
VinePPO: Refining Credit Assignment in RL Training of LLMs

Amirhossein Kazemnejad^{*1} Milad Aghajohari^{*1} Eva Portelance^{1,6} Alessandro Sordoni^{1,2} Siva Reddy^{1,3,4}
Aaron Courville^{†,1,4,5} Nicolas Le Roux^{†,1,4}

Abstract

Large language models (LLMs) are increasingly applied to complex reasoning tasks that require executing several complex steps before receiving any reward. Properly assigning credit to these steps is essential for enhancing model performance. Proximal Policy Optimization (PPO), a common reinforcement learning (RL) algorithm used for LLM finetuning, employs value networks to tackle credit assignment. However, recent approaches achieve strong results without it, raising questions about the efficacy of value networks in practice. In this work, we systematically evaluate the efficacy of value networks and reveal their significant shortcomings in reasoning-heavy LLM tasks, showing that they often produce poor estimate of expected return and barely outperform a random baseline when comparing alternative steps. This motivates our key question: Can improved credit assignment enhance RL training for LLMs? To address this, we propose VinePPO, a straightforward approach that leverages the flexibility of language environments to compute unbiased Monte Carlo-based estimates. Our method consistently outperforms PPO and other baselines across MATH and GSM8K datasets in less wall-clock time (up to 3.0x). Crucially, it achieves higher test accuracy for a given training accuracy, capturing more generalization signal per sample. These results emphasize the importance of accurate credit assignment in RL training of LLM.

Code available at <https://github.com/McGill-NLP/VinePPO>

^{*}Equal contribution [†]Equal advising ¹Mila ²Microsoft Research
³McGill University ⁴Canada CIFAR AI Chair ⁵Université de Montréal ⁶HEC Montréal. Correspondence to: Amirhossein Kazemnejad <amirhossein.kazemnejad@mila.quebec>, Milad Aghajohari <aghajohn@mila.quebec>.

Proceedings of the 42nd International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

1. Introduction

Reinforcement learning (RL) has become instrumental in training large language models (LLMs) to solve complex reasoning tasks such as mathematical problem solving (DeepSeek-AI et al., 2025), web navigation (Putta et al., 2024), or code generation (OpenAI, 2024). In these settings, LLMs often engage in extended reasoning steps, executing multiple actions to arrive at a solution. However, not all steps are equally impactful—some contribute significantly, while others are irrelevant or detrimental. For example, in Figure 1.a, only step s_2 provides a key insight. Indeed, most reasoning steps generated by a model do not affect the chance of it solving the problem (Figure 1.b). Identifying the contribution of each action is crucial for improving model performance. However, this is inherently difficult due to the significant delay between actions and their eventual effect. This issue, known as the *Credit Assignment (CA)* problem, is a core challenge in RL (Sutton & Barto, 1998).

Proximal Policy Optimization (PPO; Schulman et al. 2017) addresses credit assignment through a value network (or critic), a mechanism retained in its application to RL-based finetuning of LLMs (Ouyang et al., 2022). This network, typically a separate model initialized from a pretrained checkpoint, is trained during PPO finetuning to estimate the expected cumulative rewards (or value) of an intermediate action. In Figure 1.b, an ideal value network would assign high value to step s_2 and subsequent steps, where the model has a high chance of successfully solving the problem. PPO uses these value estimates to measure the *advantage* of each action and update the model accordingly.

However, recent approaches such as DPO (Rafailov et al., 2023) or GRPO (Shao et al., 2024) simplify PPO’s design by discarding fine-grained credit assignment and treating all tokens equally. Despite such simplifications, they often demonstrate strong performance (Xu et al., 2024; Chang et al., 2023). This challenges classic RL principles, where accurate CA is considered critical for optimal performance (Sutton & Barto, 1998; Greensmith et al., 2001), especially in tasks with delayed rewards. In this work, we address this apparent discrepancy by showing that PPO’s credit assignment mechanism, the value network, performs poorly in practice. Our systematic evaluation (Section 7) on tasks

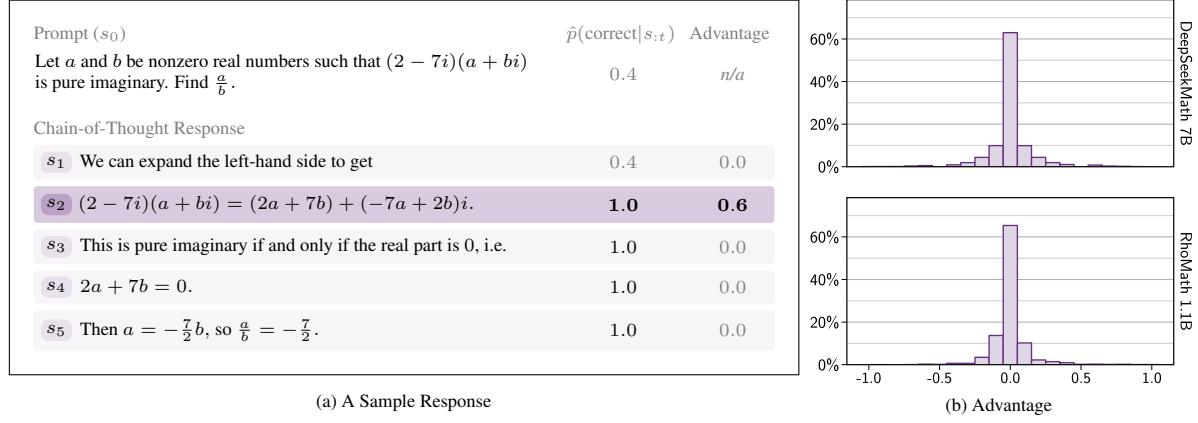


Figure 1: **(Left)** A response generated by the model. The notation $\hat{p}(\text{correct}|s_{:t})$ represents the estimated probability of successfully solving the problem at step t . Here, only step s_2 is critical; after this, the model always completes the solution correctly. **(Right)** The distribution of advantages, defined as $\hat{p}(\text{correct}|s_{:t+1}) - \hat{p}(\text{correct}|s_{:t})$, collected over a subset of MATH dataset (Hendrycks et al., 2021). Most steps show little or no advantage over the preceding step.

requiring chain-of-thought reasoning reveals that the value network often provides imprecise estimates and fails to differentiate between promising and unproductive steps, which could explain why simplified approaches achieve comparable results without explicit fine-grained CA.

These findings motivate a central question: If we improve credit assignment in PPO rather than discarding it, *how much can we enhance the RL training of LLMs?* To explore this, we propose VinePPO (Figure 2), which computes *unbiased* value estimates of the intermediate states with Monte Carlo (MC) estimation instead of employing value networks. Our key insight is that language-based environments allow us to reset directly to any intermediate state simply by refeeding the partial context, enabling efficient MC rollouts without the massive overhead usually seen in generic RL environments. VinePPO preserves PPO’s overall framework but addresses the CA challenge fundamentally.

We empirically evaluate the effectiveness and computational efficiency of MC value estimation in VinePPO. Across multiple mathematical reasoning tasks, VinePPO consistently outperforms PPO and other credit assignment-free baselines. While its per-iteration runtime is generally slower due to MC sampling, VinePPO surpasses the peak performance of baselines with fewer gradient steps and ultimately less wall-clock time. Importantly, VinePPO achieves higher test accuracy for a given training accuracy, capturing more generalization signal per fitted training sample. This is critical, as genuinely challenging verifiable reasoning tasks are scarce. These results underscore the importance of CA in RL-training of LLMs and highlight VinePPO as a straightforward alternative to value network-based approaches.

Our contributions are summarized as follows:

- We analyze PPO’s value network in reasoning tasks and find it often misestimates intermediate values, barely outperforming a random chance in ranking candidate steps.
- We propose VinePPO, leveraging the flexibility of language environments to compute unbiased, MC-based value estimates without relying on a separate critic.
- We empirically highlights the benefits of refined CA. VinePPO achieves the peak performance of baselines with less wall-clock time (up to 3.0x), better KL-divergence trade-off while exhibiting better generalization slope.

2. Related Work

Credit Assignment in Post-Training of LLM PPO, as applied in RL from Human Feedback (RLHF, Ouyang et al. 2022), pioneered RL finetuning of LLMs. However, its computational overhead and hyperparameter sensitivity led to the development of simpler alternatives. RL-free methods such as DPO (Rafailov et al., 2023) operate in a bandit setting, treating the entire response as a single action. Similarly, rejection sampling methods like RestEM (Singh et al., 2024) finetune on full high-reward responses. RLOO (Ahmadian et al., 2024) and GRPO (Shao et al., 2024) abandon PPO’s value network, instead using average reward from multiple samples as a policy gradient baseline. Recent work has emphasized finer credit assignment, with Hwang et al. (2024) and Setlur et al. (2024) introducing MC-based methods to detect key errors in reasoning chains for use as ad-hoc mechanisms in DPO. Our work, by contrast, fully embraces the RL training, with the target of fixing CA in principle. Parallel efforts have also focused on building better verifiers and reward models for per-step feedback, with recent attempts to automate their data collection using MC rollouts

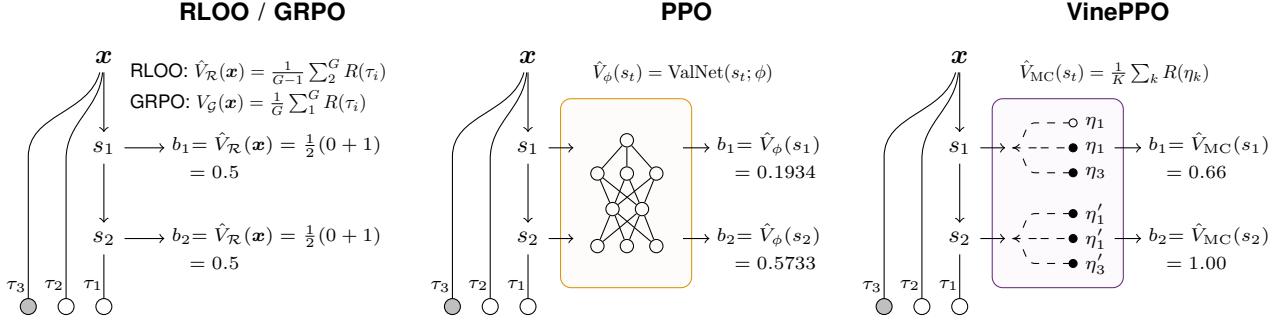


Figure 2: Comparison of credit assignment mechanisms applied on training trajectories τ_i 's, depicted for states s_1 and s_2 . **(Left)** RLOO and GRPO both treat all intermediate states equally and use the average return of trajectory group $\tau_i \sim \pi(\cdot|x)$ for the policy-gradient baselines b_1 and b_2 . GRPO additionally normalize these returns to have a unit variance. In the case of RLOO, the computed baseline could be viewed as MC estimate of value but solely for the initial state. **(Middle)** PPO trains a separate model to predict values for each state s_t . **(Right)** VinePPO generate auxiliary rollouts $\eta_k \sim \pi(\cdot|s_t)$ to obtain MC estimate of state s_t 's value. Note that η_k 's are only used for value estimation—not to update the policy directly.

(Ma et al., 2023; Uesato et al., 2022; Luo et al., 2024; Wang et al., 2024). Our method is orthogonal to these methods, operating within PPO-based training to optimize a *given* reward, instead of designing new ones.

Value Estimation in RL and Monte Carlo Tree Search (MCTS) Deep RL algorithms are typically categorized into value-based and policy-based methods. Policy-based methods like PPO usually employ critic networks for value prediction. An exception is the “Vine” variant of TRPO (Schulman et al., 2015), which uses MC samples for state value estimation. The authors, however, note that the Vine variant is limited to environments that allow intermediate state resets, rare in typical RL settings¹. However, language generation – when formulated as RL environment – enables such intermediate reset capabilities. In domains with similar reset capabilities, such as Go and Chess, MC-based methods like AlphaGo (Silver et al., 2016) and AlphaZero (Silver et al., 2017) have emerged. AlphaGo’s architecture includes a policy, trained using expert moves and self-play, and a value network that predicts game outcomes. At inference, it employs tree search guided by MC rollouts and value network to select optimal moves. AlphaZero advances this approach by distilling MCTS outcomes into the policy. Recent works have adapted AlphaZero’s principles to LLMs, employing similar search techniques for inference and trajectory distillation (Xie et al., 2024; Chen et al., 2024; Wan et al., 2024; Zhang et al., 2024; Hao et al., 2023). While this is a promising direction, our method is not an MCTS approach; it uses MC samples solely for value estimation during PPO *training* to improve credit assignment.

¹This is reflected in the design of Gym (Towers et al., 2024), which only allows resets to the initial state.

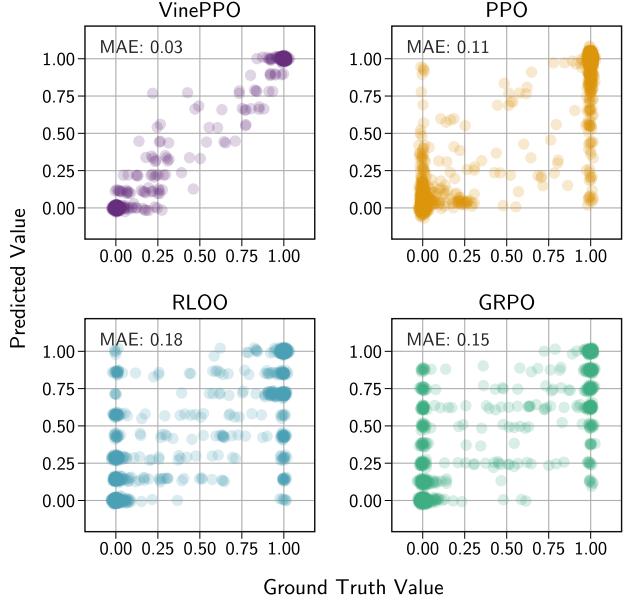


Figure 3: Distribution of predicted values for each state vs. ground truth (see Section 7 for details) for DeepSeekMath 7B on the MATH dataset, highlighting the nature of errors: PPO exhibits biased value predictions, whereas VinePPO remains unbiased. Note that RLOO/GRPO do not predict values; we plot their computed baselines against the ground truth value solely for demonstration.

3. Background

We focus on the RL tuning phase, following Ouyang et al. (2022); Shao et al. (2024). In this section, we provide an overview of actor-critic finetuning as implemented in the standard PPO framework.

RL Finetuning In this setup, the policy π_θ represents a language model that generates a response $\mathbf{y} = [y_0, \dots, y_{T-1}]$ autoregressively given an input $\mathbf{x} = [x_0, \dots, x_{M-1}]$. The goal of RL finetuning is to maximize the expected undiscounted ($\gamma = 1$) finite-horizon return, while incorporating a KL-divergence constraint to regularize the policy and prevent it from deviating too far from a reference policy π_{ref} (typically the initial supervised finetuned, SFT, model). The objective can be written as:

$$J(\theta) = \mathbb{E}_{\mathbf{x} \sim \mathcal{D}, \mathbf{y} \sim \pi(\cdot | \mathbf{x})} [\mathcal{R}(\mathbf{x}; \mathbf{y})] - \beta \text{KL}[\pi_\theta \| \pi_{\text{ref}}], \quad (1)$$

where \mathcal{D} is the dataset of prompts, $\mathcal{R}(\mathbf{x}; \mathbf{y})$ is the sequence-level reward function, and β controls the strength of the KL penalty. Note that the policy π_θ is initialized from π_{ref} .

Language Environment as an MDP Language generation is typically modeled as a token-level Markov Decision Process (MDP) in an actor-critic setting, where each response \mathbf{y} is an episode. The state at time step t , $s_t \in \mathcal{S}$, is the concatenation of the input prompt and the tokens generated up to that point: $s_t = \mathbf{x}; \mathbf{y}_{<t} = [x_0, \dots, x_{M-1}, y_0, \dots, y_{t-1}]$. At each time step, the action a_t corresponds to generating the next token y_t from fixed vocabulary. The process begins with the initial state $s_0 = \mathbf{x}$, and after each action, the environment transitions to the next state, $s_{t+1} = s_t; [a_t]$, by appending the action a_t to the current state s_t . In this case, since states are always constructed by concatenating tokens, the environment dynamics are known and the transition function is *deterministic*, i.e., $P(s_{t+1}|s_t, a_t) = 1$. During the generation process, the reward r_t is set to zero for all intermediate actions a_t 's, with the sequence-level reward $\mathcal{R}(\mathbf{x}; \mathbf{y})$ only applied at the final step when the model stops generating. A trajectory $\tau = (s_0, a_0, s_1, a_1, \dots)$ is therefore a sequence of state-action pairs, starting from the input prompt until the terminal state. Finally, we define the cumulative return of a trajectory τ as $R(\tau) = \sum_{t=0}^{T-1} r_t = r_{T-1} = \mathcal{R}(\mathbf{x}; \mathbf{y})$.

Policy Gradient Given this MDP formulation, policy gradient methods like PPO maximize Equation 1 by repeatedly sampling trajectories and taking a step in the direction of the gradient $\mathbf{g}_{\text{pg}} := \nabla_\theta J(\theta)$ to update the parameters. Policy gradient \mathbf{g}_{pg} takes the following form:

$$\mathbf{g}_{\text{pg}} = \mathbb{E}_{\tau \sim \pi_\theta} \left[\sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t | s_t) A(s_t, a_t) \right], \quad (2)$$

where $s_t = \mathbf{x}; \mathbf{y}_{<t}$, $a_t = y_t$, and $A(s_t, a_t)$ is the *advantage* function. If $A(s_t, a_t) > 0$, \mathbf{g}_{pg} will increase the probability of action a_t in state s_t , and decrease it when $A(s_t, a_t) < 0$. Intuitively, the advantage function quantifies how much better action a_t is compared to average actions taken in state

s_t under the policy. Formally, it is defined as:

$$\begin{aligned} A(s_t, a_t) &= Q(s_t, a_t) - V(s_t) \\ &= r_t + \gamma V(s_{t+1}) - V(s_t). \end{aligned} \quad (3)$$

where $Q(s_t, a_t)$ is the state-action value and $V(s_t)$ is the per-state value function². The value function, $V(s_t) : \mathcal{S} \rightarrow \mathbb{R}$, offers a long-term assessment of how desirable a particular state is under the current policy. Formally, it represents the expected cumulative reward obtained from starting in state s_t and following the policy thereafter³: $V(s_t) = \mathbb{E}_{\tau \sim \pi_\theta} [R(\tau) | s_0 = s_t]$. PPO uses the same advantage-weighted policy gradient as in Equation 2, but constrains policy updates through clipping to ensure stable training. For full details, see Appendix A.

