ASYMMETRIC PROXIMAL POLICY OPTIMIZATION: MINI-CRITICS BOOST LLM REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning (RL) has become a central paradigm for post-training large language models (LLMs) to elicit stronger reasoning. Yet, most recent RL for LLMs (RL4LLM) methods avoid explicit critics, replacing them with average advantage baselines. This shift is largely pragmatic: conventional value functions are computationally expensive to train at LLM scale and often fail under sparse rewards and long reasoning horizons. We revisit this bottleneck from an architectural perspective and introduce Asymmetric Proximal Policy Optimization (`AsyPPO`), a simple and scalable framework that restores the critic's role while remaining efficient in large-model settings. `AsyPPO` employs a set of lightweight *mini-critics*, each trained on disjoint prompt shards. This design encourages diversity while preserving calibration, reducing value-estimation bias. Beyond robust estimation, `AsyPPO` leverages inter-critic uncertainty to refine the policy update: (i) masking advantages in states where critics agree and gradients add little learning signal, and (ii) filtering high-divergence states from entropy regularization, suppressing spurious exploration. Across multiple reasoning benchmarks, `AsyPPO` consistently improves learning stability and performance over strong baselines, e.g., GRPO, achieving performance gains of $> 6\%$ on `Qwen3-4b-Base` and about 3% on `Qwen3-8b-Base` and `Qwen3-14b-Base` over classic PPO. Such results highlight the importance of architectural innovations in critics for scalable, efficient algorithms.

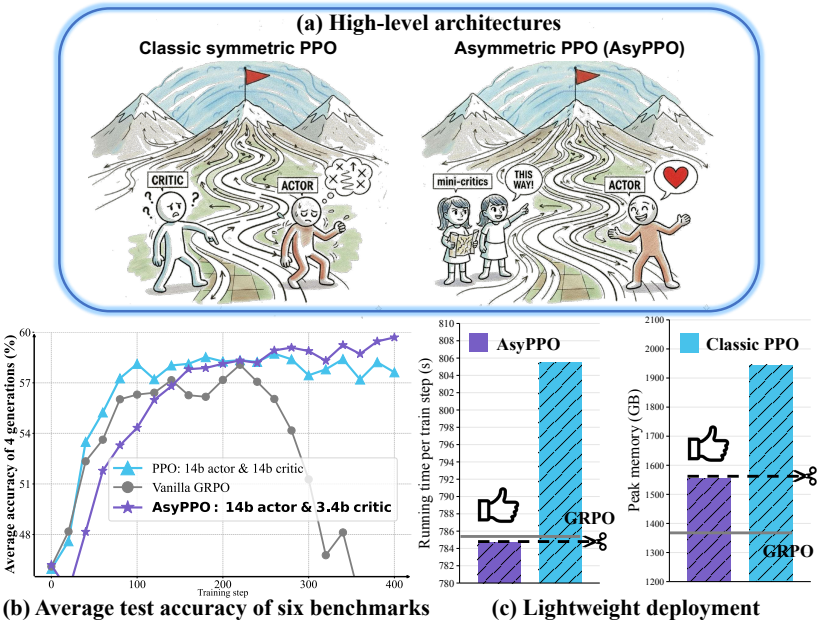


Figure 1: **(a):** The initial representational ability of the model makes asymmetric PPO possible, i.e., mini critics guide larger actors. By optimizing the ensemble critic system, `AsyPPO` achieves reliable value estimation while remaining lightweight. **(b):** Off-policy ratio=8, Report the average accuracy of 6 benchmarks, i.e., AIME 24, AIME 25, MATH-500, OlympiadBench, MinervaMath, and AMC 2023. **(c):** We set unified distributed training parameters for the algorithms on the same GPU cluster. The average clock time of the training step and the peak GPU memory usage of `AsyPPO` are significantly lower than those of the classic PPO, remain at the **GRPO level**.

1 INTRODUCTION

Proximal Policy Optimization (PPO) (Schulman et al., 2017) stands as one of the most powerful actor-critic algorithms in deep RL, and has demonstrated its potential across diverse domains such as computer games (Yu et al., 2022) and robotics control (Raj & Kos, 2024). In the realm of large-scale language models (LLMs), PPO has also proven transformative and has been widely applied in the post-training stage to stimulate the reasoning ability of LLMs (Hu et al., 2025). However, the transition from classical RL to RL4LLM introduces an unprecedented computational challenge, as LLMs operate at scales orders of magnitude larger than traditional RL environments. Directly applying PPO's default symmetric actor-critic design, where the critic is as large as the actor, creates significant computational overhead. In addition, training full critics at LLM scale is expensive and inaccurate under sparse, long-horizon rewards (Yuan et al., 2025b).

Faced with these challenges, the RL4LLM community has largely sidelined a key element of classical PPO – its critic. GRPO (He et al., 2025), and its variants, including GSPO (Zheng et al., 2025) in the Qwen series and DAPO (Yu et al., 2025), have achieved great success in replacing value functions with group sampling and average-advantage baselines for coarse-grained estimation of advantages. While effective, this paradigmatic shift abandons a key concept of RL: robust state value estimation can naturally mitigate training collapse caused by advantage bias (Wang et al., 2025b; Liu et al., 2024), especially under off-policy settings. This landscape motivates a fundamental reconsideration of architectural assumptions inherited from deep RL[1], prompting the following central question:

Can we achieve lightweight yet robust value estimation by redesigning PPO to depart from the standard symmetric actor–critic architecture, enabling stable and efficient learning?

To fill this research gap, we begin with a key insight: the initial rich representational ability inherited from pre-trained models significantly enhances the feasibility of the asymmetric actor-critic in the RL4LLM domain, unlike agents that learn from scratch in classical deep RL, painting a promising prospect for lightweight deployment and computational efficiency. Our initial experiments validate this hypothesis, where we find that a small critic, e.g., `Qwen3-0.6b-Base`, can indeed provide meaningful guidance to a much larger actor, e.g., `Qwen3-8b-Base`, demonstrating meaningful performance improvements over the base model. However, this asymmetric setup underperforms classical symmetric PPO, revealing limitations in a single small critic's value estimation capabilities.

To unlock the capabilities of the small critic, we consider critic ensembles to improve its value estimation and policy guidance. However, naive ensembles offer limited benefits for policy learning, as LLM critics start from identical pre-trained checkpoints with different heads only and are trained on the same data, leading to nearly identical behaviors that provide no corrective benefit. To tackle this critical challenge, we propose a simple yet effective non-overlapping data partitioning technique, in which each critic is trained via the subset formed by uniformly extracting responses from each prompt without overlap. This design encourages diversity among small critics and mitigates the risk of perception asynchrony among critics. Leveraging our ensemble-based value correction, small critics can provide reliable guidance to large policies despite their limited expressivity (Tint et al., 2024) (Figure 1(a)). Surprisingly, we find that double critic could be the sweet spot between correction capability and efficiency, it yields a qualitative leap in evaluation reliability while incurring the minimal redundancy needed for bias reduction. More critics increase the computation without proportional gains. Empirically, we demonstrate that two `Qwen3-1.7b-Base` critics robustly guide a larger policy, e.g., `Qwen3-14b-Base`, reducing critic over-parameterization while outperforming symmetric PPO under off-policy setting (Figure 1(b)). Notably, asymmetric architecture reduces peak memory by 20%, and accelerates training by around 20 seconds per step (see Figure 1(c)).

We further discovered that the agreement and divergence patterns in value estimates between our double critics, measured by their standard deviation, provide a useful signal for refining the policy loss objective. Value-estimation heterogeneity reflects both uncertainty and informativeness of the states. Leveraging this, we mask advantage values in states where critics strongly agree, reducing overfitting to low-quality samples and improving training stability. Conversely, we exploit divergence across critics by filtering out uncertain states from entropy regularization, since such states often correspond to low-probability continuations or spurious, reasoning-irrelevant patterns that inject

[1]Please refer to Appendix A for detailed related work.

noise into entropy measurements (Ahmed et al., 2019). Thus, restricting entropy regularization to high-confidence states promotes safer exploration and improves performance.

