1)\varepsilon = O_L(\varepsilon^2).$$

Consequently,

$$\widetilde{V}_{t-1}^{\pi,\mu} = V_{t-1}^{\pi} + O_L(\varepsilon)$$

as $\varepsilon \to 0$ for fixed L.

Proof. Conditioning first on T_t, we obtain

$$\begin{aligned}
\mathbb{E}_{\mu}[\rho_t R \mid \mathcal{F}_{t-1}] &= \sum_a \mu(a \mid \mathcal{F}_{t-1}) \frac{\pi(a \mid \mathcal{F}_{t-1})}{\mu(a \mid \mathcal{F}_{t-1})} \mathbb{E}_{\mu}[R \mid \mathcal{F}_{t-1}, T_t = a] \\
&= \sum_a \pi(a \mid \mathcal{F}_{t-1}) \mathbb{E}_{\mu}[R \mid \mathcal{F}_{t-1}, T_t = a].
\end{aligned}$$

The last expression is precisely the expected reward obtained by sampling the current token from π and using μ for the remaining continuation. This proves the hybrid-policy identity.

For the local expansion, observe that

$$W_t = \prod_{k=t}^{\tau} (1+\delta_k) = 1 + \sum_{k=t}^{\tau} \delta_k + r_t,$$

where r_t contains all products involving at least two distinct δ_k's. Since $|\delta_k| \le \varepsilon$ and $\tau - t + 1 \le L - t + 1$,

$$|r_t| \le \sum_{m=2}^{L-t+1} \binom{L-t+1}{m} \varepsilon^m = (1+\varepsilon)^{L-t+1} - 1 - (L-t+1)\varepsilon.$$

Multiplying by R, taking the conditional expectation, and using $0 \le R \le 1$ gives

$$\begin{aligned}
V_{t-1}^{\pi} &= \mathbb{E}_{\mu}[W_t R \mid \mathcal{F}_{t-1}] \\
&= \mathbb{E}_{\mu}[(1+\delta_t) R \mid \mathcal{F}_{t-1}] + \mathbb{E}_{\mu}\left[R \sum_{k=t+1}^{\tau} \delta_k \middle| \mathcal{F}_{t-1} \right] + \mathcal{R}_t \\
&= \widetilde{V}_{t-1}^{\pi,\mu} + \mathbb{E}_{\mu}\left[R \sum_{k=t+1}^{\tau} \delta_k \middle| \mathcal{F}_{t-1} \right] + \mathcal{R}_t.
\end{aligned}$$

The stated bounds follow immediately. □

For comparison, the complete first-order expansion of the continuation importance target is

$$Y_t^{\mathrm{FO}} = R\left(1 + \sum_{k=t}^{\tau} (\rho_k - 1)\right),$$

which satisfies

$$\mathbb{E}_{\mu}[Y_t^{\mathrm{FO}} \mid \mathcal{F}_{t-1}] = V_{t-1}^{\pi} + O_L(\varepsilon^2).$$

The practical target $\rho_t R$ retains the current-token component of this first-order correction and omits the remaining continuation terms. It is therefore a one-step truncation of the full policy correction. This choice avoids the multiplicative continuation weight while moving the critic target toward the updated policy.

