

ASPO: Asymmetric Importance Sampling Policy Optimization

Jiakang Wang^{1*}, Runze Liu^{1,2*}, Lei Lin¹, Wenping Hu¹, Xiu Li², Fuzheng Zhang¹, Guorui Zhou^{1†}
and Kun Gai¹

¹Kuaishou Technology, ²Tsinghua University

Abstract: Recent Large Language Model (LLM) post-training methods rely on token-level clipping mechanisms during Reinforcement Learning (RL). However, we identify a fundamental flaw in this Outcome-Supervised RL (OSRL) paradigm: the **Importance Sampling (IS) ratios of positive-advantage tokens are mismatched**, leading to unbalanced token weighting for positive and negative tokens. This mismatch suppresses the update of low-probability tokens while over-amplifying already high-probability ones. To address this, we propose Asymmetric Importance Sampling Policy Optimization (ASPO), which uses a simple yet effective strategy that **flips** the IS ratios of positive-advantage tokens, aligning their update direction with the learning dynamics of negative ones. AIS further incorporates a soft dual-clipping mechanism to stabilize extreme updates while maintaining gradient flow. Comprehensive experiments on coding and mathematical reasoning benchmarks demonstrate that ASPO significantly mitigates premature convergence, improves training stability, and enhances final performance over strong GRPO-based baselines. Our analysis provides new insights into the role of token-level weighting in OSRL and highlights the critical importance of correcting IS in LLM RL. The code and models of ASPO are available at <https://github.com/wizard-III/Archer2.0>.

1. Introduction

Reinforcement Learning from Verifiable Rewards (RLVR) has recently emerged as a powerful framework for Large Language Model (LLM) post-training (DeepSeek-AI et al., 2025; Yu et al., 2025). Among RLVR methods, Group Relative Policy Optimization (GRPO) (Shao et al., 2024) has become a widely adopted algorithm for Outcome-Supervised RL (OSRL), inspiring numerous variants and follow-up works (Yu et al., 2025; Zeng et al., 2025; He et al., 2025; Chen et al., 2025; An et al., 2025; Wang et al., 2025a; Zhang et al., 2025a).

Despite its empirical success, we find that GRPO shows a key limitation as an OSRL approach: its **token-level clipping mechanism introduces a misallocation of learning weights**. Specifically, the Importance Sampling (IS) ratio, which is originally designed as a distribution correction term, no longer behave as our expectations in OSRL, where samples are grouped and normalized at the response level. Instead, they effectively act as **token-level training weights**, determining how strongly each token contributes to the gradient update.

Our reexamination of this mechanism reveals a surprising asymmetry: for negative-advantage tokens, PPO-Clip assigns weights consistent with the desired learning dynamics, reducing weights as token probabilities increase. However, for positive-advantage tokens, the behavior is reversed: tokens that already have higher probabilities under the current policy are given **larger weights**, while those lagging behind are suppressed. This inversion creates a **weight mismatch** that distorts the learning

* Equal contribution

† Project lead

‡ Corresponding author

signal, resulting in over-updating confident tokens and under-updating weak ones. Consequently, the model exhibits entropy collapse, excessive repetition, and premature convergence to local optima.

To address these issues, we propose Asymmetric Importance Sampling Policy Optimization (ASPO), a simple yet effective grounded modification of GRPO that restores balanced token weighting. ASPO reverses the IS ratios for positive-advantage tokens, ensuring that tokens with lower probabilities under the current policy receive stronger updates, while confident tokens are down-weighted. To further enhance stability, AIS integrates a soft dual-clipping mechanism (Chen et al., 2025) that constrains extreme ratios without discarding gradients for positive tokens. Extensive experiments on both coding and mathematical reasoning tasks demonstrate that ASPO: (1) prevents overfitting and entropy collapse, (2) achieves smoother and more stable training dynamics, and (3) significantly outperforms GRPO-based baselines in final performance.

In summary, our main contributions are as follows:

- We identify a fundamental flaw in the token-level clipping design of GRPO-based OSRL methods: **the IS ratio mismatch for positive-advantage tokens.**
- We propose ASPO, which flips IS ratios for positive tokens and applies a soft dual-clip to stabilize training while preserving gradient flow.
- We provide empirical evidence that AIS mitigates entropy collapse, improves optimization stability, and enhances performance across multiple mathematical reasoning and coding benchmarks.

2. Preliminaries

2.1. Group Relative Policy Optimization

Group Relative Policy Optimization (GRPO) (Shao et al., 2024) samples a group of G rollouts for advantage estimation: $\hat{A}_t^i = \frac{R^i - \text{mean}(\{R^i\}_{i=1}^G)}{\text{std}(\{R^i\}_{i=1}^G)}$. The loss function of GRPO is defined as:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{q \sim \mathcal{D}, \{o^i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}(\cdot | q)} \left[\frac{1}{G} \sum_{i=1}^G \frac{1}{|o^i|} \sum_{t=1}^{|o^i|} \left(\min \left(r_t^i(\theta) \hat{A}_t^i, \text{clip} \left(r_t^i(\theta), 1 - \varepsilon, 1 + \varepsilon \right) \hat{A}_t^i \right) - \beta \mathbb{D}_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}}) \right) \right], \quad (1)$$

where $r_t^i = \frac{\pi_\theta(o_t^i | q, o_{\leq t}^i)}{\pi_{\theta_{\text{old}}}(o_t^i | q, o_{\leq t}^i)}$ is the Importance Sampling (IS) ratio, and β is a weight for the Kullback–Leibler (KL) divergence between the current policy π_θ and the reference policy π_{ref} .

2.2. PPO Clipping and the Improvements

The clipping mechanism is first introduced by PPO-Clip (Schulman et al., 2017). It serves as a simple yet effective constraint to prevent large policy updates. As shown in (1), GRPO (Shao et al., 2024) uses this clipping mechanism from PPO. However, Chen et al. (2025) show that this clipping mechanism clips the value and also masks the gradient of the clipped tokens.

CISPO. To address this issue, CISPO (Chen et al., 2025) proposes (2) to preserve the gradient of the clipped tokens:

$$\begin{aligned} \mathcal{J}_{\text{CISPO}}(\theta) &= \mathbb{E}_{q \sim \mathcal{D}, \{o^i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}(\cdot|q)} \\ &\quad \left[\frac{1}{\sum_{i=1}^G |o^i|} \sum_{i=1}^G \sum_{t=1}^{|o^i|} \text{sg}(\text{clip}(r_t^i(\theta), 1 - \varepsilon_{\text{low}}, 1 + \varepsilon_{\text{high}})) \hat{A}_t^i \log \pi_\theta(o_t^i | q, o_{<t}^i) \right], \end{aligned} \quad (2)$$

where $\text{sg}(\cdot)$ is the stop gradient operation, and ε_{low} and $\varepsilon_{\text{high}}$ are the lower and upper bounds of the clipping range (Yu et al., 2025), respectively. In the following sections, we refer to this CISPO-like clipping method as **soft clipping**, which clips the values but retains the gradient. In contrast, clipping both the value and the gradient is called **hard clipping**.

