

Rethinking Sample Polarity in Reinforcement Learning with Verifiable Rewards

Xinyu Tang^{1,4*}, Yuliang Zhan^{1,4*}, Zhixun Li^{2*}, Wayne Xin Zhao^{1,4†},
Zhenduo Zhang³, Zujie Wen³, Zhiqiang Zhang³, Jun Zhou³

¹Gaoling School of Artificial Intelligence, Renmin University of China

²The Chinese University of Hong Kong ³Ant Group

⁴Beijing Key Laboratory of Research on Large Models and Intelligent Governance

Abstract

Large reasoning models (LRMs) are typically trained using reinforcement learning with verifiable reward (RLVR) to enhance their reasoning abilities. In this paradigm, policies are updated using both positive and negative self-generated rollouts, which correspond to distinct *sample polarities*. In this paper, we provide a systematic investigation into how these sample polarities affect RLVR training dynamics and behaviors. We find that positive samples sharpen existing correct reasoning patterns, while negative samples encourage exploration of new reasoning paths. We further explore how adjusting the advantage values of positive and negative samples at both the polarity level and the token level affects RLVR training. Based on these insights, we propose an Adaptive and Asymmetric token-level Advantage shaping method for Policy Optimization, namely **A3PO**, that more precisely allocates advantage signals to key tokens across different polarities. Experiments across five reasoning benchmarks demonstrate the effectiveness of our approach. The code is available at <https://github.com/RUCAIBox/Sample-Polarity-of-RLVR>.

1 Introduction

Large reasoning models (DeepSeek-AI et al., 2025; Zhao et al., 2023a) have recently gained significant attention due to their impressive performance in mathematical, coding, and scientific reasoning tasks. These models typically adopt the reinforcement learning with verifiable reward (RLVR) paradigm (Yu et al., 2025; Chen et al., 2025a; Tang et al., 2026), where they generate multiple long chain-of-thought reasoning trajectories and use verifiable binary rewards to assess the correctness of the final answers. The reward signals are then used to update the model’s policy. Unlike supervised

fine-tuning (Zhang et al., 2023; Zhan et al., 2024; Li et al., 2026), which imitates external teachers by memorizing correct examples, RLVR enables models to learn from their own generated rollouts, including both positive and negative samples. This approach iteratively enhances the model’s reasoning ability by leveraging the group advantage from these samples. However, within this framework, the distinct roles of positive and negative samples, which are referred to as *sample polarity*, remain underexplored.

To explore this question, prior studies have attempted to analyze the respective contributions of positive and negative samples in RLVR. For instance, Zhu et al. (2025) decomposes RLVR into two learning paradigms: positive sample reinforcement (PSR) and negative sample reinforcement (NSR). Their findings show that training solely with negative samples yields surprising results. However, their experiments were constrained to a simple math training dataset and a small set of models, which limits the generalizability of their conclusions. Subsequent studies observe an asymmetry between positive and negative samples and propose methods to improve importance sampling (Wang et al., 2025a), advantage shaping (Zhu et al., 2025), and clipping mechanisms (Hao et al., 2025; Xi et al., 2025). Nevertheless, a thorough analysis of how positive and negative samples influence RLVR training dynamics remains incomplete.

In this paper, we systematically analyze the roles of positive and negative samples in RLVR by applying PSR and NSR to three different base LLMs. We find that positive samples sharpen the model’s existing correct reasoning paths, reduce entropy, and result in shorter outputs. In contrast, negative samples promote the discovery of new reasoning patterns, increase entropy, and encourage longer responses. However, using only one sample polarity impairs reasoning performance and boundary, demonstrating that both types are important for

* Equal contribution.

† Corresponding author.

RLVR. Given these distinct roles, a crucial question arises: *how can we modulate their influence to steer the training process?*

To answer this, we further investigate how adjusting the influence of positive and negative samples at different granularities affects RLVR training. At the polarity level, assigning higher weights to positive samples accelerates reward improvement but narrows exploration diversity, whereas emphasizing negative samples encourages broader exploration at the expense of slower reward progress. To examine the training process in finer granularity, we perform token-level advantage shaping to determine which specific tokens in positive and negative samples contribute more to the training dynamics. Our results indicate that weighting tokens based on their entropy and probability has distinct effects for each polarity. Building on these findings, we propose an **A**daptive and **A**symmetric token-level **A**dvantage shaping method for **P**olicy **O**ptimization, namely **A3PO**. This approach dynamically adjusts the advantages of high-probability tokens in negative samples and low-probability tokens in positive samples, enabling finer-grained advantage allocation. Experiments across three LLMs and five reasoning benchmarks validate the effectiveness of **A3PO**.

Our contributions are summarized as follows:

- Through a comprehensive analysis of sample polarity in RLVR, we identify distinct training dynamics between them and observe that both sample types synergistically promote RLVR training.
- We investigate how varying the influence of positive and negative samples at different granularities affects RLVR training through polarity-level and token-level advantage shaping.
- We propose an adaptive and asymmetric token-level advantage shaping method, which enables finer-grained advantage allocation and leads to more effective and stable RLVR training.

2 Related Work

2.1 Reinforcement Learning with Verifiable Rewards

Reinforcement learning with verifiable rewards (RLVR) effectively improves the reasoning ability of large language models (Zhan et al., 2026; Wang et al., 2025c; Tang et al., 2025). Under this paradigm, an LLM acts as a policy model that generates multiple long chain-of-thought reasoning paths (Li et al., 2025a; Liu et al., 2026). The model

is optimized using binary outcome-based rewards, which removes the need for a learned reward model. As a representative algorithm, group relative policy optimization (GRPO) (DeepSeek-AI et al., 2025) computes advantages directly from groups of roll-outs, avoiding reliance on a learned value network and enabling scalable reasoning through zero-RL. Following GRPO, subsequent studies have refined the algorithm by introducing enhanced techniques for advantage estimation (Chen et al., 2025b; Yan et al., 2026), loss aggregation (Zhao et al., 2025; Zheng et al., 2025), importance sampling (Chen et al., 2025a), and sampling strategies (Li et al., 2025c; Chen et al., 2026).

2.2 Sample Polarity in RLVR

In RLVR, both positive and negative samples are important for policy optimization. Positive samples reinforce correct reasoning paths, while negative samples allow models to learn from their mistakes. Prior work has examined the distinct effects of these two sample types (Zhu et al., 2025) and proposed methods that treat them differently (Wang et al., 2025a; Hao et al., 2025; Xi et al., 2025). However, a systematic analysis of how positive and negative samples fundamentally shape training dynamics and how to best modulate their influence remains lacking. To address this, we provide a more thorough investigation of how different sample polarities affect training dynamics and analyze their respective contributions to RLVR. In addition, we conduct a finer-grained analysis of different sample polarities via polarity-level and token-level advantage shaping. A detailed comparison with previous work is provided in Appendix E.

3 Rethinking the Role of Positive and Negative Samples in RLVR

In this section, we analyze how positive and negative samples affect RLVR training across different base LLMs. Specifically, we study reinforcement using only positive and negative samples and compare their training dynamics and model behaviors.

