PRIVY TO THE FOIL: RECASTING VALUE ESTIMATION WITH A SELF-PRIVILEGED CRITIC FOR RLVR

Anonymous authors
Paper under double-blind review

ABSTRACT

Assigning credit to intermediate steps remains a central challenge in training Large Language Models (LLMs) on multi-step reasoning tasks with sparse terminal rewards, and actor-critic methods such as PPO address this by learning value functions to construct token-level advantages. Their effectiveness, however, hinges on reliable value estimation, a difficult task requiring the critic to both assess progress toward a correct solution and anticipate an evolving policy's future behavior; errors in either can compromise credit assignment and destabilize online training. In this paper, we revisit the standard state-only formulation of value estimation and propose **πPPO**, a self-privileged actor-critic framework. By reusing verified same-prompt rollouts as contrastive evidence, πPPO helps the critic assess intermediate reasoning against successful and failed attempts, while preserving standard policy optimization and the deployment interface. Experiments show that πPPO consistently improves value-estimation quality by a substantial margin and outperforms representative actor-critic and critic-free RLVR baselines on challenging mathematical reasoning benchmarks, while remaining effective even when paired with substantially smaller asymmetric critics.

1 INTRODUCTION

On-policy feedback-based learning methods stand out in LLM post-training for their effectiveness, especially on incentivizing long-form reasoning capabilities on verifiable tasks, providing a promising way for scaling (Guo et al., 2025a; Team et al., 2025; Yang et al., 2025). Despite the empirical success of recent value-free methods, explicit value estimation remains a fundamental optimization paradigm in reinforcement learning. Beyond reducing variance, it provides a state-conditioned baseline, enabling principled credit assignment and readily accommodates diverse rollout structures (Williams, 1992; Greensmith et al., 2004; Schulman et al., 2015). Therefore, advancing value estimation offers a promising path toward improving policy optimization in LLM post-training.

However, accurately estimating the value of a partial response remains inherently challenging: the value network must both understand the space of correct solutions—the very task the policy itself is learning—and predict the policy's future behavior (Shan et al., 2026; Zhang et al., 2026). This difficulty is further exacerbated in long-form reasoning, where rewards are sparse and often observable only at the ends of exceptionally long trajectories, leaving a vast space of intermediate states with learning targets that are both weakly informative and highly uncertain. In actor–critic training, inaccurate value estimates can distort token-level advantages and misguide policy updates, undermining training stability and final performance (Schulman et al., 2015; Yuan et al., 2025).

Existing approaches improve critic learning through better initialization and return estimation (Yuan et al., 2025; Yue et al., 2025; Hou et al., 2026), or obtain more informative value targets through additional continuation sampling and intermediate-state modeling (Kazemnejad et al., 2024; Guo et al., 2025b; Chen et al., 2026b). Nevertheless, whether by strengthening the critic or enriching its targets, these solutions inherit the same largely unquestioned premise: both demands placed on the critic must be met by the critic itself, using only the information available from its prescribed vantage point; as we show later, however, this need not be taken as given.

To this end, we propose **P**rivileged-**I**nformation **PPO** (**πPPO**), which turns verified trajectories into online references for state-level value prediction. We adapt PPO to sample groups of rollouts for each prompt, giving each trajectory a dual role: a sample for policy learning and a potential

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.

reference for evaluating partial responses in other trajectories. For each target rollout, we construct a leave-target-out critic context from its verified siblings, preserving a valid baseline while leaving the policy's conditioning unchanged. These references evolve with the policy and require no additional generation beyond the rollout groups used for policy updates.

As shown in Section 4, πPPO consistently improves value-estimation quality across two model backbones on mathematical reasoning tasks, with corresponding gains in end-to-end policy performance over representative RL baselines. This design also enables πPPO to use substantially smaller critics while outperforming actor–critic baselines with actor-sized value models.

Our contributions are as follows:

- **Conceptual.** We recast value estimation in RLVR as an information problem characterizing inherent properties of value prediction under binary verifiable rewards that highlight potential limitations of the conventional state-only formulation.
- **Methodological.** We propose πPPO, which supplies online evidence on the critic side, conditioning each value estimation on a leave-target-out context of verified same-prompt rollouts already collected for policy updates, thus preserving a valid baseline while leaving policy inference unchanged, all without additional rollout generation.
- **Empirical.** We show that πPPO consistently improves value-estimation quality and outperforms representative actor–critic and critic-free RLVR baselines on mathematical reasoning benchmarks. It remains effective even with substantially smaller asymmetric critics.

2 PRELIMINARY

Notation. In this paper, we define an LLM parameterized by θ as a policy π_θ. Let x denote a query and $\mathcal{D}$ the set of queries. Given a response y to the query x, its likelihood under the policy π_θ is expressed as $\pi_\theta(y \mid x) = \prod_{t=1}^{|y|} \pi_\theta(y_t \mid x, y_{<t})$, where $|y|$ denotes the number of tokens in y. A query-response pair (x, y) is scored by a rule-based outcome reward $r(x, y) \in \{0, 1\}$, indicating whether the response y aligns with the ground truth of x.

Proximal Policy Optimization (PPO). PPO (Schulman et al., 2017) follows the actor–critic framework, jointly learning a policy model (actor) π_θ and a value model (critic) V_ϕ parameterized by ϕ, updating the policy using the following clipped policy surrogate (we omit the KL regularization term hereinafter for brevity):

$$\mathcal{J}_{\text{PPO}}(\theta) = \mathbb{E}_{x\sim\mathcal{D},\, y\sim\pi_{\theta_{\text{old}}}(\cdot|x)} \left[\frac{1}{|y|} \sum_{t=1}^{|y|} \min\left(\rho_t(\theta)\widehat{A}_t,\ \text{clip}\left(\rho_t(\theta), 1-\varepsilon, 1+\varepsilon\right) \widehat{A}_t \right) \right], \tag{1}$$

where $\rho_t(\theta) = \frac{\pi_\theta(y_t|x, y_{<t})}{\pi_{\theta_{\text{old}}}(y_t|x, y_{<t})}$ denotes the token-level probability ratio at step t, and ε is the clipping coefficient. The learned critic enters the policy update through the advantage $\widehat{A}_t$, which is commonly constructed using Generalized Advantage Estimation (GAE) (Schulman et al., 2015). Since the critic scores every state $s_t = (x, y_{<t})$ individually, the objective thereby enables dense, token-level credit assignment even when the environment returns a single terminal reward.

Group Relative Policy Optimization (GRPO). GRPO (Shao et al., 2024) bypasses the value model by computing the relative advantage of each response within a group of G responses to the same query. Given $\{y_i\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}(\cdot \mid x)$, it averages the clipped surrogate over the group, using

$$\widehat{A}_{i,t} = \widehat{A}_i = \frac{r(x, y_i) - \text{mean}\big(\{r(x, y_j)\}_{j=1}^{G}\big)}{\text{std}\big(\{r(x, y_j)\}_{j=1}^{G}\big)}. \tag{2}$$

All tokens in y_i share the same advantage $\widehat{A}_i$.

Privileged Information (PI). Learning using privileged information augments a standard input $x \in \mathcal{X}$ with task-relevant auxiliary information $x^\star \in \mathcal{X}^\star$ available during training but absent from the standard inference input (Vapnik & Vashist, 2009). For LLMs, both the query and privileged information can be represented as token sequences over the model vocabulary $\mathcal{V}$, i.e., $x, \mathcal{I} \in \mathcal{V}^*$, and incorporated by conditioning the policy as $\pi_\theta(\cdot \mid x, \mathcal{I})$, where $\mathcal{I}$ may encode hints, verified answers, or reference rationales (Qu et al., 2026; Zhao et al., 2026).

3 METHODOLOGY

In this section, we introduce a new perspective on incorporating privileged information into policy optimization: enriching the critic's information while preserving the policy's original objective of maximizing environment-defined rewards. We instantiate this perspective using information available within the training process, without additional generation or external expert knowledge.