Estimating Advantage via Value Networks In practice, the advantage $A(s_t, a_t)$ is not known beforehand and is typically estimated by first using a value network \hat{V}_ϕ to approximate the *true value function* $V(s_t)$, then substituting the learned values into Equation 3 or alternative methods like GAE (Schulman et al., 2016). The value network is parameterized by ϕ and trained alongside the policy network π_θ . The training objective for the value network minimizes the mean squared error between the predicted value and the empirical return:

$$\mathcal{L}_V(\phi) = \mathbb{E}_{\tau \sim \pi_\theta} \left[\frac{1}{T} \sum_t \frac{1}{2} (\hat{V}_\phi(s_t) - G_t)^2 \right], \quad (4)$$

where $G_t = \sum_{t'=t}^{T-1} r_{t'}$ is the empirical return from state s_t . PPO uses the same objective for \hat{V}_ϕ but applies clipping for training stability (see Appendix A.1 for details). In RL-tuning of LLMs, the value network is often initialized using the initial SFT policy π_{ref} (or the reward model when available), with the language modeling head swapped out for a scalar head to predict values (Zheng et al., 2023). This setup leverages the prior knowledge of the pretrained model.

4. Accurate Credit Assignment with VinePPO

As outlined in Section 3, a step in the PPO gradient update aims to increase the probability of better-than-average actions while decreasing the probability of those that perform worse—a process quantified by the advantage $A(s_t, a_t)$. However, the true advantage is generally unknown and must be estimated, typically by substituting estimates from a value network into Equation 3. As we will elaborate in Section 7, value networks are often inaccurate and result in biased value computation. Fortunately, the language environment as an MDP (Section 3) offers a useful property that

²Such derivation is possible as the language environment is deterministic.

³We drop the dependency on π_θ for brevity.

allows for unbiased estimation of $V(s_t)$. Since states are simply concatenated tokens, we can prompt the language model π_θ to generate continuations from any intermediate state. This flexibility allows us to explore alternative future paths from arbitrary points in a generation.

Specifically, computing the advantage requires access to $V(s_t) = \mathbb{E}[R(\tau) | s_0 = s_t]$. VinePPO obtain an MC estimation of this expectation by randomly sampling continuations and averaging their returns. That is, for each state s_t in a training trajectory τ , we utilize the resetting property and re-feed the partial context corresponding to s_t to the current policy to sample K auxiliary rollouts $\eta_1, \dots, \eta_K \sim \pi_\theta(\cdot | s_t)$. The empirical mean of returns across these rollouts serves as the value estimate:

$$\hat{V}_{\text{MC}}(s_t) = \frac{1}{K} \sum_{k=1}^K R(\eta_k). \quad (5)$$

Critically, η_k 's are used exclusively for value estimation and do not contribute directly to policy gradient updates as we lack CA on them. Once the value $\hat{V}_{\text{MC}}(s_t)$ is computed, we estimate the advantages of each action using Equation 3:

$$\hat{A}_{\text{MC}}(s_t, a_t) = r(s_t, a_t) + \gamma \hat{V}_{\text{MC}}(s_{t+1}) - \hat{V}_{\text{MC}}(s_t). \quad (6)$$

For any $K \geq 1$, the policy gradient computed using the advantage estimator \hat{A}_{MC} is an unbiased estimate of the gradient of expected return g_{pg} . PPO framework then uses \hat{A}_{MC} to update the policy on trajectory τ .

Variance and computational efficiency represent core trade-offs in every Monte Carlo estimation. Here, the sampling parameter K control such tradeoff—increasing K reduces estimator variance at the expense of increased sampling demands. In Section 6, we rigorously characterize these properties for VinePPO.

To enhance the efficiency of \hat{A}_{MC} , we group states within a reasoning step and compute a single advantage, which is assigned to all tokens in that step (examples in Appendix B). This trades off granularity for efficiency, allowing finer resolution with more compute, or coarser estimates with limited resources. Furthermore, modern LLM inference engines (Kwon et al., 2023; Zheng et al., 2024) enable rapid on-the-fly generation⁴, making our MC-based approach computationally practical at scale.

By restricting modifications only to the advantage computation stage of PPO, our approach also isolates the effects of improved credit assignment, revealing how unbiased advantage estimation fundamentally alters policy optimization dynamics compared to value-network baselines.

⁴Achieving up to 5K tokens/second on a single Nvidia A100 GPU for 7B LLMs in bfloat16.

5. Experimental Setup

Datasets and Pretrained LLMs We conduct experiments on publicly available LLMs and datasets to ensure reproducibility. We use base versions of DeepSeekMath 7B (Shao et al., 2024) and RhoMath 1.1B (Lin et al., 2024) which are pretrained on mathematical and natural language corpora. We chose mathematical reasoning datasets MATH (Hendrycks et al., 2021), *competition-level* mathematical problems, and GSM8K (Cobbe et al., 2021), simpler *grade-school level* math word problems. Both datasets are well-established and present a range of difficulty levels. For each dataset, we finetune the base LLM on its respective training set to obtain the initial SFT policy (π_{ref}). Throughout the paper, model names refer ones initialized from these SFT checkpoints. We employ *full-parameter finetuning* to leverage the models' full capacity (Biderman et al., 2024).

Baselines Our main baseline is the standard PPO framework (Ouyang et al., 2022; Huang et al., 2024), which VinePPO builds on and improves through better credit assignment. We also compare against PPO variants that forego the credit assignment: RLOO (Ahmadian et al., 2024) and GRPO (Shao et al., 2024). For RL-free alternatives, we include RestEM (Singh et al., 2024), a form of iterative rejection finetuning (Yuan et al., 2023; Anthony et al., 2017), and DPO⁺ (Pal et al., 2024), a working variant of DPO with strong performance on reasoning. Except VinePPO and standard PPO, all other baselines omit explicit credit assignment by design: i.e. they assign the same weight to all the tokens of a response. All methods use the same SFT checkpoint to ensure fair comparison. For each experiment, we choose the best checkpoint based on a held-out validation set. We compare all methods by accuracy (Pass@1) on test sets, measuring the correctness of final answers.

Training Details and Hyperparameters We adopt a binary task reward \mathcal{R} that evaluates final answer correctness against ground truth, following previous work (Pal et al., 2024; Singh et al., 2024). To ensure fair comparison, all methods consume the same number of episodes during training: for each question, we sample eight episodes and go over the dataset 8 times, yielding 64 episodes per question across all methods. For PPO, we first conduct an extensive hyperparameter search (such as KL penalty coefficient, batch size, minibatch size, GAE λ , number of epochs per iteration) and rigorously implement all established best practices and well-known techniques (Huang et al., 2024; Ivison et al., 2024) (Refer to Appendix C.2 for the full list). This ensures our evaluation reflects PPO's state-of-the-art configuration and its full potential. VinePPO inherits PPO's *exact hyperparameters* and only modifies the advantage estimation, keeping the rest unchanged. This design allows us to isolate the effect of refined credit assignment. For PPO

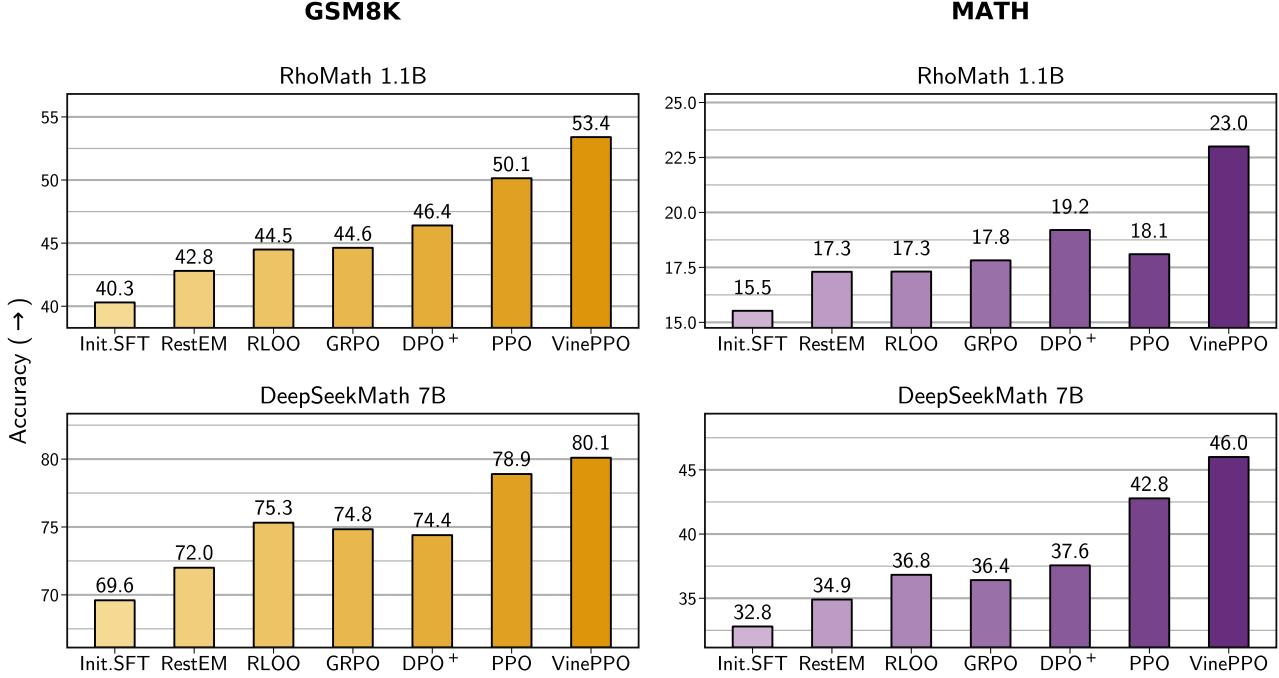


Figure 4: VinePPO outperforms standard PPO, GRPO, RLOO, and other RL-free baselines on Pass@1 performance on MATH and GSM8K datasets, while also exhibiting scalability across different model sizes.

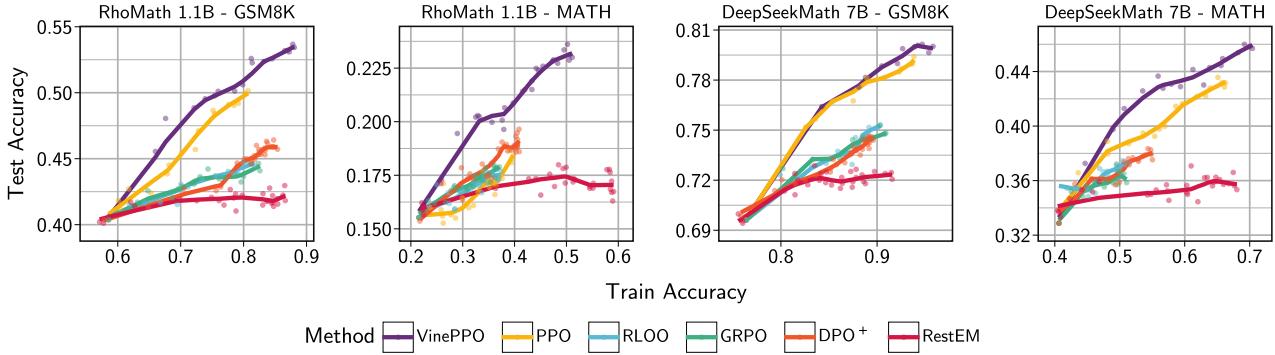


Figure 5: Generalization slope improves with improved credit assignment. VinePPO has steepest generalization: making the highest generalization gains than baselines when fitting the same amount of training data. On the other end of CA spectrum, RestEM overfits its training data.

variants (RLOO, GRPO), we closely follow their HuggingFace implementations. For these, we initialize with PPO’s hyperparameters but perform additional tuning to stabilize training while maintaining the same episode budget. For RL-free baselines (RestEM, DPO+), we strictly adhere to their original implementations (Singh et al., 2024; Pal et al., 2024) and match their sample consumption to other RL methods. For \hat{V}_{MC} in VinePPO, we conduct a full ablation study on K in Section 6.1, with $K = 9$ used as the default setting unless otherwise specified. To ensure a fair comparison of compute efficiency, we conduct controlled experiments in Section 6.2, where all methods are evaluated under identical

hardware and parallelization protocols. Full implementation details, including hyperparameters and training procedures, are documented in Appendix C.6 to ensure reproducibility.

6. Results

In this section, we evaluate the effect of better CA on task performance, efficiency, and generalization dynamics.

6.1. Task Performance

VinePPO consistently outperforms standard PPO throughout training (Figure C.4) and other baselines (Figure 4) achiev-

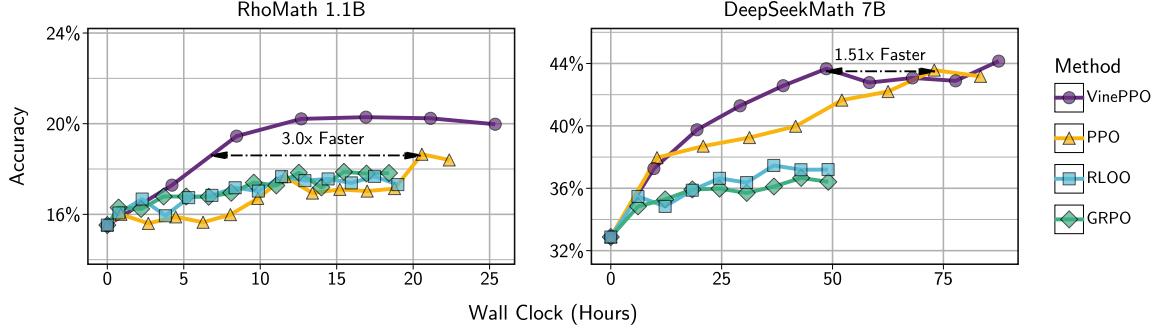


Figure 6: Accuracy vs. Wall Clock Time for both methods measured on the same hardware (shown only up to PPO’s final performance). Despite VinePPO taking longer per iteration (up to 2x for 7B and 5x for 1.1B models), it passes PPO’s peak performance in fewer iterations and less overall time.

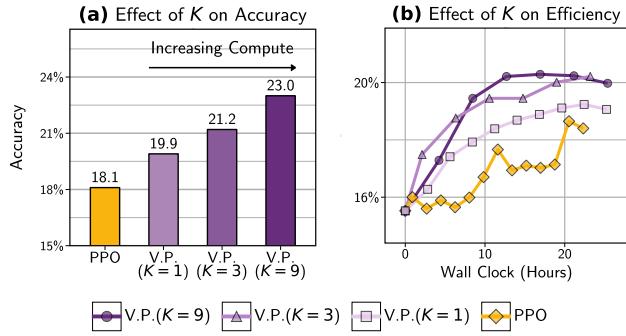


Figure 7: (a) Effect of the number of auxiliary rollouts K for estimating $\hat{V}_{MC}(s_t)$ on RhoMath 1.1B and MATH (see Figure D.11 for GSM8K). Increasing K consistently improves accuracy. (b) Wall-clock time for the same experiments. While increased sampling makes each iteration slower, the reduced variance leads to faster overall convergence.

ing the highest test accuracy on both models and datasets. Notably, the performance gap widens in MATH which is more challenging than GSM8K. To confirm that PPO’s limitations are not due to undertrained value networks, we measured their *explained variance*, a standard metric for value function quality, which ranged between 0.7–0.9 across tasks (Figure D.5), indicating a well-trained critic. Because the PPO and VinePPO runs only differ in their value estimation, comparing these two isolates the effect of CA. As shown in Figure D.8, VinePPO reaches higher test accuracy given a limited KL budget. Additionally, VinePPO is more robust to higher sampling temperatures (Figure D.10).

6.2. Computational Efficiency

Training on a single trajectory in GRPO, RLOO, RestEM and DPO⁺, involves a forward and backward pass. PPO and VinePPO have extra computations of different types. PPO uses double GPU memory — the value network needs

112GB for a 7B LLM, considering both model and its optimizer. Additionally, PPO requires a forward pass for value prediction and a forward-backward pass for value network training. VinePPO replaces the value network with MC samples. Since generation is expensive, each step of VinePPO is slower (up to 5x for RhoMath 1.1B and 2x for DeepSeekMath 7B compared to PPO). VinePPO compensates for slower iterations by making each one more effective through better CA. Under the same hardware, it achieves higher test accuracy faster than baselines (Figure 6). Specifically, VinePPO matches PPO’s peak accuracy in *fewer gradient steps and less wall-clock time*. Figure 6 shows RhoMath 1.1B and DeepSeekMath 7B require about 3.0x and 1.51x less time and 9x and 2.8x fewer steps compared to PPO. This improvement occurs despite all hyperparameters being tuned for PPO. Therefore, switching to VinePPO could enhance the performance within the same compute budget.

6.3. Generalization Slope

High-quality and challenging reasoning tasks are scarce, making generalization a key challenge. Once a training instance is fitted, it provides no further signal for generalization. Thus, algorithms that maximize generalization efficiency are superior—achieving higher test accuracy for a given train accuracy. As shown in Figure 5, VinePPO demonstrates the strongest generalization gains compared to all other baselines. Notably, RestEM overfits near the end. This aligns with recent findings that RL generalizes while SFT primarily memorizes (Chu et al., 2025). Overall, allocating more compute to refining credit assignment, rather than brute-force data fitting, leads to stronger generalization.