Overall, we refer to the above components as Asymmetric Proximal Policy Optimization (AsyPPO). The contributions of AsyPPO can be summarized in three main aspects:

1. **Robust Estimation:** Prompt-level data partitioning enhances ensemble reliability and yields consistent performance improvements. (§3.1)
2. **Lightweight Architecture:** The asymmetric design mitigates critic over-parameterization and opens a new direction for RL4LLM. (§3.1)
3. **Objective Refinement:** We introduce two uncertainty-aware modifications to the PPO objective that improve sample efficiency and enable safer exploration. (§3.2)

2 PRELIMINARIES

Proximal Policy Optimization (PPO). PPO (Schulman et al., 2017) is a widely used actor-critic algorithm in the policy gradient family. It improves the stability by optimizing a *clipped surrogate objective*, which limits how much the updated policy π_θ can deviate from the old policy $\pi_{\theta_{\text{old}}}$ at each update step. The objective is defined as:

$$\mathcal{J}_{\text{PPO}}(\theta) = \mathbb{E}_{\left[q\sim P(Q),\ o\sim\pi_{\theta_{\text{old}}}(O|q)\right]} \frac{1}{|o|}\sum_{t=1}^{|o|}\min\left(\frac{\pi_\theta(o_t|q,o_{<t})}{\pi_{\theta_{\text{old}}}(o_t|q,o_{<t})}A_t,\ \text{clip}\left(\frac{\pi_\theta(o_t|q,o_{<t})}{\pi_{\theta_{\text{old}}}(o_t|q,o_{<t})},\ 1-\epsilon,\ 1+\epsilon\right)A_t\right), \tag{1}$$

where π_θ and $\pi_{\theta_{\text{old}}}$ denote the current and previous policy, respectively. Here q is a sampled *question* and o the generated *output sequence*,, with o_t the t-th token. ϵ is the clipping hyperparameter that constrains the update ratio. A_t is the advantage estimate at step t, typically computed with Generalized Advantage Estimation (GAE) (Schulman et al., 2015).

Generalized Advantage Estimation (GAE). GAE addresses the bias–variance trade-off in advantage estimation by combining multi-step returns with exponentially decaying weights:

$$\hat{A}_t^{\text{GAE}(\gamma,\lambda)} = \sum_{l=0}^{\infty}(\gamma\lambda)^l\,\delta_{t+l}, \qquad \delta_t = r_t + \gamma V(s_{t+1}) - V(s_t).$$

Here $V(s)$ is the value function, $\gamma \in [0,1]$ is the discount factor, and $\lambda \in [0,1]$ is the GAE parameter that balances bias and variance. Setting $\lambda = 0$ recovers the low-variance, high-bias $TD(0)$ estimator, while $\lambda = 1$ corresponds to the high-variance, low-bias Monte Carlo estimator. In practice, PPO leverages GAE together with the clipped objective, yielding stable training and improved sample efficiency. The choice of γ and λ critically influences the temporal horizon and smoothness of the advantage estimates, and thus the convergence of the policy.

3 ASYMMETRIC PROXIMAL POLICY OPTIMIZATION

We begin by empirically examining the potential of the asymmetric actor-critic framework while highlighting the limitations of naive ensemble critics in LLM reasoning. By analyzing key differences between classical deep RL and RL4LLM, we propose a group-level non-overlapping data division strategy that enables lightweight mini-critics to provide reliable value estimation (§3.1). Building on this, we investigate the role of divergence and agreement among the mini-critics and find that uncertainty in their value estimates carries strong representational power for measuring sample quality. Leveraging this insight, we incorporate value uncertainty as a signal into the policy optimization objective, reformulating the loss function and refining the entropy regularization to improve sample efficiency and exploration capability of the policy (§3.2).

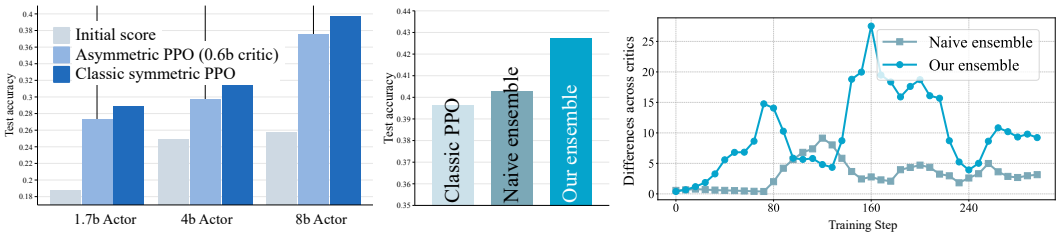


Figure 2: **Left:** The single mini-critic parameterized by `Qwen3-0.6b-Base` can effectively guide policies across model scales. **Middle:** There are significant differences in the guiding ability of the two ensemble critics for policies. Actors uniformly use `Qwen3-8B-Base`, while critics use `Qwen3-0.6B-Base`. **Right:** Our ensemble method intensifies the cognitive differences among mini-critics. The y-axis represents the standard deviation between the values calculated by the two mini-critics. We train on 5,000 questions sampled from DeepMath-103K (He et al., 2025) and evaluate policies on five challenging math benchmarks: AIME 2024, MATH-500, OlympiadBench, MinervaMath, and AMC 2023. For each question, we report the average of 4 generations.

3.1 Towards Lightweight Value Estimation

In LLM reasoning, the policy inherits expressive capabilities from the pre-trained model at initialization. As shown in Figure 2 (Left), even without critic warm-up, a small critic, i.e., `Qwen3-0.6B-Base` (Yang et al., 2025), can provide useful guidance, demonstrating the potential of an asymmetric architecture. However, due to sparse rewards and the small critic's limited familiarity with long-tail reasoning trajectories favored by larger models (Li et al., 2025), its value estimates are often inaccurate. This leads to suboptimal policy guidance compared to symmetric PPO.

Starting from the ensemble system. To strengthen mini-critic perceptual capacity, we first adopt an ensemble of critics, a standard technique in classical deep RL for reducing estimation bias (Chen et al., 2021). In practice, we add a second critic based on the same base model and average their predictions for value estimation. These corrected values are then used in advantage computation via GAE. However, as Figure 2 (Middle) shows, this naive ensemble approach yields limited improvement. The reason becomes clear in Figure 2 (Right), the two mini-critics exhibit nearly identical behavior, failing to provide the diversity that ensembles rely on. In classical RL, critics are initialized randomly, ensuring parameter diversity and differentiated value estimates, which is essential for ensemble effectiveness. By contrast, in RL4LLM, critics are typically initialized from the same pre-trained model, which accelerates learning but reduces diversity. This motivates a key question: *under homogeneous initialization, can ensemble critics be adapted to remain effective in LLM reasoning?*

Group level non-overlap data division. Beyond explicitly increasing parameter differences through initialization, another promising approach is to provide differentiated optimization signals for each critic during training. Intuitively, training critics on non-overlapping subsets of data encourages them to learn from distinct trajectories and reward distributions, steering their updates in different directions and promoting functional diversity. However, in practice, randomly partitioning the training data can lead to asynchronous perception at the prompt level, where critics encounter inconsistent reasoning patterns from different questions. This imbalance increases the risk of overfitting to specific response types, resulting in unstable discrepancies in value estimates. In extreme cases, such divergence may cause policy collapse. To mitigate this, we uniformly divide the data into disjoint subsets at the prompt level, ensuring that each critic receives an equal share of responses within every prompt (or group). This design maintains perceptual synchrony across critics within each question while creating differentiated rewards and