3.1 INTERPRETING A STATE: THE "MISSING" CONTEXT IN VALUE ESTIMATION

Standard critics fit a single value predictor across states, predicting each state's expected return from its representation. We view this shared-prediction setup through the lens of fully amortized optimization (Shu, 2017; Amos, 2025), in which a generic model serves all prediction instances without instance-specific adaptation or auxiliary evidence. We first examine the resulting limitation of this formulation and then characterize the underlying information structure for value estimation under binary verifiable rewards.

Proposition 1 (Privileged conditioning for amortized value estimation)**.**
Fix a policy π with undiscounted terminal return $R \in \{0,1\}$, and define $V^{\pi}(s) = \mathbb{E}[R \mid s]$ and $V^{\pi,+}(s,\mathcal{I}) = \mathbb{E}[R \mid s,\mathcal{I}]$. Assume finite squared error and that contextual critics $V_{\phi}^{+}(s,\mathcal{I})$ can represent every state-only critic $V_{\phi}(s)$. Under a common joint distribution of $(s,\mathcal{I},R)$,

$$\underbrace{\inf_{\phi} \mathbb{E}[(R - V_{\phi}^{+}(s,\mathcal{I}))^2]}_{\textit{contextual prediction risk}} = \mathbb{E}_{s,\mathcal{I}}[\mathrm{Var}(R \mid s,\mathcal{I})] + \inf_{\phi} \mathbb{E}[(V_{\phi}^{+}(s,\mathcal{I}) - V^{\pi,+}(s,\mathcal{I}))^2]$$

$$\leq \mathbb{E}_{s}[\mathrm{Var}(R \mid s)] + \inf_{\phi} \mathbb{E}[(V_{\phi}(s) - V^{\pi}(s))^2] = \underbrace{\inf_{\phi} \mathbb{E}[(R - V_{\phi}(s))^2]}_{\textit{state-only prediction risk}}.$$

Thus, the residual error is not solely an optimization artifact: part of it reflects return uncertainty unresolved by the target state alone. We interpret this gap as *informational*: a state-conditioned predictor captures only the return variation predictable from s, while additional evidence $\mathcal{I}$ can reduce the Bayes error insofar as it explains residual variation beyond s.

Under binary verifiable rewards, this return uncertainty concerns whether a trajectory will succeed or fail. We next examine value prediction through this outcome distinction, relating prediction error to the contrast between the critic's predictions on successful and failed trajectories.

Proposition 2 (Success–failure contrast in value estimation)**.**
Fix a policy π with undiscounted terminal return $R \in \{0,1\}$ and $\mathrm{Var}(R) > 0$. All expectations are taken under the common joint distribution of (s,R). For any square-integrable $V_{\phi}(s)$, define

$$\Delta_{\phi} = \mathbb{E}[V_{\phi}(s) \mid R = 1] - \mathbb{E}[V_{\phi}(s) \mid R = 0].$$

Its prediction risk satisfies

$$\mathbb{E}[(R - V_{\phi}(s))^2] \geq \underbrace{\mathrm{Var}(R)(1 - \Delta_{\phi})^2}_{\textit{success–failure contrast error}}.$$

Consequently, any critic outperforming the optimal constant predictor $V_0 = \mathbb{E}[R]$ (with risk $\mathrm{Var}(R)$) must exhibit positive success–failure contrast, i.e., $\Delta_{\phi} > 0$.

The bound specifies a necessary condition on the contrast between the critic's predictions on successful and failed trajectories for achieving a given level of prediction accuracy.

Taken together, we therefore rethink the evidence also as *contrastive*: by providing positive and negative references for the outcome distinction that the critic must infer, we reduce the uncertainty left by state-only value estimation insofar as the evidence clarifies the target prefix's relative support under the successful and failed trajectory processes.

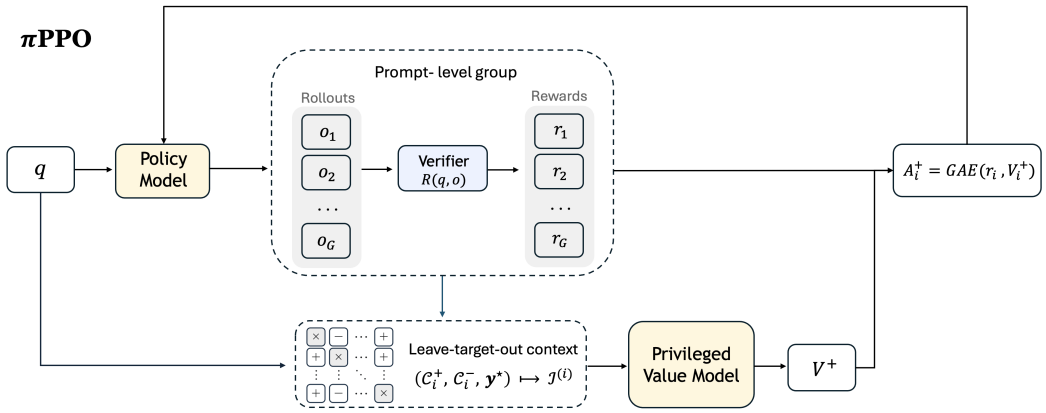


Figure 1: Overview of πPPO. For each prompt, the policy generates a group of rollouts, and the verifier assigns rewards according to their outcomes. The labeled rollouts are then reused to construct privileged context $\mathcal{I}^{(i)}$ for each target rollout, excluding the target itself. The privileged value model uses this context to estimate values along the target rollout. These estimates, together with the verifier rewards, are used to compute advantages for the PPO policy update.

3.2 πPPO

Motivated by the analysis above, we propose πPPO (Figure 1), a critic-side extension of PPO that reuses verified same-prompt rollouts as contrastive evidence for value estimation. By constructing a leave-target-out context pool via an adaptation of PPO to group rollouts, we enable each trajectory to serve both as a training target to be evaluated and as a potential labeled reference for its siblings, allowing the critic to assess a prefix in relation to successful and failed attempts at the same problem.

Leave-target-out contrastive context. For a prompt $\boldsymbol{x}$, the old policy samples G rollouts $\boldsymbol{y}^{(j)} \sim \pi_{\theta_{\text{old}}}(\cdot \mid \boldsymbol{x})$, which the verifier labels. The context of a target $\boldsymbol{y}^{(i)}$ is drawn from the remaining rollouts, partitioned by label,

$$\mathcal{C}_i^\sigma = \{\boldsymbol{y}^{(j)} : j \neq i,\ r(\boldsymbol{x}, \boldsymbol{y}^{(j)}) = \mathbf{1}_{\{\sigma=+\}}\}, \qquad \sigma \in \{+,-\}. \tag{3}$$

To use trajectories as evidence for prefix-level prediction, we separate the target being evaluated from the trajectories supplying its context. Let $\mathcal{S}_k(\mathcal{C})$ denote k reference rollouts randomly sampled from $\mathcal{C}$, together with their verifier labels. Writing $\|$ for concatenation and $\boldsymbol{y}^\star$ for the supplied ground-truth answer to $\boldsymbol{x}$, we construct

$$\mathcal{I}^{(i)} = \begin{cases} \mathcal{S}_1(\mathcal{C}_i^+) \| \mathcal{S}_1(\mathcal{C}_i^-), & \mathcal{C}_i^+ \neq \varnothing,\ \mathcal{C}_i^- \neq \varnothing, \\ \mathcal{S}_2(\mathcal{C}_i^+), & \mathcal{C}_i^- = \varnothing, \\ \mathcal{S}_2(\mathcal{C}_i^-) \| [\boldsymbol{y}^\star], & \mathcal{C}_i^+ = \varnothing. \end{cases} \tag{4}$$

When only one outcome is represented among the remaining rollouts, we use two siblings of that outcome; the ground-truth answer is additionally supplied only in the all-negative case and is identified separately from the sampled reference attempts.