6.4. Effect of K

We assess the impact of K , the number of MC samples, by running an ablation on RhoMath 1.1B, varying K from 1 to 3 and 9. As shown in Figure 7, VinePPO improves with

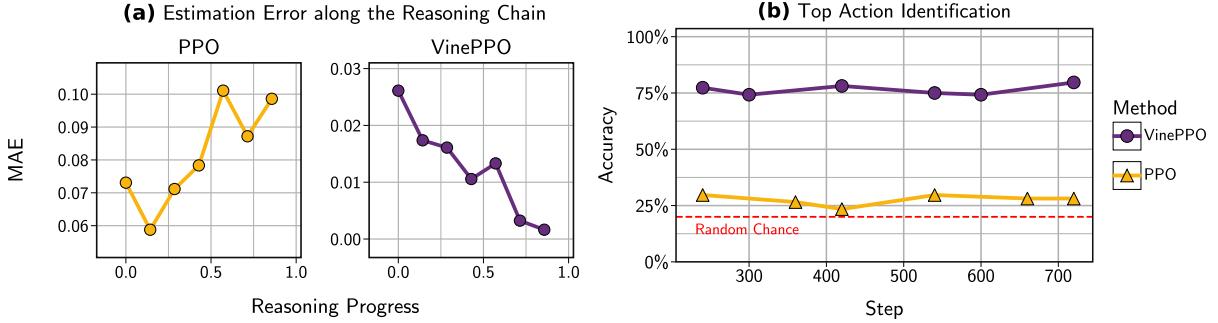


Figure 8: (a) Visualizing the Mean Absolute Error (MAE) of the value predictions at different point of the reasoning chain. Value Network in PPO fails to generalize as the reasoning chain progresses, while VinePPO’s value estimates become more accurate as the model become more deterministic. (b) Accuracy of identifying the top action in a set of five possible next states. VinePPO consistently outperforms the value network.

higher K since more MC samples reduce the variance of \hat{A}_{MC} . While high variance of MC estimation could theoretically hinder training, our results show that even small K values work well in this setting. Interestingly, increasing K also improves compute efficiency. Although each iteration takes longer, it becomes more effective. This suggests that increasing K provides a practical way to leverage additional computational resources for better performance.

7. Why and How Value Networks Fail

In this section, we analyze the performance gap between PPO and VinePPO by focusing on their value predictions—their only difference. First, We establish a “*ground truth*” value at each reasoning step within trajectories by running 256 MC samples and averaging the returns. Next, We compare the value predictions against this ground truth⁵. We present the results for DeepSeekMath 7B , our biggest model, on the MATH dataset (all results in Appendix D.5).

Accuracy Figure 3 presents the distribution of value predictions at each reasoning step. VinePPO’s estimates are unbiased, with variance peaking at 0.5 and dropping to zero at 0 and 1. PPO’s value network shows high bias and often misclassifies bad states (ground truth near 0) as good and vice versa. We define a prediction as “correct” if it is within 0.05 of the ground truth. As shown in Figure D.12 PPO’s value network starts with low accuracy, gradually improving to 65%. In contrast, VinePPO consistently achieves 70-90% accuracy throughout training.

Top Action Identification In value-based RL, accurately ranking actions is more important than accurate value esti-

mates. While PPO, a policy-based method, depends heavily on accurate value estimates, it raises an interesting question: Can PPO’s value network still rank actions correctly? We tested this by sampling five possible next steps from a shared initial state and measuring whether the method predicted the next step with the highest ground truth value by assigning it the highest predicted value. As shown in Figure 8.b, PPO’s value network performs near chance levels for most of the training, improving only slightly over time. In contrast, VinePPO consistently identifies the top action with high accuracy throughout.

Error Per Reasoning Step To understand value estimation dynamics, we plot value estimation error against reasoning step position (normalized; 3rd of 10 steps = 0.3). As shown in Figure 8.a, PPO performs worse as reasoning progresses. We hypothesize this is because early steps resemble training data, allowing the value network to rely on memorization. Later steps are more diverse and value network struggles to generalize. VinePPO’s prediction error decreases with reasoning progression. We attribute this to greater determinism in later steps, as the model conditions on a longer context. This stability improves value estimation from the same number of MC samples.

8. Discussion

We showed that better credit assignment improves RL training of LLMs. VinePPO is a stepping stone to identify and fix PPO’s broken credit assignment. It also opens two future research directions. VinePPO is the first RL post-training algorithm that scales generalization slope by scaling post-training compute. Algorithms that have better generalization trends are valuable given the limited resource of truly challenging and verifiable reasoning tasks. Second, VinePPO highlights the value of reconsidering the implicit assumptions behind default algorithm implementations borrowed

⁵The return of sampled trajectory starting from a step is a bernoulli random variable. Let θ denote the true success probability. Our estimator, $\bar{X} = \frac{1}{256} \sum_{i=1}^{256} X_i$, has variance $\frac{\theta(1-\theta)}{256}$. At $\theta = 0.5$ (maximal variance), this becomes $\frac{0.25}{256} \approx 0.001$.

from Deep RL. In Deep RL, we typically start with a random policy, making it crucial to quickly improve the model's performance. In this context, it's more effective to allocate compute toward gathering additional environment samples rather than perfecting each gradient update. However, with an already capable LLM, it is better to spend more compute to make sure we steer its weights carefully. Overall, we hope VinePPO inspires the community to develop more effective RL training algorithms for LLMs.

Impact Statement

Our work aims to improve the ability of large language models to perform complex reasoning tasks, potentially contributing to advances in fields such as education, scientific research, and software development. At the same time, more capable reasoning systems could be used irresponsibly, for instance, by automating sophisticated misinformation or other harmful applications. We therefore encourage researchers and practitioners to employ appropriate safeguards when applying our methods. Overall, this paper advances fundamental techniques in machine learning; its societal impact will depend on responsible deployment and continued ethical considerations by the community.

Acknowledgements

We thank Matheus Pereira for his efforts on facilitating experimentation. AC and NR are supported by CIFAR AI Chair. SR is supported by a Facebook CIFAR AI Chair and NSERC Discovery Grant program. We thank Mila IDT team and Digital Research Alliance of Canada for the compute provided for experimentation.

References

- Ahmadian, A., Cremer, C., Gallé, M., Fadaee, M., Kreutzer, J., Pietquin, O., Üstün, A., and Hooker, S. Back to Basics: Revisiting REINFORCE-style Optimization for Learning from Human Feedback in LLMs. In Ku, L., Martins, A., and Srikumar, V. (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL 2024, pp. 12248–12267, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/V1/2024.ACL-LONG.662. URL <https://doi.org/10.18653/v1/2024.acl-long.662>.
- Anthony, T., Tian, Z., and Barber, D. Thinking Fast and Slow with Deep Learning and Tree Search. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA*, pp. 5360–5370, USA, 2017. URL <https://proceedings.neurips.cc/paper/2017/hash/d8e1344e27a5b08cdfd5d027d9b8d6de-Abstract.html>.
- Biderman, D., Ortiz, J. J. G., Portes, J., Paul, M., Greengard, P., Jennings, C., King, D., Havens, S., Chiley, V., Frankle, J., Blakeney, C., and Cunningham, J. P. LoRA Learns Less and Forgets Less. *CoRR*, abs/2405.09673, 2024. doi: 10.48550/ARXIV.2405.09673. URL <https://doi.org/10.48550/arXiv.2405.09673>.
- Chang, J. D., Brantley, K., Ramamurthy, R., Misra, D., and Sun, W. Learning to generate better than your llm. *arXiv preprint arXiv:2306.11816*, 2023.
- Chen, G., Liao, M., Li, C., and Fan, K. AlphaMath Almost Zero: process Supervision without process. *CoRR*, abs/2405.03553, 2024. doi: 10.48550/ARXIV.2405.03553. URL <https://doi.org/10.48550/arXiv.2405.03553>.
- Chu, T., Zhai, Y., Yang, J., Tong, S., Xie, S., Schuurmans, D., Le, Q. V., Levine, S., and Ma, Y. Sft memorizes, rl generalizes: A comparative study of foundation model post-training, 2025. URL <https://arxiv.org/abs/2501.17161>.
- Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., and Schulman, J. Training Verifiers to Solve Math Word Problems. *CoRR*, abs/2110.14168, 2021. URL <https://arxiv.org/abs/2110.14168>.
- DeepSeek-AI, Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL <https://arxiv.org/abs/2501.12948>.
- Greensmith, E., Bartlett, P. L., and Baxter, J. Variance Reduction Techniques for Gradient Estimates in Reinforcement Learning. In Dietterich, T. G., Becker, S., and Ghahramani, Z. (eds.), *Advances in Neural Information Processing Systems 14 [Neural Information Processing Systems: Natural and Synthetic, NIPS 2001, December 3-8, 2001]*, pp. 1507–1514, Vancouver, British Columbia, Canada, 2001. MIT Press. URL <https://proceedings.neurips.cc/paper/2001/hash/584b98aac2dddf59ee2cf19ca4ccb75e-Abstract.html>.

- Hao, S., Gu, Y., Ma, H., Hong, J. J., Wang, Z., Wang, D. Z., and Hu, Z. Reasoning with Language Model is Planning with World Model. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023*, pp. 8154–8173, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/V1/2023.EMNLP-MAIN.507. URL <https://doi.org/10.18653/v1/2023.emnlp-main.507>.
- Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., and Steinhardt, J. Measuring Mathematical Problem Solving With the MATH Dataset. In Vanschoren, J. and Yeung, S. (eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021*, 2021. URL <https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/be83ab3ecd0db773eb2dc1b0a17836a1-Abstract-round2.html>.
- Huang, S., Noukhovitch, M., Hosseini, A., Rasul, K., Wang, W., and Tunstall, L. The N+ Implementation Details of RLHF with PPO: A Case Study on TL;DR Summarization. *CoRR*, abs/2403.17031, 2024. doi: 10.48550/ARXIV.2403.17031. URL <https://doi.org/10.48550/arXiv.2403.17031>.
- Hwang, H., Kim, D., Kim, S., Ye, S., and Seo, M. Self-explore to Avoid the Pit: Improving the Reasoning Capabilities of Language Models with Fine-grained Rewards. *CoRR*, abs/2404.10346, 2024. doi: 10.48550/ARXIV.2404.10346. URL <https://doi.org/10.48550/arXiv.2404.10346>.
- Ivison, H., Wang, Y., Liu, J., Wu, Z., Pyatkin, V., Lambert, N., Smith, N. A., Choi, Y., and Hajishirzi, H. Unpacking DPO and PPO: Disentangling Best Practices for Learning from Preference Feedback. *CoRR*, abs/2406.09279, 2024. doi: 10.48550/ARXIV.2406.09279. URL <https://doi.org/10.48550/arXiv.2406.09279>.
- Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C. H., Gonzalez, J., Zhang, H., and Stoica, I. Efficient Memory Management for Large Language Model Serving with PagedAttention. In Flinn, J., Seltzer, M. I., Druschel, P., Kaufmann, A., and Mace, J. (eds.), *Proceedings of the 29th Symposium on Operating Systems Principles, SOSP 2023*, pp. 611–626, Koblenz, Germany, 2023. ACM. doi: 10.1145/3600006.3613165. URL <https://doi.org/10.1145/3600006.3613165>.
- Lewkowycz, A., Andreassen, A., Dohan, D., Dyer, E., Michalewski, H., Ramasesh, V. V., Slone, A., Anil, C., Schlag, I., Gutman-Solo, T., Wu, Y., Neyshabur, B., Gur-Ari, G., and Misra, V. Solving Quantitative Reasoning Problems with Language Models. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022*, New Orleans, LA, USA, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/18abbeef8cfe9203fdf9053c9c4fe191-Abstract-Conference.html.
- Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., Leike, J., Schulman, J., Sutskever, I., and Cobbe, K. Let's Verify Step by Step. In *The Twelfth International Conference on Learning Representations, ICLR 2024*, Vienna, Austria, 2024. OpenReview.net. URL <https://openreview.net/forum?id=v8L0pN6EOi>.
- Lin, Z., Gou, Z., Gong, Y., Liu, X., Shen, Y., Xu, R., Lin, C., Yang, Y., Jiao, J., Duan, N., and Chen, W. Rho-1: Not All Tokens Are What You Need. *CoRR*, abs/2404.07965, 2024. doi: 10.48550/ARXIV.2404.07965. URL <https://doi.org/10.48550/arXiv.2404.07965>.
- Luo, L., Liu, Y., Liu, R., Phatale, S., Lara, H., Li, Y., Shu, L., Zhu, Y., Meng, L., Sun, J., and Rastogi, A. Improve Mathematical Reasoning in Language Models by Automated Process Supervision. *CoRR*, abs/2406.06592, 2024. doi: 10.48550/ARXIV.2406.06592. URL <https://doi.org/10.48550/arXiv.2406.06592>.
- Ma, Q., Zhou, H., Liu, T., Yuan, J., Liu, P., You, Y., and Yang, H. Let's reward step by step: Step-level reward model as the Navigators for Reasoning. *CoRR*, abs/2310.10080, 2023. doi: 10.48550/ARXIV.2310.10080. URL <https://doi.org/10.48550/arXiv.2310.10080>.
- OpenAI. OpenAI o1 System Card, 2024. URL <https://api.semanticscholar.org/CorpusID:272684752>.
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P. F., Leike, J., and Lowe, R. Training language models to follow instructions with human feedback. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022*, New Orleans, LA, USA, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html.

- Pal, A., Karkhanis, D., Dooley, S., Roberts, M., Naidu, S., and White, C. Smaug: Fixing Failure Modes of Preference Optimisation with DPO-positive. *CoRR*, abs/2402.13228, 2024. doi: 10.48550/ARXIV.2402.13228. URL <https://doi.org/10.48550/arXiv.2402.13228>.
- Putta, P., Mills, E., Garg, N., Motwani, S., Finn, C., Garg, D., and Rafailov, R. Agent q: Advanced reasoning and learning for autonomous ai agents, 2024. URL <https://arxiv.org/abs/2408.07199>.
- Qwen. Qwen2.5-Math: The world's leading open-sourced mathematical LLMs. <https://qwenlm.github.io/blog/qwen2.5-math/>, 2024. Accessed: 2024-09-23.
- Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023*, New Orleans, LA, USA, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/a85b405ed65c6477a4fe8302b5e06ce7-Abstract-Conference.html.
- Schulman, J. Notes on the KL-divergence Approximation. <http://joschu.net/blog/kl-approx.html>, 2020. Accessed: 2024-09-23.
- Schulman, J., Levine, S., Abbeel, P., Jordan, M. I., and Moritz, P. Trust Region Policy Optimization. In Bach, F. R. and Blei, D. M. (eds.), *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pp. 1889–1897, Lille, France, 2015. JMLR.org. URL <http://proceedings.mlr.press/v37/schulman15.html>.
- Schulman, J., Moritz, P., Levine, S., Jordan, M. I., and Abbeel, P. High-dimensional Continuous Control Using Generalized Advantage Estimation. In Bengio, Y. and LeCun, Y. (eds.), *4th International Conference on Learning Representations, ICLR 2016 Proceedings*, San Juan, Puerto Rico, 2016. URL <http://arxiv.org/abs/1506.02438>.
- Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal Policy Optimization Algorithms. *CoRR*, abs/1707.06347, 2017. URL <http://arxiv.org/abs/1707.06347>.
- Setlur, A., Garg, S., Geng, X., Garg, N., Smith, V., and Kumar, A. RL on Incorrect Synthetic Data Scales the Efficiency of LLM Math Reasoning by Eight-fold. *CoRR*, abs/2406.14532, 2024. doi: 10.48550/ARXIV.2406.14532. URL <https://doi.org/10.48550/arXiv.2406.14532>.
- Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Zhang, M., Li, Y. K., Wu, Y., and Guo, D. DeepSeek-Math: Pushing the Limits of Mathematical Reasoning in Open Language Models. *CoRR*, abs/2402.03300, 2024. doi: 10.48550/ARXIV.2402.03300. URL <https://doi.org/10.48550/arXiv.2402.03300>.
- Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T. P., Leach, M., Kavukcuoglu, K., Graepel, T., and Hassabis, D. Mastering the game of Go with deep neural networks and tree search. *Nat.*, 529(7587):484–489, 2016. doi: 10.1038/NATURE16961. URL <https://doi.org/10.1038/nature16961>.
- Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T. P., Simonyan, K., and Hassabis, D. Mastering Chess and Shogi by Self-play with a General Reinforcement Learning Algorithm. *CoRR*, abs/1712.01815, 2017. URL <http://arxiv.org/abs/1712.01815>.
- Singh, A., Co-Reyes, J. D., Agarwal, R., Anand, A., Patil, P., Garcia, X., Liu, P. J., Harrison, J., Lee, J., Xu, K., Parisi, A. T., Kumar, A., Alemi, A. A., Rizkowsky, A., Nova, A., Adlam, B., Bohnet, B., Elsayed, G. F., Sedghi, H., Mordatch, I., Simpson, I., Gur, I., Snoek, J., Pennington, J., Hron, J., Kenealy, K., Swersky, K., Mahajan, K., Culp, L., Xiao, L., Bileschi, M. L., Constant, N., Novak, R., Liu, R., Warkentin, T., Qian, Y., Bansal, Y., Dyer, E., Neyshabur, B., Sohl-Dickstein, J., and Fiedel, N. Beyond Human Data: Scaling Self-training for Problem-solving with Language Models. *Transactions on Machine Learning Research*, 2024, 2024. URL <https://openreview.net/forum?id=1NAyUngGFK>.
- Sutton, R. S. and Barto, A. G. Introduction to Reinforcement Learning. In *Introduction to Reinforcement Learning*, 1998. URL <https://api.semanticscholar.org/CorpusID:261579713>.
- Sutton, R. S., McAllester, D. A., Singh, S., and Mansour, Y. Policy Gradient Methods for Reinforcement Learning with Function Approximation. In Solla, S. A., Leen, T. K., and Müller, K. (eds.),