GSPO. GSPO (Zheng et al., 2025) further improves clipping with a sequence-level IS approach:

$$\mathcal{J}_{\text{GSPO}}(\theta) = \mathbb{E}_{q \sim \mathcal{D}, \{o^i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}(\cdot|q)} \left[\frac{1}{G} \sum_{i=1}^G \min \left(s^i(\theta) \hat{A}_t^i, \text{clip}(s^i(\theta), 1 - \varepsilon, 1 + \varepsilon) \hat{A}_t^i \right) \right], \quad (3)$$

where $s^i(\theta) = \left(\frac{\pi_\theta(o^i|q)}{\pi_{\theta_{\text{old}}}(o^i|q)} \right)^{\frac{1}{|o^i|}} = \exp \left(\frac{1}{|o^i|} \sum_{t=1}^{|o^i|} \log \frac{\pi_\theta(o_t^i|q, o_{<t}^i)}{\pi_{\theta_{\text{old}}}(o_t^i|q, o_{<t}^i)} \right)$ is the sequence-level IS ratio.

3. Importance Sampling is not Important

3.1. Motivation

Importance Sampling (IS) is a fundamental technique for estimating expectations under one probability distribution while sampling from another (Precup et al., 2000). In RL, this method is particularly useful because the distribution of trajectories from the current policy often differs from that of older policies. By adjusting returns with importance ratios, data from previous policies can be reused, improving sample efficiency in off-policy learning (Schulman et al., 2017).

In OSRL for LLMs, such as GRPO (Shao et al., 2024), all tokens within a response share the same advantage value. However, when examining an individual token’s contribution to the final correctness, the advantage assigned by GRPO may suffer from two issues: (1) it may be numerically inaccurate, as different tokens contribute unequally to the final answer’s correctness (Wang et al., 2025b,a), and (2) the update direction could be misleading, as a response with a correct final answer might still include incorrect intermediate steps (Zhang et al., 2025b; Zhao et al., 2025).

The original motivation of IS is to correct distribution mismatch, improving token-level return estimation. This leads to the following question:

*If the reward of each token is already inaccurate due to outcome-based advantage estimation,
how important is it to further adjust the distribution using IS weights?*

3.2. Experimental Validation

3.2.1. Setup

To evaluate the practical effects of IS weights in OSRL, we compare two methods. Both methods use the same token-masking strategy, but differ in their IS weights: (1) GRPO with original IS,

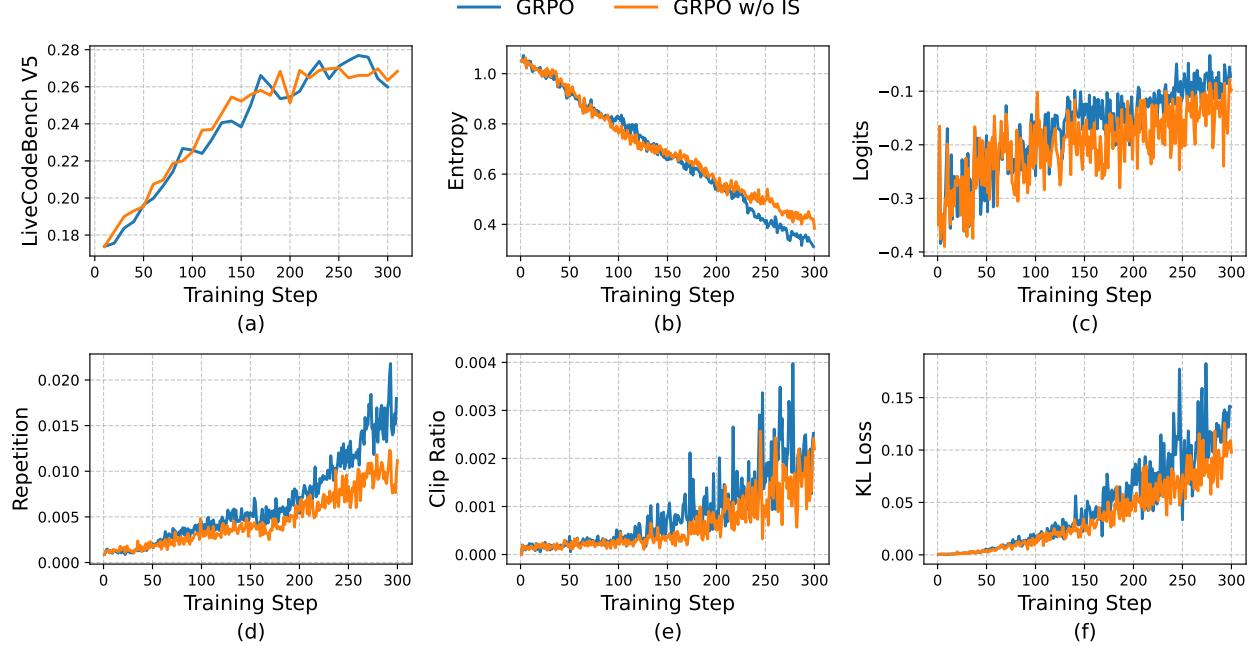


Figure 1: Comparison of test accuracy and training dynamics between DAPO and DAPO without IS.

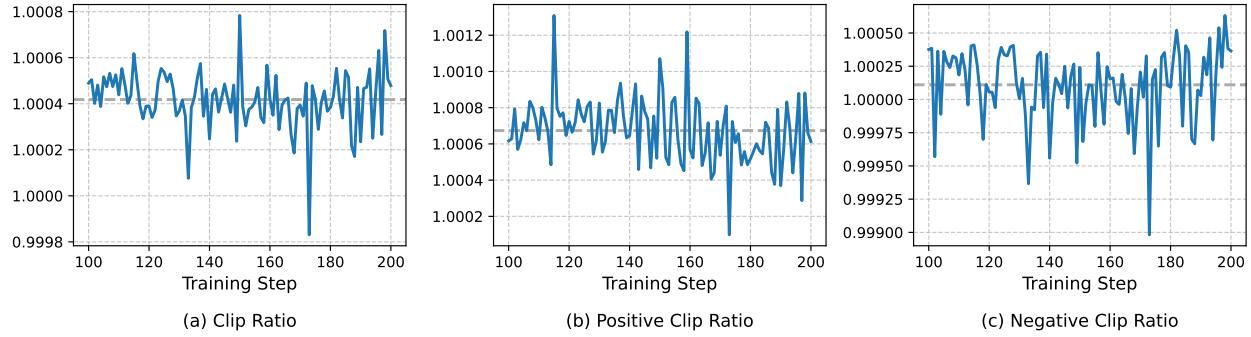


Figure 2: Curves of response-level IS ratios throughout DAPO training. The average IS ratios are shown in gray dashed lines.

and (2) GRPO without IS (where all IS weights are fixed to 1.0). To better observe the effects of IS, the experiments involve a relatively high number of off-policy updates: `ppo_epochs` = 3, and `train_batch_size / mini_batch_size` = 4. All subsequent experiments follow this configuration.