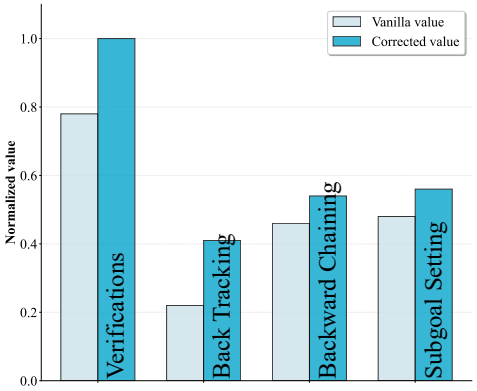


Figure 3: **Our ensemble critics achieve positive estimation of the state involving key reasoning pattens.** We follow Gandhi et al. (2025), identify the reasoning behavior via GPT4-o (Hurst et al., 2024), and hire trained `Qwen3-0.6b-Base` as mini-critics and `Qwen3-8b-Base` as the vanilla critic.

observations. Our ensemble critic system training process can then be formalized as:

$$\mathcal{L}_{\text{critic}}(\boldsymbol{\phi}) = \sum_{m=1}^{M} \mathcal{L}_{\text{critic}}^{(m)}(\phi_m) = \sum_{m=1}^{M} \mathbb{E}_{(s_t,R_t)\sim\mathcal{D}_m}\left[\left(V(s_t;\phi_m) - R_t\right)^2\right], \tag{2}$$

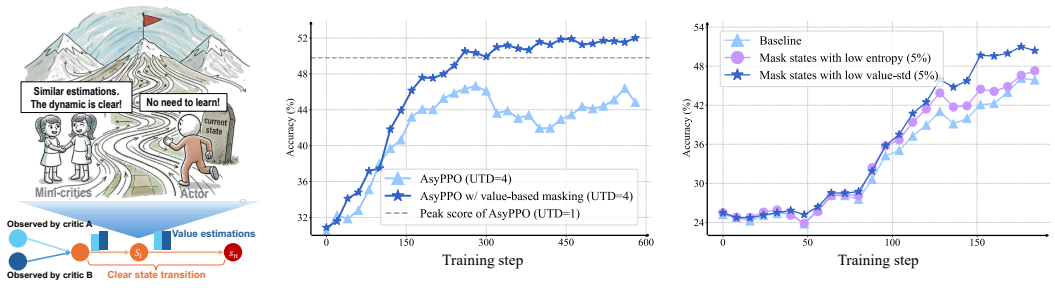
M denotes the number of mini-critic with parameters $\{\phi_m\}_{m=1}^{M}$. Each critic aim to fit the return R_t based on its assigned subset $\mathcal{D} = \bigcup_{m=1}^{M} \mathcal{D}_m, \mathcal{D}_i \cap \mathcal{D}_j = \emptyset$. Corrected advantage $\bar{A}$ can be obtained:

$$\bar{A}_t(\gamma,\lambda) = \sum_{l=0}^{T-t-1} (\gamma\lambda)^l \delta_{t+l},\ \delta_t = r_t + \gamma\bar{V}(s_{t+1}) - \bar{V}(s_t);\ \bar{V}(s_t) = \frac{1}{M}\sum_{m=1}^{M} V_m(s_t;\phi_m) \tag{3}$$

The results in Figure 2 (Middle, Right) demonstrate that critics trained under our ensemble strategy exhibit clearly differentiated behaviors. Statistical analysis from a linguistic perspective (Figure 3) reveals that the corrected values from our ensemble framework significantly encourage the policy to acquire core reasoning patterns. Overall, our method effectively unlocks the efficiency of asymmetric PPO and points to a promising new direction for RL4LLM algorithm design.

Takeaway 1

Optimizing the ensemble critic design enhances the learning capacity of the asymmetric actor–critic while significantly reducing computational overhead.



(a) Motivation of our advantage masking (b) Ablation study of our advantage masking (c) Comparison of two masking mechanisms

Figure 4: **(a):** Agreement among critics implies the state's downstream dynamics are well modeled by the policy, making these samples low-value for learning and best avoided for overfitting. **(b):** In the high data-reuse setting (UTD=4), masking the bottom 20% (by value-std) boosts AsyPPO's learning efficiency, **yields an improvement of about 6 points**. The accuracy records of the six benchmarks follow Figure 1 (b). **(c):** We evaluated two 5% masking mechanisms on vanilla AsyPPO (baseline), i.e., entropy vs. value-std. The value-std masking produced the strongest learning efficiency benefit. Actors use `Qwen3-4B-Base`, while critics use `Qwen3-0.6B-Base`.

3.2 Policy Loss Reconstruction

Beyond enabling robust value estimation, we conjecture that ensemble mini-critics can further enhance policy learning efficiency. Intuitively, the degree of agreement among critics' value estimates for a given state can serve as a meaningful signal for policy optimization. This insight arises from our analysis of value fitting dynamics (Lee et al., 2021): when critics produce similar value estimates for a state s_i, it often indicates that s_i is **low-informative**. Such states are frequently encountered across trajectories, and the rewards they yield exhibit low variance, causing critics to converge in their predictions, as visualized in Figure 4 (a). Analysis in Appendix D shows the positive correlation between value-std and the policy gradient, empirically supporting the above speculation.

Advantage masking based on the value agreement. Recent studies show that preventing the policy from overfitting to low-information samples can substantially improve learning efficiency (Liu et al., 2025b). Since the degree of agreement across critics reflects state informativeness, where high agreement implies low uncertainty and limited learning potential, we use the standard deviation of

critics' outputs to quantify the benefit of optimizing a given state. Specifically, we identify the top k percentage of states with the highest agreement (i.e., lowest standard deviation) and mask their corresponding advantages in the policy loss. This suppresses gradient updates from low-informative transitions, filtering out noisy or redundant learning signals directing policy optimization toward higher-value data. The resulting policy loss objective is:

$$\mathcal{J}_{\text{PPO}}(\theta) = \mathbb{E}\frac{1}{|o|}\sum_{t=1}^{|o|}\mathbb{I}^{A}\cdot\min\left(\mathcal{IS}_t\cdot\bar{A}_t,\ \text{clip}\left(\mathcal{IS}_t,\ 1-\epsilon,\ 1+\epsilon\right)\bar{A}_t\right);\ \mathbb{I}_t^{\text{A}} = \begin{cases}0, & \text{if } \sigma_t \in Low_k(\sigma)\\ 1, & \text{otherwise}\end{cases} \tag{4}$$

Here, $\sigma_t = \text{std}\left(\{V(s_t;\phi_m)\}_{m=1}^{M}\right)$ denotes the agreement of value estimates for state s_t. Important sampling is defined as $\mathcal{IS}_t = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}$. Figure 4(b) shows that, masking the advantages corresponding to the 20% of states with high critic convergence, the policy exhibits stable learning dynamics even under high sample reuse (update-to-data ratio (UTD) = 4, i.e., each sample was used for training four times) and significantly improves sample efficiency. We further compared value-std (critic-side uncertainty) with entropy (policy-side uncertainty) (Wang et al., 2025a; Rahn et al., 2024; Cui et al., 2025b) by masking an equal fraction of states per step according to each metric. Figure 4(c) shows that value-std–based masking consistently delivers stronger learning benefits. This observation echoes classic RL findings (Osband et al., 2016), where ensemble-based value uncertainty acts as a proxy for learning dynamics. Figure 5 reveals that low value-std states consistently align with low entropy, suggesting that value-std is a more precise uncertainty metric than entropy.