- Advances in Neural Information Processing Systems 12, [NIPS Conference], pp. 1057–1063, Denver, Colorado, USA, 1999. The MIT Press.* URL <http://papers.nips.cc/paper/1713-policy-gradient-methods-for-reinforcement-learning-with-function-approximation>.
- Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Canton-Ferrer, C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., et al. Llama 2: Open Foundation and Fine-tuned Chat Models. *CoRR*, abs/2307.09288, 2023. doi: 10.48550/ARXIV.2307.09288. URL <https://doi.org/10.48550/arXiv.2307.09288>.
- Towers, M., Kwiatkowski, A., Terry, J., Balis, J. U., De Cola, G., Deleu, T., Goulão, M., Kallinteris, A., Krimmel, M., KG, A., et al. Gymnasium: A standard interface for reinforcement learning environments. *arXiv preprint arXiv:2407.17032*, 2024.
- Trung, L. Q., Zhang, X., Jie, Z., Sun, P., Jin, X., and Li, H. ReFT: Reasoning with Reinforced Fine-tuning. In Ku, L., Martins, A., and Srikumar, V. (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024*, pp. 7601–7614, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/V1/2024.ACL-LONG.410. URL <https://doi.org/10.18653/v1/2024.acl-long.410>.
- Uesato, J., Kushman, N., Kumar, R., Song, H. F., Siegel, N. Y., Wang, L., Creswell, A., Irving, G., and Higgins, I. Solving math word problems with process-and outcome-based feedback. *CoRR*, abs/2211.14275, 2022. doi: 10.48550/ARXIV.2211.14275. URL <https://doi.org/10.48550/arXiv.2211.14275>.
- Wan, Z., Feng, X., Wen, M., McAlleer, S. M., Wen, Y., Zhang, W., and Wang, J. AlphaZero-like Tree-search can Guide Large Language Model Decoding and Training. In *Forty-first International Conference on Machine Learning, ICML 2024*, Vienna, Austria, 2024. OpenReview.net. URL <https://openreview.net/forum?id=C4OpREezgj>.
- Wang, P., Li, L., Shao, Z., Xu, R. X., Dai, D., Li, Y., Chen, D., Wu, Y., and Sui, Z. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. *CoRR*, abs/2406.06592, 2024. URL <https://arxiv.org/abs/2312.08935>.
- Xie, Y., Goyal, A., Zheng, W., Kan, M., Lillicrap, T. P., Kawaguchi, K., and Shieh, M. Monte Carlo Tree Search Boosts Reasoning via Iterative Preference Learning. *CoRR*, abs/2405.00451, 2024. doi: 10.48550/ARXIV.2405.00451. URL <https://doi.org/10.48550/arXiv.2405.00451>.
- Xu, S., Fu, W., Gao, J., Ye, W., Liu, W., Mei, Z., Wang, G., Yu, C., and Wu, Y. Is DPO Superior to PPO for LLM Alignment? A Comprehensive Study. In *Forty-first International Conference on Machine Learning, ICML 2024*, Vienna, Austria, 2024. OpenReview.net. URL <https://openreview.net/forum?id=6XH8R7YrSk>.
- Yuan, Z., Yuan, H., Li, C., Dong, G., Tan, C., and Zhou, C. Scaling Relationship on Learning Mathematical Reasoning with Large Language Models. *CoRR*, abs/2308.01825, 2023. doi: 10.48550/ARXIV.2308.01825. URL <https://doi.org/10.48550/arXiv.2308.01825>.
- Zhang, D., Zhoubian, S., Yue, Y., Dong, Y., and Tang, J. ReST-MCTS*: LLM Self-training via Process Reward Guided Tree Search. *CoRR*, abs/2406.03816, 2024. doi: 10.48550/ARXIV.2406.03816. URL <https://doi.org/10.48550/arXiv.2406.03816>.
- Zheng, L., Yin, L., Xie, Z., Sun, C., Huang, J., Yu, C. H., Cao, S., Kozyrakis, C., Stoica, I., Gonzalez, J. E., Barrett, C., and Sheng, Y. Sglang: Efficient execution of structured language model programs. *CoRR*, abs/2312.07104, 2024. URL <https://arxiv.org/abs/2312.07104>.
- Zheng, R., Dou, S., Gao, S., Hua, Y., Shen, W., Wang, B., Liu, Y., Jin, S., Liu, Q., Zhou, Y., Xiong, L., Chen, L., Xi, Z., Xu, N., Lai, W., Zhu, M., Chang, C., Yin, Z., Weng, R., Cheng, W., Huang, H., Sun, T., Yan, H., Gui, T., Zhang, Q., Qiu, X., and Huang, X. Secrets of RLHF in Large Language Models Part I: PPO. *CoRR*, abs/2307.04964, 2023. doi: 10.48550/ARXIV.2307.04964. URL <https://doi.org/10.48550/arXiv.2307.04964>.
- Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., Christiano, P. F., and Irving, G. Fine-tuning Language Models from Human Preferences. *CoRR*, abs/1909.08593, 2019. URL <https://arxiv.org/abs/1909.08593>.

A. Reviewing PPO

PPO, as used in RL tuning of LLMs, formulates language generation as token-level MDP (Section 3), where each response \mathbf{y} is an episode. The state at time step t , $s_t \in \mathcal{S}$, is the concatenation of the prompt and the tokens generated so far: $s_t = \mathbf{x}; \mathbf{y}_{<t} = [x_0, \dots, x_{M-1}, y_0, \dots, y_{t-1}]$. The action a_t corresponds to generating the next token y_t from the model's vocabulary. Given a prompt \mathbf{x} , an episode of this MDP starts from the initial state $s_0 = \mathbf{x}$, and with each action taken, the environment moves to a subsequent state, $s_{t+1} = s_t; [a_t]$, by adding the action a_t to the existing state s_t . In the language environment, because states are always formed by concatenating tokens, the environment dynamics are fully known, and the transition function is *deterministic*, meaning $P(s_{t+1}|s_t, a_t) = 1$. Throughout the generation process, the reward r_t is set to zero for all intermediate actions a_t , with the sequence-level reward $\mathcal{R}(\mathbf{x}; \mathbf{y})$ applied only at the final step when the model stops the generation. That is:

$$r_t = r(s_t, a_t) = \begin{cases} \mathcal{R}(\mathbf{x}; \mathbf{y}) & \text{if } t = T - 1, \text{ where } s_{t+1} = \mathbf{y} \text{ is terminal,} \\ 0 & \text{otherwise.} \end{cases} \quad (7)$$

A trajectory $\tau = (s_0, a_0, s_1, a_1, \dots)$ thus represents a sequence of state-action pairs that begins at the input prompt and continues until reaching the terminal state. Finally, the cumulative return of a trajectory τ is defined as $R(\tau) = \sum_{t=0}^{T-1} r_t = r_{T-1} = \mathcal{R}(\mathbf{x}; \mathbf{y})$.

The goal of RL tuning is to maximize the expected return of the model's responses to prompts in the dataset, as defined by the reward function \mathcal{R} (Equation 1). PPO, similar to other policy gradient methods, achieves this goal by repeatedly sampling trajectories for a batch of prompt sampled from \mathcal{D} and taking multiple optimization steps in the direction of the gradient \mathbf{g}_{ppo} to update the parameters. PPO gradient \mathbf{g}_{ppo} is defined as the gradient of the following loss:

$$\mathcal{L}_{\text{ppo}}(\theta) = \mathbb{E}_{\tau \sim \pi_{\theta_k}} \left[\sum_{t=0}^{T-1} \min \left(\frac{\pi_\theta(a_t | s_t)}{\pi_{\theta_k}(a_t | s_t)} A_t^{\theta_k}, \text{clip}(\theta) A_t^{\theta_k} \right) - \beta \text{KL}[\pi_\theta \| \pi_{\text{ref}}] \right] \quad (8)$$

where π_{θ_k} is the policy at the previous iteration, ϵ is the clipping parameter, β is the KL penalty coefficient, $A_t^{\theta_k} = A^{\theta_k}(s_t, a_t)$ is the advantage estimate for policy π_{θ_k} , and the $\text{clip}(\theta)$ function is:

$$\text{clip}(\theta) = \text{clip} \left(\frac{\pi_\theta(a_t | s_t)}{\pi_{\theta_k}(a_t | s_t)}, 1 - \epsilon, 1 + \epsilon \right). \quad (9)$$

Note that the KL penalty could be also added to the reward function \mathcal{R} . We follow the more recent implementations (Shao et al., 2024; Qwen, 2024), where it is added to the loss function. The KL term can be computed using the following unbiased estimator (Schulman, 2020):

$$\hat{\text{KL}}(\theta) = \frac{\pi_{\text{ref}}(a_t | s_t)}{\pi_\theta(a_t | s_t)} - \log \frac{\pi_{\text{ref}}(a_t | s_t)}{\pi_\theta(a_t | s_t)} - 1, \quad (10)$$

where π_{ref} denotes the reference model (initial SFT).

A.1. Value Network

In addition to the policy π_θ , PPO also trains a separate value network \hat{V}_ϕ to obtain an estimate the true values $V(s_t)$ of states s_t . Parameterized by ϕ , the value network is trained alongside the policy network π_θ using the following loss:

$$\mathcal{L}_{\text{ValNet}}(\phi) = \frac{1}{2} \mathbb{E}_{\tau \sim \pi_\theta} \left[\frac{1}{T} \sum_{t=0}^{T-1} \max \left(\left\| \hat{V}_\phi(s_t) - G_t \right\|^2, \left\| \text{clip}(\phi) - G_t \right\|^2 \right) \right] \quad (11)$$

where \hat{V}_{ϕ_k} is the value network at the previous iteration, $G_t = \sum_{t'=t}^{T-1} \gamma^{t'-t} r_{t'}$ is the empirical return from state s_t , ϵ' is a value clipping parameter, and the $\text{clip}(\phi)$ is defined as:

$$\text{clip}(\phi) = \text{clip} \left(\hat{V}_\phi(s_t), \hat{V}_{\phi_k}(s_t) - \epsilon', \hat{V}_{\phi_k}(s_t) + \epsilon' \right). \quad (12)$$

In RL-tuning of LLMs, the value network is typically initialized from the initial policy π_{ref} (or the reward model, if available), replacing the language modeling head with a scalar output head to predict values (Zheng et al., 2023). This approach takes advantage of the base model's prior knowledge for value estimation.

Advantage Estimation Once the estimated values $\hat{V}_\phi(s_t)$ are obtained, the advantages $A(s_t, a_t)$ are computed using the GAE (Schulman et al., 2016):

$$A(s_t, a_t) \approx \hat{A}^{\text{GAE}}(s_t, a_t) \quad (13)$$

$$= (1 - \lambda) \left(\hat{A}_t^{(1)} + \lambda \hat{A}_t^{(2)} + \lambda^2 \hat{A}_t^{(3)} + \dots \right) \quad (14)$$

$$= \sum_{l=0}^{\infty} (\gamma \lambda)^l \delta_{t+l} \quad (15)$$

$$= \sum_{l=0}^{\infty} (\gamma \lambda)^l \left(r_{t+l} + \gamma \hat{V}_\phi(s_{t+l+1}) - \hat{V}_\phi(s_{t+l}) \right) \quad (16)$$

where $\delta_t = r_t + \gamma \hat{V}_\phi(s_{t+1}) - \hat{V}_\phi(s_t)$ is the temporal difference error, λ is the GAE parameter, and γ is the discount factor. Also, we have:

$$\hat{A}_t^{(k)} := \sum_{l=0}^{k-1} \gamma^l \delta_{t+l} = r_t + \gamma r_{t+1} + \dots + \gamma^{k-1} r_{t+k-1} + \gamma^k \hat{V}_\phi(s_{t+k}) - \hat{V}_\phi(s_t). \quad (17)$$

Adjusting the GAE parameter λ allows for a trade-off between bias and variance in the advantage estimates. However, as we discuss in Appendix C.6, we found that $\lambda = 1$ works best in our experiments (similar to the findings of Trung et al. (2024) and Ahmadian et al. (2024)). In this case, the GAE simplifies to the following form (assuming $\gamma = 1$): $\hat{A}^{\text{GAE}}(s_t, a_t) = \sum_{t'=t}^{T-1} r_{t'} - \hat{V}_\phi(s_t)$.

B. Reasoning Step Separation Examples

In this section, we outline the methodology used to segment solutions into discrete reasoning steps for the MATH and GSM8K datasets, as illustrated in Figures B.1 and B.2.

For the MATH dataset, we begin by splitting solutions based on clear natural boundaries such as newline characters or punctuation marks (e.g., periods or commas). Care is taken to avoid splitting within mathematical expressions, ensuring that mathematical formulas remain intact. After this initial segmentation, if any resulting step exceeds 100 characters, we further try to divide it by identifying logical breakpoints, such as equal signs (=) within math mode.

For the GSM8K dataset, we take a simpler approach, segmenting the reasoning steps by newlines alone as with this task newlines already serve as natural delimiters.

C. Experimental Details

C.1. Datasets

We focus on mathematical reasoning datasets that require step-by-step solutions and are widely used to evaluate the reasoning capabilities of LLMs. Below is a brief overview of the datasets used in our experiments:

MATH (Hendrycks et al., 2021) The MATH dataset contains problems from high school math competitions, covering a wide range of topics such as algebra, geometry, and probability. For our experiments, we use the OpenAI split provided by Lightman et al. (2024), which consists of 500 problems for testing and 12,500 problems for training. We further divide the training set into 11,500 problems for training and 500 problems for validation. Each problem includes a step-by-step solution, ending in a final answer marked by `\boxed{}` in the solution (e.g., “..so the smallest possible value of c is $\boxed{\pi}$ ”). This marking allows for verification of the correctness of model-generated responses by comparing the final answer to the ground truth. We use the scripts provided by Lewkowycz et al. (2022), Lightman et al. (2024), and Shao et al. (2024) to extract and compare the final answers to the ground truth.

GSM8K (Cobbe et al., 2021) The GSM8K dataset comprises high-quality grade-school math problems, requiring basic arithmetic or elementary algebra to solve. Although simpler than the MATH dataset, GSM8K is still widely used to assess the reasoning capabilities of LLMs. It contains 1,319 problems for testing and 7,473 for training. To create a validation set, we further split the training set into 7,100 problems for training and 373 for validation. Verifying the correctness of model responses is straightforward, as the final answer is typically an integer, marked by `####` in the solution.

Prompt

If we express $3x^2 + x - 4$ in the form $a(x - h)^2 + k$, then what is k ?

Response

We complete the square.[†] First, we factor 3 out of the terms $3x^2 + x$ to get $3(x^2 + \frac{x}{3})$.[†] We can square $x + \frac{1}{6}$ to get $x^2 + \frac{x}{3} + \frac{1}{36}$, so[†]

$$\begin{aligned} 3\left(x^2 + \frac{x}{3}\right) &= 3\left[\left(x + \frac{1}{6}\right)^2 - \frac{1}{36}\right] \\ &= 3\left(x + \frac{1}{6}\right)^2 - \frac{3}{36} \\ &= 3\left(x + \frac{1}{6}\right)^2 - \frac{1}{12}, \end{aligned}$$

and

$$\begin{aligned} 3\left(x^2 + \frac{x}{3}\right) - 4 &= 3\left(x + \frac{1}{6}\right)^2 - \frac{1}{12} - 4 \\ &= 3\left(x + \frac{1}{6}\right)^2 - \frac{49}{12}. \end{aligned}$$

We see that $k = \boxed{-\frac{49}{12}}$.[†]

Prompt

Suppose the roots of the polynomial $x^2 - mx + n$ are positive prime integers (not necessarily distinct). Given that $m < 20$, how many possible values of n are there?

Response

Let p and q be the prime roots.[†] Then, we know that $m = p + q$ and $n = pq$.[†] Since $m < 20$, the primes p and q must both be less than 20.[†]

The primes less than 20 are 2, 3, 5, 7, 11, 13, 17, 19.[†]

Now we list all possible pairs (p, q) such that $p + q < 20$,[†] remembering to also include the cases in which $p = q$:[†]

$$\begin{aligned} &(2, 2), (2, 3), (2, 5), (2, 7), (2, 11), (2, 13), (2, 17) \\ &(3, 3), (3, 5), (3, 7), (3, 11), (3, 13) \\ &(5, 5), (5, 7), (5, 11), (5, 13) \\ &(7, 7), (7, 11) \end{aligned}$$

There are $7 + 5 + 4 + 2 = 18$ pairs in total.[†] Each pair produces a value for n , and furthermore,[†] these values are all distinct, because every positive integer has a unique prime factorization.[†] Thus, there are 18 possible values for n .[†]

Figure B.1: Examples of solutions separated into its reasoning steps on the MATH dataset. Steps are highlighted using distinct colors. [†] denotes the reasoning step boundary.

Prompt

Tobias is buying a new pair of shoes that costs \$95. He has been saving up his money each month for the past three months. He gets a \$5 allowance a month. He also mows lawns and shovels driveways. He charges \$15 to mow a lawn and \$7 to shovel. After buying the shoes, he has \$15 in change. If he mows 4 lawns, how many driveways did he shovel?

Response

He saved up \$110 total because $95 + 15 = 110$ [†]

He saved \$15 from his allowance because $3 \times 5 = 15$ [†]

He earned \$60 mowing lawns because $4 \times 15 = 60$ [†]

He earned \$35 shoveling driveways because $110 - 60 - 15 = 35$ [†]

He shoveled 5 driveways because $35 / 7 = 5$.

5[†]

Prompt

Tim rides his bike back and forth to work for each of his 5 workdays. His work is 20 miles away. He also goes for a weekend bike ride of 200 miles. If he can bike at 25 mph how much time does he spend biking a week?

Response

He bikes $20 \times 2 = 40$ miles each day for work.[†]

So he bikes $40 \times 5 = 200$ miles for work[†]

That means he bikes a total of $200+200 = 400$ miles for work[†]

So he bikes a total of $400 / 25 = 16$ hours

16[†]

Figure B.2: Examples of solutions separated into its reasoning steps on the GSM8K dataset. Steps are highlighted using distinct colors. [†] denotes the reasoning step boundary.

C.2. PPO Implementation

To ensure our PPO implementation is robust, and our evaluation reflects its full potential, we have applied a set of well-established techniques and best practices from the literature (Huang et al., 2024; Ivison et al., 2024; Zheng et al., 2023). Below, we outline the key implementation details that were most effective in our experiments:

- **Advantage Normalization:** After calculating the advantages, we normalize them to have zero mean and unit variance,

Table 1: Summary of PPO hyperparameters used in the experiments.