3.2.2. Results

As shown in Figure 1, the test accuracy of the two methods shows little difference. Standard GRPO reaches its peak earlier and then begins to decline, while GRPO without IS converges more smoothly and maintains values near its peak without showing a downward trend. Although the peak of GRPO is slightly higher, the difference is only around 0.4 points, which is within the range of random fluctuations in evaluation. Regarding training dynamics, such as entropy, repetition rate, truncation rate, and KL divergence, those of GRPO without IS change more gradually than those of standard GRPO, particularly in the later stages of training.

3.2.3. Analysis

Entropy drop is mainly driven by low-entropy tokens with positive advantages. As shown in Figure 2, the average IS weight of a response is typically slightly greater than 1.0 (around 1.0004) for GRPO-based algorithms. Furthermore, the average weight of positive samples is slightly higher than that of negative samples. While this difference may seem small, its cumulative effect over tens of thousands of tokens is significant. This subtle discrepancy between positive and negative sample weights causes GRPO training to **prioritize fitting positive samples rather than suppressing negative ones**, and this gap widens throughout training, leading to entropy decay. Since increasing the weight of positive samples accelerates entropy decay, it often results in slightly poorer final performance. Combining this observation with prior research on entropy mechanisms (Cui et al., 2025), we conclude that changes in training entropy during GRPO are likely driven by **low-entropy tokens**, and this entropy may not accurately reflect exploration on high-entropy tokens.

Why the average weight of positive samples higher than that of negative samples? This aligns with the model’s expected learning behavior: after several training rounds, the current policy assigns higher generation probabilities to positive samples than the old policy, resulting in the clip ratio greater than 1. Conversely, negative samples fall below the old policy, shifting the clip ratio toward the lower-right region. This indicates that the model is learning in the correct direction, gradually distinguishing the predicted probabilities of positive and negative samples more effectively.

What happens without IS weights? With all IS weights set to 1.0, on one hand, this makes the overall weights slightly lower than in standard GRPO. On the other hand, it *eliminates the weight difference between positive and negative samples*. While these two factors slow down the learning speed compared to standard GRPO, they do not compromise the final performance.

Based on the above analysis, we draw the following conclusions:

Takeaways for Importance Sampling

- Removing IS almost does not degrade final performance. Instead, it yields more stable behavior, whereas GRPO with IS shows performance decline near the end.
- The analysis of training dynamics does not support the role of IS as a distribution-correction mechanism. Instead, the findings highlight the importance of the **token weights** in training.
- Positive samples have larger IS ratios than negative samples, leading to entropy drop.
- When GRPO-based algorithms are applied to LLM training, the **token-masking mechanism** in PPO-Clip is the key factor, with IS weights having little to no practical impact.

4. Importance Sampling Ratios of Positive Tokens are Mismatched

4.1. Rethinking the Role of Token Clipping

Our preliminary experiments and analyses indicate that IS is not a decisive factor for GRPO-based algorithms in LLMs. Instead, evidence suggests that these weights function less as classical IS ratios and more as **token-level training weights**: some tokens receive larger weights during updates, while others receive smaller ones. We therefore rethink the role of these weights in the original PPO-Clip design from the perspective of token weighting.

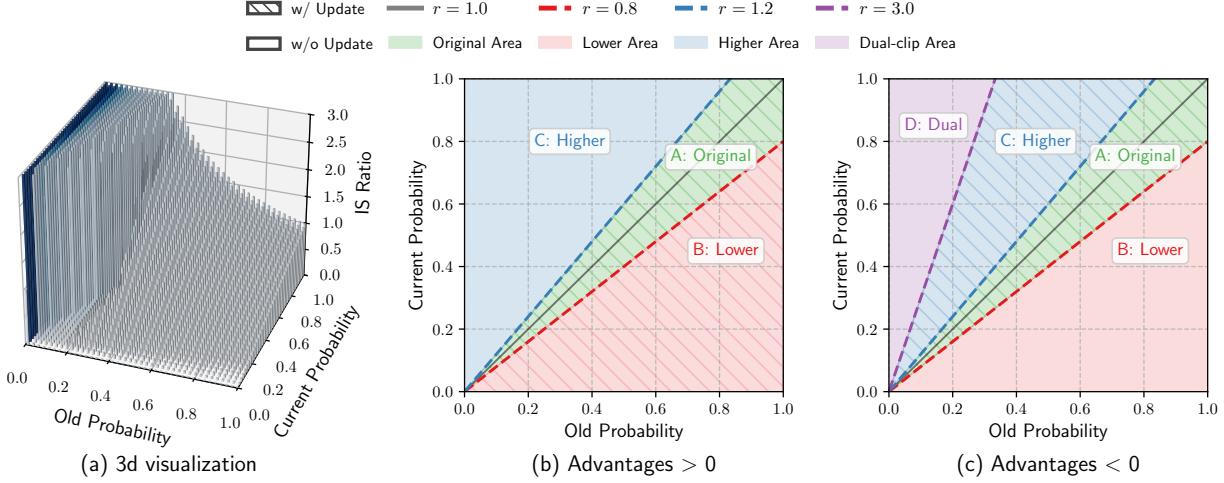


Figure 3: 3D and 2D visualization of IS weights in PPO-Clip.

The design principle of PPO-Clip is to ensure training stability by preventing tokens that already have a strong advantage in the update direction from dominating the update. This avoids overly aggressive parameter changes that could push the model too far from the old policy. An ideal weighting scheme might look like this: along the update direction of the advantage, the lower a token’s probability is relative to the old policy, the larger its training weight should be. Conversely, the higher its probability, the smaller the weight should be. There are two reasons for this: (1) Assigning higher weight to tokens that are lagging behind accelerates their learning progress. (2) Such tokens are already far from the old policy, so updates pose less risk of destabilization. To make this more intuitive, we visualize IS weights in a three-dimensional coordinate plot, where the z-axis represents the original IS ratio, as shown in Figure 3(a). It can be seen that when the current probability is low, the corresponding IS weight is also small, resulting in insufficient training for these tokens.

4.2. Weight Misallocation of Positive Tokens in PPO-Clip

As shown in Figure 3(c), for negative-advantage samples, the weight distribution behaves as expected: weights decrease gradually from the top-left to the bottom-right of the figure. However, for positive-advantage samples, shown in Figure 3(b), the allocation is the opposite of our intuition. Tokens in the top-left region, whose probabilities under the current policy are already much larger than under the old policy, are given larger weights, while tokens in the bottom-right region, with lower current probabilities, are assigned very small weights.