3.1 Experimental Setup

We conduct experiments with three different types of LLMs: a math-enhanced LLM Qwen2.5-7B-Math (Yang et al., 2024), a general pretrained LLM Qwen3-8B-Base (Yang et al., 2025), and a distilled LLM after supervised fine-tuning DeepSeek-R1-Distill-Qwen-7B (DeepSeek-AI et al., 2025). Fol-

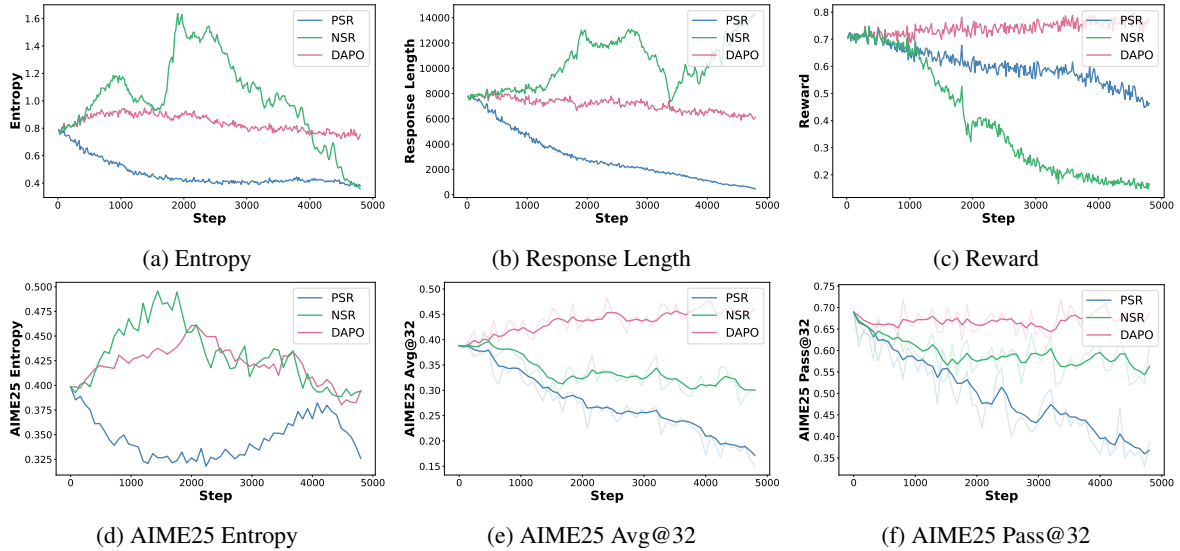


Figure 1: RLVR training dynamics under three training paradigms on Deepseek-R1-Distilled-Qwen-7B.

lowing [Zhu et al. \(2025\)](#), we perform reinforcement separately using only positive and negative samples for each LLM, and include DAPO, which utilizes both types of samples, for comparison. More details on positive and negative sample reinforcement are included in Appendix B.1, and experimental setups are provided in Appendix A.

3.2 Different Training Dynamics of Positive and Negative Sample Reinforcement

Positive samples reduce entropy, negative samples maintain it. As shown in Figure 1a and 1d, reinforcement with only positive samples leads to a rapid decline in model entropy, while reinforcement with only negative samples helps maintain higher entropy levels on both training and validation data. This occurs because positive reinforcement amplifies the logits of tokens that appear in correct solutions, making the model more confident in a narrow set of high-probability predictions. In contrast, negative reinforcement reduces the logits of tokens present in incorrect solutions and indirectly boosts alternatives, thus preserving greater exploration diversity and higher entropy.

Positive samples produce shorter responses, negative samples yield longer ones. Figure 1b shows that models trained with positive samples alone generate increasingly shorter responses, whereas those trained with negative samples produce longer outputs. This is because positive reinforcement rewards the most efficient path to correct answers, implicitly penalizing extra reasoning steps. On

the other hand, negative reinforcement suppresses incorrect tokens without encouraging brevity, allowing models to explore longer reasoning chains.

Using only one sample polarity harms reasoning abilities and boundaries. As shown by the training reward in Figure 1c and validation performance in Figure 1e and 1f, training with only positive or negative samples damages the model’s reasoning ability and boundary, with further degradation over training. It is worth noting that although [Zhu et al. \(2025\)](#) suggests that negative-only reinforcement can improve reasoning boundaries, we find that it only maintains Pass@32 performance comparable to DAPO on Qwen2.5-7B-MATH, indicating that such a conclusion is limited to certain models. This further confirms that both positive and negative samples are essential in RLVR.

Negative samples are key to preserving generalization. As illustrated in Figure 1c and Figure 1e, reward on the training set declines faster or grows slower with negative sample reinforcement compared to positive sample reinforcement. However, models trained with negative samples achieve better performance on the validation set. This suggests that negative samples are crucial for maintaining the model’s generalization ability in RLVR.

3.3 Different Training Dynamics across Base LLMs

Figure 2 illustrates how reward changes during RLVR training across various base LLMs. For Qwen2.5-7B-Math, using only positive or negative

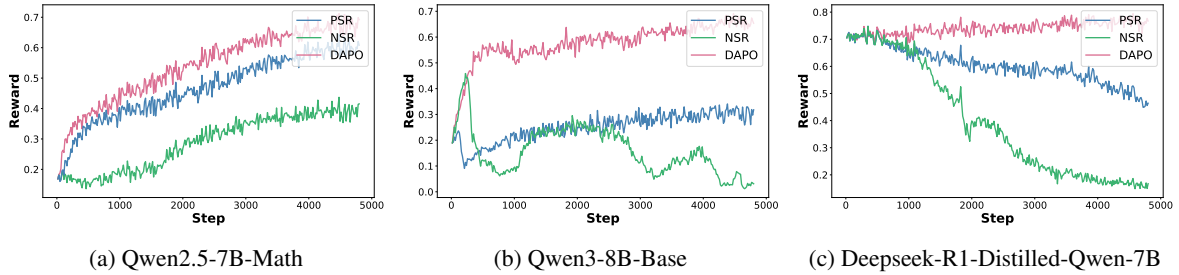


Figure 2: RLVR training reward across different training paradigms and base LLMs.

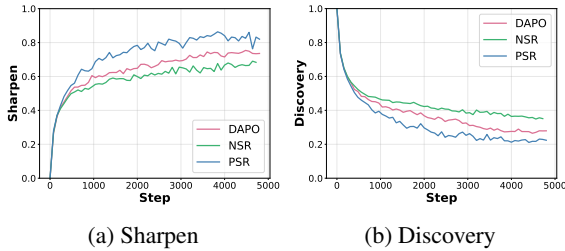


Figure 3: Training behaviors of different paradigms.

samples can improve reward, but is less effective than using both together. Here, both types of samples jointly accelerate RL training and lead to better final performance. In contrast, for Deepseek-R1-Distilled-Qwen-7B, training with only one polarity damages reasoning ability. Only when both positive and negative samples are combined does reward improve consistently.

When training Qwen3-8B-Base with only positive samples, the reward initially drops sharply and then recovers in later stages. We observe that the model exhibits reward hacking, where it learns to guess answers directly rather than perform step-by-step reasoning. On the other hand, using only negative samples results in reward fluctuation without steady progress. This occurs because negative sample reinforcement continuously shifts probability away from high-probability tokens to others, which increases the likelihood of generating irrelevant tokens and ultimately leads to mojibake output. Detailed analyses of accuracy changes on validation samples are presented in Appendix F.

3.4 Positive Samples Encourage Sharpen, Negative Samples Help Discovery

There are two prevailing views on RLVR: sharpening and discovery, which appear to be in direct opposition (Zhang et al., 2025). The sharpening view (Yue et al., 2025a) posits that RLVR does not create genuinely new patterns, but instead refines and reweights correct responses already available in the base model. In contrast, the discovery

Method	Step 25	Step 50	Step 75	Step 100
PSR	18.3%	13.3%	13.3%	10.0%
NSR	25.0%	28.3%	30.0%	38.3%
DAPO	23.3%	26.7%	26.7%	20.0%

Table 1: Percentage of new reasoning paths for PSR, NSR, and DAPO across training steps on AIME problems.

view (Liu et al., 2025) suggests that RLVR can uncover new reasoning patterns not acquired during pre-training and not generated through repeated sampling. To investigate how sample polarities contribute to each perspective, we examine the model’s generated rollouts from an n-gram perspective. Here, we define two metrics (as detailed in Appendix C):

- **Sharpening:** The proportion of n-grams in the current rollout that have appeared in previously correct rollouts, which measures how much the model reinforces existing correct patterns.
- **Discovery:** The proportion of n-grams in the current rollout that have never appeared before, reflecting the model’s exploration of new paths.