Privileged advantage estimation. By constructing the self-elicited privileged information above, we obtain a critic that no longer predicts from the prefix alone but performs a contextual forward pass $V_\phi^+(\boldsymbol{x}, \mathcal{I}^{(i)}, \boldsymbol{y}_{<t}^{(i)})$, while the actor stays conditioned on $(\boldsymbol{x}, \boldsymbol{y}_{<t}^{(i)})$ alone. These context-conditioned values enter the standard TD residuals and GAE:

$$\begin{aligned} \delta_t^{+(i)} &= r_t^{(i)} + \gamma V_\phi^+(s_{t+1}^{(i)}, \mathcal{I}^{(i)}) - V_\phi^+(s_t^{(i)}, \mathcal{I}^{(i)}), \\ \widehat{A}_t^{+(i)} &= \sum_{\ell=0}^{T_i - t} (\gamma\lambda)^\ell \delta_{t+\ell}^{+(i)}. \end{aligned} \tag{5}$$

Here $T_i = |\boldsymbol{y}^{(i)}|$, and the terminal value is set to zero. The actor is then updated by substituting $\widehat{A}_t^{+(i)}$ for the advantage in the PPO clipped objective, while retaining its original conditioning.

Notably, the πPPO design yields two favorable properties:

- **Valid policy gradient baseline under privileged value estimation.** Let $s_t = (\boldsymbol{x}, \boldsymbol{y}_{<t})$, and let the privileged information $\mathcal{I}$ be available only to the critic. Because $V_\phi^+(s_t, \mathcal{I})$ is independent of the current action a_t, it remains a valid baseline:

$$\mathbb{E}_{a_t \sim \pi_\theta(\cdot|s_t)} \left[\nabla_\theta \log \pi_\theta(a_t \mid s_t) V_\phi^+(s_t, \mathcal{I}) \right] = V_\phi^+(s_t, \mathcal{I}) \nabla_\theta \sum_{a_t} \pi_\theta(a_t \mid s_t) = 0. \quad (6)$$

  Thus, privileged value estimation introduces no additional bias into the expected actor gradient and leaves the underlying policy objective unchanged.

- **Training–inference policy consistency.** The privileged context enters only the critic, so the actor uses the same conditioning at both training and inference:

$$\pi_\theta^{\text{train}}(a_t \mid s_t) = \pi_\theta^{\text{infer}}(a_t \mid s_t) = \pi_\theta(a_t \mid s_t), \qquad s_t = (\boldsymbol{x}, \boldsymbol{y}_{<t}). \quad (7)$$

Overall, πPPO incorporates verified contrastive rollout evidence into policy optimization through context-conditioned value estimates and resulting advantages. Since the rollout group collected for each PPO update supplies the critic context, the evidence is refreshed on the fly as the policy evolves.

4 Experiments

4.1 Experimental Settings

In this section, we provide a brief overview of the key experimental setup, including training procedures, baselines, and evaluation details. Additional information can be found in Appendix C.

Training details. We primarily conduct experiments on Qwen3-4B and Qwen3-8B (Yang et al., 2025) with thinking mode off, using DAPO-17K (Yu et al., 2026). We build on the VeRL codebase (Sheng et al., 2024) and follow the standard PPO recipe with clip-higher enabled. All methods use the same hyperparameters (batchsize 256, actor learning rate 1e-6) and rollout settings (temperature 1.0, top-p 1.0, top-k disabled, 8 rollouts per prompt). Following Yuan et al. (2025), we adopted 50 steps pretrain for the value model warmup as adjustment for RLVR tasks for all critic-based methods, further implementation details can be found in the Appendix C.1.

Baselines. We compare πPPO against representative RL baselines. We use vanilla PPO (Schulman et al., 2017) as the standard actor–critic baseline and include VAPO (Yue et al., 2025) as an RLVR-adapted improvement over PPO. We further compare against GRPO (Shao et al., 2024) and DAPO (Yu et al., 2026), a widely adopted GRPO variant with competitive performance and robust training on RLVR tasks. Together, these experiments assess the benefit of augmenting the PPO critic with verified sibling context and compare πPPO with RLVR-adapted actor–critic and strong value-free baselines on reasoning tasks. We additionally examine the training behavior of the self-distillation family in Appendix E.2.

Benchmarks and evaluation metrics. We evaluate our method on challenging in-domain math benchmarks: AIME 2024 (AI-MO, 2024), AIME 2025 (OpenCompass, 2025), BeyondAIME (ByteDance-Seed, 2025), HMMT25 (February) (Balunović et al., 2025), and HMMT25 (November) (Balunović et al., 2025). Following Guo et al. (2025a), we generate 8–16 outputs per problem, depending on test-set size, and report $Pass@1$ accuracy. All methods are evaluated using identical prompts and decoding configurations; details are provided in Appendix C.2. We further assess out-of-distribution generalization, with results reported in Appendix E.1.

For assessing the critic, we mainly focus on explained variance (EV) and value-function loss. Let $\widehat{G}_{i,t}$ and $\widehat{V}_{i,t}$ denote the value-regression target and critic prediction at token position t in rollout i, respectively. Over the valid response-token positions $\mathcal{M}$, we compute

$$\text{EV} = 1 - \frac{\text{Var}_{(i,t)\in\mathcal{M}} \left(\widehat{G}_{i,t} - \widehat{V}_{i,t} \right)}{\text{Var}_{(i,t)\in\mathcal{M}} \left(\widehat{G}_{i,t} \right) + \epsilon}, \qquad \mathcal{L}_V = \frac{1}{|\mathcal{M}|} \sum_{(i,t)\in\mathcal{M}} \left(\widehat{G}_{i,t} - \widehat{V}_{i,t} \right)^2. \quad (8)$$

Together, EV and $\mathcal{L}_V$ provide complementary views of across-sample target-variance capture and pointwise regression-target fitting, respectively.

Table 1: Mathematical reasoning accuracy (%). "Avg@k" denotes mean accuracy over k random generations (i.e., pass@1); **Overall** is the mean over the five benchmarks. Best result within each model size is in bold.

Method	Mathematical Reasoning						Tokens
	AIME 24 Avg@16	AIME 25 Avg@16	Beyond AIME Avg@8	HMMT_{Feb} Avg@16	HMMT_{Nov} Avg@16	Overall Avg.	Gen./ Train 10^9
(a) Qwen3-4B							
Initial policy	21.7	18.9	10.8	12.3	7.3	14.2	–
Critic-free							
GRPO	60.4	54.8	34.5	34.6	41.5	45.2	2.33 / 2.38
DAPO	63.1	56.7	34.5	34.6	41.9	46.2	4.42 / 2.64
Actor–critic							
PPO	57.9	54.2	32.4	32.7	40.4	43.5	2.42 / 2.48
VAPO	63.3	52.1	33.3	33.8	43.9	45.3	2.07 / 2.14
πPPO (ours)	**66.7**	**61.5**	**39.0**	**37.5**	**46.9**	**50.3**	2.53 / 2.59
(b) Qwen3-8B							
Initial policy	25.0	20.2	13.1	11.5	10.0	16.0	–
Critic-free							
GRPO	69.4	58.3	36.1	32.5	45.2	48.3	2.25 / 2.30
DAPO	71.3	57.1	37.8	33.3	46.9	49.3	4.34 / 2.59
Actor–critic							
PPO	68.1	55.2	36.3	32.3	47.7	47.9	2.35 / 2.42
VAPO	67.3	55.8	34.3	33.5	47.1	47.6	1.95 / 2.02
πPPO (ours)	**73.5**	**61.5**	**40.0**	**34.4**	**48.5**	**51.6**	2.57 / 2.63

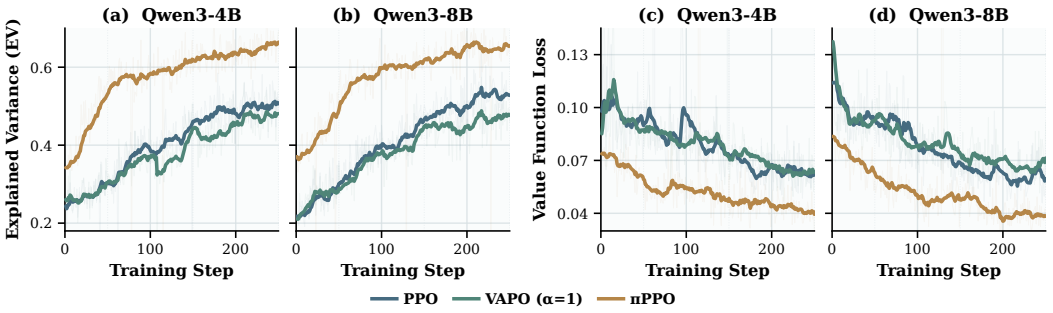


Figure 2: Critic explained variance (EV; a,b) and value-function loss (c,d) during training. Across Qwen3-4B and Qwen3-8B, πPPO consistently achieves higher EV and lower value-function loss than PPO and VAPO, with improvements evident from the early training steps.