This mismatch causes two problems: (1) tokens in the bottom-right region, which are clearly lagging behind the old policy, are suppressed further by excessively low weights. For example, if the old policy probability is 0.9 and the current policy probability is 0.1, the assigned weight is merely $1/9$, resulting in a weak update signal and insufficient learning. (2) for tokens in the top-left region that already have a significant advantage, they are assigned disproportionately high weights. On the one hand, the model is more likely to deviate from the old policy, undermining training stability. On the other hand, the model overfits on positive samples, further amplifying their probabilities after the update. In the next update, their weights increase even more, forming a **self-reinforcing loop**. This mechanism underlies the previously observed phenomena such as entropy collapse and increased output repetition. For more details on how to distinguish healthy convergence and local optima, please refer to Appendix B.

4.3. Experimental Validation

4.3.1. Setup

To empirically validate the preceding analysis, we conduct a controlled comparison between the original DAPO baseline and a modified variant. In this variant, we replace the token-level IS ratios of **positive-advantage samples** with their **response-level average IS ratios**, while keeping the negative-advantage samples unchanged. This design isolates the impact of the mismatched IS weights identified in Section 4.2, ensuring that any observed behavioral differences arise solely from the reweighting of positive tokens.

If our hypothesis is correct, the modified setup should partially alleviate the weight mismatch issue. Specifically, we expect to observe (1) a slower and smoother decline in entropy, indicating reduced overfitting; (2) slower growth in repetition rate and clip ratio, suggesting improved stability; and (3) comparable or even improved task performance despite reduced update aggressiveness.

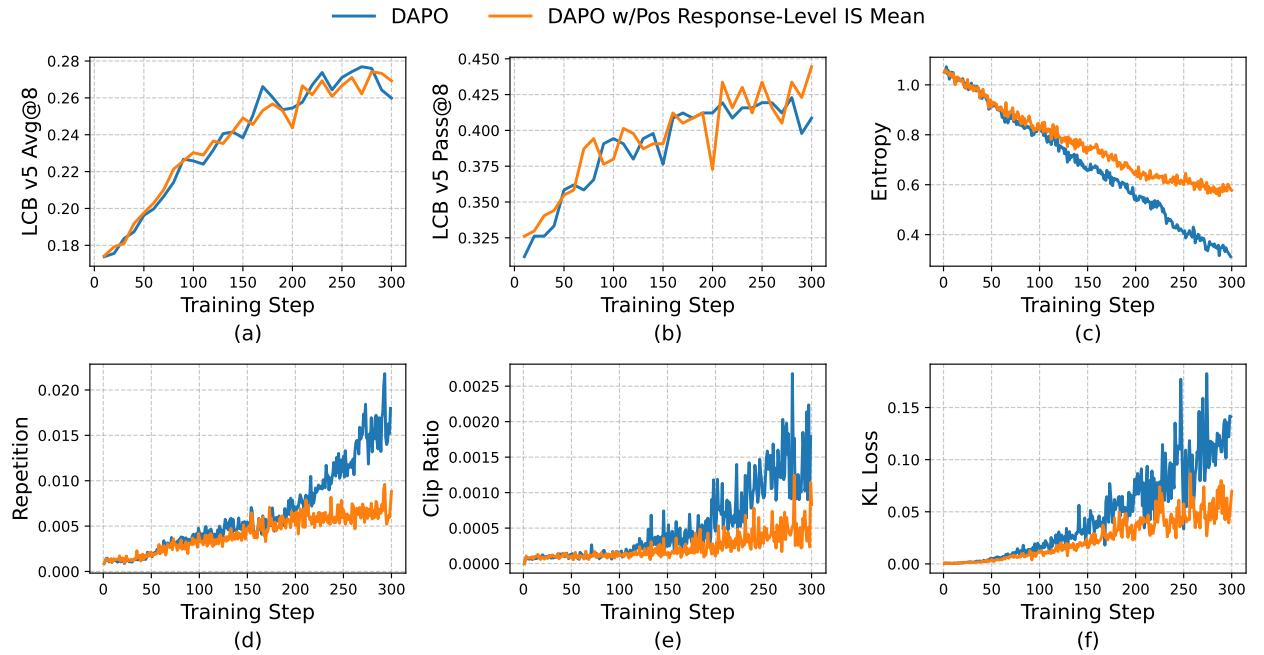


Figure 4: Comparison of DAPO and DAPO with positive samples using response-level IS weights.

4.3.2. Results

The training dynamics of both models are presented in Figure 4. We summarize the observations as follows.

Smoother training dynamics. After replacing the positive-token IS weights with response-level means, all training curves become substantially smoother. The entropy decline slows down (Figure 4(c)), and the increases in repetition rate (d), clip ratio (e), and KL divergence loss (f) are noticeably moderated without the accelerating trends observed in DAPO. These results provide strong empirical evidence that the original IS ratio design indeed leads to unstable optimization through excessive weighting of high-probability tokens.

Comparable performance with stable training. As shown in Figure 5(a), the modified method achieves performance improvements at a rate comparable to DAPO, demonstrating that stable training does not compromise convergence speed. In the later training stages, while DAPO exhibits degradation in multiple metrics, the modified variant remains steady, indicating a higher optimization ceiling and greater robustness to local instability.

Improved exploration. An intriguing observation appears in the pass@8 metric (Figure 4(b)): the modified approach achieves a clear performance advantage in the later training phase. This suggests that mitigating overfitting on positive samples not only improves stability but also prevents premature convergence to local optima, thereby encouraging more diverse and exploratory behavior in the policy.

Takeaways for IS Weights of Positive Tokens

- The standard PPO-Clip design introduces a **token-weight mismatch for positive-advantage samples**, where high-probability tokens receive disproportionately large update weights.
- This imbalance leads to overfitting, entropy collapse, and increased repetition, ultimately pushing the policy toward a local optimum and limiting its capacity for further improvement.
- Reweighting positive samples with response-level IS ratios effectively mitigates these issues, confirming the validity of our preliminary analysis and motivating the design of our proposed method in the next section.

5. Method

5.1. Asymmetric Importance Sampling

Based on the analysis in Section 3 and 4, we have identified the mismatched IS weights of positive tokens and the consequence of unstable training. To tackle these issues, we propose Asymmetric Importance Sampling (AIS), which is a simple yet effective approach for clipping and IS ratio computation. AIS inverts the weights of positive samples, aligning their update behavior with that of negative samples. In other words, tokens whose current policy probabilities are lower than the old policy should be assigned higher learning weights, while those with higher probabilities should receive lower weights.

Specifically, the implementation can be divided into three steps:

Step 1: Token Masking. We retain the original clipping mechanism in GRPO. The gradient of tokens that satisfy: (1) $r_t^i(\theta) < 1 - \varepsilon_{\text{low}}$ ($\hat{A}_t^i < 0$) or (2) $r_t^i(\theta) > 1 - \varepsilon_{\text{high}}$ ($\hat{A}_t^i > 0$) will be masked in a hard clipping manner.