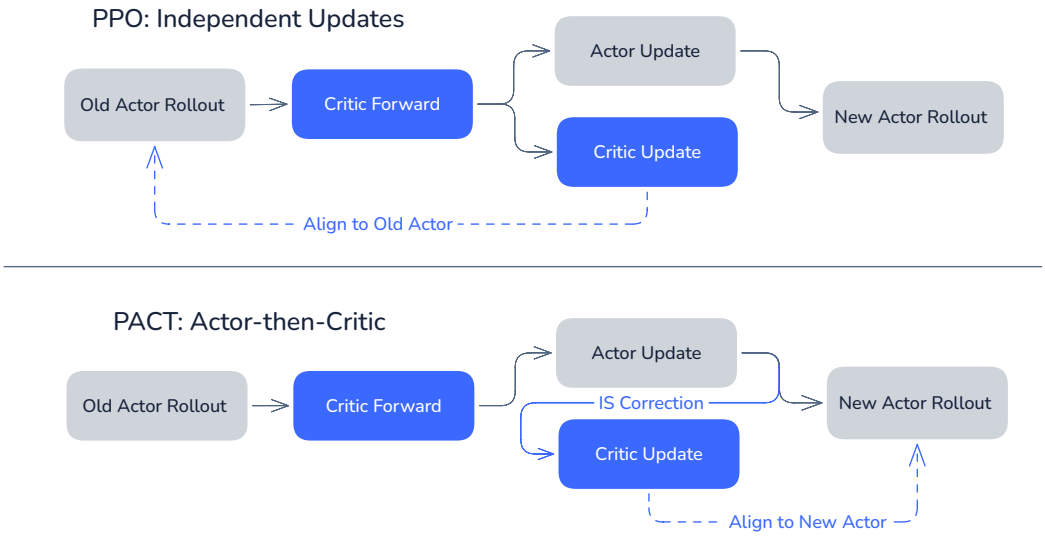


Figure 2: PPO and PACT training workflows.

C Policy Aligned Critic Training Procedure

This section provides the detailed training procedure for PACT. Figure 2 compares the PPO and PACT workflows. Algorithm 1 summarizes PACT. Here, ActorUpdate denotes its policy optimization step. All importance ratios and critic targets are computed token-wise.

Notation: actor policy π_{θ_0}; probabilistic critic $\widehat{V}_{\phi_0}(\mathcal{F}) = \sigma(z_{\phi_0}(\mathcal{F}))$; number of iterations K; actor/critic steps B; importance-ratio acceptance bounds $\rho_{\min}, \rho_{\max}$

Algorithm 1 Policy Aligned Critic Training

Require: $\pi_{\theta_0}, \widehat{V}_{\phi_0}, K, B, \rho_{\min}, \rho_{\max}$
  for $k = 0, \ldots, K-1$ **do**
    $\mathcal{D}_k, R, \ell^{\text{old}} \leftarrow \textsc{Rollout}(\pi_{\theta_k})$
    $\widehat{V} \leftarrow \mathrm{sg}\big(\widehat{V}_{\phi_k}(\mathcal{D}_k)\big)$
    $\widehat{C} \leftarrow \textsc{Credit}(R, \widehat{V})$
    $\theta \leftarrow \theta_k$
    for $b = 1, \ldots, B$ **do**
      $\theta \leftarrow \textsc{ActorUpdate}(\theta, \mathcal{D}_k, \widehat{C}, \ell^{\text{old}})$
    end for
    $\theta_{k+1} \leftarrow \theta$
    $\ell^{\text{new}} \leftarrow \textsc{LogProb}(\pi_{\theta_{k+1}}, \mathcal{D}_k)$
    $\rho \leftarrow \mathrm{sg}\big[\exp\big(\ell^{\text{new}} - \ell^{\text{old}}\big)\big]$
    $\mathsf{m} \leftarrow \mathbb{I}\big\{\rho_{\min} \le \rho \le \rho_{\max}\big\}$
    $Y \leftarrow \rho \odot R$
    $\phi \leftarrow \phi_k$
    for $b = 1, \ldots, B$ **do**
      $\mathcal{L}_{\text{PACT}}(\phi) \leftarrow \dfrac{\sum \mathsf{m} \odot \mathrm{BCE}\big(\widehat{V}_\phi(\mathcal{D}_k), Y\big)}{\sum \mathsf{m}}$
      $\phi \leftarrow \phi - \eta_\phi \nabla_\phi \mathcal{L}_{\text{PACT}}(\phi)$
    end for
    $\phi_{k+1} \leftarrow \phi$
  end for
  return $\pi_{\theta_K}, \widehat{V}_{\phi_K}$