4.2 Main Results

❶ πPPO consistently improves end-to-end policy performance. As shown in Table 1, πPPO achieves the highest accuracy on all five mathematical reasoning benchmarks at both model scales. Its overall scores reach 50.3% on Qwen3-4B and 51.6% on Qwen3-8B, exceeding the strongest baseline, DAPO, by 4.1 and 2.3 percentage points, respectively. Compared with vanilla PPO, the gains are 6.8 and 3.7 points. These results show that privileged value estimation translates into consistent policy gains over both actor–critic and critic-free baselines.

❷ Privileged context substantially improves value estimation. Figure 2 compares the critic explained variance and value-function loss throughout training. With either backbone model, the two metrics reveal a consistent pattern: πPPO starts with a higher explained variance upon entering training, improves sharply during the early updates, and subsequently remains at a high level, while its value-function loss quickly decreases and stays consistently lower. In contrast, the state-only critics of PPO and VAPO improve more gradually and plateau at weaker levels, providing less informative value estimates throughout training.

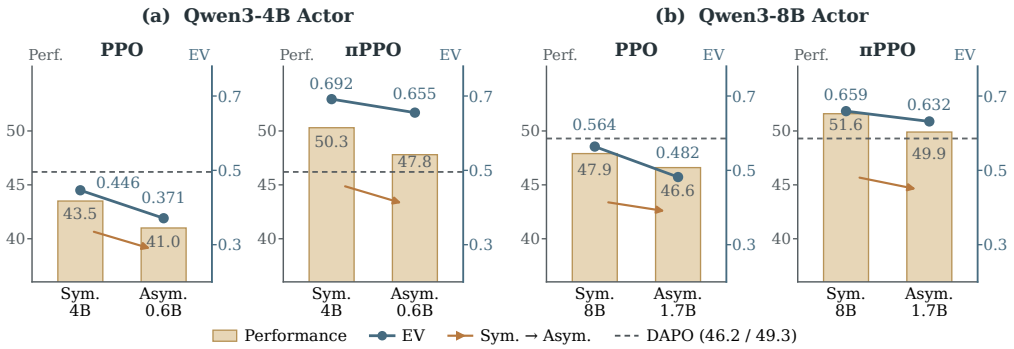


Figure 3: Performance (bars, left axis) and explained variance (EV; curves, right axis) of PPO and πPPO. Sym. and Asym. denote critics matching the actor size and smaller critics, respectively. Dashed lines indicate DAPO, the best-performing baseline.

4.3 πPPO ADMITS SMALLER ASYMMETRIC CRITICS

Another commitment of standard actor–critic implementations is to instantiate the actor and critic with the same model backbone. We next ask whether the value estimation advantage in πPPO can be traded for critic capacity. To test this hypothesis, we pair Qwen3-4B and Qwen3-8B actors with their corresponding 0.6B and 1.7B critics from the same family. For both PPO and πPPO, we compare these asymmetric configurations against their actor-sized symmetric counterparts.

❶ **πPPO achieves strong overall performance with substantially smaller critics.** As shown in Figure 3, πPPO achieves overall accuracies of 47.8% and 49.9% under asymmetric actor–critic configurations, still exceeding the strongest baseline on both backbones. Notably, it also outperforms actor–critic baselines with substantially larger, actor-sized critics, showing that privileged evidence can unlock stronger policy optimization with compact critics.

❷ **πPPO retains its value estimation advantage at a fraction of the critic size.** Reducing the critics to 0.6B and 1.7B cuts their parameter counts by approximately 85% and 79%, yet EV decreases by only 0.037 and 0.027. These declines are smaller than in PPO, and the resulting EVs of 0.655 and 0.632 still exceed the 0.446 and 0.564 achieved by PPO with actor-sized critics. Overall accuracy decreases by only 2.5 and 1.7 percentage points relative to the symmetric configurations.

Taken together, these results suggest that privileged value estimation offers a rationale for revisiting actor–critic scale matching: the improved estimation quality motivates asymmetric designs that retain strong policy performance with substantially smaller critics.

5 ANALYSIS

5.1 CHARACTERIZING PRIVILEGED VALUE ESTIMATION

To investigate the performance gap between PPO and πPPO, we examine their critic predictions, which provide advantage estimates for an otherwise unchanged PPO update. All analyses use the training data. Since ground-truth state values are not directly observable, we construct an empirical reference by running 256 MC samples for each prefix state s_t and averaging the returns. We compare both critics against this MC reference to assess the accuracy of their value predictions.

Accuracy of value estimates. We first group the prefix states by their MC-estimated values and compare the distributions of predictions from the two critics in Figure 4(a). Accurate estimates should concentrate near the diagonal, whereas systematic deviations indicate over- or underestimation. PPO exhibits widely dispersed predictions, particularly intermediate value states, while πPPO generally produces tighter distributions that more closely follow the MC reference. Figure 4(b) further shows that πPPO achieves lower MSE at every evaluated prefix position. As generation progresses, PPO's error generally increases, whereas that of πPPO remains comparatively stable. At the final evaluated position, the MSE is approximately 0.05 for πPPO versus 0.15 for PPO.

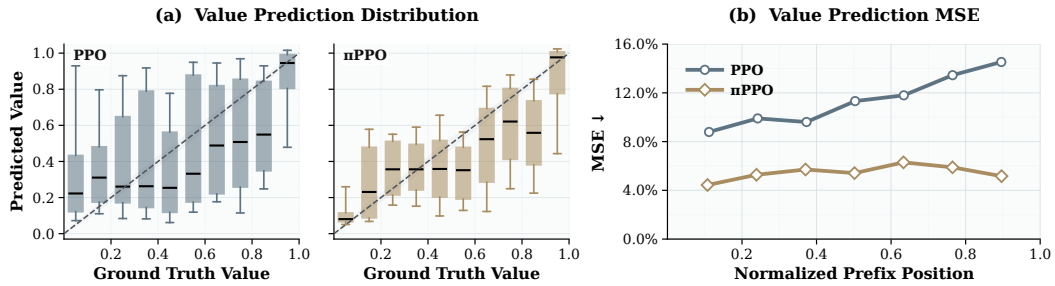


Figure 4: **(a)** Critic predictions grouped by MC reference value. Boxes show the interquartile range, black bars indicate the mean, and whiskers mark the 10th and 90th percentiles. The dashed diagonal denotes exact agreement with the MC reference. **(b)** Mean squared error against the MC reference across normalized prefix positions.