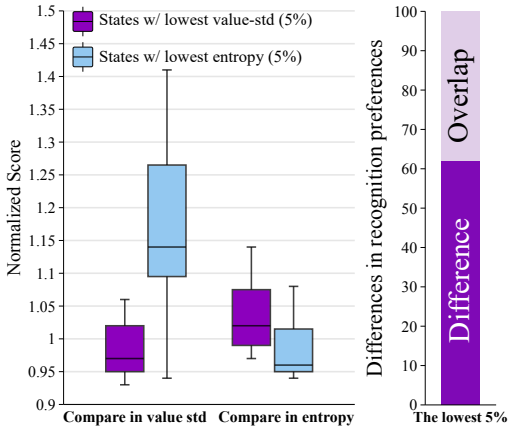


Figure 5: **Left:** States with low value-std maintain low entropy (left box group), but states with low entropy may have a high value-std (right box group). **Right:** States with low entropy and states with low value-std show obvious differences. We sampled the 5% lowest entropy states and the 5% lowest value-std states at step 150.

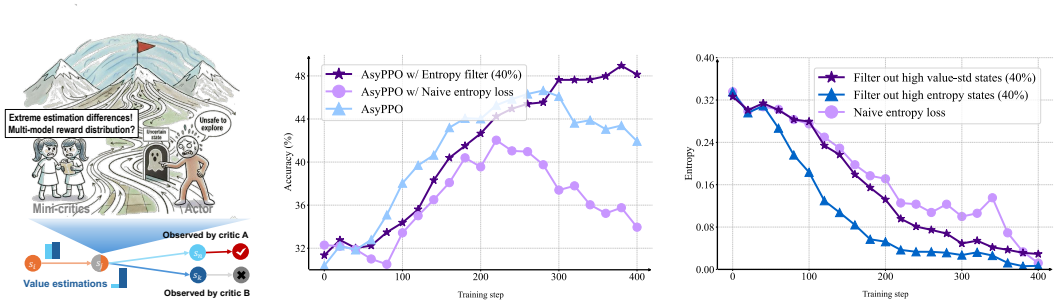
Takeaway 2

Agreement among critics provides a reliable measure of the learning benefit of the states.

Entropy filtering based on value divergence. When critics exhibit significant divergence in their evaluation of a state s_j, reflected in a high standard deviation, it may indicate that s_j is *reasoning-independent*. For instance, different critics may encounter divergent reward distributions for trajectories passing through s_j, due to factors such as inference-irrelevant tokens or inherent semantic patterns in model generations. With a large λ, the dispersion in returns distribution propagates back to each state, amplifying disagreement among critics. In such cases, persistent exploration at s_j is meaningless, as it does not correspond to an actionable decision state (Figure 6 (a)). To promote meaningful exploration while avoiding wasteful updates on noisy or non-decision states, we introduce a safe entropy regularization weighted by β. Specifically, we filter out states with high value standard deviation when computing entropy $\mathcal{H}$. Complete policy loss can be rewritten as:

$$\mathcal{J}_{\text{PPO}}(\theta) = \mathbb{E}_{[q\sim P(Q),\ o\sim\pi_{\theta_{\text{old}}}(O|q)]}\frac{1}{|o|}\sum_{t=1}^{|o|}\left[\mathbb{I}_t^{\text{A}}\cdot\min\left(\mathcal{IS}_t\cdot\bar{A}_t,\ \text{clip}\left(\mathcal{IS}_t,\ 1-\epsilon,\ 1+\epsilon\right)\bar{A}_t\right)\right.$$
$$\left. + \beta\cdot\mathbb{I}_t^{\mathcal{H}}\cdot\mathcal{H}\left[\pi_\theta(\cdot|s_t)\right]\right];\ \mathbb{I}_t^{\mathcal{H}} = \begin{cases}0, & \text{if } \sigma_t \in top_h(\sigma)\\ 1, & \text{otherwise}\end{cases}. \tag{5}$$

Figure 6 (b) shows that, unlike naive entropy loss, which can yield suboptimal learning, our entropy regularization mitigates entropy collapse and stabilizes policy learning, avoiding spurious exploration while guiding the policy toward better convergence with higher returns. We also compare filtering



(a) Motivation of our entropy filtering (b) Ablation study of our entropy filtering (c) Comparison of two filtering mechanisms

Figure 6: **(a):** When critics diverge, the state is weakly coupled to the final outcome and has complex future dynamics; exploration in such non-critical states should be avoided. **(b):** Excluding states with high value-estimate standard deviation from the entropy loss prevents policy collapse induced by naive entropy regularization and **yields a roughly 7 percentage-point improvement**. The setup follows the settings in Figure 1 (b). **(c):** Excluding the top 40% of high value-std states from the entropy loss preserves policy entropy at levels comparable to naive entropy guidance, whereas filtering the same percentage of states with the highest entropy collapse. The settings are consistent with Figure 4.

based on value-std versus entropy. As shown in Figure 6 (c), the overlap between the two sets is minimal. Even after filtering the top 40% of hight value-std, policy entropy remains stable, while filtering the same fraction of high-entropy states causes entropy collapse. Statistical analysis of filtered tokens (Appendix E) further confirms that removed words are typically adverbs, interjections that are irrelevant to decision-making. Algorithm 1 summarizes the full execution process of AsyPPO.

Takeaway 3

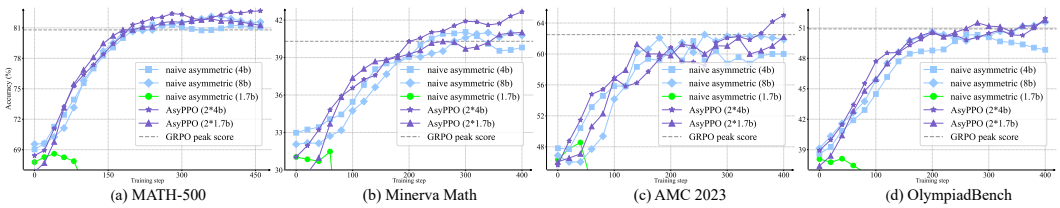
Divergence among value estimations indicates the cost-effectiveness of exploring the states.

Algorithm 1 Asymmetric PPO with two mini-critic

π_θ: actor. $V_{\phi_{\{1,2\}}}$: mini-critics, $o_t \in O$: generation up to step t in response o under prompt q, O denotes the total response in the batch. $\sigma(O)$: value estimation std across the critics. $\bar{A}$: corrected advantage. $\mathbb{I}^A$: The index for advantage masking. $\mathbb{I}^{\mathcal{H}}$: The index for entropy filtering.
while training step $<$ maximum step **do**
  # Rollout under a batch of prompts Q
  $O \leftarrow \pi_\theta(Q)$
  # Train critics
  Build training subsets for each critic, and update $V_{\phi_{\{1,2\}}}$. ▷ Eq.2
  # Correct the advantage
  $\bar{A} \leftarrow GAE(\bar{V}, r)$, $\bar{V} \leftarrow mean(V_{\phi_1}(Q,O), V_{\phi_2}(Q,O))$ ▷ Eq.3
  # Train the policy
  Generate masking vector $\mathbb{I}^A \leftarrow Low_k(\sigma(O))$ and filtering vector $\mathbb{I}^{\mathcal{H}} \leftarrow Top_h(\sigma(O))$.
  Update π_θ via reconstructed PPO loss. ▷ Eq.5
end while

4 Experiments

In §3, we described the architecture and training pipeline of AsyPPO and, through controlled ablation studies, demonstrated its efficacy on 4B and 8B LLMs (**refer to Appendix B for detailed results**). Building on that, this section examines AsyPPO more broadly through a suite of experiments. We organize the subsequent studies around three research questions: **RQ1:** Can AsyPPO and naive asymmetric PPO unlock general reasoning in larger LLM? **RQ2:** How sensitive is AsyPPO to the size and number of critics? **RQ3:** What setups are effective for advantage masking and entropy filtering?