Parameter	Value	
TRAINING		
Optimizer	AdamW	
Adam Parameters (β_1, β_2)	(0.9, 0.999)	
Learning rate	1×10^{-6}	
Weight Decay	0.0	
Max Global Gradient Norm for Clipping	1.0	
Learning Rate Scheduler	Polynomial	
Warm Up	3% of training steps	
# Train Steps For MATH dataset	1000 steps (around 8 dataset epochs)	
# Train Steps For GSM8K dataset	650 steps (around 8 dataset epochs)	
GENERAL		
Maximum Response Length	1024 tokens	
Maximum Sequence Length for RhoMath 1.1B	2048 tokens	
Maximum Sequence Length for DeepSeekMath 7B	2500 tokens	
PPO		
# Responses per Prompt	8	Search Space: {8, 16, 32}
# Episodes per PPO Step	512	Search Space: {256, 512}
# Prompts per PPO Step	$512/8 = 64$	
Mini-batch Size	64	
# Inner epochs per PPO Step	2	Search Space: {1, 2}
Sampling Temperature	0.6	Search Space: {0.6, 0.8, 1.0}
Discount Factor γ	1.0	
GAE Parameter λ	1.0	Search Space: [0.95 – 1.0]
KL Penalty Coefficient β	1e-4	Search Space: {1e-1, 1e-2, 3e-3, 1e-4}
Policy Clipping Parameter ϵ	0.2	
Value Clipping Parameter ϵ'	0.2	

not only across the batch but also across data parallel ranks. This normalization step is applied consistently in both our PPO and VinePPO implementations.

- **Reward Normalization:** We follow [Ivison et al. \(2024\)](#) and do not normalize the rewards, as the reward structure in our task is already well-defined within the range of [0, 1]. Specifically, correct responses are assigned a reward of 1, while incorrect responses receive 0.
- **End-of-Sequence (EOS) Trick:** As detailed in [Appendix A](#), rewards are only applied at the final token of a response, which corresponds to the EOS token when the response is complete. For responses that exceed the maximum length, we truncate the response to the maximum length and apply the reward to the last token of the truncated sequence. We also experimented with penalizing truncated responses by assigning a negative reward (-1), but this did not lead to performance improvements.
- **Dropout Disabling:** During the RL tuning phase, we disable dropout across all models. This ensures that the log probabilities remain consistent between different forward passes, thereby avoiding stochastic effects that could hurt training stability.
- **Fixed KL Coefficient** We use a constant coefficient for the KL penalty. Although the original PPO implementation for finetuning language models ([Ziegler et al., 2019](#)) utilized an adaptive KL controller, more recent implementations typically do not use this approach ([Ouyang et al., 2022; Touvron et al., 2023; Xu et al., 2024](#)).

C.3. SFT Models

To ensure a systematic and reproducible evaluation, we create our SFT models π_{ref} by finetuning the *base pretrained LLMs* (as opposed to their “Instruct” version) on the training splits of the respective datasets. Specifically, we produce four distinct SFT models: two base LLM (DeepSeekMath 7B and RhoMath 1.1B) across two datasets (MATH and GSM8K). The base models are finetuned using the Adam optimizer without weight decay. We employ a learning rate warm-up over 6% of

Table 2: Summary of RLOO and GRPO hyperparameters used in the experiments.

Parameter	Value
TRAINING	
Optimizer	AdamW
Adam Parameters (β_1, β_2)	(0.9, 0.999)
Learning rate	1×10^{-6}
Weight Decay	0.0
Max Global Gradient Norm for Clipping	1.0
Learning Rate Scheduler	Polynomial
Warm Up	3% of training steps
# Train Steps For MATH dataset	1000 steps (around 8 dataset epochs)
# Train Steps For GSM8K dataset	650 steps (around 8 dataset epochs)
GENERAL	
Maximum Response Length	1024 tokens
Maximum Sequence Length for RhoMath 1.1B	2048 tokens
Maximum Sequence Length for DeepSeekMath 7B	2500 tokens
RL ALGORITHM	
# Responses per Prompt	8
# Episodes per PPO Step	512
# Prompts per PPO Step	$512/8 = 64$
Mini-batch Size	64
# Inner epochs per PPO Step	2
Sampling Temperature	0.6
Discount Factor γ	1.0
KL Penalty Coefficient β	3e-3
Policy Clipping Parameter ϵ	0.2
Search Space: {1e-2, 3e-3, 1e-3, 3e-4, 1e-4}	

the total training steps. Each model is trained for three epochs with a batch size of 64, and the best checkpoint is selected based on validation accuracy. For each SFT model, we conduct a hyperparameter sweep over learning rates in the range $\{1 \times 10^{-7}, 3 \times 10^{-7}, 1 \times 10^{-6}, 3 \times 10^{-6}, 1 \times 10^{-5}, 3 \times 10^{-5}, 8 \times 10^{-5}, 1 \times 10^{-4}\}$ to ensure optimal performance. We then use these SFT models as the initial checkpoint for training the methods mentioned in our paper.

C.4. Evaluation

We evaluate each method’s performance on the test sets of each dataset. For example, when we report that PPO achieves 42.8% accuracy on the MATH dataset for the DeepSeekMath 7B model, this means the PPO training was initialized with the SFT model specific to DeepSeekMath 7B on the MATH dataset, and accuracy was measured on the MATH test set. Our primary evaluation metric is accuracy, specifically Pass@1, which reflects the percentage of correctly answered problems on the first attempt. This metric is crucial because it represents a realistic user interaction, where the model is expected to deliver a correct answer without the need for multiple tries. For each evaluation, we sample a response from the model for a given prompt, using a maximum token length of 1024 and a temperature of 0.35. A response is considered correct if its final answer matches the ground truth final answer, as detailed in Appendix C.1. Furthermore, each accuracy score is averaged over 16 evaluation rounds, each conducted with different random seeds. This will ensure a robust and low variance assessment of model performance.

C.5. Other Baselines

GRPO (Shao et al., 2024) and RLOO (Ahmadian et al., 2024) GRPO replaces PPO’s value network with a policy gradient baseline computed from the average return of a group of responses to the same input. For each training question x , all algorithms generate G responses, yielding training trajectories $\tau_1, \tau_2, \dots, \tau_G \sim \pi(\cdot|x)$ with corresponding returns R_1, R_2, \dots, R_G . Note that in the case of GRPO, we need to have $G > 1$. Then, GRPO computes the empirical mean $\mu_x = \frac{1}{G} \sum_{i=1}^G R_i$ and standard deviation σ_x of these returns. For each trajectory τ_i , the advantage $A(s, a)$ for all state-action

Table 3: Summary of RestEM hyperparameters used in the experiments.

Parameter	Value
TRAINING	
Optimizer	AdamW
Adam Parameters (β_1, β_2)	(0.9, 0.999)
Learning rate	1×10^{-6}
Weight Decay	0.0
Max Global Gradient Norm for Clipping	1.0
Learning Rate Scheduler	Polynomial
Warm Up	3% of training steps
RESTEM	
# iterations	10
# Sampled Responses per Prompt	8
Sampling Temperature	0.6
Checkpoints every # iteration	500 step
Checkpoint Selection	until validation improves Search Space: {until validation improves, best validation}

Table 4: Summary of DPO-Positive hyperparameters used in the experiments.

Parameter	Value
TRAINING	
Optimizer	AdamW
Adam Parameters (β_1, β_2)	(0.9, 0.999)
Learning rate	1×10^{-6}
Weight Decay	0.0
Max Global Gradient Norm for Clipping	1.0
Learning Rate Scheduler	Polynomial
Warm Up	3% of training steps
DPO-POSITIVE	
# DPO- β	0.1 for MATH, 0.3 for GSM8K
# DPO-Positive- λ	50.
# Epochs	3
# Sampled Responses per Prompt	64
# Pairs per prompt	64
Sampling Temperature	0.6

pairs $(s, a) \in \tau_i$ is defined as:

$$A(s, a) = \frac{R_i - \mu_{\mathbf{x}}}{\sigma_{\mathbf{x}}}.$$

Notably, this introduces bias in policy gradient estimation because the return R_i of the current trajectory is used in computing its own baseline. RLOO addresses this bias by employing a leave-one-out strategy for baseline computation. Specifically, for each trajectory τ_i , the baseline is computed using the returns of all other trajectories in the group, excluding R_i . Let $\mu_{\mathbf{x}}^{(i)}$ denote the empirical mean of $\{R_j\}_{j \neq i}$. The advantage for all state-action pairs in τ_i is then computed as:

$$A(s, a) = R_i - \mu_{\mathbf{x}}^{(i)}.$$

This modification ensures that the baseline for each trajectory is independent of its own return, yielding an unbiased policy gradient estimate.

DPO⁺ (DPO-Positive) (Pal et al., 2024) The original DPO method has a failure mode when the edit distance between positive (correct) and negative (incorrect) responses is small. In these cases, the probability of both responses tends to decrease. This issue is especially common in reasoning and mathematical tasks, where multiple solution paths may involve similar equations or steps. Although DPO achieves its goal by reducing the probability of the incorrect response more

than the correct one, it ultimately still lowers the likelihood of generating the correct response. This undermines model performance, making it a failure mode despite partially fulfilling the DPO objective. (Pal et al., 2024; Hwang et al., 2024). While previous methods mitigated this issue by maintaining a high edit distance between positive and negative response pairs, DPO-Positive (Pal et al., 2024) addresses it more effectively. It introduces an additional term to the DPO objective, penalizing any reduction in the probability of the correct response below its probability under the reference model. This ensures that the correct response is not overly suppressed, even when the edit distance is small. The final objective of DPO-Positive is::

$$\begin{aligned} \mathcal{L}_{\text{DPO-Positive}}(\pi_\theta; \pi_{\text{ref}}) = & -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\log \sigma \left(\underbrace{\beta \left(\log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)} \right)}_{\text{DPO Original term}} \right. \right. \\ & \left. \left. - \lambda \cdot \max \left(0, \log \frac{\pi_{\text{ref}}(y_w|x)}{\pi_\theta(y_w|x)} \right) \right) \right] \end{aligned} \quad (18)$$

where λ is a hyperparameter controlling the weight of the additional term keeping the probabilities of correct responses high. We chose DPO-Positive as a baseline due to its strong performance in (Setlur et al., 2024).

RestEM (Singh et al., 2024) RestEM is an iterative method where, in each iteration, the base model is trained on correct, self-generated responses from the chosen checkpoint of the previous iteration. RestEM takes gradient steps to maximize this objective until the fine-tuned model’s accuracy drops on a validation split. The objective of the fine-tuning process is to maximize the log-likelihood of correct responses. Training the model with a maximum likelihood objective on correct responses is mathematically equivalent to training the model with REINFORCE (Sutton et al., 1999), without a baseline, where the entire response is treated as a single action. The reward is 1 when the response is correct, and 0 otherwise. Specifically, we have:

$$\underbrace{\mathbb{E}_{\mathbf{x} \sim \mathcal{D}, \mathbf{y} \sim \pi(\cdot|\mathbf{x}), \mathcal{R}(\mathbf{x}; \mathbf{y})=1} [\nabla_\theta \log P_\theta(\mathbf{y}|\mathbf{x})]}_{\text{max log-likelihood on correct responses}} = \underbrace{\mathbb{E}_{\mathbf{x} \sim \mathcal{D}, \mathbf{y} \sim \pi(\cdot|\mathbf{x})} [\nabla_\theta \log P_\theta(\mathbf{y}|\mathbf{x}) \mathcal{R}(\mathbf{x}; \mathbf{y})]}_{\text{REINFORCE}} \quad (19)$$

Therefore, maximizing log-likelihood training on correct responses is equivalent to train with policy gradient without precise credit assignment, such as without advantages for specific actions. In our experiments, we observe the impact of this limitation in both Figure C.3 and Figure 5 where RestEM overfits on the training data.

C.6. Hyperparameters

In this section, we present a comprehensive overview of the hyperparameters used in our experiments. It’s important to note that the number of training episodes was carefully selected to ensure that the amount of training data remained consistent across all methods.

PPO Finetuning LLMs using PPO is known to be sensitive to hyperparameter selection, and finding the optimal settings is critical for achieving strong performance. To ensure the robustness of our study, we explored hyperparameter values reported in recent studies (Shao et al., 2024; Zheng et al., 2023; Ivison et al., 2024; Huang et al., 2024) and conducted various sweeps across a wide range of values to identify the best configuration for our tasks and models. Specifically, we find the set of hyperparameters that perform best across both MATH and GSM8K using RhoMath 1.1B model. Then, we employ the optimal set of parameters for the rest of our experiments. The full set of hyperparameters, along with their respective search spaces, is detailed in Table 1.

VinePPO We utilized the same hyperparameter setup as in the PPO implementation (Table 1) for VinePPO.

RLOO and GRPO Since policy optimization in RLOO and GRPO is similar to PPO, we initialize their hyperparameters from PPO. This not only ensure we start from a strong set of values, but also allows for a systematic comparison among these algorithms. We further tune their KL coefficient for stable training. Note that lack of credit assignment mechanism could lead to high variance policy gradient update, resulting in unstable training (Greensmith et al., 2001). See Table 2 for the full list.

Table 5: Average time spent per each training step for different methods and models measured for MATH dataset

Method	Model	Hardware	Average Training Step Time (s)
PPO	RhoMath 1.1B	4 × Nvidia A100 80GB	80
VinePPO	RhoMath 1.1B	4 × Nvidia A100 80GB	380
PPO	DeepSeekMath 7B	8 × Nvidia H100 80GB	312
VinePPO	DeepSeekMath 7B	8 × Nvidia H100 80GB	583

RestEM To ensure fair comparison we equalize the number of sampled responses for training between our RestEM run and our PPO runs. Therefore, in each RestEM iteration we sample 8 responses per prompt and train for 8 epochs on the correct responses. To enhance RestEM’s performance, we also conducted a sweep of other reasonable parameters(Table 3). This included increasing the number of samples to expand the training data and reducing the number of correct responses per question to minimize overfitting. However, we observed no significant improvement .

DPO⁺ (DPO-Positive) We adopted the same hyperparameters as those used by Setlur et al. (2024). In addition, we conducted a search for the optimal value of β to see if using the same β as in our PPO experiments would yield better performance than the values they recommended. To maintain a fair comparison, we ensured that the number of training samples in our DPO⁺ runs matched those in our PPO run where we trained for eight epochs, with each epoch consisting of training on eight responses per question. To match this, we generated 64 pairs of positive and negative responses given 64 self-generated responses from the base model. (Table 4)

C.7. Compute

All experiments were conducted using multi-GPU training to efficiently handle the computational demands of large-scale models. For the RhoMath 1.1B model, we utilized a node with $4 \times$ Nvidia A100 80GB GPUs to train both PPO and VinePPO. For the larger DeepSeekMath 7B model, we employed a more powerful setup, using a node with $8 \times$ Nvidia H100 80GB GPUs. Additionally, for training DeepSeekMath 7B models with the RestEM approach, we used a node with $4 \times$ Nvidia A100 80GB GPUs. The average training step time for each method on the MATH dataset is presented in Table 5.

C.8. Software Stack

Both PPO and VinePPO require a robust and efficient implementation. For model implementation, we utilize the Huggingface library. Training is carried out using the DeepSpeed distributed training library, which offers efficient multi-GPU support. Specifically, we employ DeepSpeed ZeRO stage 0 (vanilla data parallelism) for RhoMath 1.1B and ZeRO stage 2 (shared optimizer states and gradients across GPUs) for DeepSeekMath 7B . For trajectory sampling during RL training, we rely on the vLLM library (Kwon et al., 2023), which provides optimized inference for LLMs. Additionally, VinePPO leverages vLLM to generate Monte Carlo samples for value estimation. Specifically, after each RL training iteration, the current policy’s checkpoint is loaded into vLLM. Then, we use vLLM’s serving API to sample new trajectories and also Monte Carlo Samples for VinePPO’s value estimation. In our setup, we spawn a separate vLLM engine on each GPU rank. This would allow for data parallelism during both sample generation and training. This software stack ensures that our experiments are both efficient and reproducible. For instance, during VinePPO training, we achieve an inference speed of up to 30K tokens per second using $8 \times$ Nvidia H100 GPUs with the DeepSeekMath 7B model.

C.9. Reproducibility

In this study, all experiments were conducted using open-source libraries, publicly available datasets, and open-weight LLMs. To ensure full reproducibility, we will release both Singularity and Docker containers, equipped with all dependencies and libraries, enabling our experiments to be run on any machine equipped with NVIDIA GPUs, now or in the future. Additionally, we will make our codebase publicly available on GitHub at <https://github.com/McGill-NLP/VinePPO>

D. Full Results

D.1. Training Plots

In this section, we present additional training plots for both PPO and VinePPO on the GSM8K dataset, as shown in Figure D.6. Figure D.7 further illustrates the trade-off between accuracy and KL divergence, while Figure D.9 highlights the computational efficiency of the models⁶.

We observe consistent patterns with the results reported in Section 6. Although the performance gap for the DeepSeekMath 7B model is narrower on GSM8K, VinePPO still higher accuracy with significantly lower KL divergence and faster per-iteration time (this happens because responses to GSM8K problems are typically shorter, making MC estimation quite fast).

D.2. Explained Variance and Mean Absolute Error (MAE) of Value Prediction During Training

To ensure healthy training runs, we assess value prediction accuracy using explained variance and mean absolute error (MAE). Explained variance quantifies how much of the variance in ground-truth values is captured by the estimator:

$$\text{ExplainedVariance} = 1 - \frac{\sum_{g=1}^n (v_g - \hat{v}_g)^2}{\sum_{g=1}^n (v_g - \bar{v})^2},$$

where v_g are ground-truth values, \hat{v}_g are predictions, and $\bar{v} = \frac{1}{n} \sum_{g=1}^n v_g$ is the mean of the ground-truth values. The mean absolute error (MAE) is given by:

$$\text{MAE} = \frac{1}{n} \sum_{g=1}^n |v_g - \hat{v}_g|.$$

As shown in Figure D.5, PPO shows improving explained variance and decreasing MAE, indicating stable training. VinePPO achieves the highest explained variance and lowest MAE. RLOO and GRPO are included solely for demonstration, illustrating the deviation of their baselines from ground truth value estimates.

D.3. KL Divergence

The RL objective (Equation 1) balances maximizing task performance while constraining deviations from the initial policy π_{ref} , measured by KL divergence. We analyze how VinePPO and PPO navigate this trade-off by plotting task accuracy against KL divergence $\text{KL}[\pi_\theta \| \pi_{\text{ref}}]$ throughout training (Figure D.8). Results show VinePPO consistently achieves higher accuracy at same KL divergence, indicating more efficient use of the “KL budget.” This efficiency stems from VinePPO’s more precise credit assignment. As shown in Figure 1, many advantages are zero, and VinePPO excludes these steps from the loss. By avoiding unnecessary updates on non-contributing tokens, VinePPO reduces non-essential parameter adjustments that would inflate KL.