The results are shown in Figure 3. We observe that as training proceeds, the model increasingly reinforces previously correct reasoning processes while reducing the frequency of exploration. For sharpening, the ranking is PSR > DAPO > NSR. For discovery, the ranking is NSR > DAPO > PSR. These findings indicate that positive samples help models exploit and strengthen previously correct trajectories, while negative samples facilitate exploration of unseen reasoning paths.

To further capture finer-grained information from rollouts, we conduct a semantic-level analysis using LLM-as-a-judge (*i.e.*, GPT-5.1). We evaluate PSR, NSR, and DAPO on 60 problems from AIME24 and AIME25 (8 rollouts each), measuring the emergence of new reasoning paths at training steps 25, 50, 75, and 100 compared to initial roll-

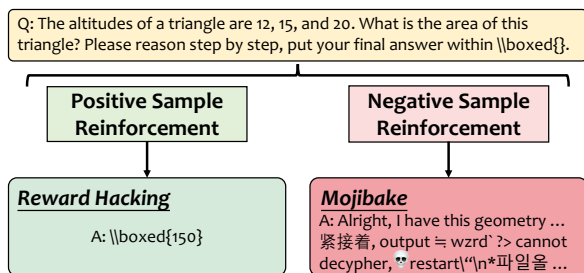


Figure 4: Case study of positive and negative sample reinforcement on Qwen3-8B-Base.

outs using DeepSeek-R1-Distilled-Qwen-7B. As illustrated in Table 1, the results further confirm that negative samples are more likely to produce new reasoning paths, and this gap widens as training progresses.

3.5 Case Study

In this part, we present a case study examining the distinct behaviors of positive sample reinforcement and negative sample reinforcement on Qwen3-8B-Base. The results are shown in Figure 4. We observe that **continuous positive sample reinforcement** leads the model to strengthen its existing correct reasoning paths. Over time, this causes the model to progressively shorten its responses. Eventually, this results in reward hacking, where the model outputs only the final answer without step-by-step reasoning. In contrast, **continuous negative sample reinforcement** encourages the model to repeatedly learn from its own mistakes, which drives it to explore alternative reasoning paths more broadly. As a result, the model ventures into low-probability regions of the output space, which sometimes leads to garbled or nonsensical outputs.

4 Impacts of Advantage Shaping with Different Sample Polarities at Varying Granularities on RLVR Training

Our previous analyses show that training with only one sample polarity impairs performance, confirming the importance of both positive and negative samples in RLVR. In this section, we further investigate how adjusting the influence of each polarity at different granularities affects RLVR training.

4.1 Polarity-level Advantage Shaping

In this part, we conduct polarity-level advantage shaping using Qwen2.5-7B-Math. Specifically, we scale the advantage values of one sample type by

factors of 0.2, 0.5, 2, and 5, while keeping the other type fixed. More details on polarity-level advantage shaping are provided in Appendix B.2. Standard RLVR training ($1 \times$ for both) is included for comparison. The results are presented in Figure 5.

Higher positive advantage speeds up reward improvement but limits exploration diversity. Increasing the advantage values of positive samples accelerates reward growth on the training set, as the model learns more quickly from correct examples. However, this also makes the model more confident and focused on reinforcing existing successful patterns, thereby limiting exploration diversity. Consequently, the model produces responses with lower entropy and shorter lengths.

Higher negative advantage encourages exploration but slows reward improvement. Conversely, assigning higher advantages to negative samples encourages the model to avoid mistakes and explore alternatives. This leads to higher entropy and longer responses, as the model tests various reasoning paths. While this maintains diversity, it slows reward improvement on the training set because the model spends more time exploring rather than directly learning from previous successes.

The relative ratio between positive and negative advantage values determines training dynamics. Our results show that training dynamics depend primarily on the relative ratio between positive and negative advantage values, not their absolute values. For example, settings P2N1 and P1N0.5 have the same relative ratio and exhibit similar training trends. Besides, we also find that excessively high positive advantage causes overfitting to familiar patterns and limits exploration, while overly high negative advantage makes the model overly cautious and slows learning. Among all settings, a positive-to-negative advantage ratio of 0.5 achieves the best performance on the validation set. This balanced ratio enables effective learning from both positive and negative samples, which maintains exploration and ensures steady reward improvement.

4.2 Token-level advantage shaping

To better understand which specific tokens in positive and negative samples contribute more to RLVR training dynamics, we perform token-level advantage shaping. Specifically, we adjust the advantages assigned to tokens with different entropy and probability distributions, and observe the changes

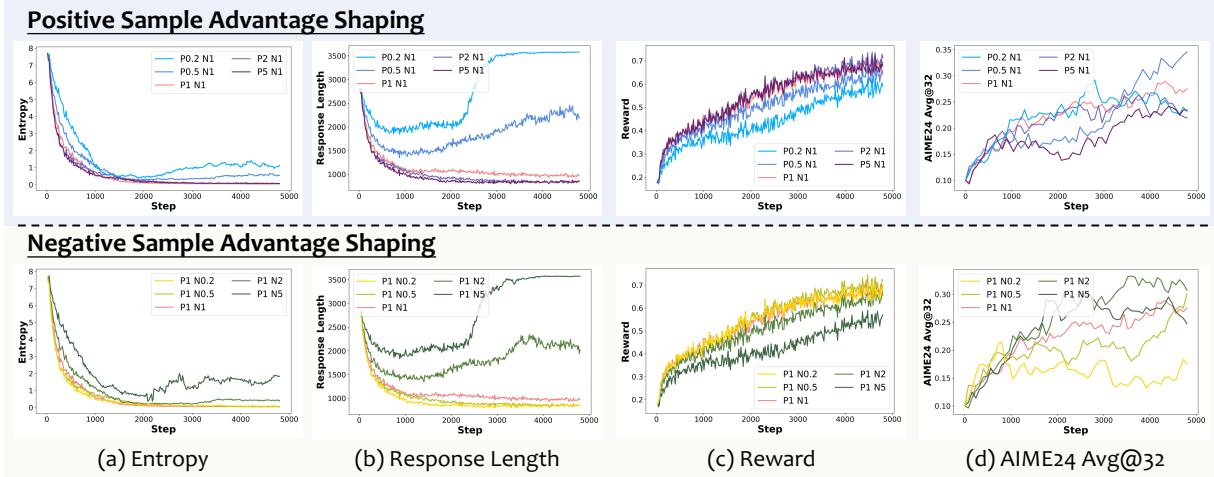


Figure 5: Polarity-level advantage shaping results on Qwen2.5-7B-Math. Each label is formatted as “PXNY”, where “X” and “Y” represent the advantage scaling factors for positive and negative samples. For example, “P1N5” denotes positive sample weight $\times 1$ and negative sample weight $\times 5$.

in RLVR training dynamics. Following prior work (Wang et al., 2025b), we amplify the advantages of tokens in the top and bottom 20% based on either entropy or probability using scaling factors of 0.2 and 5, respectively. More details on token-level advantage shaping are provided in Appendix B.3. This scaling value amplifies the training dynamics, as larger scaling factors produce more pronounced effects. Additionally, we also explore different token ratio settings in Appendix G and find that varying the proportion of weighted tokens does not change the overall training trends.

Entropy-based token-level advantage shaping. The experimental results of entropy-based token-level advantage shaping are presented in Figure 6.