Step 2: Weight Flipping. For tokens with negative advantage values, AIS ratio is the same with that of GRPO, i.e., $\hat{r}_t^i = r_t^i$. For tokens with $\hat{A}_t^i > 0$, we use the reciprocal of their IS weights and the AIS ratio is computed as:

$$\hat{r}_t^i = \frac{\pi_{\theta_{\text{old}}}(o_t^i | q, o_{<t}^i) \pi_{\theta}(o_t^i | q, o_{<t}^i)}{\text{sg}(\pi_{\theta}^2(o_t^i | q, o_{<t}^i))}, \quad (4)$$

where $\text{sg}(\cdot)$ denotes stop gradient operation.

Step 3: Dual Clipping. PPO-Clip usually uses a dual-clip mechanism (Ye et al., 2020) to handle cases where, for $\hat{A} < 0$, extremely small or large ratios could lead to weight explosion, destabilizing training. Originally, for the $\hat{A} > 0$ region, this problem was naturally avoided by the masking mechanism. However, since we now invert the weights for positive samples, extreme cases shift to the right-hand side of the $\hat{A} > 0$ region (region B in Figure 3(b)). Therefore, when using AIS, positive-sample tokens also require dual-clip. Specifically, this dual-clip is implemented using a **soft clipping** manner, which only clips the values but retains the gradient.

It is important to note that tokens clipped by dual-clip are fundamentally different from tokens masked in the first step. The latter are blocked because they already have sufficient advantage in the update direction, whereas the former are tokens that lag significantly behind the old policy but need their weight magnitude constrained due to abnormal computation. We still want these tokens to participate in training, so we use the soft clipping proposed by CISPO (Chen et al., 2025) for these dual-clipped tokens.

5.2. Gradient Analysis

The gradient of original GRPO is as follows:

$$\begin{aligned}\nabla_{\theta} \mathcal{J}(\theta) &= \nabla_{\theta} \mathbb{E}_{q \sim \mathcal{D}, \{o^i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}(\cdot|q)} \left[\frac{1}{G} \sum_{i=1}^G r_t^i(\theta) \hat{A}_t^i \right] \\ &= \mathbb{E}_{q \sim \mathcal{D}, \{o^i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}(\cdot|q)} \left[\frac{1}{G} \sum_{i=1}^G \frac{\nabla_{\theta} \pi_{\theta}(o_t^i | q, o_{<t}^i)}{\pi_{\theta_{\text{old}}}(o_t^i | q, o_{<t}^i)} \hat{A}_t^i \right] \\ &= \mathbb{E}_{q \sim \mathcal{D}, \{o^i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}(\cdot|q)} \left[\frac{1}{G} \sum_{i=1}^G \frac{\pi_{\theta}(o_t^i | q, o_{<t}^i)}{\pi_{\theta_{\text{old}}}(o_t^i | q, o_{<t}^i)} \nabla_{\theta} \log \pi_{\theta}(o_t^i | q, o_{<t}^i) \hat{A}_t^i \right]\end{aligned}\tag{5}$$

Then, we derive the gradient of ASPO as follows:

$$\begin{aligned}\nabla_{\theta} \mathcal{J}(\theta) &= \nabla_{\theta} \mathbb{E}_{q \sim \mathcal{D}, \{o^i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}(\cdot|q)} \left[\frac{1}{G} \sum_{i=1}^G \hat{r}_t^i(\theta) \hat{A}_t^i \right] \\ &= \mathbb{E}_{q \sim \mathcal{D}, \{o^i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}(\cdot|q)} \left[\frac{1}{G} \sum_{i=1}^G \frac{\pi_{\theta_{\text{old}}}(o_t^i | q, o_{<t}^i) \nabla_{\theta} \pi_{\theta}(o_t^i | q, o_{<t}^i)}{\text{sg}(\pi_{\theta}^2(o_t^i | q, o_{<t}^i))} \hat{A}_t^i \right] \\ &= \mathbb{E}_{q \sim \mathcal{D}, \{o^i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}(\cdot|q)} \left[\frac{1}{G} \sum_{i=1}^G \frac{\pi_{\theta_{\text{old}}}(o_t^i | q, o_{<t}^i) \pi_{\theta}(o_t^i | q, o_{<t}^i)}{\pi_{\theta}^2(o_t^i | q, o_{<t}^i)} \nabla_{\theta} \log \pi_{\theta}(o_t^i | q, o_{<t}^i) \hat{A}_t^i \right] \\ &= \mathbb{E}_{q \sim \mathcal{D}, \{o^i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}(\cdot|q)} \left[\frac{1}{G} \sum_{i=1}^G \frac{\pi_{\theta_{\text{old}}}(o_t^i | q, o_{<t}^i)}{\pi_{\theta}(o_t^i | q, o_{<t}^i)} \nabla_{\theta} \log \pi_{\theta}(o_t^i | q, o_{<t}^i) \hat{A}_t^i \right]\end{aligned}\tag{6}$$

It is worth noting that the gradient of ASPO in (6) differs from the original gradient of GRPO in (5) at the highlighted red term. From the above derivation, we can observe that the gradient of ASPO is positively correlated with $\frac{1}{\pi_{\theta}}$, indicating that the gradient becomes larger when the probability of a token is lower.

Table 1: Evaluation results on mathematical benchmarks. The results of ASPO are shaded and the highest values are bolded.

Method	AIME24		AIME25		AMC23		MATH-500		Minerva		Olympiad		Avg.
	avg@64	pass@64	avg@64	pass@64	avg@64	pass@64	avg@4	pass@4	avg@8	pass@8	avg@4	pass@4	
DeepSeek-R1-1.5B	30.6	80.0	23.5	63.3	70.7	100.0	83.6	92.4	27.6	48.2	44.6	59.4	46.8
↳ DAPO	42.1	80.0	28.6	56.7	80.3	97.5	87.6	94.6	29.2	46.3	53.2	65.8	53.5
↳ DeepScaleR-1.5B	42.0	83.3	29.0	63.3	81.3	100.0	87.7	93.6	30.3	51.1	50.7	61.0	53.5
↳ FastCuRL-1.5B-V3	48.1	80.0	32.7	60.0	86.4	95.0	89.8	94.0	33.6	50.0	55.3	64.3	57.7
↳ Nemotron-1.5B	48.0	76.7	33.1	60.0	86.1	97.5	90.6	93.6	35.3	47.8	59.2	66.8	58.7
↳ ASPO-Math-1.5B	49.0	80.0	35.1	70.0	87.2	95.0	90.5	94.4	35.1	50.4	58.8	66.9	59.3

Table 2: Evaluation results on code benchmarks. The results of ASPO are shaded and the highest values are bolded.