D.4. Temperature Tolerance

Sampling temperature is a critical hyperparameter controlling the randomness of sampled trajectories. At higher temperatures models generates more diverse trajectories, accelerating early training through increased exploration. However, this diversity challenges PPO’s value network, requiring generalization over a wider range of states. We compared VinePPO and PPO using temperatures $T \in \{0.6, 0.8, 1.0\}$ over the initial third of training steps. Figure D.10 shows VinePPO consistently benefits from higher temperatures, achieving faster convergence. Conversely, PPO fails to benefit from increased exploration and even diverges at $T = 1.0$, where trajectories are most diverse.

D.5. Value Prediction Analysis

In this section, we provide additional plots for value analysis. Specifically, Figures D.13 to D.16 demonstrates these plots for on the MATH dataset, and Figures D.17 to D.20 on the GSM8K dataset.

Furthermore, we present the prediction error per step in Figures D.21 to D.24.

⁶For GSM8K runs of RhoMath 1.1B , different hardware was used, making direct comparison of wall-clock time not feasible.

E. More Examples of Advantages in VinePPO

In addition to Figure 1, we provide more examples of effective advantage computation of VinePPO in Figures E.25 to E.27.

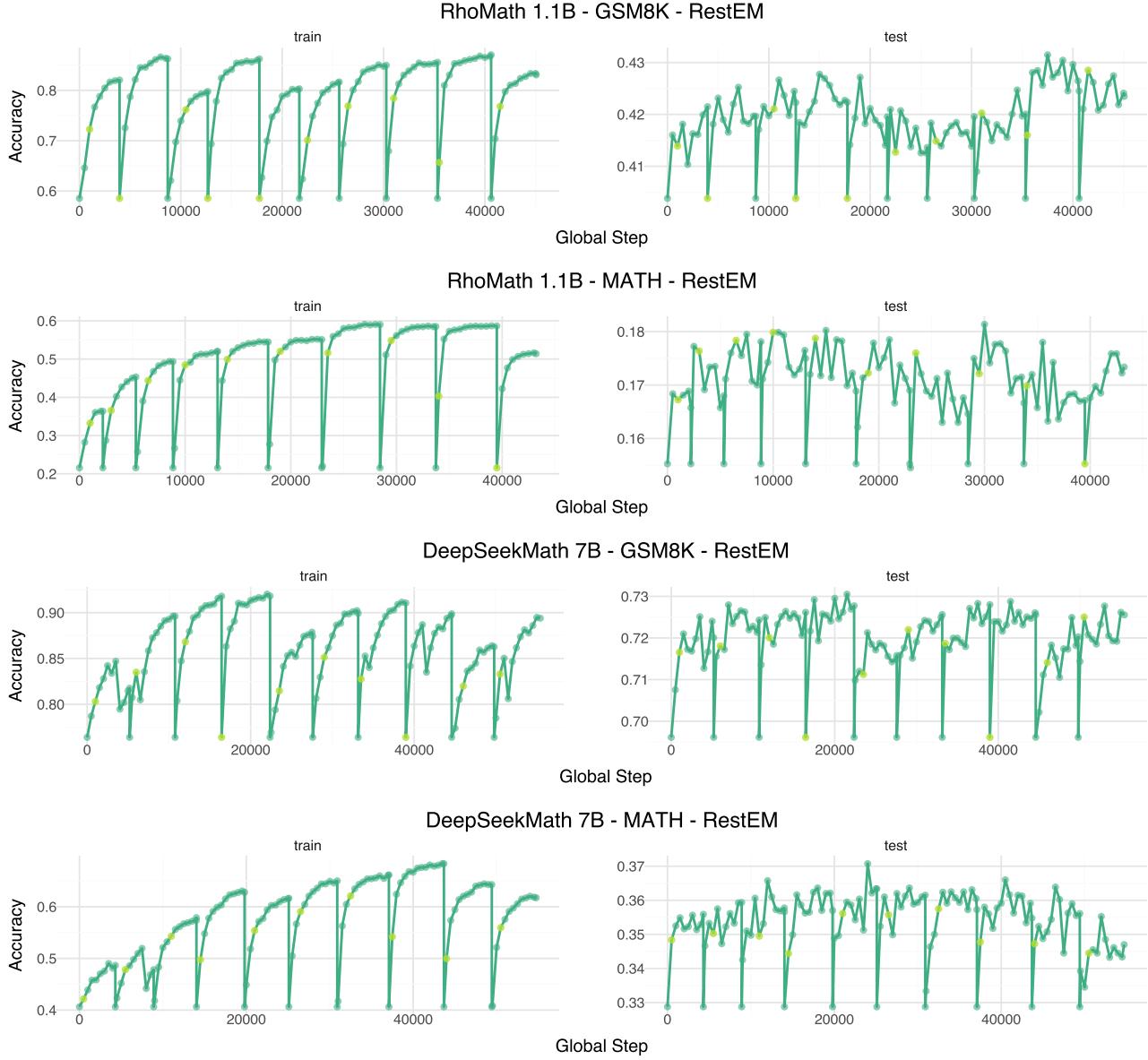


Figure C.3: Performance comparisons across different models and datasets: (a) RhoMath 1.1B on GSM8K, (b) RhoMath 1.1B on MATH, (c) DeepSeekMath 7B on GSM8K, and (d) DeepSeekMath 7B on MATH. The yellow points are chosen checkpoints based on the RestEM rule. Within each iteration, we train on the generated data of the chosen checkpoint for eight epochs and then we choose the first place where performance on a validation split drops following Singh et al. (2024)

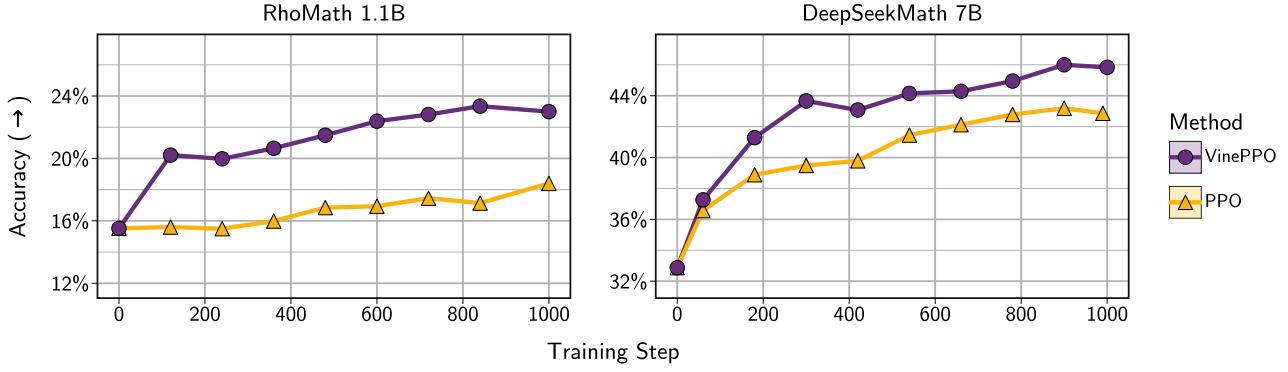


Figure C.4: Comparison of the training behavior between VinePPO and PPO. VinePPO demonstrates consistently higher accuracy (as measured on the test set of MATH dataset) throughout the training. Refer to Appendix D for more detailed plots.

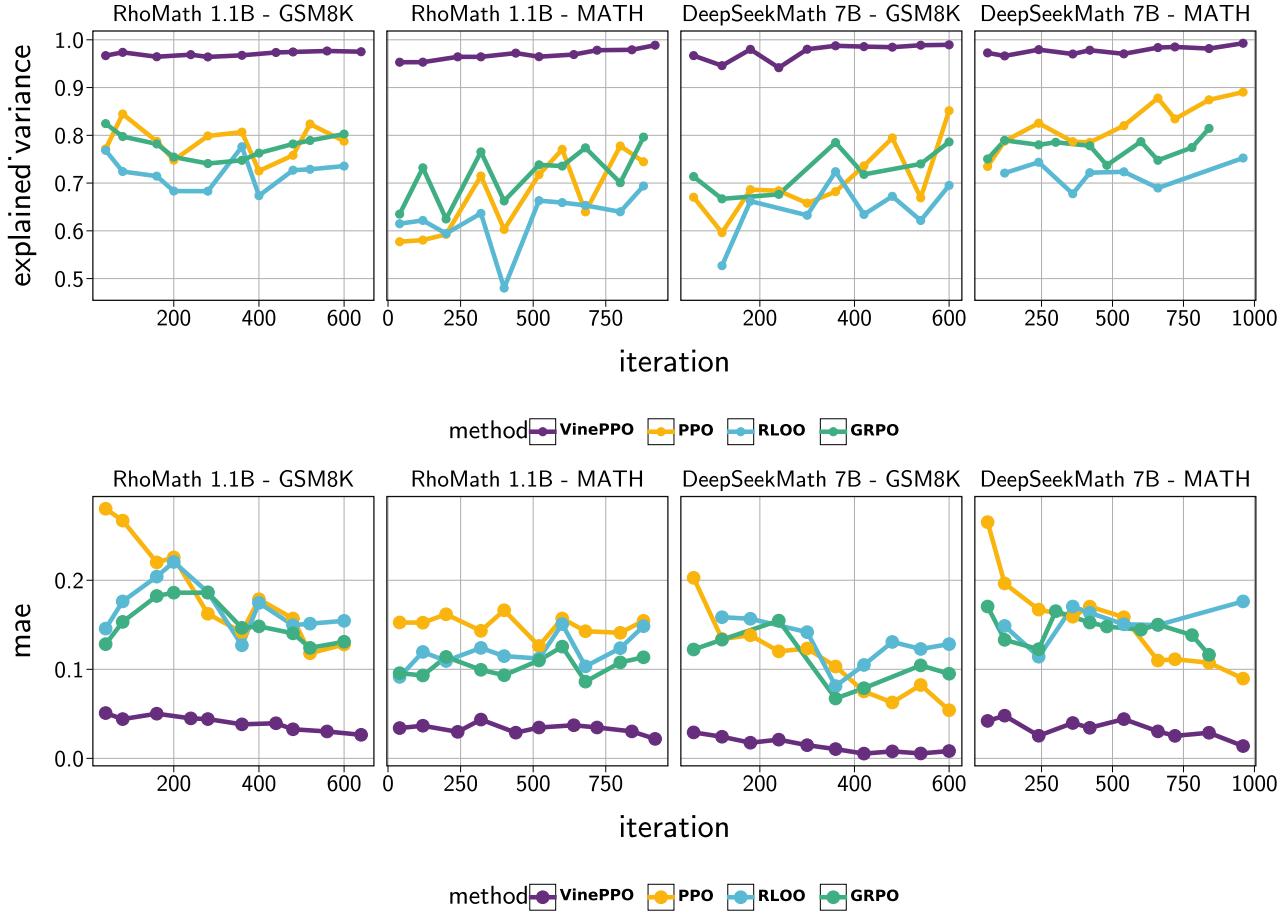


Figure D.5: **Explained Variance and Mean Absolute Error of values.** VinePPO demonstrates higher explained variance in value predictions and lower mean absolute error compared to RLOO, GRPO, and PPO across both datasets. Additionally, PPO's value predictions show non-negative explained variance values close to one, indicating stable and effective training. Note that RLOO and GRPO are included solely for demonstration, illustrating the deviation of their baselines from ground truth value estimates.

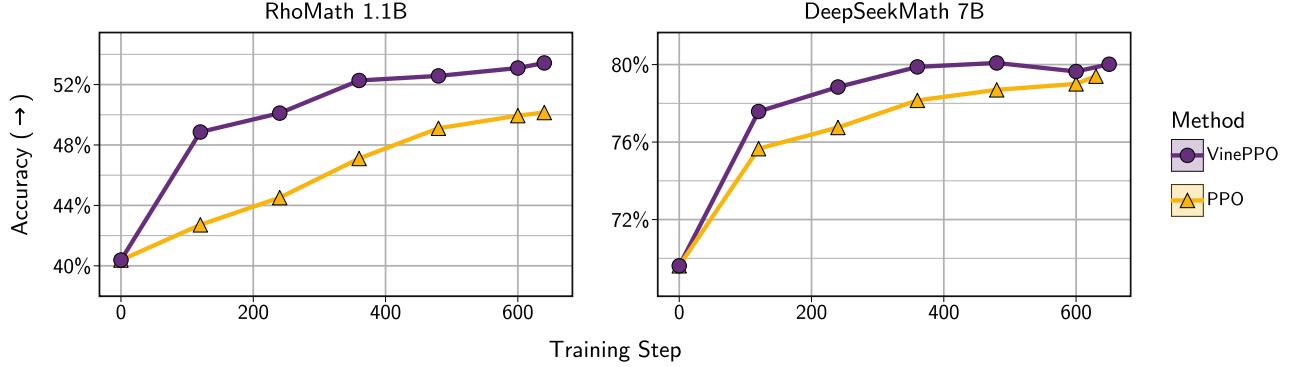


Figure D.6: Comparison of the training behavior between VinePPO and PPO. VinePPO demonstrates consistently higher accuracy throughout the training on the GSM8K dataset. Refer to Figure C.4 for MATH dataset.

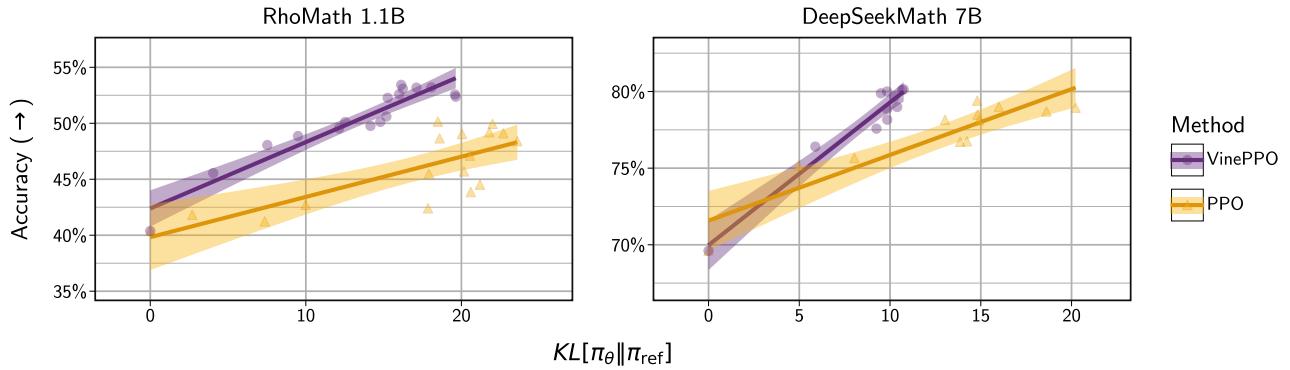


Figure D.7: Task accuracy as a function of KL divergence during training on the GSM8K dataset. VinePPO significantly higher accuracy per KL. Refer to Figure D.8 for MATH dataset.

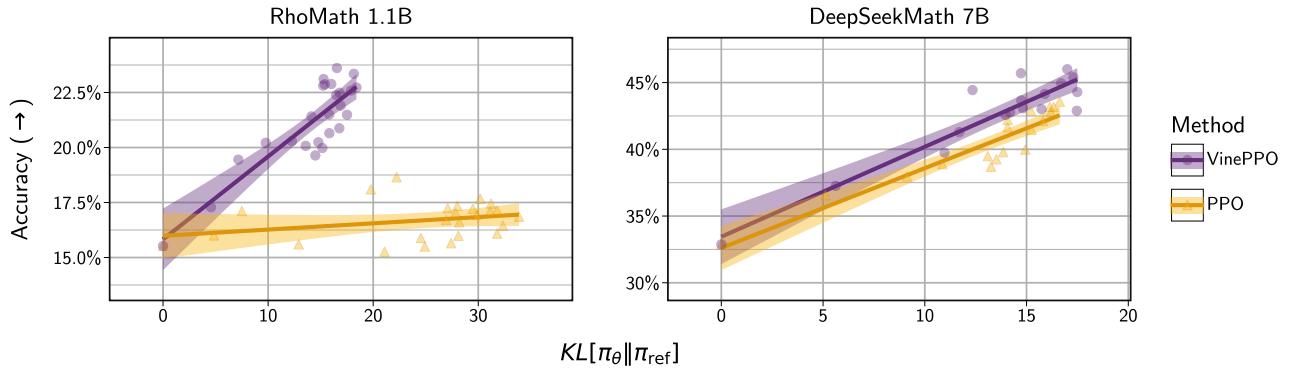


Figure D.8: Task accuracy as a function of KL divergence during training on the MATH dataset. VinePPO achieves higher accuracy, reflecting more efficient credit assignment and focused updates.

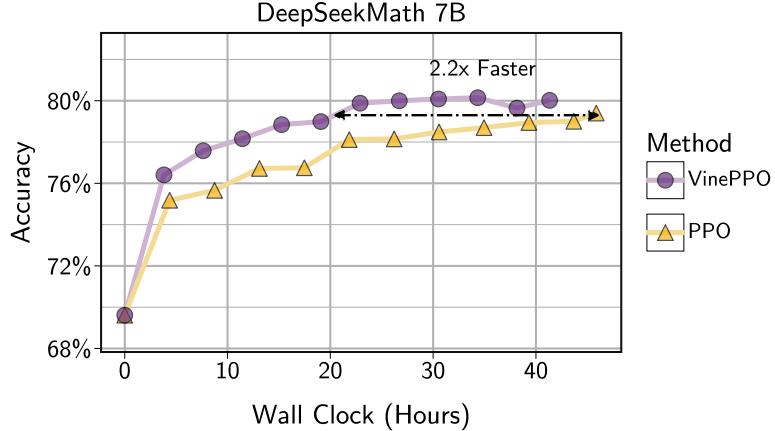


Figure D.9: Accuracy vs. Wall Clock Time for both methods measured on the same hardware throughout the entire training. Since the responses to GSM8K problems are short, VinePPO is even faster per-iteration in our setup and it reaches PPO’s peak performance in fewer iterations and less overall time.