(a) MATH-500 (b) Minerva Math (c) AMC 2023 (d) OlympiadBench

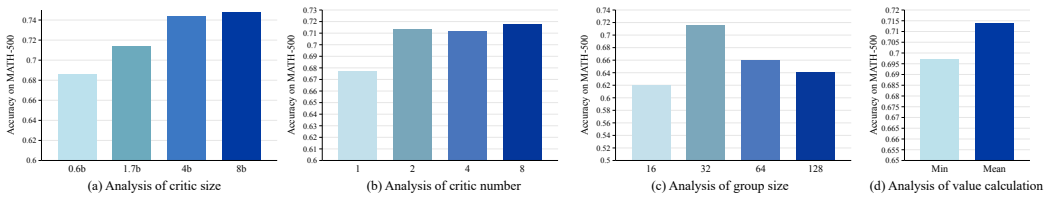
Figure 7: AsyPPO improves accuracy by **an average of about 3 + points** compared to GRPO, and achieves more than **20% lighter weight** than symmetrical PPO. Our naive asymmetric PPO still works on the 14b policy, but fails under the 1.7b critic setting. However, AsyPPO unlocks the 1.7b critic's ability to guide the 14b actor.

4.1 Generalization to Large Models

Setup. To ensure consistency with prior research, we fix the global batch size to 1024, with a maximum response length of 8192 tokens. The learning rate is set to $1e-6$. For text generation, we use a top_p value = 0.99, and top_k value = 100, temperature 0.99, $UTD = 4$ (also referred to as PPO_epoch, result in off-policy). The actor is `Qwen3-14b-Base`, while critics vary in size from the `Qwen3-Base` family. To ensure reproducibility and fairness, we exclusively use open-source datasets. We use the hard training dataset from Liu et al. (2025d); Zeng et al. (2025), which exposes clear performance gaps across algorithms in long-tail reasoning tasks. We report the average@4 across 4 challenging benchmarks, i.e., MATH-500 (Lightman et al., 2023), OlympiadBench (He et al., 2024), MinervaMath (Lewkowycz et al., 2022), and AMC 2023 (Xue et al., 2025).

Baselines. For all algorithms, actors are initialized using `Qwen3-14b-Base`. Naive asymmetric PPO uses a single critic, i.e., `Qwen3-1.7b-Base`, `Qwen3-4b-Base` and `Qwen3-8b-Base`, and optimize with the vanilla PPO optimization objective. AsyPPO employs two mini-critics with advantage masking at 20% and entropy filtering at 20%. We use the setting of GRPO recommended by Liu et al. (2025d). Full hyperparameter details are provided in Appendix C.2.

Results. Figure 7 shows that AsyPPO with two 4b critics achieves the strongest results across all tasks. Compared to GRPO, AsyPPO improves accuracy by an average of about 3 points. For naive asymmetric PPO (a single mini-critic guiding a large actor), we observe a clear critic-capacity threshold: single `Qwen3-1.7b-Base` critic cannot reliably guide 14b actors, despite successfully guiding an 8B actor; upgrading to a 4B critic restores effective learning. By contrast, AsyPPO lowers this requirement, 1.7b critics deliver substantial reasoning gains. Combined with the lightweight deployability in Figure 1(c), AsyPPO establishes an efficient and practical RL4LLM design.



(a) Analysis of critic size (b) Analysis of critic number (c) Analysis of group size (d) Analysis of value calculation

Figure 8: **(a):** The increase in the size of the critic further enhances the effectiveness of AsyPPO, which can be regarded as the marginal benefit brought by the parameter scaling up. We initialize the actor using `Qwen3-8b-Base` and initialize the double mini critic using four sizes of the Qwen3 Base model. **(b):** A qualitative improvement in performance can be achieved by using two mini critics. **(c):** A suitable group size for AsyPPO is 32. **(d):** Using the mean of the critic's estimated value can achieve better correction of the value than using min. For (b,c,d), we initialize the actor using `Qwen3-8b-Base` and initialize the mini critics using `Qwen3-1.7b-Base`.

4.2 Ablation Study

The preceding results show that AsyPPO consistently enhances reasoning in base models across scales. We provide a module-wise analysis to characterize the algorithm from multiple perspectives.

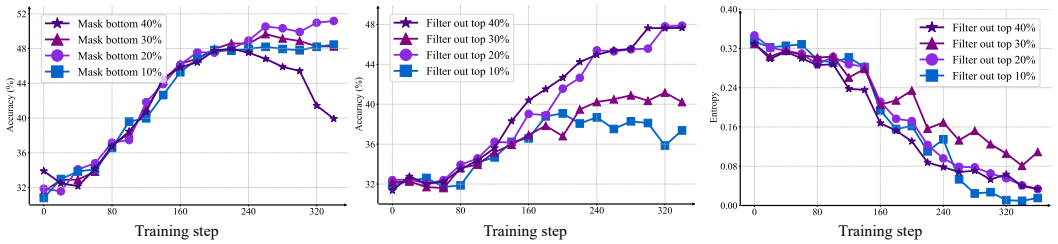


Figure 9: **Left**: Average test score on six benchmarks under various advantage masking setup. **Middle**: Average test score under various filtering out setup. **Right**: entropy curves during training. All experiments were based on `Qwen3-8b-Base` actor and `Qwen3-1.7b-Base` critic. The accuracy calculation follow Figure 2.

Ensemble critic system. Figure 8 (a) shows a scaling-law–like trend: increasing critic size steadily raises the policy's peak score. We recommend using the largest critic model that fits in GPU memory to maximize AsyPPO's optimization capacity. However, we do not see comparable gains from increasing the number of critics: Figure 8 (b) shows that two mini-critics are sufficient for a clear step-change in performance. Varying the GRPO group size (trajectories per prompt) under our non-overlapping group setup while keeping other parameters at their defaults (Figure 8 (c)), and found 32 to be a robust setting. Comparing ensemble value aggregation (Figure 8 (d)), the mean of values outperforms min value, suggesting overestimation is not a dominant issue in RL4LLM.

Value-convergent-based advantage masking. To identify a robust advantage-masking percentage, we adopt the main experiment settings with `Qwen3-8b-Base` as the policy and two `Qwen3-1.7b-Base` critics. Figure 9 (Left) shows that masking 20% of low-value-std states provides the strongest gains.

Value-divergence-based entropy filter. To find an appropriate filtering percentage, we follow the same setup as for advantage masking. We test masking 10%, 20%, 30%, and 40% of the highest-value-std states from the entropy loss. As shown in Figure 9 (Middle, Right), larger masks induce entropy collapse, while 20% strikes the best exploration–exploitation balance.

5 Conclusion

We reframed the critic bottleneck in RL4LLM as an architectural rather than a purely algorithmic or optimization issue. Our proposed Asymmetric Proximal Policy Optimization (AsyPPO) reinstates the critic's role via double lightweight mini-critics trained on disjoint prompt-level data, yielding diverse yet calibrated value estimates. Beyond improving value estimation robustly, we showed that inter-critic uncertainty provides an actionable signal for policy optimization: masking advantages for low-informativeness states and filtering high-divergence states from entropy regularization both reduce overfitting and promote safer, more effective exploration. Across standard LLM reasoning benchmarks, AsyPPO consistently improves general reasoning for models of varied sizes, empirically supporting asymmetric actor–critic design as a viable and efficient direction for RL4LLM. AsyPPO mitigates critic over-parameterization while improving the sample and compute efficiency of PPO.

Limitations. To ensure fairness and reliability under limited GPU resources, all experiments initialized both actor and critic models from the widely used Qwen3 series. Evaluation on additional model families (e.g., Llama (Grattafiori et al., 2024)) is left for future work. Following (Liu et al., 2025d), we fixed the maximum generation length to 8k tokens, a common academic setting that balances inference coverage while avoiding inference-cost blowups. We plan to assess the algorithm's generalization under ultra-long inference budgets and adopt classical RL practice of using a more diverse set of random seeds to further strengthen the robustness of our conclusions.