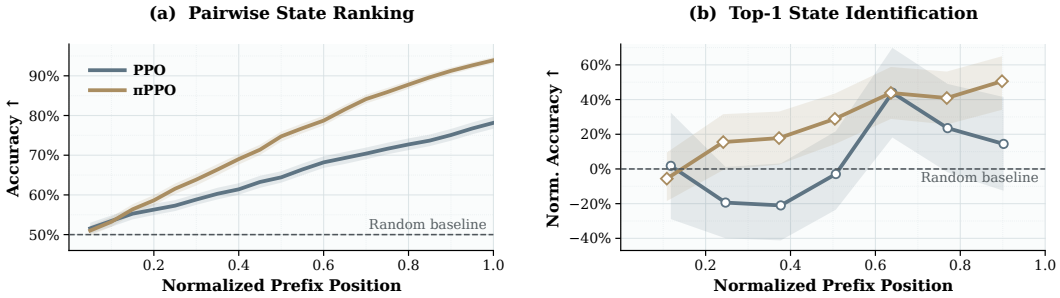


Figure 5: Critic ranking and state identification across normalized prefix positions. **(a)** Probability of assigning a higher value to a prefix from a successful trajectory than to one from an unsuccessful trajectory. **(b)** Chance-normalized top-1 accuracy for identifying the state with the highest MC reference value among prefixes sharing the same prompt and reasoning-node rank. Dashed lines indicate random-selection baselines.

State ranking and identification. Pairwise ranking measures how often a critic assigns a higher value to a prefix from a successful trajectory than to one from an unsuccessful trajectory. Figure 5(a) shows that πPPO's advantage widens as generation progresses, reaching approximately 94% pairwise accuracy compared with 78% for PPO at the final position. We further form candidate sets from different trajectories of the same prompt at each of seven reasoning-node ranks, and test whether the critic identifies the state with the highest MC reference value. πPPO achieves higher chance-normalized top-1 accuracy at most evaluated positions, reaching approximately 50% versus 15% for PPO at the final position (Figure 5(b)).

Reasoning progress decomposition. Beyond value accuracy and state ranking, we examine whether the critic tracks local value changes across reasoning functions using the ThinkARM taxonomy (Li et al., 2026). This is assessed by correlating stepwise changes in critic predictions with changes in MC reference values and breaking down the results by reasoning function. Figure 6 shows higher correlations for πPPO overall and within Analyze and Implement, which together account for approximately 90% of the evaluated nodes. PPO's correlations are near zero in both functions, while πPPO reaches approximately 0.2. These results suggest improved tracking of local value changes during the dominant intermediate reasoning activities.

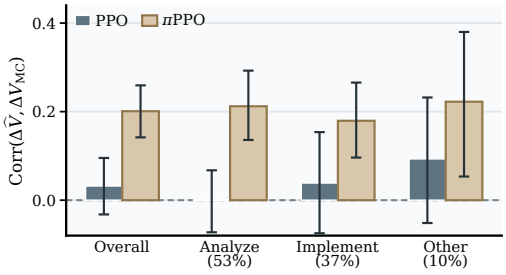


Figure 6: Pearson correlation between predicted and MC-estimated value changes. Error bars show trajectory-cluster 95% CIs; percentages indicate the shares of evaluated nodes.

5.2 THE ROLE OF CONTRASTIVE EVIDENCE

We test alternative forms of privileged information to identify which aspects contribute to the gains: removing correctness labels while retaining the selected references (*w/o labels*), replacing opposite-outcome pairs with same-outcome references (*w/ same polarity*), and using only the ground-truth final answer (*w/ GT only*). Removing labels or opposite-outcome pairing erases most gains over baselines; the offline GT-only variant trails full πPPO by 4.7 and 3.1 points on Qwen3-4B and Qwen3-8B, respectively. These results link the gains to the contrastive organization of online trajectory evidence.

Method	Overall	
	Qwen3-4B Avg.	Qwen3-8B Avg.
πPPO	**50.3**	**51.6**
w/o labels	44.1	48.8
w/ same polarity	43.6	48.6
w/ GT only	45.6	48.5

Table 2: Effects of correctness labels, reference polarity, and ground-truth-only context on average accuracy (%) across five benchmarks.

6 RELATED WORK

RL for LLM through the lens of value estimation. RL algorithms used for LLM post-training are predominantly policy-gradient methods, with REINFORCE (Williams, 1992) as their prototype. Prominent approaches in LLM post-training adopt either an actor–critic (e.g., PPO-based RLHF (Schulman et al., 2017; Ouyang et al., 2022), VC-PPO (Yuan et al., 2025), VAPO (Yue et al., 2025)) or critic-free formulation (e.g., bandit-like algorithms such as GRPO (Shao et al., 2024), DAPO (Yu et al., 2026), DR.GRPO (Liu et al., 2025)). Despite differing in their use of a learned critic, both formulations can be examined through the lens of value estimation. From this perspective, a central question underlying their relative effectiveness is whether value estimation is sufficiently reliable to improve the bias–variance trade-off in policy-gradient updates. πPPO addresses this question from the perspective of information available to the critic, introducing training-time privileged information to improve value estimation for policy optimization.

Privileged information for policy optimization. Privileged information (PI) was formalized in learning using privileged information (LUPI), where a teacher supplies additional information available only during training to improve learning and generalization (Vapnik & Vashist, 2009); how such information is incorporated, however, remains flexible. In LLM post-training, a direct approach is hint-based RL, adding PI from external experts into the input or supplied as a response prefix to guide sampling toward successful trajectories (Chen et al., 2026a; Qu et al., 2026), training in an off-policy way, also causing a distributional mismatch between training and inference time. Recent on-policy self-distillation uses PI to induce a stronger in-context distribution that supervises the policy on its own rollouts (Zhao et al., 2026; Shenfeld et al., 2026; Tang et al., 2026). This provides dense token-level supervision, but model-induced targets may not align with the original reward-maximization objective (Kaur et al., 2026; Nguyen et al., 2026). πPPO places PI within the optimization process through critic-side value estimation, improving the information available for policy updates while preserving the actor's conditioning and the environmental reward objective.

7 CONCLUSION

In this paper, we recast value estimation in RLVR as an information problem, showing that a critic scoring each state in isolation can be inherently limited by its formulation, and that the evidence beyond target state can lower this floor. Building on this view, we propose πPPO, which conditions the critic on a self-elicited, leave-target-out context of verified same-prompt siblings while preserving the actor's original conditioning and environment-defined reward objective, with no additional generation. Across mathematical reasoning benchmarks with Qwen3-4B and Qwen3-8B backbones, πPPO outperforms its state-only actor–critic counterparts in critic quality, consistently achieving higher explained variance and lower value loss, while delivering the strongest end-to-end accuracy among all evaluated actor–critic and critic-free baselines—even staying ahead when we further exploit capacity asymmetry with a smaller critic model. Overall, πPPO provides a complementary design view on developing more effective RL training algorithms for LLMs.

Reproducibility Statement

To ensure reproducibility, we provide a detailed description of our method in Section 3 and present all experimental setups in Section 4 and Appendix C, including training configurations and evaluation protocols. All evaluation datasets, the training set, the training framework, as well as the baselines used in our experiments, are publicly available and properly referenced.

References

AI-MO. Aime 2024. https://huggingface.co/datasets/AI-MO/aimo-validation-aime, 2024.

Brandon Amos. Tutorial on amortized optimization, 2025. URL https://arxiv.org/abs/2202.00665.

Mislav Balunović, Jasper Dekoninck, Ivo Petrov, Nikola Jovanović, and Martin Vechev. Matharena: Evaluating llms on uncontaminated math competitions, February 2025. URL https://matharena.ai/.

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

Justin Chen, Xiangyu Peng, Prafulla Kumar Choubey, Kung-Hsiang Huang, Jiaxin Zhang, Mohit Bansal, and Chien-Sheng Wu. Nudging the boundaries of llm reasoning. In *International Conference on Learning Representations*, 2026a.

Zizhe Chen, Jiqian Dong, Yizhou Tian, Garry Yang, Yongqiang Chen, Zhitang Chen, and James Cheng. Hista and numca: Estimate state value effectively for llm reinforcement learning, 2026b. URL https://arxiv.org/abs/2605.29782.

Evan Greensmith, Peter L Bartlett, and Jonathan Baxter. Variance reduction techniques for gradient estimates in reinforcement learning. *Journal of Machine Learning Research*, 5(Nov):1471–1530, 2004.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025a. doi: 10.1038/s41586-025-09422-z.