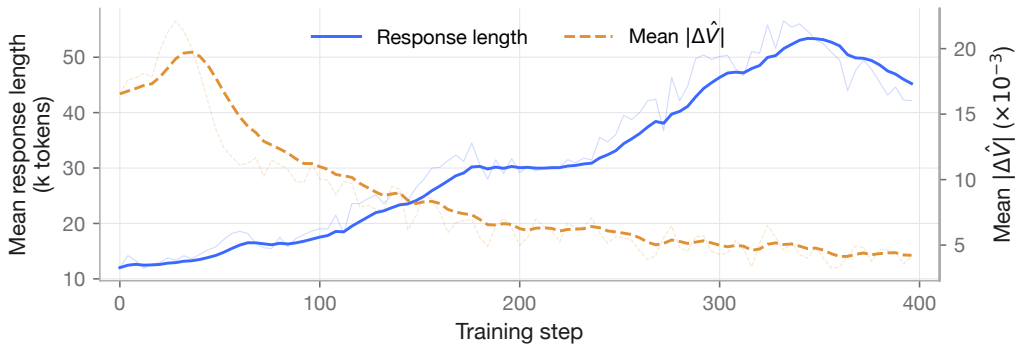


Figure 3: Dynamics of response length and mean $|\widehat{V}_t - \widehat{V}_{t-1}|$ during PACT training.

D Experimental Details

D.1 Prompt of Agentic Reasoning for Math

After constructing the training subset, we convert each selected DAPO-Math-17k example and each evaluation example from a direct-answer problem into an agentic tool-use episode. We preserve the original problem statement and ground-truth final answer, while adding neither reference solutions nor tool-use demonstrations. The ground-truth answer is stored separately and is never included in the policy input.

Specifically, we prepend an instruction that informs the model of the available Python sandbox and explicitly encourages tool use for calculation, case enumeration, and verification. We also append a formatting instruction requiring the final answer to appear in a `\boxed{}` expression on the last line. The resulting prompt template is:

> Solve the following math problem. You are in a sandbox with a working Python interpreter—USE it: write and run Python code to do the calculations, enumerate cases, and verify your result, instead of computing by hand.
>
> `original problem statement`
>
> Please put your final answer in `\boxed{}`. The last line of your response must have the form `Answer: \boxed{answer}`.

D.2 Further Empirical Evidence for Theorem 4

As shown in Figure 3, responses become progressively longer during PACT training, while the mean $|\widehat{V}_t - \widehat{V}_{t-1}|$ decreases. This trend is consistent with the intuition behind Theorem 4.

D.3 Training Dynamics

Figures 4 and 5 show the evolution of mean trajectory rewards during training on mathematical reasoning and agentic coding, respectively.

D.4 Key Training Hyperparameters

Table 3 summarizes the key training hyperparameters for mathematical reasoning and coding.

PPO, GRPO, and SAO use rollout log-probabilities for actor updates, whereas PACT applies TIS with a ratio range of $[0, 2]$ to account for the mismatch between training and rollout inference. PACT w/o critic IS retains all other settings of PACT, including actor-side corrections, the BCE critic objective, and the Actor-then-Critic update order.