Future work. AsyPPO opens new avenues for RL4LLM design and raises several interesting questions. For example, do ensemble critic systems composed of different model families and sizes exhibit performance differences? Do variations in critic hyperparameter settings affect calibration and uncertainty estimates? Promising directions also include confidence-weighted ensemble critics to improve value estimation and analyze the relationship between value uncertainty and entropy.

Reproducibility Statement

To ensure reproducibility, we conduct experiments with an open-source RL training framework (Wang et al., 2025c) and initialize both the actor and critic from widely used open-source Qwen3 models. The policy is trained on an open-source dataset (DeepMath-103K). Moreover, we provide a complete specification of training hyperparameters and the evaluation protocol, including the computation of test accuracy, in Appendix C.2.

Ethics statement

This paper presents work whose goal is to advance the field of Machine Learning, and RL4LLM in particular. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Zafarali Ahmed, Nicolas Le Roux, Mohammad Norouzi, and Dale Schuurmans. Understanding the impact of entropy on policy optimization. In *International conference on machine learning*, pp. 151–160. PMLR, 2019.

Aili Chen, Aonian Li, Bangwei Gong, Binyang Jiang, Bo Fei, Bo Yang, Boji Shan, Changqing Yu, Chao Wang, Cheng Zhu, et al. Minimax-m1: Scaling test-time compute efficiently with lightning attention. *arXiv preprint arXiv:2506.13585*, 2025.

Xinyue Chen, Che Wang, Zijian Zhou, and Keith Ross. Randomized ensembled double q-learning: Learning fast without a model. *arXiv preprint arXiv:2101.05982*, 2021.

Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, Qixin Xu, Weize Chen, Jiarui Yuan, Huayu Chen, Kaiyan Zhang, Xingtai Lv, Shuo Wang, Yuan Yao, Xu Han, Hao Peng, Yu Cheng, Zhiyuan Liu, Maosong Sun, Bowen Zhou, and Ning Ding. Process reinforcement through implicit rewards. *ArXiv*, abs/2502.01456, 2025a. URL https://api.semanticscholar.org/CorpusID:276107672.

Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, et al. The entropy mechanism of reinforcement learning for reasoning language models. *arXiv preprint arXiv:2505.22617*, 2025b.

Tiantian Fan, Lingjun Liu, Yu Yue, Jiaze Chen, Chengyi Wang, Qiying Yu, Chi Zhang, Zhiqi Lin, Ruofei Zhu, Yufeng Yuan, et al. Truncated proximal policy optimization. *arXiv preprint arXiv:2506.15050*, 2025.

Wei Fu, Jiaxuan Gao, Xujie Shen, Chen Zhu, Zhiyu Mei, Chuyi He, Shusheng Xu, Guo Wei, Jun Mei, Jiashu Wang, et al. Areal: A large-scale asynchronous reinforcement learning system for language reasoning. *arXiv preprint arXiv:2505.24298*, 2025.

Kanishk Gandhi, Ayush Chakravarthy, Anikait Singh, Nathan Lile, and Noah D Goodman. Cognitive behaviors that enable self-improving reasoners, or, four habits of highly effective stars. *arXiv preprint arXiv:2503.01307*, 2025.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, et al. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. *arXiv preprint arXiv:2402.14008*, 2024.

Zhiwei He, Tian Liang, Jiahao Xu, Qiuzhi Liu, Xingyu Chen, Yue Wang, Linfeng Song, Dian Yu, Zhenwen Liang, Wenxuan Wang, et al. Deepmath-103k: A large-scale, challenging, decontaminated, and verifiable mathematical dataset for advancing reasoning. *arXiv preprint arXiv:2504.11456*, 2025.

Jian Hu. Reinforce++: A simple and efficient approach for aligning large language models. *arXiv preprint arXiv:2501.03262*, 2025.

Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, Xiangyu Zhang, and Heung-Yeung Shum. Open-reasoner-zero: An open source approach to scaling up reinforcement learning on the base model. *arXiv preprint arXiv:2503.24290*, 2025.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Hojoon Lee, Dongyoon Hwang, Donghu Kim, Hyunseung Kim, Jun Jet Tai, Kaushik Subramanian, Peter R. Wurman, Jaegul Choo, Peter Stone, and Takuma Seno. Simba: Simplicity bias for scaling up parameters in deep reinforcement learning. *ArXiv*, abs/2410.09754, 2024. URL https://api.semanticscholar.org/CorpusID:273346233.

Kimin Lee, Michael Laskin, Aravind Srinivas, and Pieter Abbeel. Sunrise: A simple unified framework for ensemble learning in deep reinforcement learning. In *International conference on machine learning*, pp. 6131–6141. PMLR, 2021.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. Solving quantitative reasoning problems with language models. *Advances in neural information processing systems*, 35:3843–3857, 2022.

Zhong-Zhi Li, Xiao Liang, Zihao Tang, Lei Ji, Peijie Wang, Haotian Xu, Haizhen Huang, Weiwei Deng, Ying Nian Wu, Yeyun Gong, et al. Tl; dr: Too long, do re-weighting for effcient llm reasoning compression. *arXiv preprint arXiv:2506.02678*, 2025.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*, 2023.

Jiacai Liu, Chaojie Wang, Chris Yuhao Liu, Liang Zeng, Rui Yan, Yiwen Sun, Yang Liu, and Yahui Zhou. Improving multi-step reasoning abilities of large language models with direct advantage policy optimization. *arXiv preprint arXiv:2412.18279*, 2024.

Jiashun Liu, Johan Samir Obando Ceron, Aaron Courville, and Ling Pan. Neuroplastic expansion in deep reinforcement learning. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=20qZK2T7fa.

Jiashun Liu, Johan Obando-Ceron, Pablo Samuel Castro, Aaron Courville, and Ling Pan. The courage to stop: Overcoming sunk cost fallacy in deep reinforcement learning. *arXiv preprint arXiv:2506.13672*, 2025b.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding r1-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*, 2025c.

Zihe Liu, Jiashun Liu, Yancheng He, Weixun Wang, Jiaheng Liu, Ling Pan, Xinyu Hu, Shaopan Xiong, Ju Huang, Jian Hu, et al. Part i: Tricks or traps? a deep dive into rl for llm reasoning. *arXiv preprint arXiv:2508.08221*, 2025d.

Guozheng Ma, Lu Li, Zilin Wang, Li Shen, Pierre-Luc Bacon, and Dacheng Tao. Network sparsity unlocks the scaling potential of deep reinforcement learning. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=mIomqOskaa.

Olya Mastikhina, Dhruv Sreenivas, and Pablo Samuel Castro. Optimistic critics can empower small actors. *arXiv preprint arXiv:2506.01016*, 2025.

Siddharth Mysore, Bassel El Mabsout, Renato Mancuso, and Kate Saenko. Honey. i shrunk the actor: A case study on preserving performance with smaller actors in actor-critic rl. *2021 IEEE Conference on Games (CoG)*, pp. 01–08, 2021.

Ian Osband, Charles Blundell, Alexander Pritzel, and Benjamin Van Roy. Deep exploration via bootstrapped dqn. *Advances in neural information processing systems*, 29, 2016.

Nate Rahn, Pierluca D'Oro, and Marc G Bellemare. Controlling large language model agents with entropic activation steering. *arXiv preprint arXiv:2406.00244*, 2024.

Ravi Raj and Andrzej Kos. Intelligent mobile robot navigation in unknown and complex environment using reinforcement learning technique. *Scientific Reports*, 14, 2024. URL https://api.semanticscholar.org/CorpusID:273024254.