Method	LCB v5 (2024.08.01-2025.02.01)		LCB v6 (2025.02.01-2025.05.01)		Avg.
	avg@8	pass@8	avg@16	pass@16	
DeepSeek-R1-1.5B	16.7	29.0	17.2	34.4	17.0
↳ DAPO	26.0	40.5	27.6	43.5	26.8
↳ DeepCoder-1.5B	23.3	39.1	22.6	42.0	23.0
↳ Nemotron-1.5B	26.1	35.5	29.5	42.8	27.8
↳ ASPO-Code-1.5B	31.5	47.0	30.5	46.0	31.0

6. Experiments

6.1. Setup

Models and Baselines. We conduct experiments using DeepSeek-R1-Distill-Qwen-1.5B ([DeepSeek-AI et al., 2025](#)) as the base model and compare ASPO against several representative baselines: (1) Base Model, (2) DAPO ([Yu et al., 2025](#)), (3) DeepScaleR-1.5B ([Luo et al., 2025b](#)), (4) DeepCoder-1.5B ([Luo et al., 2025a](#)), (5) FastCuRL-1.5B-V3 ([Song et al., 2025](#)), and (6) Nemotron-1.5B ([Liu et al., 2025a](#)). Details of these baselines are provided in Appendix A.

Evaluation. We evaluate models on both mathematical and coding domains. For math, we use six challenging datasets: AIME24 ([MAA, 2024](#)), AIME25 ([MAA, 2025](#)), AMC23 ([MAA, 2023](#)), MATH-500 ([Lightman et al., 2024](#)), Minerva Math ([Lewkowycz et al., 2022](#)), and OlympiadBench ([He et al., 2024](#)). For coding, we adopt LiveCodeBench v5 (2024.08.01–2025.02.01) and v6 (2025.02.01–2025.05.01) ([Jain et al., 2025](#)). Inference is performed using vLLM ([Kwon et al., 2023](#)) with a maximum output length of 32,768 tokens and a temperature of 0.8. We report both *avg@K* and *pass@K* for each benchmark.

Implementation Details. We implement ASPO based on DAPO ([Yu et al., 2025](#)) with the *verl* framework ([Sheng et al., 2025](#)). The training batch size is set to 64, with a mini-batch size of 32. The learning rate is 1×10^{-6} . For each prompt, 16 responses are sampled with a maximum response length of 32,768 and a temperature of 1.0. Additional implementation details are provided in Appendix A.

6.2. Main Results

The results in Table 1 and Table 2 show that ASPO improves upon the base model by 12.5% and 17.0% on math and coding tasks, respectively. Moreover, ASPO consistently outperforms DAPO and several

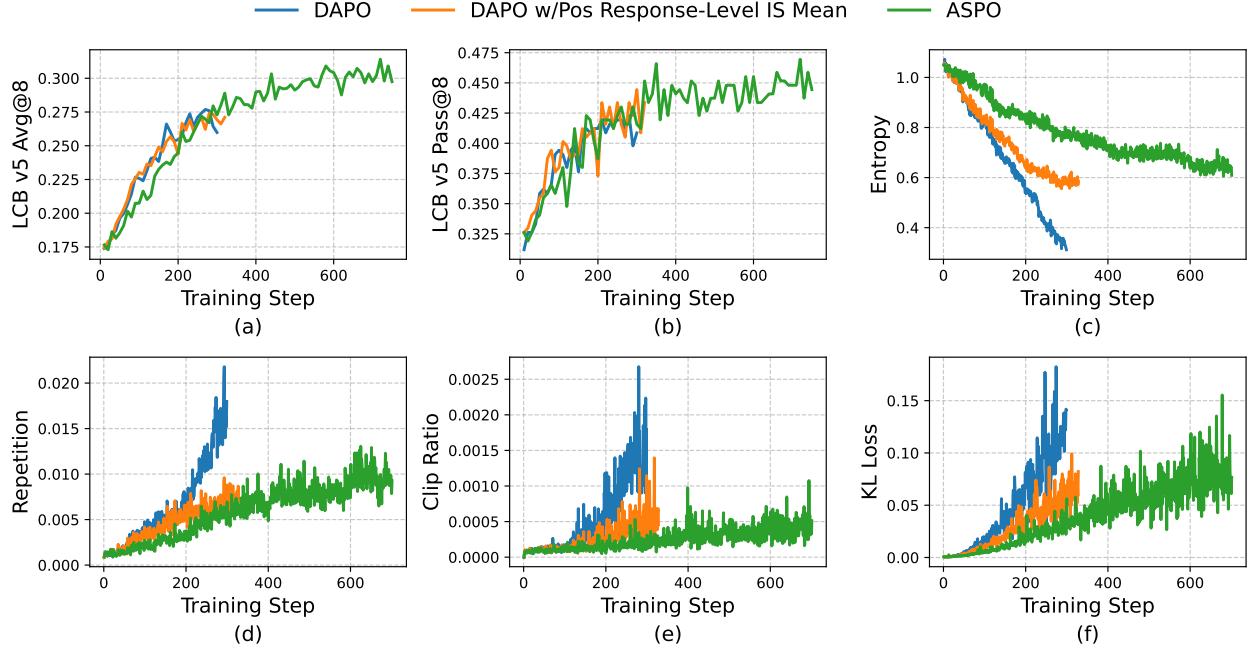


Figure 5: Comparison among DAPO, DAPO with response-level IS weights for positive samples, and ASPO.

strong OSRL methods averaged across all evaluated benchmarks, demonstrating the effectiveness of ASPO.

6.3. Analysis

As shown in Figure 5, compared with DAPO, introducing AIS yields trends that align with our earlier findings, but with even stronger improvements. Specifically, entropy decreases more gradually (c), the repetition rate (d) and token clipping ratio (e) grow more slowly, and all metrics eventually stabilize, showing characteristics of **healthy convergence** as discussed in Appendix B.

More importantly, since entropy declines more smoothly and remains at a higher level, the model avoids premature collapse and continues learning effectively. As training progresses, performance steadily improves, significantly surpassing the best results achieved by GRPO-based training.

It is also noteworthy that in the early training stages, models trained with AIS exhibit slightly slower improvement on test metrics (Figure 5(a)) compared to other variants. This occurs because inverting positive-sample weights reduces the overall average weight, leading to slower initial fitting of positive samples. However, as training continues, the performance not only catches up but ultimately surpasses the other approaches.

7. Related Work

7.1. RL for Large Language Models

Reinforcement Learning from Human Feedback (RLHF) (Christiano et al., 2017; Liu et al., 2022) has achieved remarkable success in aligning LLMs with human values. Recently, DeepSeek-R1 (Shao et al., 2024) and the GRPO algorithm (Shao et al., 2024) have demonstrated that RLVR effectively

enhances the reasoning capabilities of LLMs. Subsequent works have extended the GRPO algorithm for RLVR (Luo et al., 2025b; Yu et al., 2025; Yue et al., 2025; He et al., 2025; Wang et al., 2025a; Liu et al., 2025c). Our method follows this GRPO-based line of research but introduces a new approach of the clipping and IS mechanism to address inherent limitations in GRPO.