- Reinforcing **positive samples with high-entropy tokens** accelerates entropy reduction, as these tokens often represent critical decision points where the model explores multiple reasoning paths.
- Reinforcing **positive samples with low-entropy tokens** has little effect on training dynamics, since these tokens typically reflect familiar reasoning patterns where the model is already highly confident.
- Reinforcing **negative samples with high-probability tokens** slows the decrease in entropy, as the model remains uncertain about alternatives even when the current path is incorrect, which helps preserve exploration capacity.
- Reinforcing **negative samples with low-entropy tokens** speeds up entropy reduction, enabling the model to quickly identify and suppress confident but incorrect reasoning patterns, thereby

increasing the confidence of models.

Probability-based token-level advantage shaping. The results of Probability-based token-level advantage shaping are illustrated in Figure 7.

- Reinforcing **high-probability positive tokens** accelerates entropy reduction, as these tokens represent correct reasoning paths the model has already mastered, which sharpens the policy distribution around these established patterns.
- Reinforcing **low-probability positive tokens** leads to entropy increase, because encouraging these low-confidence correct alternatives widens the policy distribution and promotes exploration.
- Reinforcing **high-probability negative tokens** raises entropy. This is because penalizing confidently wrong predictions reduces the model’s certainty in high-probability outcomes, encouraging it to reconsider and diversify its predictions.
- Reinforcing **low-probability negative tokens** reduces entropy, as further suppressing already unlikely incorrect paths reinforces avoidance of these tokens and narrows the policy distribution.

5 Adaptive and Asymmetric Advantage Shaping for Policy Optimization

After analyzing how sample polarity influences RLVR training dynamics, we further explore how to leverage this property to enhance the reasoning capabilities of LLMs. To this end, we propose an adaptive and asymmetric token-level advantage shaping method to achieve stable and effective RLVR optimization. In this section, we first introduce the method, then describe the experiment

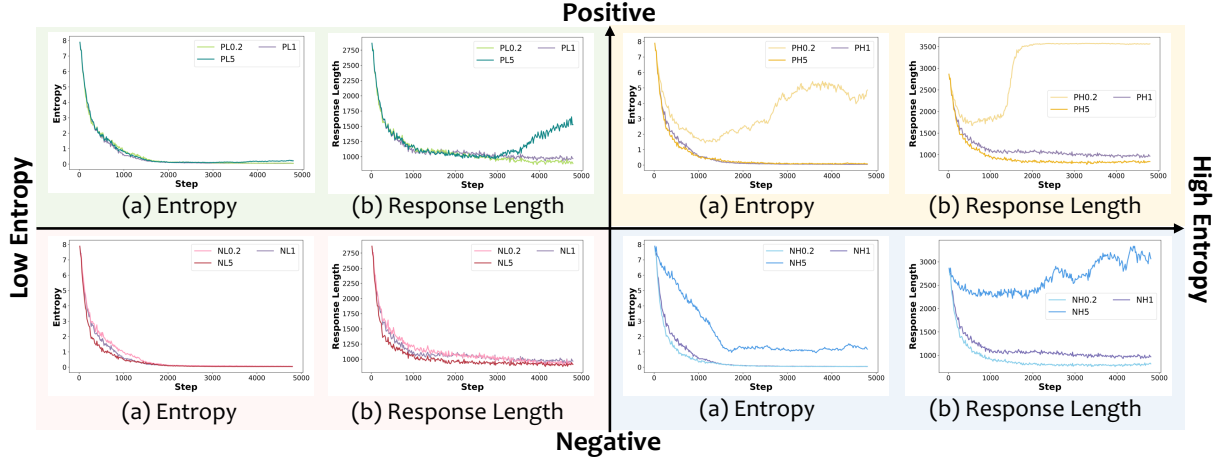


Figure 6: Token-level entropy-based advantage shaping. The x-axis indicates the entropy of shaped tokens (right: high entropy “H”; left: low entropy “L”). The y-axis shows shaped token polarity (top: positive “P”; bottom: negative “N”). Each label follows the format [Polarity][Entropy][Scaling Factor], where the first letter denotes token polarity, the second indicates entropy level, and the numeric value specifies the scaling factor applied to the advantage of those tokens. In the figure, lines with darker colors correspond to amplifying the advantage values of these tokens, while lighter colors indicate reducing their advantage values.

setup, and finally present the results.

5.1 Method

In our previous analysis, we identified two token types that play important roles in the early stages of RLVR training: low-probability tokens from positive samples and high-probability tokens from negative samples. These tokens help maintain higher entropy, which encourages continued exploration and prevents premature convergence. Based on this finding, we propose an adaptive and asymmetric token-level advantage shaping method that dynamically adjusts the weighting of different token categories during training. Our approach assigns larger advantage values to the above token types early in training to actively encourage exploration. However, keeping such asymmetric weighting for too long can cause training-inference engine mismatch and performance collapse (See Appendix H). Therefore, we gradually reduce these weights in a controlled manner as training progresses, allowing the optimization to smoothly transition to a standard training regime. Our method builds on the DAPO (Yu et al., 2025) framework with a modified objective function:

$$\mathcal{J}_{\text{A3PO}}(\theta) = \mathbb{E}_{q \sim D, o \sim \pi_{\theta_{\text{old}}}(\cdot|q)} \left\{ \sum_{t=1}^{|o|} \min \left[r_t \hat{A}_t, \text{clip}(r_t, 1 - \varepsilon_{\text{low}}, 1 + \varepsilon_{\text{high}}) \hat{A}_t \right] \right\}, \quad (1)$$

where $r_t = \frac{\pi_{\theta}(o_t|q, o_{<t})}{\pi_{\theta_{\text{old}}}(o_t|q, o_{<t})}$ denotes the ratio between the current and old policies, \hat{A}_t is our shaped advantage, and ε_{low} and $\varepsilon_{\text{high}}$ are clipping bounds that constrain policy updates. The asymmetric and adaptive advantage shaping is defined as:

$$\hat{A}_t = \begin{cases} A_t \cdot \max(\rho^+ - \alpha^+ s, 1) & A_t > 0, p_t \leq \tau_o^+ \\ A_t \cdot \max(\rho^- - \alpha^- s, 1) & A_t < 0, p_t \geq \tau_o^- \\ A_t & \text{else.} \end{cases} \quad (2)$$

Here, A_t is the normalized accuracy across groups, τ_o^+ is the threshold for the lowest-probability token in a positive rollout, and τ_o^- is the threshold for the highest-probability token in a negative rollout. ρ^+ and ρ^- denote the initial advantage scaling factors, α^+ and α^- control their decay coefficients for positive and negative samples, respectively. s denotes the current training step.

5.2 Experimental Setup

We run our experiments on three LLMs (*i.e.*, Qwen2.5-7B-Math (Yang et al., 2024)), Qwen3-8B-Base (Yang et al., 2025), and Deepseek-R1-Distill-Qwen-7B (DeepSeek-AI et al., 2025)). For comparison, we include GRPO (DeepSeek-AI et al., 2025), DAPO (Yu et al., 2025), polarity-level advantage shaping method (*i.e.*, W-REINFORCE (Zhu et al., 2025)), and token-level advantage shaping methods (*i.e.*, w/ Fork Tokens (Wang et al., 2025b) and Lp-Reg (Huang et al., 2025)) as baselines. Detailed descriptions of the baselines are provided in Appendix D. To evaluate the reasoning ability of the