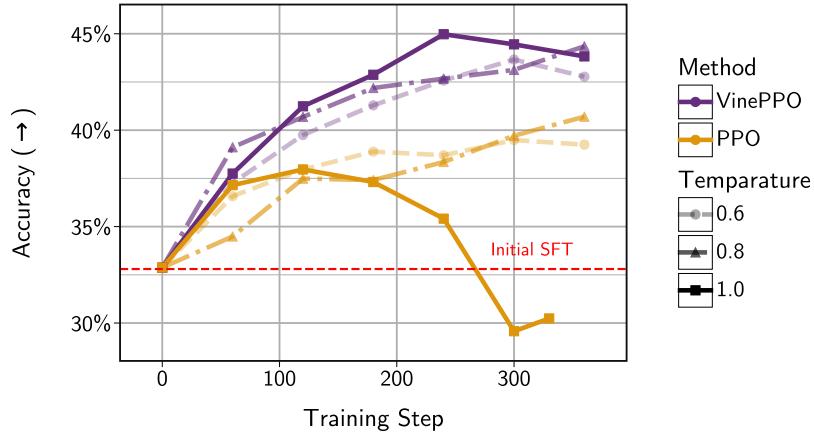


Figure D.10: Test set accuracy during training with higher temperature presented for DeepSeekMath 7B and MATH dataset. VinePPO can tolerate higher temperatures.

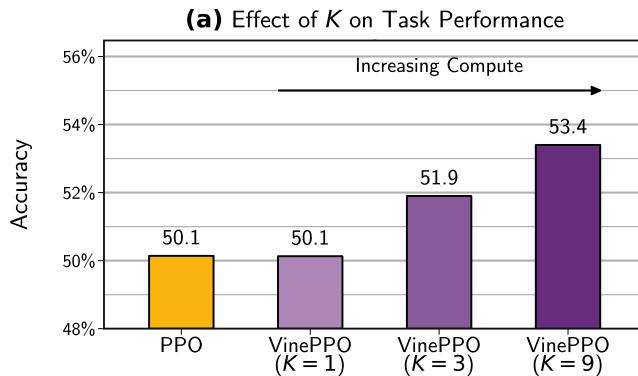


Figure D.11: Ablating the number of auxiliary trajectories K for estimating $\hat{V}_{MC}(s_t)$ on RhoMath 1.1B and GSM8K. Increasing K consistently improves task performance. (see Figure 7 for MATH dataset)

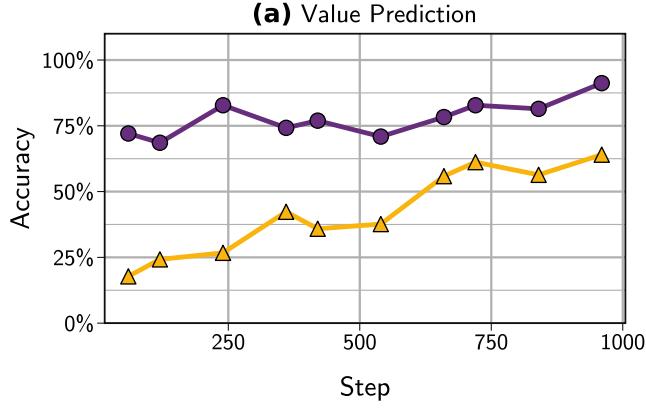


Figure D.12: Value prediction accuracy formulated as a classification problem, where a prediction is considered correct if it falls within 0.05 of the ground truth.

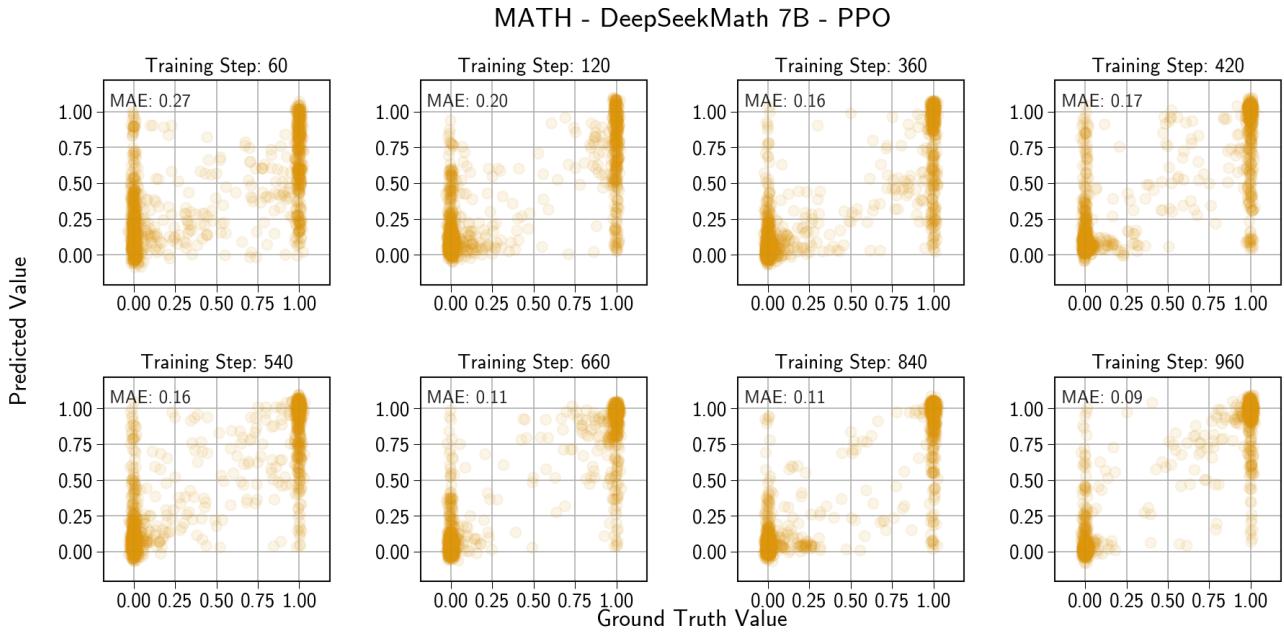


Figure D.13: Distribution of predicted values for each state vs. ground truth (computed using 256 MC samples) during training. MAE denotes the Mean Absolute Error (MAE).

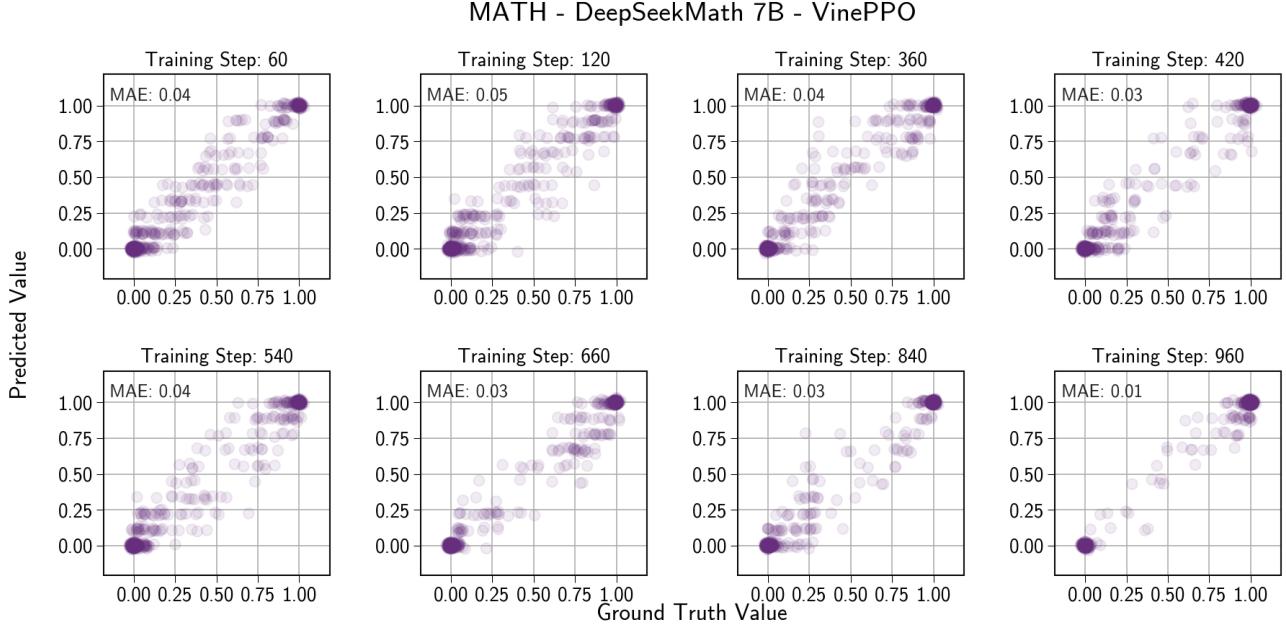


Figure D.14: Distribution of predicted values for each state vs. ground truth (computed using 256 MC samples) during training. MAE denotes the Mean Absolute Error (MAE).

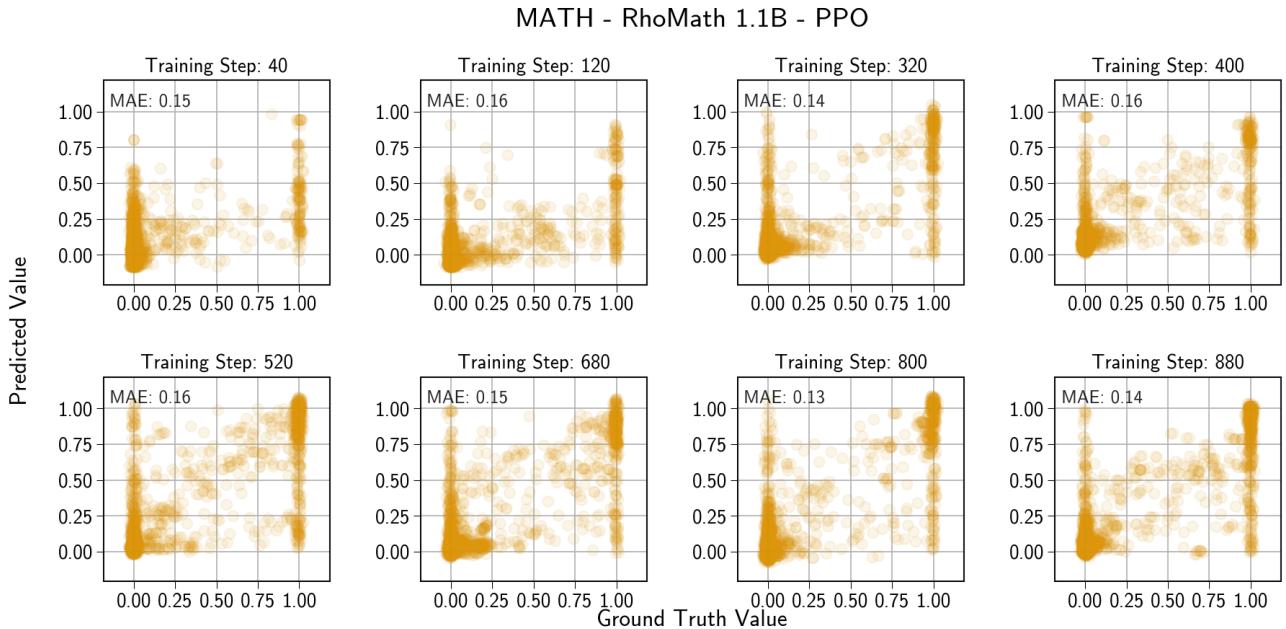


Figure D.15: Distribution of predicted values for each state vs. ground truth (computed using 256 MC samples) during training. MAE denotes the Mean Absolute Error (MAE).

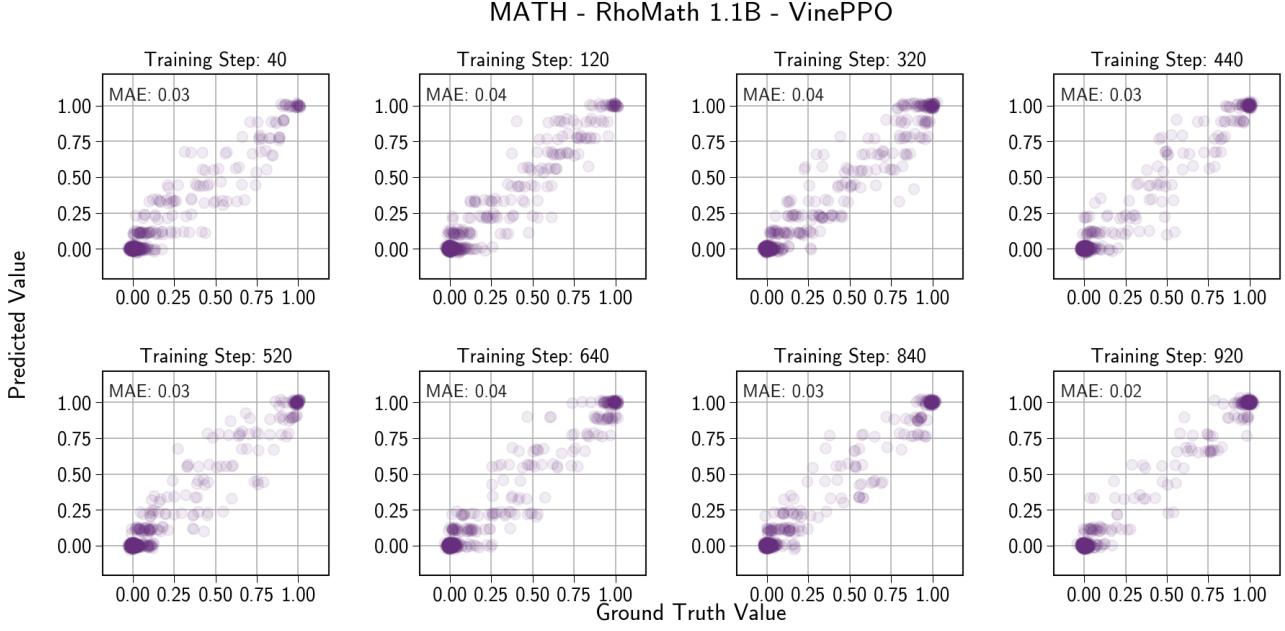


Figure D.16: Distribution of predicted values for each state vs. ground truth (computed using 256 MC samples) during training. MAE denotes the Mean Absolute Error (MAE).

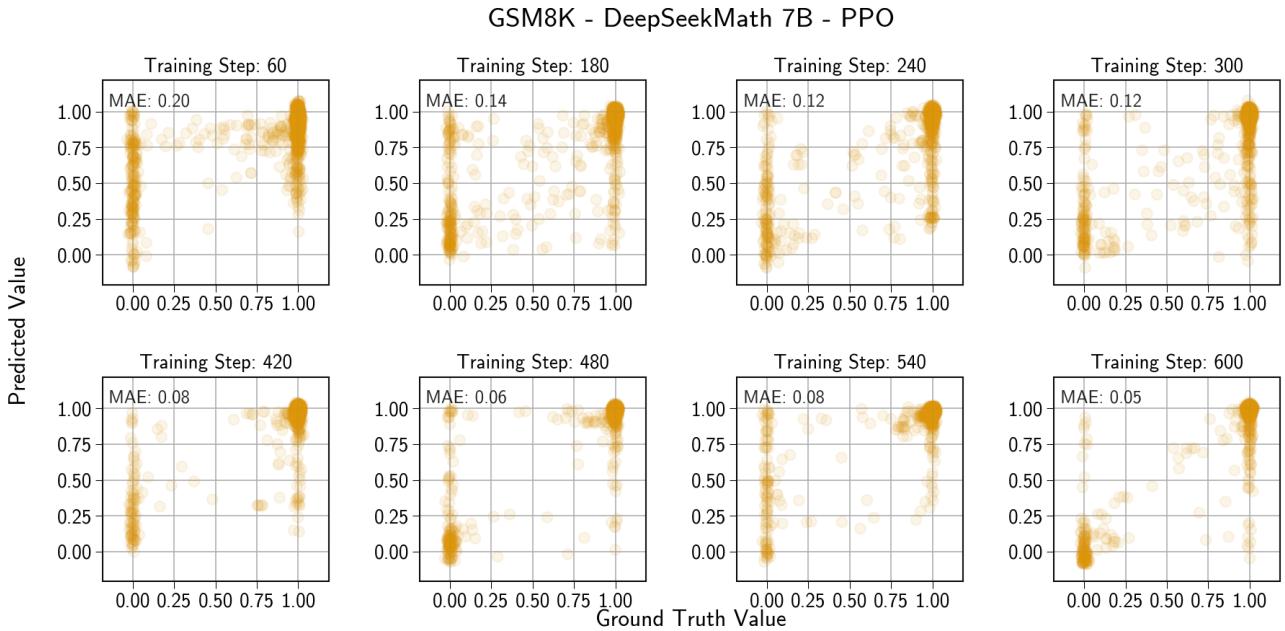


Figure D.17: Distribution of predicted values for each state vs. ground truth (computed using 256 MC samples) during training. MAE denotes the Mean Absolute Error (MAE).

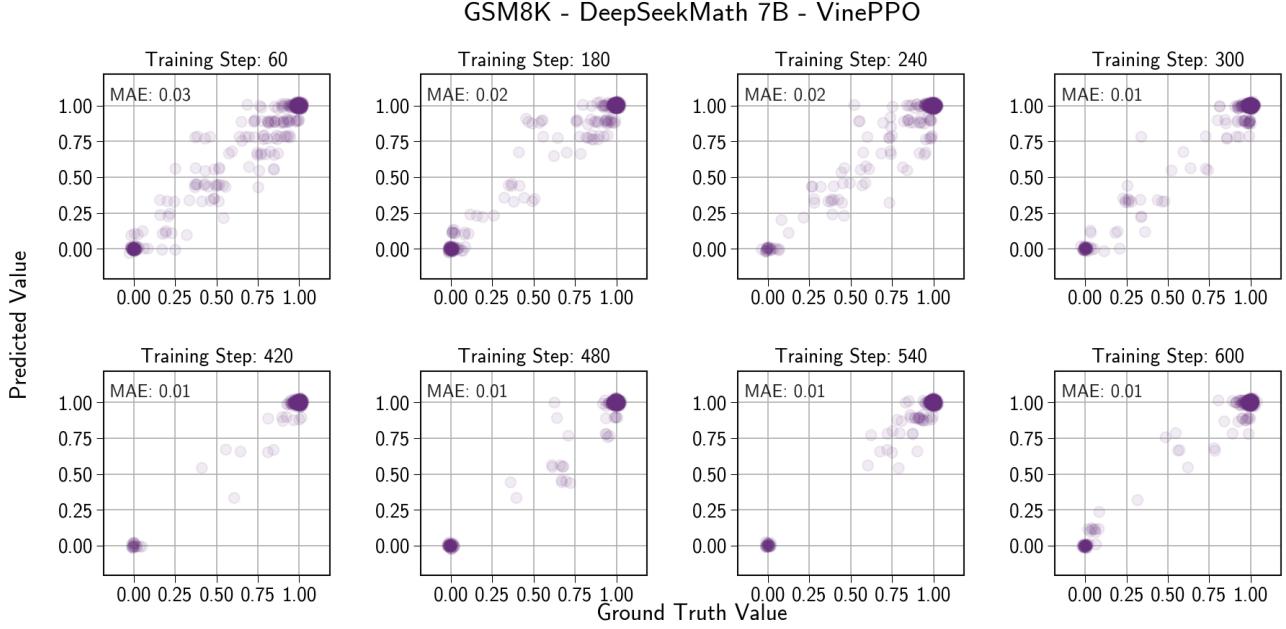


Figure D.18: Distribution of predicted values for each state vs. ground truth (computed using 256 MC samples) during training. MAE denotes the Mean Absolute Error (MAE).