Yiran Guo, Lijie Xu, Jie Liu, Dan Ye, and Shuang Qiu. Segment policy optimization: Effective segment-level credit assignment in rl for large language models, 2025b. URL https://arxiv.org/abs/2505.23564.

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

Simran Kaur, Narutatsu Ri, Yinghui He, Liam Fowl, and Sanjeev Arora. Rethinking on-policy self-distillation for thinking models. *arXiv preprint arXiv:2607.05184*, 2026.

Amirhossein Kazemnejad, Milad Aghajohari, Eva Portelance, Alessandro Sordoni, Siva Reddy, Aaron Courville, and Nicolas Le Roux. Vineppo: Unlocking rl potential for llm reasoning through refined credit assignment. 2024.

Ming Li, Chenrui Fan, Yize Cheng, Soheil Feizi, and Tianyi Zhou. Schoenfeld's anatomy of mathematical reasoning by language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 32773–32802, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1513. URL https://aclanthology.org/2026.acl-long.1513/.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding r1-zero-like training: A critical perspective. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=5PAF7PAY2Y.

Xuan-Phi Nguyen, Shrey Pandit, Yiran Zhao, Anurag Koul, Zeyu Liu, and Shafiq Joty. Privileged likelihood is not automatically value: Three checks for token credit in on-policy self-distillation. *arXiv preprint arXiv:2608.09263*, 2026.

OpenCompass. Aime 2025. https://huggingface.co/datasets/opencompass/AIME2025, 2025.

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.

Leyi Pan, Shuchang Tao, Yunpeng Zhai, Lingzhe Zhang, Zhaoyang Liu, Bolin Ding, Aiwei Liu, and Lijie Wen. Rlcsd: Reinforcement learning with contrastive on-policy self-distillation. *arXiv preprint arXiv:2606.11709*, 2026.

Yuxiao Qu, Amrith Setlur, Virginia Smith, Ruslan Salakhutdinov, and Aviral Kumar. Pope: Learning to reason on hard problems via privileged on-policy exploration. *arXiv preprint arXiv:2601.18779*, 2026.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*, 2024.

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.

Zikang Shan, Han Zhong, Liwei Wang, and Li Zhao. Bringing value models back: Generative critics for value modeling in llm reinforcement learning, 2026. URL https://arxiv.org/abs/2604.10701.

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.

Idan Shenfeld, Mehul Damani, Jonas Hübotter, and Pulkit Agrawal. Self-distillation enables continual learning. *arXiv preprint arXiv:2601.19897*, 2026.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv: 2409.19256*, 2024.

Rui Shu. Amortized optimization. https://ruishu.io/2017/11/07/amortized-optimization/, 2017.

Chenming Tang, Hsiu-Yuan Huang, Weijie Liu, Clive Bai, Saiyong Yang, and Yunfang Wu. Do not step into the same river twice: Learning to reason from trial and error. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8018–8032, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.363. URL https://aclanthology.org/2026.acl-long.363/.

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.

Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural networks*, 22(5-6):544–557, 2009.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290, 2024.

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

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Chenxu Yang, Chuanyu Qin, Qingyi Si, Minghui Chen, Naibin Gu, Dingyu Yao, Zheng Lin, Weiping Wang, Jiaqi Wang, and Nan Duan. Self-distilled rlvr. *arXiv preprint arXiv:2604.03128*, 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, 2025. URL https://arxiv.org/abs/2503.01491.

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.

Yi-Kai Zhang, Zhiyuan Yao, Hongyan Hao, Yueqing Sun, Qi Gu, Hui Su, Xunliang Cai, De-Chuan Zhan, and Han-Jia Ye. v_0: A generalist value model for any policy at state zero, 2026. URL https://arxiv.org/abs/2602.03584.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. *arXiv preprint arXiv:2601.18734*, 2026.

A THE USE OF LARGE LANGUAGE MODELS

We used LLMs solely to improve the grammar and wording of the manuscript. Their use did not extend to research ideation or full-manuscript writing.

B THEORETICAL PROOF

B.1 VALUE ESTIMATION UNDER PRIVILEGED CONDITIONING

Proof of Proposition 1. All expectations are taken under the fixed joint distribution of $(s, \mathcal{I}, R)$ induced by the policy and the context construction. The binary return is square-integrable, and we consider only critics with finite squared error.

Step 1: Risk decomposition. For any fixed contextual critic, write

$$R - V_\phi^+(s,\mathcal{I}) = \big(R - V^{\pi,+}(s,\mathcal{I})\big) + \big(V^{\pi,+}(s,\mathcal{I}) - V_\phi^+(s,\mathcal{I})\big).$$

By the definition of conditional expectation,

$$\mathbb{E}[R - V^{\pi,+}(s,\mathcal{I}) \mid s,\mathcal{I}] = 0.$$

Since $V^{\pi,+}(s,\mathcal{I}) - V_\phi^+(s,\mathcal{I})$ is a function of $(s,\mathcal{I})$, the cross term vanishes:

$$\begin{aligned}&\mathbb{E}\Big[\big(R - V^{\pi,+}(s,\mathcal{I})\big)\big(V^{\pi,+}(s,\mathcal{I}) - V_\phi^+(s,\mathcal{I})\big)\Big] \\ &\quad = \mathbb{E}\Big[\big(V^{\pi,+}(s,\mathcal{I}) - V_\phi^+(s,\mathcal{I})\big)\mathbb{E}[R - V^{\pi,+}(s,\mathcal{I}) \mid s,\mathcal{I}]\Big] = 0.\end{aligned}$$

Squaring the decomposition and taking expectations therefore gives

$$\mathbb{E}[(R - V_\phi^+)^2] = \mathbb{E}_{s,\mathcal{I}}[\mathrm{Var}(R \mid s,\mathcal{I})] + \mathbb{E}[(V_\phi^+ - V^{\pi,+})^2]. \tag{9}$$

The variance term does not depend on ϕ, so it can be taken outside the infimum over the shared critic parameters. This proves the first equality in the proposition.

Similarly, conditioning on s gives

$$\mathbb{E}[(R - V_\phi(s))^2] = \mathbb{E}_s[\mathrm{Var}(R \mid s)] + \mathbb{E}[(V_\phi(s) - V^\pi(s))^2]. \tag{10}$$

Taking the infimum over ϕ establishes the final equality.

Step 2: Optimal risk comparison. We next compare the two optimal risks. By assumption, for every state-only parameter vector ϕ there exists a contextual parameter vector ψ (e.g., with the contribution of the additional context $\mathcal{I}$ effectively set to zero) such that

$$V_\psi^+(s,\mathcal{I}) = V_\phi(s) \qquad \text{almost surely.}$$

Consequently, for every ϕ,

$$\inf_\psi \mathbb{E}[(R - V_\psi^+(s,\mathcal{I}))^2] \le \mathbb{E}[(R - V_\phi(s))^2].$$

Taking the infimum on the right gives

$$\inf_\psi \mathbb{E}[(R - V_\psi^+(s,\mathcal{I}))^2] \le \inf_\phi \mathbb{E}[(R - V_\phi(s))^2]. \tag{11}$$

Taking infima in Eqs. (9) and (10), whose variance terms are independent of the critic parameters, and applying Eq. (11) proves the proposition. □

B.2 SUCCESS–FAILURE CONTRAST IN VALUE ESTIMATION

Proof of Proposition 2. All expectations are taken under the fixed joint distribution of (s, R) specified in Proposition 2. Since R is binary and the critic has finite squared error, all second moments below are finite.