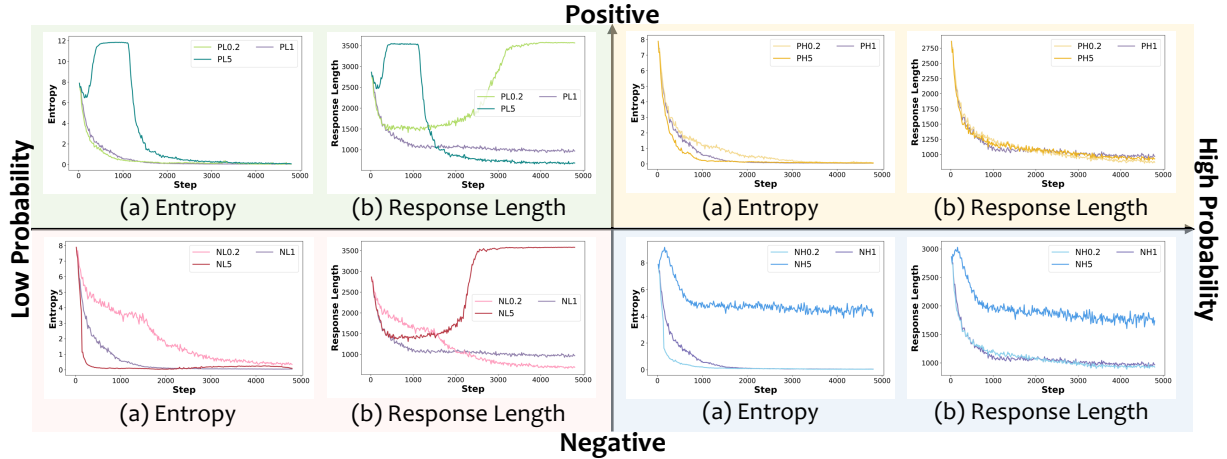


Figure 7: Token-level probability-based advantage shaping. The x-axis indicates the probability of shaped tokens (right: high probability “H”; left: low probability “L”). The y-axis shows shaped token polarity (top: positive “P”; bottom: negative “N”). Each label follows the format [Polarity][Probability][Scaling Factor], where the first letter denotes shaped token polarity, the second indicates their probability level, and the number is the scaling factor applied to the advantage for these tokens. In the figure, lines with darker colors correspond to amplifying the advantage values of these tokens, while lighter colors indicate reducing their advantage values.

methods, we test them on three mathematical (*i.e.*, AIME24, AIME25, and MATH500 (Hendrycks et al., 2021)) and two other reasoning benchmarks (*i.e.*, GPQA (Rein et al., 2023) and LiveCodeBench (Jain et al., 2025)). Detailed experimental setups are presented in Appendix A.

5.3 Main Results

Figure 2 compares the training dynamics of DAPO and A3PO. We observe that A3PO maintains higher entropy and longer responses throughout training, suggesting that the model preserves a richer probability distribution and avoids premature convergence to a narrow output mode. Although A3PO shows a slightly slower growth in training reward compared to DAPO, it achieves higher validation accuracy, with the performance gap widening as training progresses. These results suggest that the policy learned by A3PO generalizes better, allowing the model to acquire more general reasoning capabilities rather than merely memorizing patterns in the training data.

The main results are presented in Table 2. We observe that DAPO further improves performance over GRPO, which can be attributed to its clip higher mechanism, as it retains low-probability positive tokens and helps the model learn novel reasoning paths. The polarity-level advantage shaping method further boosts performance by assigning higher advantage values to negative samples. Additionally, DAPO w/Fork Tokens and Lp-Reg yield

gains by assigning higher weights to high-entropy tokens and regularizing low-probability tokens, respectively. However, these methods do not account for the opposing effects that high-entropy and low-probability tokens can have in positive versus negative samples on RLVR training dynamics. Treating all tokens uniformly may partially counteract their respective contributions. To address this question, we propose an adaptive and asymmetric token-level advantage shaping method, which dynamically adjusts the advantage values of high-probability tokens in negative samples and low-probability tokens in positive samples. This finer-grained allocation of advantages enables more stable and effective RLVR training, ultimately achieving the best performance. More detailed analyses of A3PO are presented in Appendix I.

5.4 Ablation Study

To evaluate the effectiveness of each component in our method, we conduct ablation studies on three math benchmarks using Qwen3-8B-Base. As shown in Table 3, removing any component leads to performance degradation, confirming that all components are essential. We observe that shaping advantages for both positive low-probability tokens and negative high-probability tokens improve performance, as both help maintain entropy and encourage exploration. In addition, the adaptive strategy ensures stable RLVR training. For instance, without this strategy, training instability

Model	Method	AIME24	AIME25	MATH500	GPQA	LiveCodeBench	Average
Qwen2.5-7B-Math	GRPO	26.4±1.0	19.3±0.5	83.8±0.7	33.7±0.7	11.6±1.1	35.0±0.8
	DAPO	27.6±1.0	21.4±1.2	85.4±0.6	34.6±0.8	12.4±0.5	36.3±0.8
	DAPO w/ Fork Tokens	28.6±0.7	22.5±0.9	86.8±0.5	36.5±0.7	14.3±1.0	37.7±0.8
	W-REINFORCE	28.3±0.9	21.4±0.7	87.3±1.0	36.2±1.1	13.8±0.5	37.4±0.8
	Lp-Reg	29.2±1.1	22.2±1.1	87.1±0.8	36.9±0.6	13.8±1.3	37.8±1.0
	A3PO	31.5*±0.8	24.8*±0.6	90.4*±0.6	39.1*±1.2	16.4*±1.0	40.4*±0.8
Qwen3-8B-Base	GRPO	32.4±0.9	23.1±0.5	82.3±0.7	45.3±0.6	29.4±0.9	42.5±0.7
	DAPO	34.2±0.9	26.1±1.0	84.5±0.6	45.8±0.8	29.7±0.5	44.1±0.8
	DAPO w/ Fork Tokens	35.4±0.6	25.7±0.8	86.2±0.5	47.2±0.6	31.2±0.9	45.1±0.7
	W-REINFORCE	35.3±0.8	26.3±0.6	86.9±0.9	47.4±1.0	30.4±0.5	45.3±0.8
	Lp-Reg	35.9±1.0	25.8±0.9	87.4±0.7	47.8±0.6	30.9±1.1	45.6±0.9
	A3PO	37.8*±0.7	30.4*±0.6	91.3*±0.6	50.2*±1.0	33.8*±0.9	48.7*±0.8
Deepseek-R1-Distill-Qwen-7B	GRPO	59.4±0.5	49.2±0.5	95.2±0.5	48.4±0.3	42.5±0.7	58.9±0.5
	DAPO	60.8±0.7	50.8±0.6	95.5±0.3	50.2±0.5	43.2±0.4	60.1±0.5
	DAPO w/ Fork Tokens	61.2±0.3	51.6±0.8	95.8±0.8	50.4±0.2	44.1±0.7	60.6±0.6
	W-REINFORCE	61.6±0.6	51.3±0.4	95.7±0.4	51.4±0.6	44.6±0.5	60.9±0.5
	Lp-Reg	61.9±0.6	52.0±0.6	96.2±0.3	51.2±0.7	44.7±0.8	61.2±0.6
	A3PO	65.2*±0.8	54.1*±0.6	96.9*±0.2	53.8*±0.2	47.2*±0.3	63.4*±0.4

Table 2: Performance comparison of different methods on various reasoning benchmarks. We highlight the best performance across different RLVR methods. Numbers marked with * indicate that the improvement is statistically significant compared with baselines (t-test with p-value < 0.05).

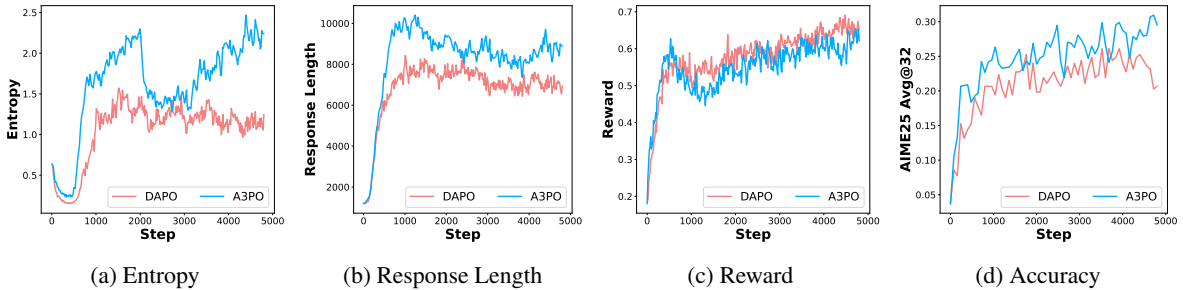


Figure 8: RLVR training dynamics of DAPO and A3PO on Qwen3-8B-Base.