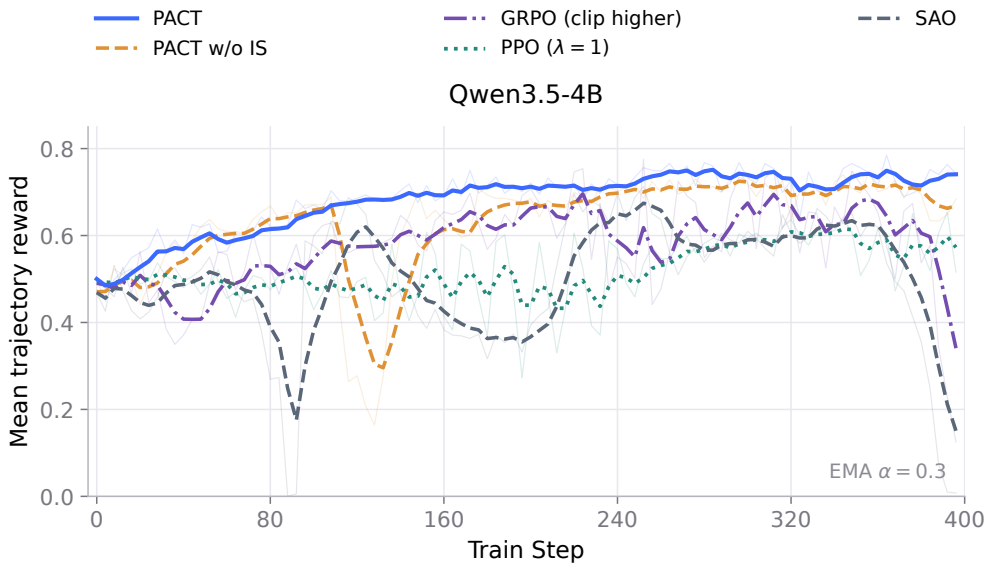


Figure 4: Training reward on agentic mathematical reasoning with Qwen3.5-4B. Faint lines show raw mean trajectory rewards, while bold lines show exponential moving averages with $\alpha = 0.3$. Each rollout round corresponds to four training steps.

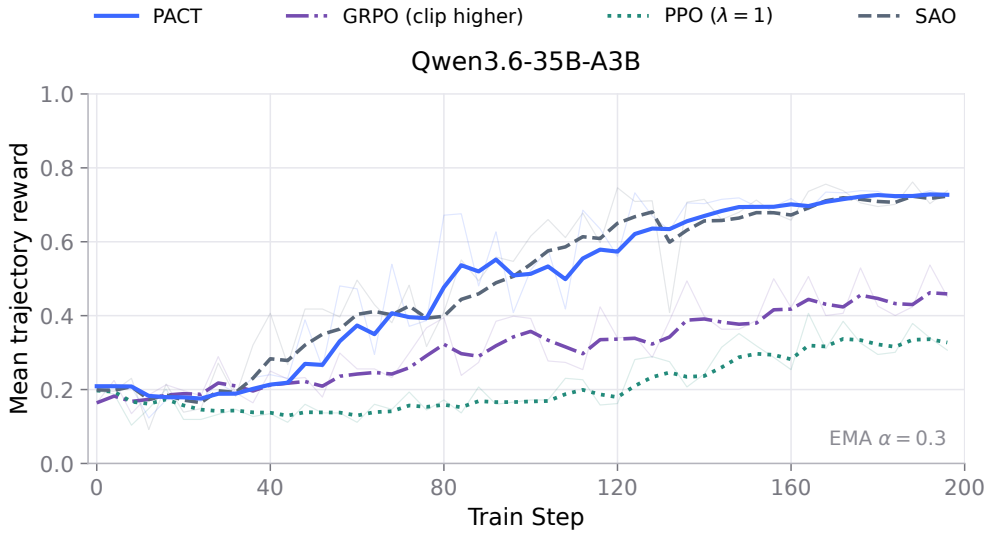


Figure 5: Training reward on agentic coding with Qwen3.6-35B-A3B. Faint lines show raw mean trajectory rewards, while bold lines show exponential moving averages with $\alpha = 0.3$.

E LIMITATIONS

Our characterization of credits does not imply that all possible notions of credit must take the form derived in this work. The uniqueness result is conditional on the three proposed regularity conditions. Just as replacing Euclid's parallel postulate leads to different but internally consistent geometries, adopting different requirements for credit may lead to different representations.

Second, the notion of credit studied here is statistical rather than causal. It is defined through conditional expectations and does not characterize the counterfactual causal effect of replacing an individual token or action.

Table 3: Key training hyperparameters.

Hyperparameter	Mathematical Reasoning	Coding
Actor learning rate	$1e-6$	$1e-6$
Critic learning rate	$5e-6$	$5e-6$
PACT actor update	DIS	PPO clipping
PACT actor DIS range	$[0.7, 6.0]$	N/A
PACT critic IS acceptance range	$[0, 6]$	$[0, 6]$

Finally, this work characterizes what token-level credit should be under the proposed conditions, but does not solve the problem of estimating it exactly in practical LLM reinforcement learning. Developing accurate and efficient estimators of token-level credit remains an important direction for future work.