Step 1: Outcome-conditioned risk decomposition. By the definition of Δ_ϕ and the binary nature of R,

$$\begin{aligned}\mathbb{E}[R - V_\phi(s) \mid R] &= (1 - \Delta_\phi)R - \mathbb{E}[V_\phi(s) \mid R = 0],\\ \mathrm{Var}(R - V_\phi(s) \mid R) &= \mathrm{Var}(V_\phi(s) \mid R).\end{aligned}$$

Applying the law of total variance to $R - V_\phi(s)$ therefore gives

$$\begin{aligned}\mathbb{E}[(R - V_\phi(s))^2] &= \big(\mathbb{E}[V_\phi(s)] - \mathbb{E}[R]\big)^2 + \mathrm{Var}(R - V_\phi(s))\\ &= \big(\mathbb{E}[V_\phi(s)] - \mathbb{E}[R]\big)^2 + \mathbb{E}_R[\mathrm{Var}(V_\phi(s) \mid R)]\\ &\quad + \mathrm{Var}(R)(1 - \Delta_\phi)^2.\end{aligned} \tag{12}$$

Step 2: Contrast lower bound. Let

$$\varepsilon_\phi = \big(\mathbb{E}[V_\phi(s)] - \mathbb{E}[R]\big)^2 + \mathbb{E}_R[\mathrm{Var}(V_\phi(s) \mid R)] \geq 0.$$

Equation (12) then yields

$$\begin{aligned}\mathbb{E}[(R - V_\phi(s))^2] &= \mathrm{Var}(R)(1 - \Delta_\phi)^2 + \varepsilon_\phi\\ &\geq \mathrm{Var}(R)(1 - \Delta_\phi)^2.\end{aligned} \tag{13}$$

For any constant predictor c,

$$\mathbb{E}[(R - c)^2] = \mathrm{Var}(R) + (c - \mathbb{E}[R])^2,$$

so the optimal constant prediction risk is $\mathrm{Var}(R)$, attained at $c = \mathbb{E}[R]$. If $\Delta_\phi \leq 0$, then $(1-\Delta_\phi)^2 \geq 1$, and Eq. (13) implies $\mathbb{E}[(R - V_\phi(s))^2] \geq \mathrm{Var}(R)$. Thus, any critic improving upon the optimal constant prediction risk must satisfy $\Delta_\phi > 0$. □

C Implementation Details

In this section, we provide the details of our main experiments in Section 4.

C.1 Training Details

We build our implementation on the VeRL framework (Sheng et al., 2024). All methods are trained for 250 steps with a batch size of 256, utilizing learning rates of 10^{-6} for the actor and 10^{-5} for the critic. Following DAPO and VAPO (Yu et al., 2026; Yue et al., 2025), we adopt Clip-Higher with $\varepsilon_{\text{low}} = 0.2$ and $\varepsilon_{\text{high}} = 0.28$, omitting KL regularization. For critic-based methods, we set $\gamma = \lambda = 1.0$ and adopt fixed-policy value pretraining with Monte Carlo targets following VC-PPO (Yuan et al., 2025). Specifically, we perform 50 pretraining steps with the actor frozen, linearly warming up the critic learning rate from 10^{-7} to 10^{-5}.

During policy training, we generate $G = 8$ responses per prompt with temperature 1.0, top-p 1.0, and top-k disabled. For πPPO, references are selected randomly subject to the leave-target-out construction in Eq. (4). The maximum response length and the privileged-context truncation length are both set to 8192 tokens per sample. During value pretraining, we sample one response per prompt online, and set the batchsize to 2048 instead; πPPO obtains its reference context from a cache of the most recent verified rollouts from previous epochs. The remaining pretraining settings are shared with the standard critic. Algorithm 1 summarizes the overall procedure.

Baselines. All baselines share the common training, rollout, and evaluation settings described above. For a fair comparison, all methods use Clip-Higher and generate $G = 8$ responses per prompt. For DAPO (Yu et al., 2026), we implement dynamic sampling with a rejection pool size of 512, twice the training batch size. For VAPO (Yue et al., 2025), we set the length-adaptive GAE coefficient α to 1.0 (functioning only on the actor side).

Context ablations. All ablations modify the critic's privileged context and otherwise follow the common settings. For *w/o labels*, we preserve the selected reference pair, replace correctness headers with ordinal labels, and shuffle the reference order. For *w/ same polarity*, we supply two correct or two incorrect sibling references with their correctness labels, otherwise falling back to a contrastive pair. For *w/ GT only*, we remove all rollout references and provide only the verified ground-truth final answer.

Algorithm 1: πPPO: Privileged-Information PPO

Input: Actor π_θ; warmed-up privileged critic V_ϕ^+; prompt dataset $\mathcal{D}$ with ground-truth answers; verifier R; group size G; discount γ; GAE parameter λ

for *each training iteration* **do**
- $\theta_{\text{old}} \leftarrow \theta$
- Sample a prompt batch $\mathcal{B} \subset \mathcal{D}$
- **foreach** *prompt* $\boldsymbol{x} \in \mathcal{B}$ **do**
  - Retrieve its ground-truth answer $\boldsymbol{y}^\star$
  - `// Collect and verify same-prompt rollouts`
  - Independently sample $\{\boldsymbol{y}^{(i)}\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}(\cdot \mid \boldsymbol{x})$
  - Evaluate $R(\boldsymbol{x}, \boldsymbol{y}^{(i)}) \in \{0, 1\}$ and construct token rewards $\{r_t^{(i)}\}_{t=1}^{T_i}$, where $T_i = |\boldsymbol{y}^{(i)}|$
  - **for** $i = 1, \ldots, G$ **do**
    - `// Exclude the target before selecting references`
    - $\mathcal{C}_i^+ \leftarrow \{\boldsymbol{y}^{(j)} : j \neq i,\ R(\boldsymbol{x}, \boldsymbol{y}^{(j)}) = 1\}$
    - $\mathcal{C}_i^- \leftarrow \{\boldsymbol{y}^{(j)} : j \neq i,\ R(\boldsymbol{x}, \boldsymbol{y}^{(j)}) = 0\}$
    - **if** $\mathcal{C}_i^+ \neq \varnothing$ and $\mathcal{C}_i^- \neq \varnothing$ **then**
      - $\mathcal{I}^{(i)} \leftarrow \mathcal{S}_1(\mathcal{C}_i^+) \| \mathcal{S}_1(\mathcal{C}_i^-)$
    - **else if** $\mathcal{C}_i^- = \varnothing$ **then**
      - $\mathcal{I}^{(i)} \leftarrow \mathcal{S}_2(\mathcal{C}_i^+)$
    - **else**
      - $\mathcal{I}^{(i)} \leftarrow \mathcal{S}_2(\mathcal{C}_i^-) \| [\boldsymbol{y}^\star]$
    - `// Estimate values under a fixed context`
    - $s_t^{(i)} \leftarrow (\boldsymbol{x}, \boldsymbol{y}_{<t}^{(i)})$
    - Compute and store $\widehat{V}_{i,t} \leftarrow V_\phi^+(s_t^{(i)}, \mathcal{I}^{(i)})$, with zero value at terminal states
    - **for** $t = T_i, \ldots, 1$ **do**
      - $\delta_t^{+(i)} \leftarrow r_t^{(i)} + \gamma \widehat{V}_{i,t+1} - \widehat{V}_{i,t}$
      - $\widehat{A}_t^{+(i)} \leftarrow \delta_t^{+(i)} + \gamma\lambda \widehat{A}_{t+1}^{+(i)}$, with $\widehat{A}_{T_i+1}^{+(i)} = 0$
    - Compute and store value-regression targets $\{\widehat{G}_{i,t}\}_{t=1}^{T_i}$ using the PPO return-target rule
- **for** *each PPO optimization epoch* **do**
  - Update ϕ by value regression on $(s_t^{(i)}, \mathcal{I}^{(i)}, \widehat{G}_{i,t})$
  - Update θ using the clipped PPO objective with $\widehat{A}_t^{+(i)}$ and $\rho_t^{(i)}(\theta) = \dfrac{\pi_\theta(y_t^{(i)} \mid \boldsymbol{x}, \boldsymbol{y}_{<t}^{(i)})}{\pi_{\theta_{\text{old}}}(y_t^{(i)} \mid \boldsymbol{x}, \boldsymbol{y}_{<t}^{(i)})}$

C.2 Evaluation and Analysis Details

We use the same vanilla evaluation prompts and decoding settings across methods for each backbone. Responses are sampled with temperature 0.6, top-p 0.95, top-k 20, and a maximum length of 32768 tokens. We generate 16 responses per problem on AIME 2024, AIME 2025, HMMT February, and HMMT November, and 8 responses per problem on BeyondAIME. We report the average correctness across sampled responses as an estimate of pass@1, with Overall computed as the mean over the five benchmarks.