Dataset	AIME 24	AIME 25	MATH500	Average
A3PO	37.8	30.4	91.3	53.2
w/o PL adv	36.5	27.8	88.1	50.8
w/o NH adv	35.8	29.1	87.4	50.8
w/o both	34.2	26.1	84.5	48.3
w/o adaptive	37.5	27.1	90.9	51.8

Table 3: Ablation study on three math benchmarks.

due to training–inference mismatch limits further improvements on AIME25.

6 Conclusion

In this paper, we systematically analyzed the roles of positive and negative samples in RLVR, demonstrating their distinct contributions to training dynamics. Our findings showed that positive samples sharpen correct patterns, while negative samples promote exploration, and both are essential for training. Based on these findings, we proposed an adaptive and asymmetric token-level advantage

shaping method that allowed more precise allocation of advantages and led to stable and improved RLVR training. Experiments across multiple models validate the effectiveness of our approach.

7 Limitations

In this paper, we provide a comprehensive analysis of Reinforcement Learning with Verifiable Rewards (RLVR) from the perspectives of sample polarity, revealing the different roles of positive and negative samples during RLVR training. One limitation of this work is that our experiments are conducted only on the text-based reasoning tasks. In future work, we plan to extend our analysis and methods to other model families, including vision–language or omni-modal models. Additionally, due to constraints in computational resources and budget, we have not evaluated our analysis and approach in agent-based scenarios, such as deep research or software engineering agents.

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China No. 92470205 and Beijing Major Science and Technology Project under Contract No. Z251100008425002. This work was also supported by Ant Group Research Fund. Xin Zhao is the corresponding author.

References

- Aili Chen, Aonian Li, Bangwei Gong, Binyang Jiang, Bo Fei, Bo Yang, Boji Shan, Changqing Yu, Chao Wang, Cheng Zhu, Chengjun Xiao, Chengyu Du, Chi Zhang, Chu Qiao, Chunhao Zhang, Chunhui Du, Congchao Guo, Da Chen, Deming Ding, Dianjun Sun, Dong Li, Enwei Jiao, Haigang Zhou, Haimo Zhang, Han Ding, Haohai Sun, Haoyu Feng, Huaiguang Cai, Haichao Zhu, Jian Sun, Jiaqi Zhuang, Jiaren Cai, Jiayuan Song, Jin Zhu, Jingyang Li, Jinhao Tian, Jinli Liu, Junhao Xu, Junjie Yan, Junteng Liu, Junxian He, Kaiyi Feng, Ke Yang, Kecheng Xiao, Le Han, Leyang Wang, Lianfei Yu, Liheng Feng, Lin Li, Lin Zheng, Linge Du, Lingyu Yang, Lunbin Zeng, Minghui Yu, Mingliang Tao, Mingyuan Chi, Mozhi Zhang, Mujie Lin, Nan Hu, Nongyu Di, Peng Gao, Pengfei Li, Pengyu Zhao, Qibing Ren, Qidi Xu, Qile Li, Qin Wang, Rong Tian, Ruitao Leng, Shaoxiang Chen, Shaoyu Chen, Shengmin Shi, Shitong Weng, Shuchang Guan, Shuqi Yu, Sichen Li, Songquan Zhu, Tengfei Li, Tianchi Cai, Tianrun Liang, Weiyu Cheng, Weize Kong, Wenkai Li, Xiancai Chen, Xiangjun Song, Xiao Luo, Xiao Su, Xiaobo Li, Xiaodong Han, Xinzhu Hou, Xuan Lu, Xun Zou, Xuyang Shen, Yan Gong, Yan Ma, Yang Wang, Yiqi Shi, Yiran Zhong, and Yonghong Duan. 2025a. Minimax-m1: Scaling test-time compute efficiently with lightning attention. *CoRR*, abs/2506.13585.
- Zhipeng Chen, Xiaobo Qin, Youbin Wu, Yue Ling, Qinghao Ye, Wayne Xin Zhao, and Guang Shi. 2025b. Pass@k training for adaptively balancing exploration and exploitation of large reasoning models. *CoRR*, abs/2508.10751.
- Zhipeng Chen, Xiaobo Qin, Wayne Xin Zhao, Youbin Wu, and Ji-Rong Wen. 2026. Adaptive ability decomposing for unlocking large reasoning model effective reinforcement learning. *CoRR*, abs/2602.00759.
- 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, Fulli 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, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, and S. S. Li. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948.
- Zhezhen Hao, Hong Wang, Haoyang Liu, Jian Luo, Jiarui Yu, Hande Dong, Qiang Lin, Can Wang, and Jiawei Chen. 2025. Rethinking entropy interventions in rlvr: An entropy change perspective. *Preprint*, arXiv:2510.10150.
- Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the MATH dataset. In *NeurIPS Datasets and Benchmarks*.
- Guanhua Huang, Tingqiang Xu, Mingze Wang, Qi Yi, Xue Gong, Siheng Li, Ruibin Xiong, Kejiao Li, Yuhao Jiang, and Bo Zhou. 2025. Low-probability tokens sustain exploration in reinforcement learning with verifiable reward. *CoRR*, abs/2510.03222.
- Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. 2025. Livecodebench: Holistic and contamination free evaluation of large language models for code. In *ICLR*. OpenReview.net.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *SOSP*, pages 611–626. ACM.
- Yifan Li, Zhenghao Chen, Ziheng Wu, Kun Zhou, Ruipu Luo, Can Zhang, Zhentao He, Yufei Zhan, Wayne Xin Zhao, and Minghui Qiu. 2025a. Unleashing perception-time scaling to multimodal reasoning models. *CoRR*, abs/2510.08964.
- Yifan Li, Yukai Gu, Yingqian Min, Zikang Liu, Yifan Du, Kun Zhou, Min Yang, Wayne Xin Zhao, and Minghui Qiu. 2025b. Beyond the last frame: Process-aware evaluation for generative video reasoning. *arXiv preprint arXiv:2512.24952*.
- Yifan Li, Kun Zhou, Xin Zhao, Lei Fang, and Jirong Wen. 2026. Analyzing and mitigating object hallucination: A training bias perspective. In *AAAI*, pages 6636–6643. AAAI Press.