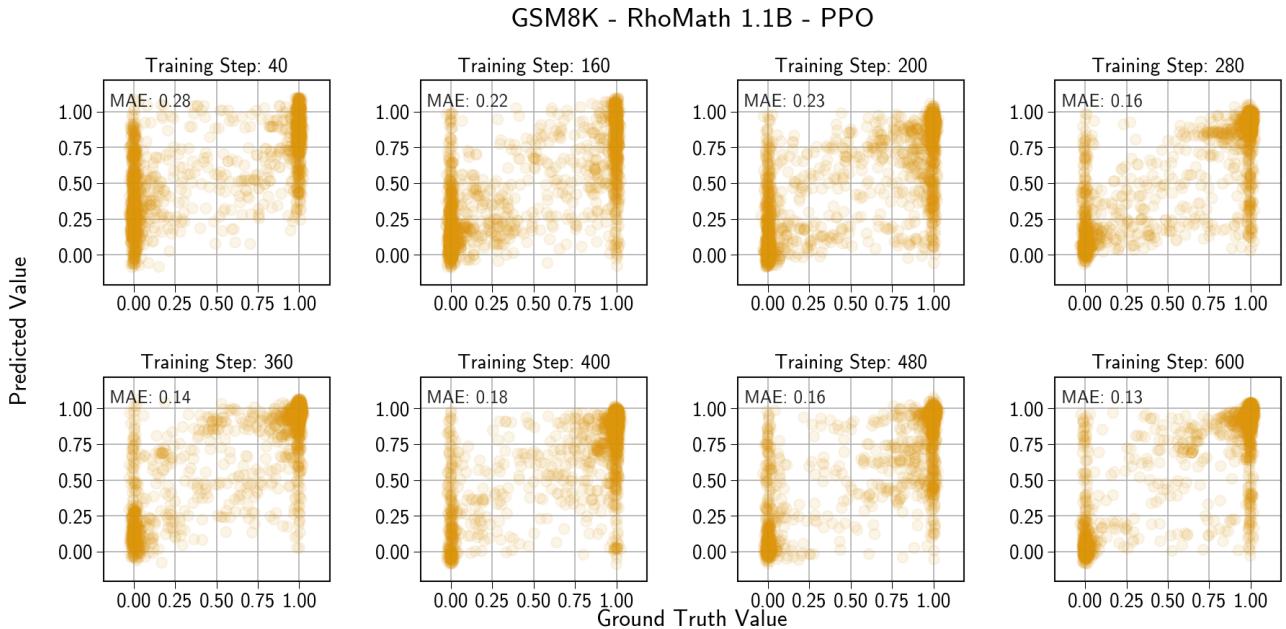


Figure D.19: Distribution of predicted values for each state vs. ground truth (computed using 256 MC samples) during training. MAE denotes the Mean Absolute Error (MAE).

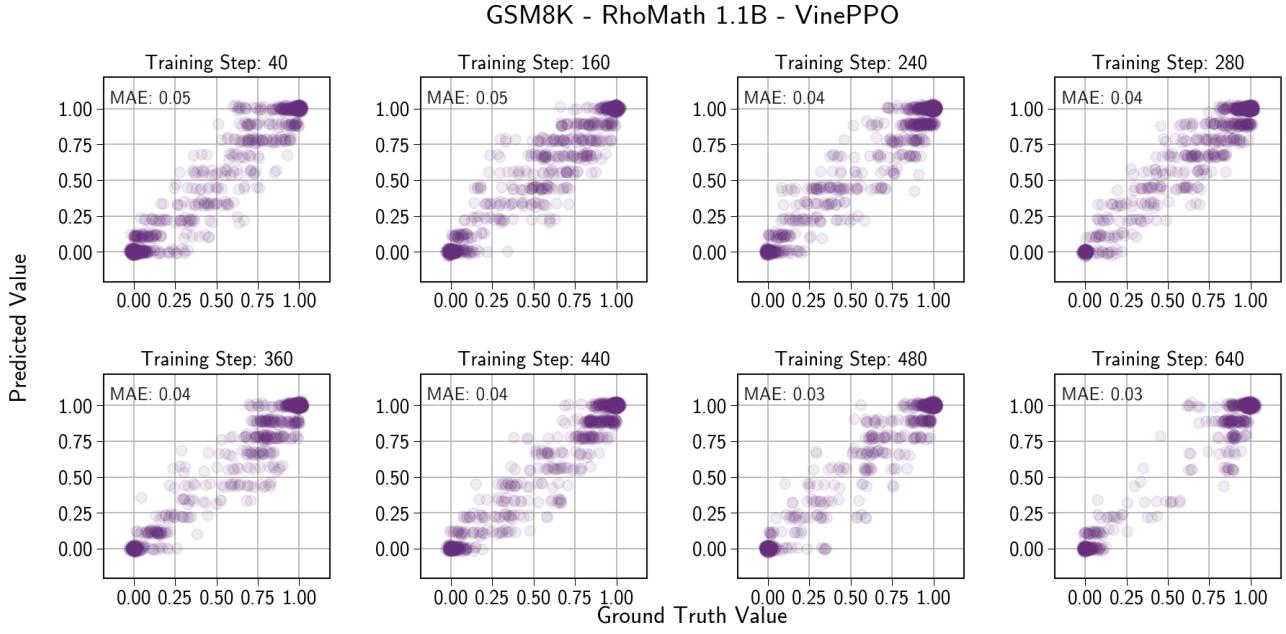


Figure D.20: Distribution of predicted values for each state vs. ground truth (computed using 256 MC samples) during training. MAE denotes the Mean Absolute Error (MAE).

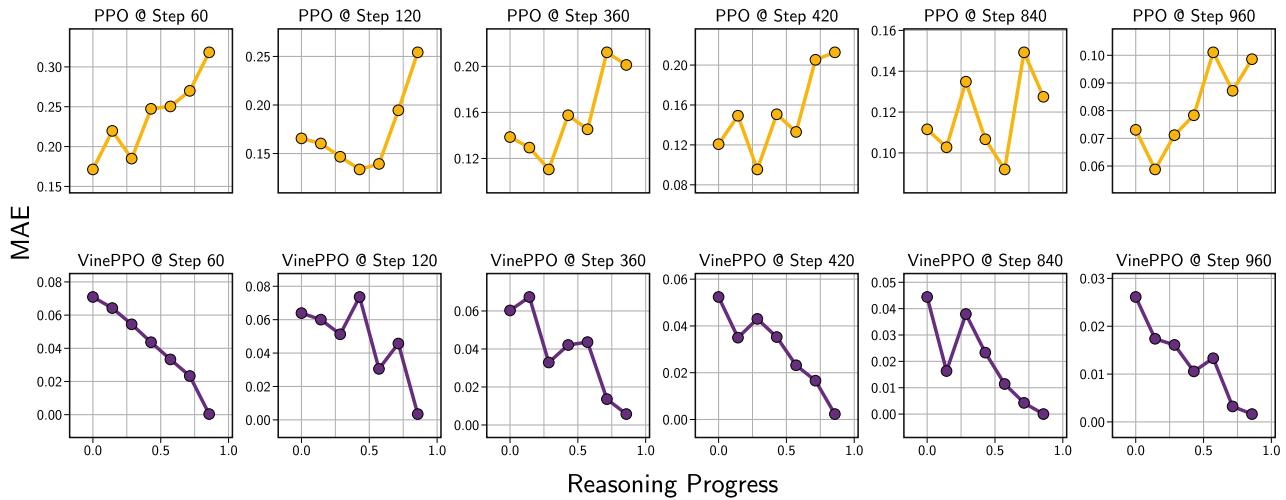


Figure D.21: Visualizing the Mean Absolute Error (MAE) of the value predictions in different point of reasoning chain, plotted for DeepSeekMath 7B on MATH dataset.

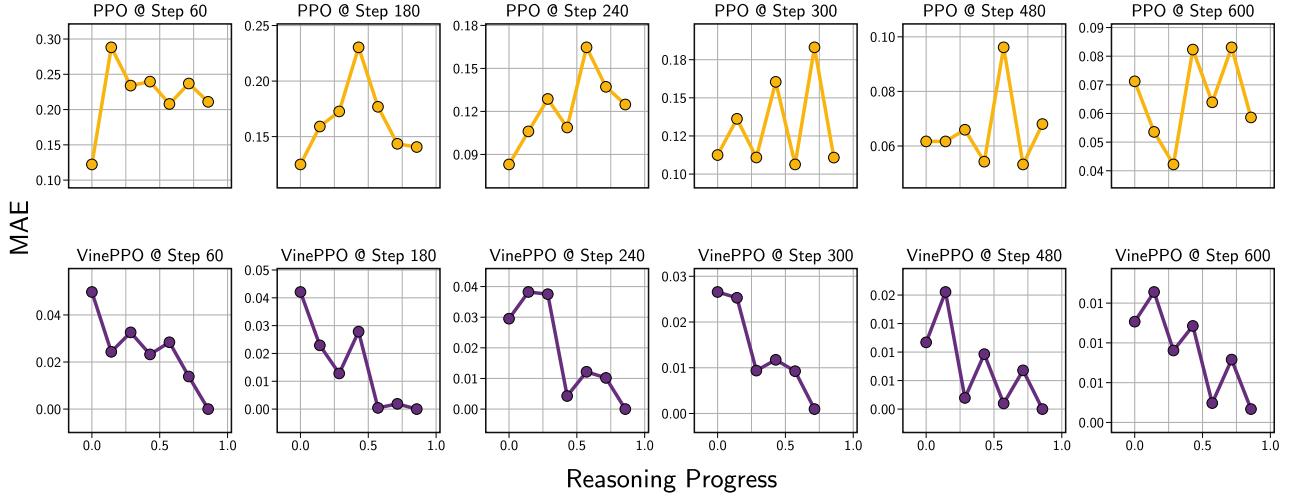


Figure D.22: Visualizing the Mean Absolute Error (MAE) of the value predictions in different point of reasoning chain, plotted for DeepSeekMath 7B on GSM8K dataset.

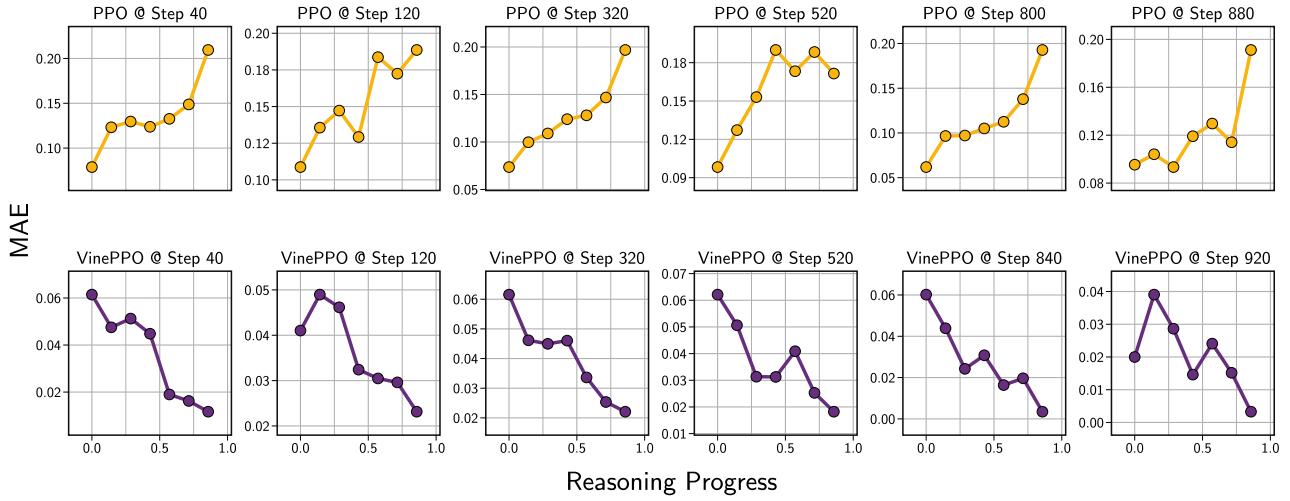


Figure D.23: Visualizing the Mean Absolute Error (MAE) of the value predictions in different point of reasoning chain, plotted for RhoMath 1.1B on MATH dataset.

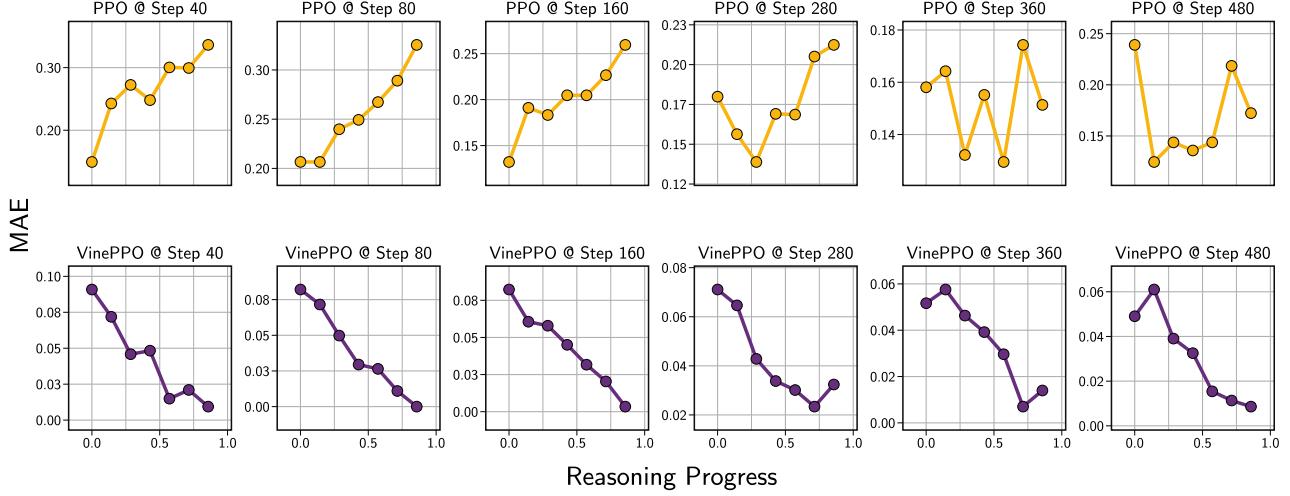


Figure D.24: Visualizing the Mean Absolute Error (MAE) of the value predictions in different point of reasoning chain, plotted for RhoMath 1.1B on GSM8K dataset.

Prompt (s_0)	Advantage
The vertex of the parabola described by the equation $y = -3x^2 - 30x - 81$ is (m, n) . What is n ?	
Response	
s_1 We can re-write the given equation as $y = -3(x^2 + 10x) - 81$.	0.0
s_2 To complete the square, we add $(-10/2)^2 = 25$ inside the parentheses:	0.2
s_3 $y = -3(x^2 + 10x + 25) - 81 - (-3)(25)$	0.5
s_4 $= -3(x + 5)^2 - 81 + 75$	0.0
s_5 $= -3(x + 5)^2 - 6$.	0.0
s_6 The vertex of a parabola in the form $a(x - h)^2 + k$ is at the point (h, k) ,	0.0
s_7 so the vertex of the given parabola is at the point $(-5, -6)$.	0.0
s_8 Therefore, $n = \boxed{-6}$.	0.0

Example of Advantages Computed by VinePPO

Figure E.25: Steps s_2 and s_3 provide a critical insight for advancing the solution since they provide a way to organize the equation.

Prompt (s_0)	Advantage
Let $f(x) = 2x - 3$ and $g(f(x)) = 5 - 4x$. Find $g(4)$.	0.4
Response	
s_1 We know that $g(f(x)) = 5 - 4x$.	0.0
s_2 So, if we can figure out what $f(x)$ is when $x = 4$,	-0.6
s_3 we can plug that value into $g(f(x))$ to find $g(4)$.	0.0
s_4 Since $f(x) = 2x - 3$, we have $f(4) = 2(4) - 3 = 5$.	0.0
s_5 Therefore, $g(f(4)) = g(5) = 5 - 4(5) = 5 - 20 = \boxed{-15}$.	0.0

Example of Advantages Computed by VinePPO

Figure E.26: This example demonstrates an error in step s_2 where the correct step would be ...what x is when $f(x) = 4$. This mistake prevents the model to successfully solve the question after s_2 .

Prompt (s_0)	Advantage
Write the expression $\frac{4+3c}{7} + 2$ as a single fraction.	
Response	
s_1 We get a common denominator of 7:	0.0
s_2 $\frac{4+3c}{7} + 2 = \frac{4+3c}{7} + \frac{2 \cdot 7}{7}$	0.0
s_3 $= \frac{4+3c}{7} + \frac{14}{7}$	-0.3
s_4 $= \frac{4+3c+14}{7}$	0.4
s_5 $= \boxed{\frac{18+3c}{7}}$.	0.0

Example of Advantages Computed by VinePPO

Figure E.27: Steps s_3 and s_4 , despite impacting the probability of solving the question, they are not meaningfully incorrect or insightful.