7.2. Clipping Mechanism in RL

PPO (Schulman et al., 2017) introduces clipping as an alternative to KL divergence to constrain policy updates relative to the reference policy, and GRPO (Shao et al., 2024) adopts this clipping loss for LLM RL. However, CISPO (Chen et al., 2025) observes that gradients for clipped tokens are masked and proposes preserving them based on the PPO-Clip objective (Schulman et al., 2017) but without the conservative min operation. GSPO (Zheng et al., 2025) argues that the optimization objective should match the reward’s granularity, proposing sequence-level clipping and IS ratio computation. DCPO (Yang et al., 2025) employs dynamic-adaptive clipping ranges instead of fixed bounds. Our method also targets the clipping term in GRPO but differs primarily by focusing on the IS ratio, incorporating a reciprocal weight for positive tokens.

8. Conclusion

In this paper, we focus on the clipping mechanism in RL and identify a fundamental issue in the clipping mechanism of GRPO-based methods. To this end, we propose ASPO with flipped IS ratios of positive tokens to stabilize OSRL for LLMs. Experimental results across mathematical and coding domains demonstrate that our method effectively alleviates token-level weight mismatch, mitigates entropy collapse, and improves training stability, leading to superior model performance than the baselines.

Limitations

Due to the limit of computational resources, the experiments are only conducted on the 1.5B scale model. Also, the conclusions of current experiments may only apply to GRPO-based algorithms for LLMs. For PPO-based algorithms with token-level IS ratios and advantages or step-level process-supervised methods (Liu et al., 2025b,c), whether importance sampling is not important still requires further investigation.

References

- Chenxin An, Zhihui Xie, Xiaonan Li, Lei Li, Jun Zhang, Shansan Gong, Ming Zhong, Jingjing Xu, Xipeng Qiu, Mingxuan Wang, and Lingpeng Kong. Polaris: A post-training recipe for scaling reinforcement learning on advanced reasoning models, 2025. URL <https://hkunlp.github.io/blog/2025/Polaris>.
- 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.
- Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing*

Systems, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/d5e2c0adad503c91f91df240d0cd4e49-Paper.pdf.

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*, 2025.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhua Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanja Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. OlympiadBench: A challenging benchmark for promoting AGI with olympiad-level bilingual multimodal scientific problems. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3828–3850, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.211. URL <https://aclanthology.org/2024.acl-long.211>.

Jujie He, Jiacai Liu, Chris Yuhao Liu, Rui Yan, Chaojie Wang, Peng Cheng, Xiaoyu Zhang, Fuxiang Zhang, Jiacheng Xu, Wei Shen, et al. Skywork open reasoner 1 technical report. *arXiv preprint arXiv:2505.22312*, 2025.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=chfJJYC3iL>.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, SOSP ’23, page 611–626, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400702297. doi: 10.1145/3600006.3613165. URL <https://doi.org/10.1145/3600006.3613165>.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramaresh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. Solving Quantitative Reasoning Problems with Language Models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 3843–3857. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/18abbeef8cfe9203fdf9053c9c4fe191-Paper-Conference.pdf.

Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, Thomas Hubert, Peter Choy, Cyprien de Masson d’Autume, Igor Babuschkin, Xinyun Chen, Po-Sen Huang, Johannes Welbl, Sven Gowal, Alexey Cherepanov, James Molloy, Daniel J. Mankowitz, Esme Sutherland Robson, Pushmeet Kohli, Nando de Freitas, Koray Kavukcuoglu, and Oriol Vinyals. Competition-level code generation with alphacode. *Science*, 378(6624):1092–1097, 2022. doi: 10.1126/science.abq1158. URL <https://www.science.org/doi/abs/10.1126/science.abq1158>.

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*, 2024. URL <https://openreview.net/forum?id=v8L0pN6E0i>.

Mingjie Liu, Shizhe Diao, Ximing Lu, Jian Hu, Xin Dong, Yejin Choi, Jan Kautz, and Yi Dong. Prorl: Prolonged reinforcement learning expands reasoning boundaries in large language models. *arXiv preprint arXiv:2505.24864*, 2025a.

Runze Liu, Fengshuo Bai, Yali Du, and Yaodong Yang. Meta-reward-net: Implicitly differentiable reward learning for preference-based reinforcement learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 22270–22284. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/8be9c134bb193d8bd3827d4df8488228-Paper-Conference.pdf.

Runze Liu, Junqi Gao, Jian Zhao, Kaiyan Zhang, Xiu Li, Biqing Qi, Wanli Ouyang, and Bowen Zhou. Can 1b llm surpass 405b llm? rethinking compute-optimal test-time scaling. *arXiv preprint arXiv:2502.06703*, 2025b.

Runze Liu, Jiakang Wang, Yuling Shi, Zhihui Xie, Chenxin An, Kaiyan Zhang, Jian Zhao, Xiaodong Gu, Lei Lin, Wenping Hu, Xiu Li, Fuzheng Zhang, Guorui Zhou, and Kun Gai. Attention as a compass: Efficient exploration for process-supervised rl in reasoning models. *arXiv preprint arXiv:2509.26628*, 2025c.

Michael Luo, Sijun Tan, Roy Huang, Ameen Patel, Alpay Ariyak, Qingyang Wu, Xiaoxiang Shi, Rachel Xin, Colin Cai, Maurice Weber, Ce Zhang, Li Erran Li, Raluca Ada Popa, and Ion Stoica. Deepcoder: A fully open-source 14b coder at o3-mini level. <https://pretty-radio-b75.notion.site/DeepCoder-A-Fully-Open-Source-14B-Coder-at-O3-Mini-Level-1cf81902c14680b3bee5eb349> 2025a. Notion Blog.

Michael Luo, Sijun Tan, Justin Wong, Xiaoxiang Shi, William Y. Tang, Manan Roongta, Colin Cai, Jeffrey Luo, Li Erran Li, Raluca Ada Popa, and Ion Stoica. Deepscaler: Surpassing o1-preview with a 1.5b model by scaling rl. <https://pretty-radio-b75.notion.site/DeepScaleR-Surpassing-O1-Preview-with-a-1-5B-Model-by-Scaling-RL-19681902c1468005b2025b>. Notion Blog.

MAA. American mathematics contest 12 (amc 12), November 2023. URL https://artofproblemsolving.com/wiki/index.php/AMC_12_Problems_and_Solutions.

MAA. American invitational mathematics examination (aime), February 2024. URL https://artofproblemsolving.com/wiki/index.php/AIME_Problems_and_Solutions.

MAA. American invitational mathematics examination (aime), February 2025. URL https://artofproblemsolving.com/wiki/index.php/AIME_Problems_and_Solutions.

Guilherme Penedo, Anton Lozhkov, Hynek Kydlíček, Loubna Ben Allal, Edward Beeching, Agustín Pi-queres Lajarín, Quentin Gallouédec, Nathan Habib, Lewis Tunstall, and Leandro von Werra. Codeforces. <https://huggingface.co/datasets/open-r1/codeforces>, 2025.