- Yizhi Li, Qingshui Gu, Zhoufutu Wen, Ziniu Li, Tian-shun Xing, Shuyue Guo, Tianyu Zheng, Xin Zhou, Xingwei Qu, Wangchunshu Zhou, Zheng Zhang, Wei Shen, Qian Liu, Chenghua Lin, Jian Yang, Ge Zhang, and Wenhao Huang. 2025c. Treepo: Bridging the gap of policy optimization and efficacy and inference efficiency with heuristic tree-based modeling. *CoRR*, abs/2508.17445.
- Mingjie Liu, Shizhe Diao, Ximing Lu, Jian Hu, Xin Dong, Yejin Choi, Jan Kautz, and Yi Dong. 2025. Prorl: Prolonged reinforcement learning expands reasoning boundaries in large language models. *CoRR*, abs/2505.24864.
- Yurou Liu, Mingyang Li, Xinyuan Zhu, Rui Jiao, Yiming Dong, Xinyu Tang, Yang Liu, Jieping Ye, Bing Su, and Zheng Wang. 2026. Drugtrail: Explainable drug discovery via structured reasoning and druggability-tailored preference optimization. In *The Fourteenth International Conference on Learning Representations*.
- Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *ICLR (Poster)*. OpenReview.net.
- Yongyu Mu, Jiali Zeng, Bei Li, Xinyan Guan, Fandong Meng, Jie Zhou, Tong Xiao, and Jingbo Zhu. 2025. Dissecting long-chain-of-thought reasoning models: An empirical study. *Preprint*, arXiv:2506.04913.
- David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Driani, Julian Michael, and Samuel R. Bowman. 2023. GPQA: A graduate-level google-proof q&a benchmark. *CoRR*, abs/2311.12022.
- Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2025. Hybridflow: A flexible and efficient RLHF framework. In *EuroSys*, pages 1279–1297. ACM.
- Xinyu Tang, Xiaolei Wang, Zhihao Lv, Yingqian Min, Xin Zhao, Binbin Hu, Ziqi Liu, and Zhiqiang Zhang. 2025. Unlocking general long chain-of-thought reasoning capabilities of large language models via representation engineering. In *ACL (1)*, pages 6832–6849. Association for Computational Linguistics.
- Xinyu Tang, Zhenduo Zhang, Yurou Liu, Xin Zhao, zujie wen, Zhiqiang Zhang, and JUN ZHOU. 2026. Towards high data efficiency in reinforcement learning with verifiable reward. In *The Fourteenth International Conference on Learning Representations*.
- Jiakang Wang, Runze Liu, Lei Lin, Wenping Hu, Xiu Li, Fuzheng Zhang, Guorui Zhou, and Kun Gai. 2025a. Aspo: Asymmetric importance sampling policy optimization. *Preprint*, arXiv:2510.06062.
- Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, Yuqiong Liu, An Yang, Andrew Zhao, Yang Yue, Shiji Song, Bowen Yu, Gao Huang, and Junyang Lin. 2025b. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for LLM reasoning. *CoRR*, abs/2506.01939.
- Yuhao Wang, Ruiyang Ren, Yucheng Wang, Jing Liu, Xin Zhao, Hua Wu, and Haifeng Wang. 2026. Beera: Balanced entropy engineering for retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 33737–33745.
- Yuhao Wang, Ruiyang Ren, Yucheng Wang, Wayne Xin Zhao, Jing Liu, Hua Wu, and Haifeng Wang. 2025c. Reinforced informativeness optimization for long-form retrieval-augmented generation. *CoRR*, abs/2505.20825.
- Yuhao Wang, Ruiyang Ren, Yucheng Wang, Wayne Xin Zhao, Jing Liu, Hua Wu, and Haifeng Wang. 2025d. Unveiling knowledge utilization mechanisms in llm-based retrieval-augmented generation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1262–1271.
- Zhiheng Xi, Xin Guo, Yang Nan, Enyu Zhou, Junrui Shen, Wenxiang Chen, Jiaqi Liu, Jixuan Huang, Zhihao Zhang, Honglin Guo, Xun Deng, Zhikai Lei, Miao Zheng, Guoteng Wang, Shuo Zhang, Peng Sun, Rui Zheng, Hang Yan, Tao Gui, Qi Zhang, and Xu-anjing Huang. 2025. Bapo: Stabilizing off-policy reinforcement learning for llms via balanced policy optimization with adaptive clipping. *Preprint*, arXiv:2510.18927.
- Yuchen Yan, Liang Jiang, Jin Jiang, Shuaicheng Li, Zujie Wen, Zhiqiang Zhang, Jun Zhou, Jian Shao, Yueting Zhuang, and Yongliang Shen. 2026. Inftythink+: Effective and efficient infinite-horizon reasoning via reinforcement learning. *CoRR*, abs/2602.06960.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jian Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 technical report. *CoRR*, abs/2505.09388.
- An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang

- Ren, and Zhenru Zhang. 2024. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement. *CoRR*, abs/2409.12122.
- Qiyang Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiase Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Weinan Dai, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. 2025. DAPO: an open-source LLM reinforcement learning system at scale. *CoRR*, abs/2503.14476.
- Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. 2025a. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *CoRR*, abs/2504.13837.
- Yu Yue, Yufeng Yuan, Qiyang Yu, Xiaochen Zuo, Ruofei Zhu, Wenyuan Xu, Jiase Chen, Cheng-Xiang Wang, Tiantian Fan, Zhengyin Du, Xiangpeng Wei, Xiangyu Yu, Gaohong Liu, Juncai Liu, Lingjun Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Ru Zhang, Xin Liu, Mingxuan Wang, Yonghui Wu, and Lin Yan. 2025b. VAPO: efficient and reliable reinforcement learning for advanced reasoning tasks. *CoRR*, abs/2504.05118.
- Yu-Liang Zhan, Zhong-Yi Lu, Hao Sun, and Ze-Feng Gao. 2024. Over-parameterized student model via tensor decomposition boosted knowledge distillation. In *NeurIPS*.
- Yu-Liang Zhan, Xinyu Tang, Han Wan, Jian Li, Jirong Wen, and Hao Sun. 2026. L2v-cot: Cross-modal transfer of chain-of-thought reasoning via latent intervention. In *AAAI*, pages 12358–12366. AAAI Press.
- 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, Jiase 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. 2025. A survey of reinforcement learning for large reasoning models. *CoRR*, abs/2509.08827.
- Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and Guoyin Wang. 2023. Instruction tuning for large language models: A survey. *CoRR*, abs/2308.10792.
- Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023a. A survey of large language models. *CoRR*, abs/2303.18223.
- Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, Alban Desmaison, Can Balioglu, Pritam Damania, Bernard Nguyen, Geeta Chauhan, Yuchen Hao, Ajit Mathews, and Shen Li. 2023b. Pytorch FSDP: experiences on scaling fully sharded data parallel. *Proc. VLDB Endow.*, 16(12):3848–3860.
- Yuzhong Zhao, Yue Liu, Junpeng Liu, Jingye Chen, Xun Wu, Yaru Hao, Tengchao Lv, Shaohan Huang, Lei Cui, Qixiang Ye, Fang Wan, and Furu Wei. 2025. Geometric-mean policy optimization. *CoRR*, abs/2507.20673.
- Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. 2025. Group sequence policy optimization. *CoRR*, abs/2507.18071.
- Xinyu Zhu, Mengzhou Xia, Zhepei Wei, Wei-Lin Chen, Danqi Chen, and Yu Meng. 2025. The surprising effectiveness of negative reinforcement in LLM reasoning. *CoRR*, abs/2506.01347.

A Detailed Experimental Setup

Models. We conduct experiments on three models: Qwen2.5-Math-7B (Yang et al., 2024), Qwen3-8B-Base (Yang et al., 2025), and DeepSeek-R1-Distill-Qwen-7B (DeepSeek-AI et al., 2025). For DeepSeek-R1-Distill-Qwen-7B and Qwen3-8B-Base, we set a context length of 16384 tokens. For Qwen2.5-Math-7B, we use its maximum supported length of 4,096 tokens.

Training. Our implementation is based on the Verl (Sheng et al., 2025) pipeline, with rollouts performed using vLLM (Kwon et al., 2023). Models are trained on 16xH200 GPUs. We use the DAPO-Math dataset (Yu et al., 2025) for training. During rollouts, we set the temperature to 1 and sample 8 responses per prompt. Training follows an off-policy RL setup with a batch size of 512 and a mini-batch size of 32. Similar to prior work Yue et al. (2025b), we remove both the KL divergence loss and the entropy loss. All models are trained for 300 steps, optimized with the AdamW (Loshchilov and Hutter, 2019) optimizer using a constant learning rate of 1e-6. The actor module is trained efficiently with Fully Sharded Data Parallel (FSDP) (Zhao et al., 2023b). The chat template used is: “User: \n [question] \n Please reason step by step, and put your final answer within \boxed{ }. \n \n Assistant:”. For hyperparameter settings, we apply adaptive advantage shaping to the lowest 20% probability tokens in positive samples and the highest 20% probability tokens in negative samples. The initial advantage scaling factors ρ^+ and ρ^- are set to 2, and the decay coefficients α^+ and α^- are set to 0.005.