Critic diagnostics. For critic diagnostics, we evaluate step-250 Qwen3-8B checkpoints on 128 prompts per method and estimate values from 256 MC continuations at seven semantic prefixes, with completed responses capped at 8192 tokens. Pairwise ranking uses 20 prefix positions; top-1 identification retains candidate sets whose MC-value range exceeds twice the estimated uncertainty.

D PROMPT TEMPLATE FOR πPPO

Prompt Template used in πPPO

<|im_start|>user

/ [shared source: original user turn → actor, V_ϕ, V_ϕ^+] */*

Let x and y be real numbers with $x + y = 25$ and $xy = 126$. Find $|x - y|$.
Please reason step by step, and put your final answer within `\boxed{}`.

> */* [critic-only source: labeled reference attempts → V_ϕ^+ only] */*
>
> Below are reference attempts at the problem. Each one is explicitly labeled with whether it reached the correct final answer. They are backgrounded and are NOT part of the response.
>
> **[Reference attempt — CORRECT]:**
>
> (1) The sum and the product are given, so let $s = x + y = 25$ and $p = xy = 126$.
> (2) I want $|x - y|$, and the clean route is to obtain $(x - y)^2$ from s and p.
> (3) Since $(x-y)^2 = x^2 - 2xy + y^2$ and $(x+y)^2 = x^2 + 2xy + y^2$, subtracting gives $(x-y)^2 = (x+y)^2 - 4xy$.
> (4) Substituting, $(x - y)^2 = 625 - 4 \cdot 126 = 625 - 504 = 121$.
> (5) Therefore $|x - y| = \sqrt{121} = 11$.
> (6) Check: the numbers solve $t^2 - 25t + 126 = 0$, so they are 18 and 7; their sum is 25 and their product is 126, both as required.
> (7) The final answer is $\boxed{11}$.
>
> **[Reference attempt — INCORRECT]:**
>
> (1) The sum and the product are given, so let $s = x + y = 25$ and $p = xy = 126$.
> (2) I want $|x - y|$, and the clean route is to obtain $(x - y)^2$ from s and p.
> (3) Expanding the square of the difference, $(x - y)^2 = (x + y)^2 - 2xy$.
> (4) Substituting, $(x - y)^2 = 625 - 2 \cdot 126 = 625 - 252 = 373$.
> (5) Therefore $|x - y| = \sqrt{373} \approx 19.31$.
> (6) Check: the numbers are then about 22.16 and 2.84, and $22.16 + 2.84 = 25$, so the sum comes out right.
> (7) The final answer is $\boxed{\sqrt{373}}$.

<|im_end|>
<|im_start|>assistant

/ [shared source: actor-sampled completion → V_ϕ, V_ϕ^+] */*

(1) Let x and y be the two numbers, so $x + y = 25$ and $xy = 126$.
(2) Then x and y are the roots of $t^2 - 25t + 126 = 0$.
(3) Its discriminant is $25^2 - 4 \cdot 126 = 625 - 504 = 121$.
(4) Hence $t = \frac{25 \pm \sqrt{121}}{2} = \frac{25 \pm 11}{2}$, giving $t = 18$ and $t = 7$.
(5) So $|x - y| = 18 - 7 = 11$.
(6) The final answer is $\boxed{11}$.

<|im_end|>

Table 3: The actor and the state-only critic V_ϕ share the original user turn and actor-generated completion, whereas the privileged critic V_ϕ^+ receives the same completion with a labeled block of additional information appended to the user turn. The two reference attempts in this block are numbered step by step and agree verbatim through step (2), but diverge at step (3) in their expansion of $(x - y)^2$, using $-4xy$ and $-2xy$, respectively, forming a direct process-level contrast.

Table 4: Out-of-domain accuracy (%) of policies trained on DAPO-17K. "Avg@k" denotes mean accuracy over k random generations (i.e., pass@1); **Overall** is the mean over the two benchmarks. Asym: the method above with a smaller critic (0.6B for 4B, 1.7B for 8B). Best result within each model size is in bold.

Method	Qwen3-4B			Qwen3-8B		
	GPQA Avg@8	**MMLU** Avg@1	**Overall** Avg.	**GPQA** Avg@8	**MMLU** Avg@1	**Overall** Avg.
Initial policy	45.7	60.8	53.3	51.0	66.1	58.6
Critic-free						
GRPO	**52.9**	67.3	60.1	58.3	71.6	65.0
DAPO	52.8	67.7	60.3	56.0	**72.0**	64.0
Actor–critic						
PPO	51.5	66.3	58.9	56.3	71.3	63.8
↳ Asym	52.7	66.8	59.8	57.9	71.4	64.7
VAPO	51.4	66.0	58.7	56.9	69.7	63.3
πPPO (ours)	52.8	**68.2**	**60.5**	57.7	71.7	64.7
↳ Asym	52.6	66.6	59.6	**58.8**	71.5	**65.2**

Table 5: Comparison with RLSD on mathematical reasoning benchmarks. **Overall** averages the five benchmark scores; bold denotes the best result within each model size.

Method	Mathematical Reasoning					
	AIME 24 Avg@16	**AIME 25** Avg@16	**Beyond AIME** Avg@8	**HMMT$_{\text{Feb}}$** Avg@16	**HMMT$_{\text{Nov}}$** Avg@16	**Overall** Avg.
(a) Qwen3-4B						
Initial policy	21.7	18.9	10.8	12.3	7.3	14.2
RLSD	57.1	47.9	31.5	31.5	38.8	41.4
πPPO (ours)	**66.7**	**61.5**	**39.0**	**37.5**	**46.9**	**50.3**
(b) Qwen3-8B						
Initial policy	25.0	20.2	13.1	11.5	10.0	16.0
RLSD	67.5	50.4	35.3	28.3	38.8	44.1
πPPO (ours)	**73.5**	**61.5**	**40.0**	**34.4**	**48.5**	**51.6**

E Additional Results

E.1 Evaluation on OOD benchmarks

We additionally evaluated our method on representative multi-task benchmarks spanning diverse domains in science, humanities, and professional knowledge: GPQA-Diamond (Rein et al., 2024) and MMLU-Pro (Wang et al., 2024), for examining the out-of-distribution generalization.

As shown in Table 4, all RL-trained policies outperform their initial counterparts. At the 4B scale, πPPO achieves the best Overall score of 60.5. At the 8B scale, πPPO reaches 64.7, its asymmetric variant achieves the best 8B Overall score of 65.2 using only a 1.7B critic. These results suggest that the in-domain advantages of πPPO may carry over to other related domains.

E.2 Additional Comparison with self-distillation methods

We additionally compare with OPSD (Zhao et al., 2026) and its RL-integrated variants, RLSD (Yang et al., 2026) and RLCSD (Pan et al., 2026), since these methods can be viewed in the same general way as ours: leveraging privileged information to provide additional supervision or shape token-level rewards.

We follow the original implementations and adapt these methods to the training setup used in our main experiments. The reference configuration uses a batch size of 256, an actor learning rate of 10^{-6}, and a budget of 250 policy updates. We sample $G = 8$ responses per prompt with temperature 1.0, top-p 1.0, disabled top-k filtering, and a maximum response length of 8192 tokens. We explore multiple configurations around this setup to adapt the methods for better training stability and effectiveness.

Across the tested configurations, we consistently observe training oscillations and collapse for OPSD and RLCSD, with no effective gain observed compared to the initial model. Among these three baselines, only RLSD achieves reasonable downstream performance under our experimental setting, which follows their original setups.