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

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

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

Yi Xiang Marcus Tan, Pihe Hu, L. Pan, and Longbo Huang. Rlx2: Training a sparse deep reinforcement learning model from scratch. *ArXiv*, abs/2205.15043, 2022.

Joshua Tint, Som Sagar, Aditya Taparia, Kelly Raines, Bimsara Pathiraja, Caleb Liu, and Ransalu Senanayake. Expressivityarena: Can llms express information implicitly? *arXiv preprint arXiv:2411.08010*, 2024.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, et al. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning. *arXiv preprint arXiv:2506.01939*, 2025a.

Tao Wang, Ruipeng Zhang, and Sicun Gao. Improving value estimation critically enhances vanilla policy gradient. In *Forty-second International Conference on Machine Learning*, 2025b. URL https://openreview.net/forum?id=Nq3oz7vn3j.

Weixun Wang, Shaopan Xiong, Gengru Chen, Wei Gao, Sheng Guo, Yancheng He, Ju Huang, Jiaheng Liu, Zhendong Li, Xiaoyang Li, et al. Reinforcement learning optimization for large-scale learning: An efficient and user-friendly scaling library. *arXiv preprint arXiv:2506.06122*, 2025c.

Zhenghai Xue, Longtao Zheng, Qian Liu, Yingru Li, Xiaosen Zheng, Zejun Ma, and Bo An. Simpletir: End-to-end reinforcement learning for multi-turn tool-integrated reasoning. *arXiv preprint arXiv:2509.02479*, 2025.

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

Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. The surprising effectiveness of ppo in cooperative multi-agent games. *Advances in neural information processing systems*, 35:24611–24624, 2022.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

Lifan Yuan, Wendi Li, Huayu Chen, Ganqu Cui, Ning Ding, Kaiyan Zhang, Bowen Zhou, Zhiyuan Liu, and Hao Peng. Free process rewards without process labels. In *Forty-second International Conference on Machine Learning*, 2025a. URL https://openreview.net/forum?id=8ThnPFhGm8.

Yufeng Yuan, Yu Yue, Ruofei Zhu, Tiantian Fan, and Lin Yan. What's behind ppo's collapse in long-cot? value optimization holds the secret. *ArXiv*, abs/2503.01491, 2025b. URL https://api.semanticscholar.org/CorpusID:276766648.

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

Weihao Zeng, Yuzhen Huang, Qian Liu, Wei Liu, Keqing He, Zejun Ma, and Junxian He. Simplerl-zoo: Investigating and taming zero reinforcement learning for open base models in the wild. *arXiv preprint arXiv:2503.18892*, 2025.

Kaiyan Zhang, Yuxin Zuo, Bingxiang He, Youbang Sun, Runze Liu, Che Jiang, Yuchen Fan, Kai Tian, Guoli Jia, Peng Li, Yu Fu, Xingtai Lv, Yuchen Zhang, Sihang Zeng, Shang Qu, Hao-Si Li, Shijie Wang, Yuru Wang, Xi-Dai Long, Fangfu Liu, Xiang Xu, Jiaze Ma, Xuekai Zhu, Ermo Hua, Yihao Liu, Zonglin Li, Hua yong Chen, Xiaoye Qu, Yafu Li, Weize Chen, Zhenzhao Yuan, Junqi Gao, Dong Li, Zhiyuan Ma, Ganqu Cui, Zhiyuan Liu, Biqing Qi, Ning Ding, and Bowen Zhou. A survey of reinforcement learning for large reasoning models. 2025. URL https://api.semanticscholar.org/CorpusID:281247204.

Yuzhong Zhao, Yue Liu, Junpeng Liu, Jingye Chen, Xun Wu, Yaru Hao, Tengchao Lv, Shaohan Huang, Lei Cui, Qixiang Ye, et al. Geometric-mean policy optimization. *arXiv preprint arXiv:2507.20673*, 2025.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, et al. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025.

Dingwei Zhu, Shihan Dou, Zhiheng Xi, Senjie Jin, Guoqiang Zhang, Jiazheng Zhang, Junjie Ye, Mingxu Chai, Enyu Zhou, Ming Zhang, Caishuang Huang, Yunke Zhang, Yuran Wang, and Tao Gui. Vrpo: Rethinking value modeling for robust rl training under noisy supervision. *ArXiv*, abs/2508.03058, 2025.

A Related work

Critic-based RL4LLM algorithms Shao et al. (2024) first demonstrated that large-scale reinforcement learning (RL) with outcome-based rewards can unlock long-tail reasoning, beginning from an unaligned base model. This finding has led to numerous variations of the Proximal Policy Optimization (PPO) algorithm. As far as we know, most algorithm research is mainly based on the baseline normalized advantage calculation method (Hu, 2025; Liu et al., 2025c; Chen et al., 2025).

On the other hand, value-based algorithm innovations are relatively few, Yuan et al. (2025b) argued that the decay factor is not well-suited for complex reasoning tasks that require long chains of thought (CoT). Yue et al. (2025); Zhu et al. (2025); Zhao et al. (2025) proposed novel mechanisms to enhance the robustness of the critic model when faced with noisy reward signals. Open-Reasoner-Zero (Hu et al., 2025) argues that, within this regime, vanilla PPO without KL regularization suffices to scale training stably. T-PPO (Fan et al., 2025) uses critic to enhance the stability of policy training in the long-tail asynchronous setting (Fu et al., 2025). Another similar research line to introduce critic-like models is done with the introduction of Implicit PRM (Yuan et al., 2025a). This approach is also able to provide token-level supervision for scalable RL training. PRIME (Cui et al., 2025a) adapted a specific reward model formulation to directly generate token-level rewards. However, current mainstream RL4LLM algorithms primarily emphasize critic-free optimization (Zhang et al., 2025). In this context, our research aim to underscore the importance of the critic in RL4LLM scenarios and try to address the deployment limitations associated with critics.

Asymmetric architecture. In the realm of continuous deep RL, recent studies have investigated the potential of asymmetric network structures by reducing the capacity of the actor network. For example, Mastikhina et al. (2025); Mysore et al. (2021) suggest that the actor can function effectively with a significantly smaller capacity compared to the critic. Empirical evidence from Tan et al. (2022) supports this idea, demonstrating that sparsifying the policy network can enhance effective policy learning while significantly improving both inference and training speeds. Additionally, Liu et al. (2025a) found that pruning the actor network's topology based on trial gradients can yield better performance. Similarly, Ma et al. (2025) revealed that even random pruning of the actor network can maintain performance within the SimBa network architecture (Lee et al., 2024). These contributions highlight the adaptability of RL in accommodating asymmetric designs, providing valuable insights for our research. However, existing works primarily concentrate on reducing the actor's size within simple network frameworks. In contrast, our paper pioneers the exploration of effectively guiding a small critic to inform a larger actor by optimizing the PPO algorithm within the RL4LLM scenario.

B The performance gain of AsyPPO on the small model strategy

Policy model	Base model	Symmetric PPO	AsyPPO
Qwen3-4b-Base	30.5%	47.3%	53.1% +6.1%
Qwen3-8b-Base	31.7%	50.6%	53.8% +3.2%

Table 1: Peak accuracy comparison of Symmetric PPO and AsyPPO under high data reuse setting (UTD=4) over six benchmarks. Score calculation same as Figure 1 (b). Purple score denotes the improvement compare to Symmetric PPO.

We set both the classic symmetrical PPO and our AsyPPO to the optimal Settings. AsyPPO uniformly initializes mini-critics using the Qwen3-1.7b-Base model. AsyPPO employs two mini-critics with advantage masking at 20%. And use the open source hard training dataset in (Liu et al., 2025d), which is selected from DeepMath-103k (He et al., 2025) with sampling probability proportional to each entry's assigned difficulty level. We report the average@4 across six challenging benchmarks, i.e., MATH-500, OlympiadBench, MinervaMath, and AMC 2023, AIME 2025, AIME 2024.