Doina Precup, Richard S. Sutton, and Satinder P. Singh. Eligibility traces for off-policy policy evaluation. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, page 759–766, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc. ISBN 1558607072.

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, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, EuroSys '25, page 1279–1297, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400711961. doi: 10.1145/3689031.3696075. URL <https://doi.org/10.1145/3689031.3696075>.

Mingyang Song, Mao Zheng, Zheng Li, Wenjie Yang, Xuan Luo, Yue Pan, and Feng Zhang. Fastcurl: Curriculum reinforcement learning with stage-wise context scaling for efficient training r1-like reasoning models. *arXiv preprint arXiv:2503.17287*, 2025.

Jiakang Wang, Runze Liu, Fuzheng Zhang, Xiu Li, and Guorui Zhou. Stabilizing knowledge, promoting reasoning: Dual-token constraints for rlvr. *arXiv preprint arXiv:2507.15778*, 2025a.

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*, 2025b.

Shihui Yang, Chengfeng Dou, Peidong Guo, Kai Lu, Qiang Ju, Fei Deng, and Rihui Xin. Dcpo: Dynamic clipping policy optimization. *arXiv preprint arXiv:2509.02333*, 2025.

Deheng Ye, Zhao Liu, Mingfei Sun, Bei Shi, Peilin Zhao, Hao Wu, Hongsheng Yu, Shaojie Yang, Xipeng Wu, Qingwei Guo, Qiaobo Chen, Yinyuting Yin, Hao Zhang, Tengfei Shi, Liang Wang, Qiang Fu, Wei Yang, and Lanxiao Huang. Mastering complex control in moba games with deep reinforcement learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):6672–6679, Apr. 2020.

doi: 10.1609/aaai.v34i04.6144. URL <https://ojs.aaai.org/index.php/AAAI/article/view/6144>.

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.

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, Pengfei Li, Yu Fu, Xingtai Lv, Yuchen Zhang, Sihang Zeng, Shang Qu, Haozhan Li, Shijie Wang, Yuru Wang, Xinwei Long, Fangfu Liu, Xiang Xu, Jiaze Ma, Xuekai Zhu, Ermo Hua, Yihao Liu, Zonglin Li, Huayu 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. *arXiv preprint arXiv:2509.08827*, 2025a.

Zhenru Zhang, Chujie Zheng, Yangzhen Wu, Beichen Zhang, Runji Lin, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. The lessons of developing process reward models in mathematical reasoning. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10495–10516, Vienna, Austria, July 2025b. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.547. URL <https://aclanthology.org/2025.findings-acl.547/>.

Jian Zhao, Runze Liu, Kaiyan Zhang, Zhimu Zhou, Junqi Gao, Dong Li, Jiafei Lyu, Zhouyi Qian, Binqing Qi, Xiu Li, et al. Genprm: Scaling test-time compute of process reward models via generative reasoning. *arXiv preprint arXiv:2504.00891*, 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.

A. Experimental Details

Baselines. We conduct experiments using DeepSeek-R1-Distill-Qwen-1.5B (DeepSeek-AI et al., 2025) as the base model and compare ASPO with the following baselines:

- **Base Model:** The original model without any RL fine-tuning.
- **DAPO** (Yu et al., 2025): A strong OSRL algorithm built upon GRPO (Shao et al., 2024).
- **DeepScaleR-1.5B** (Luo et al., 2025b): A 1.5B model trained for mathematical reasoning with iterative context-length expansion.
- **DeepCoder-1.5B** (Luo et al., 2025a): A 1.5B model trained on code datasets using similar context expansion strategies as DeepScaleR.
- **FastCuRL-1.5B-V3** (Song et al., 2025): A competitive 1.5B model trained with curriculum reinforcement learning.
- **Nemotron-1.5B** (Liu et al., 2025a): A strong 1.5B reasoning model with reference policy resetting.

Evaluation. We follow standard practice (Liu et al., 2025a; Wang et al., 2025a) and report both $avg@K$ and $pass@K$ metrics. For benchmarks with fewer samples (AIME24/25 and AMC23), we set $K = 64$. For LiveCodeBench v6, we use $K = 16$; for LiveCodeBench v5 and Minerva Math, $K = 8$; and for MATH-500 and OlympiadBench, $K = 4$. To ensure fair and accurate evaluation, we adopt the official verification functions from both DeepScaleR and Math-Verify¹ for mathematical problems, following the protocol in He et al. (2025).

Implementation Details. For training data, we use a mixture of DeepScaleR-Preview-Dataset (Luo et al., 2025b), Skywork-OR1-RL-Data (He et al., 2025), and DAPO (Yu et al., 2025) for mathematical tasks. For coding, we employ DeepCoder (Luo et al., 2025a), CodeContests (Li et al., 2022), and CodeForces (Penedo et al., 2025) datasets. All datasets are cleaned and filtered following the preprocessing protocol of Wang et al. (2025a). After filtering, the mathematical dataset contains 70.8k samples, while the coding dataset contains 8.9k samples. The clipping ranges of DAPO and ASPO are set to $\varepsilon_{low} = 0.2$ and $\varepsilon_{high} = 0.28$. All experiments are conducted on 8 NVIDIA H800 GPUs, with each full training run taking approximately 7 days.

B. How to distinguish between “healthy convergence” and “local optima”?

To better contextualize our analysis, we introduce an important concept in RL training: how to distinguish between **healthy convergence** and **local optima**.

In RL training, healthy convergence typically exhibits the following characteristics:

- **Entropy** decreases gradually from a relatively high initial value and stabilizes at a small but positive level, indicating that the policy becomes more deterministic while retaining moderate exploration.
- The **reward curve** increases steadily and eventually plateaus at a stable high value.

¹<https://github.com/huggingface/Math-Verify>

- **Clip ratios** and **KL divergence loss** remain stable during later training stages, without drastic fluctuations.

In contrast, when the training becomes trapped in a local optimum, the model enters a self-reinforcing **policy-data distribution loop**, characterized by:

- **Entropy** collapses rapidly toward zero.
- The **reward curve** stagnates with no further improvement.
- Persistently high **clip ratios** without meaningful policy updates.

When training GRPO-based approaches, we may observe this phenomenon: in late-stage training, entropy collapses, repetition rates spike, clip ratios surge, and the model's test performance begins to degrade, indicating convergence to a local optimum.

In early training, **moderate entropy reduction** and a **gradual increase in clip ratio** are expected. If accompanied by an **increasing rewards**, it indicates that the policy is learning effectively. However, when later stages exhibit an abrupt entropy drop, reward stagnation, and persistently high clip ratios without further progress, it clearly signals that the model has fallen into a **local optimum**.

The underlying cause lies in the **token-level weight mismatch for positive samples in PPO-Clip** identified in Section 4, which drives the model to **overfit** certain high-probability tokens, eventually leading to entropy collapse and training degradation.