Overall, Table 1 shows that AsyPPO effectively enhances the reasoning capabilities of two small models of different sizes, achieving respective improvements of 22.6% and 22.1% over their original performance. Compared to symmetric PPO, our algorithm delivers gains of 6.1% and 3.2%, while maintaining lightweight deployment. Upon analyzing specific benchmarks, our approach demonstrates notable advancements. For instance, on AIME 2025, we observed respective increases of

approximately 4% (4B) and 6% (8B) compared to symmetric PPO. Similarly, on MATH-500, the improvements were around 3% (4B) and 2% (8B), and on MinervaMath, the gains were approximately 2% (4B) and 4% (8B). In the remaining three tasks, our method maintained performance levels comparable to those of symmetric PPO.

C Detailed Experimental Setup

C.1 Plot setup

To ensure clarity and intuitiveness in the qualitative analysis, all curves are consistently smoothed using identical parameters. Specifically, the mean values are computed using an 11-step moving window with an exponential smoothing factor of 0.6. The smooth window set as 4 and 2.

C.2 Hyperparameters

We employ ROLL, a user-friendly and efficient open-source reinforcement learning framework, to implement our pipeline. Subsequently, the key parameters observed during the training process are presented as follows. See our code config file for more details on the parameters. For the 14b policy training. We uniformly arrange the actors on (0,16) and the critics on (16,32) GPUs. For other small models, we uniformly place the actor at (0,8) and the critic at (8,16) GPU. Detailed settings can be found in next page.

```yaml
# We use below setup for 4b and 8b policy
seed: 42
max_steps: 500
save_steps: 500
logging_steps: 1
eval_steps: 1
gamma: 1.0  # discount factor
lambd: 1.0  # GAE lambda
rollout_batch_size: 64
prompt_length: 1024
response_length: 8000
value_aggregation_strategy: "mean"
gradient_mask_percentage: 0.2 # mask 20%
entropy_loss_coef: 0.01
entropy_filter_mask_percentage: 0.2 # filter out 20%
ppo_epochs: 1 # 4 is also used in main experiments
adv_estimator: "gae"
init_kl_coef: 0.0
async_generate_level: 1

actor_train:
  training_args:
    learning_rate: 1.0e-6
    weight_decay: 0
    per_device_train_batch_size: 1
    gradient_accumulation_steps: 256
    warmup_steps: 50
    num_train_epochs: 50

critic_1:
  training_args:
    learning_rate: 1.0e-5
    weight_decay: 1.0e-2
    warmup_steps: 5
    per_device_train_batch_size: 1
    gradient_accumulation_steps: 128
    warmup_steps: 5
    num_train_epochs: 50

critic_2:
  training_args:
    learning_rate: 1.0e-5
    weight_decay: 1.0e-2
    warmup_steps: 5
    per_device_train_batch_size: 1
    gradient_accumulation_steps: 128
    warmup_steps: 5
    num_train_epochs: 50
  ...

actor_infer:
  generating_args:
    max_new_tokens: ${response_length}
    top_p: 0.99
    top_k: 100
    num_beams: 1
    temperature: 0.99
    num_return_sequences: 32
  ...
```

```yaml
# We use below setup for 14b policy
seed: 42
max_steps: 500
save_steps: 500
logging_steps: 1
eval_steps: 1
gamma: 1.0  # discount factor
lambd: 1.0  # GAE lambda
value_aggregation_strategy: "mean"
gradient_mask_percentage: 0.2 # mask 20%
entropy_loss_coef: 0.01
entropy_filter_mask_percentage: 0.2  # filter out 20% or 0%
rollout_batch_size: 64
prompt_length: 1024
response_length: 8000
infer batch size: 4
ppo_epochs: 4
adv_estimator: "gae"
init_kl_coef: 0.0
async_generate_level: 1
actor_train:
  training_args:
    learning_rate: 1.0e-6
    weight_decay: 0
    per_device_train_batch_size: 2
    gradient_accumulation_steps: 6
    warmup_steps: 50
    num_train_epochs: 50
critic_1:
  training_args:
    learning_rate: 1.0e-5
    weight_decay: 1.0e-2
    warmup_steps: 5
    per_device_train_batch_size: 2
    gradient_accumulation_steps: 16
    warmup_steps: 5
    infer batch size: 4
    num_train_epochs: 50
critic_2:
  training_args:
    learning_rate: 1.0e-5
    weight_decay: 1.0e-2
    warmup_steps: 5
    per_device_train_batch_size: 2
    gradient_accumulation_steps: 16
    warmup_steps: 5
    infer batch size: 4
    num_train_epochs: 50
  ...
actor_infer:
  generating_args:
    max_new_tokens: ${response_length}
    top_p: 0.99
    top_k: 100
    num_beams: 1
    temperature: 0.99
    num_return_sequences: 32
  ...
```

C.3 PROMPT

In this work, we incorporate the following instruction into the system prompt to encourage the model to better demonstrate its reasoning process: **"Please reason step by step, and put your final answer within \boxed{}."** This setting is designed to guide the model to perform step-by-step reasoning and explicitly present the final answer in the form of \boxed{}, thereby enhancing the clarity and readability of the output.

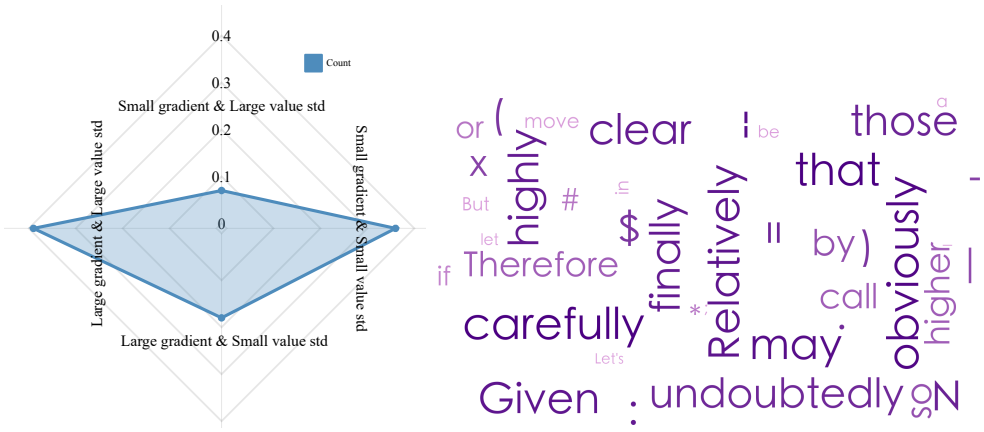


Figure 10: Left: Statistics within a mini-batch in the mid-training stage. Right: The 40 tokens that are masked most frequently in the same mini-batch.

D THE RELATIONSHIP BETWEEN VALUE STD AND STATE INFORMATION QUANTITY

Specifically, for the training scenarios of 8b actors and two 0.6b critics, we use the value-std corresponding to the global state and the median of the gradient magnitude to categorize the states into four types. Namely, large gradient & large value std, large gradient & small value std, small gradient & large value std, small gradient & small value std. The results in Figure 10 (Left) show that the vast majority of states are classified into the categories of large gradient & large value std and small gradient & small value std, thereby empirically proving the positive relationship between value std and the learning value (information quantity) of the state.

E VISUALIZATION OF WORD CLOUDS

We statistically analyzed the word clouds of the tokens with the highest mask frequency in the initial stage of `AsyPPO` training. The results in Figure 10 (Right) show that our mask mechanism tends to mask adjectives, adverbs, and some isolated symbols, with less involvement in logical transitions, except for the slightly prominent progressive word "therefore".

LLM USAGE

LLMs were used to assist paper editing and to write the code for plotting experiments.