Evaluation. We evaluate model performance on three mathematical reasoning benchmarks (i.e., AIME24, AIME25, and Math500 (Hendrycks et al., 2021)) and two additional reasoning benchmarks (i.e., GPQA (Rein et al., 2023) and LiveCodeBench (Jain et al., 2025)). Models are evaluated every 5 training steps, and we report results from the checkpoint that achieves the highest average performance across five benchmarks. All evaluations are performed in a zero-shot setting (Li et al., 2025b; Wang et al., 2026, 2025d). Following DeepSeek-AI et al. (2025), we set the temperature to 0.6 and top-k to 0.95 during inference. To ensure stable measurements, each test set is evaluated 32 times, and we report the average accuracy.

B Detailed Descriptions of Methods

In this section, we provide detailed descriptions of several methods used in the main text, including positive and negative sample reinforcement in Section 3 and the polarity-level and token-level advantage shaping method in Section 4.

B.1 Positive and Negative Sample Reinforcement

In Section 3, we follow previous work (Zhu et al., 2025) to decompose the RLVR objective into two different learning paradigms: learning from correct rollouts and learning from incorrect rollouts. This decomposition allows us to examine how positive and negative responses affect training dynamics. The RLVR objective can be expressed as the sum of two sub-objectives:

$$\mathcal{L}_{\text{RLVR}}(\theta) = \mathcal{L}_{\text{PSR}}(\theta) + \mathcal{L}_{\text{NSR}}(\theta), \quad (3)$$

where the two sub-objectives correspond to each learning paradigm:

$$\mathcal{L}_{\text{PSR}}(\theta) = -\mathbb{E}_{x \sim \mathcal{D}} \left[\sum_{y:r(x,y)=1} \pi_{\theta}(y|x) \right], \quad (4)$$

$$\mathcal{L}_{\text{NSR}}(\theta) = -\mathbb{E}_{x \sim \mathcal{D}} \left[\sum_{y:r(x,y)=0} -\pi_{\theta}(y|x) \right]. \quad (5)$$

We refer to these two learning paradigms as *positive sample reinforcement* and *negative sample reinforcement*. Positive sample reinforcement resembles supervised fine-tuning, increasing the likelihood of correct responses. In contrast, negative sample reinforcement acts like likelihood minimization, reducing the probability of incorrect responses.

B.2 Polarity-level Advantage Shaping

To explore how adjusting the influence of positive and negative samples affects RLVR training, we introduce a polarity-level advantage shaping method in Section 4.1. This approach assigns different weights to the advantage values derived from positive and negative samples, allowing us to control their relative contributions during policy optimization. Formally, the objective is defined as:

$$\mathcal{L}_{\text{Polarity-AS}}(\theta) = \beta_{\text{P}} \cdot \mathcal{L}_{\text{PSR}}(\theta) + \beta_{\text{N}} \cdot \mathcal{L}_{\text{NSR}}(\theta) \quad (6)$$

Here, β_{P} and β_{N} are scaling factors that control the advantage values for positive and negative samples,

respectively. By adjusting these scaling factors, we can study how emphasizing or de-emphasizing each sample polarity impacts RLVR training dynamics.

B.3 Token-level Advantage Shaping

To further examine the contribution of specific tokens to RLVR training, we introduce a token-level advantage shaping approach in Section 4.2. This method allows us to reweight the advantage assigned to selected tokens and observe how such adjustments affect overall training dynamics. The modified policy optimization objective can be expressed as follows:

$$\mathcal{J}_{\text{Token-AS}}(\theta) = \mathbb{E}_{q \sim \mathcal{D}, o \sim \pi_{\theta_{\text{old}}}(\cdot|q)} \left\{ \sum_{t=1}^{|o|} \min \left[r_t \hat{A}_t, \text{clip}(r_t, 1 - \varepsilon_{\text{low}}, 1 + \varepsilon_{\text{high}}) \hat{A}_t \right] \right\}, \quad (7)$$

where $r_t = \frac{\pi_{\theta}(o_t|q, o_{<t})}{\pi_{\theta_{\text{old}}}(o_t|q, o_{<t})}$ denotes the probability ratio between the current and old policies, ε_{low} and $\varepsilon_{\text{high}}$ are clipping thresholds that constrain policy updates, and \hat{A}_i represents the shaped advantage. The shaped advantage \hat{A}_i is defined based on whether a token is selected for reweighting:

$$\hat{A}_t = \begin{cases} A_t \cdot \beta_T & \text{if selected} \\ A_t & \text{else.} \end{cases} \quad (8)$$

where A_i is the original advantage computed from group rollouts, and β_T is a token-level scaling factor applied to the selected token. Tokens are selected based on specific criteria, such as entropy and probability, enabling us to study how reweighting distinct token categories influences training dynamics.

C Detailed Metrics of Sharpening and Discovery

In this part, we provide the formal definitions of the sharpening and discovery metrics introduced in Section 3.4. During the training process, we collect all self-generated trajectories produced by the model, including both correct and incorrect ones. We denote this full set of historical trajectories up to step t as $\mathcal{H}_{<t}$. For a trajectory ϕ_t generated at step t , we first split it into a set of consecutive n-grams, which we denote as $G(\phi_t)$.

Sharpening. The sharpening metric measures how much the current trajectory ϕ_t reuses reasoning

patterns that have already been seen in previous correct trajectories. To compute this, we maintain a separate set $\mathcal{C}_{<t} \subseteq \mathcal{H}_{<t}$, which contains only the correct trajectories collected before step t . Formally, it is defined as the fraction of n-grams in ϕ_t that also appear in any earlier correct trajectory from $\mathcal{C}_{<t}$:

$$S(\phi_t) = \frac{|\{g \in G(\phi_t) \mid \exists \phi_c \in \mathcal{C}_{<t} \text{ s.t. } g \in G(\phi_c)\}|}{|G(\phi_t)|}. \quad (9)$$

Here, $|\cdot|$ denotes the size of a set. The value of sharpening metric ranges from 0 to 1. A higher sharpening score indicates that the model is primarily repeating reasoning steps that have previously led to correct answers, showing a reinforcement of known successful patterns.

Discovery. Conversely, the discovery metric measures how much the current trajectory ϕ_t contains new reasoning patterns not seen in any previous trajectory. It is defined as the proportion of n-grams in ϕ_t that are absent from the entire set of historical trajectories $\mathcal{H}_{<t}$:

$$D(\phi_t) = \frac{|\{g \in G(\phi_t) \mid \forall \phi_h \in \mathcal{H}_{<t}, g \notin G(\phi_h)\}|}{|G(\phi_t)|}. \quad (10)$$

The discovery metric also lies between 0 and 1. A higher discovery score means the trajectory includes more novel reasoning steps that have never appeared before, reflecting the model’s capacity to explore genuinely new solution paths.

D Detailed Description of Baselines

In this part, we provide detailed descriptions of the baseline methods used for comparison in our experiments. Specifically, we compare our method with GRPO (DeepSeek-AI et al., 2025), DAPO (Yu et al., 2025), the polarity-level advantage shaping method (*i.e.*, W-REINFORCE (Zhu et al., 2025)), and the token-level advantage shaping method (*i.e.*, w/ Fork Tokens (Wang et al., 2025b) and Lp-Reg (Huang et al., 2025)).

• **GRPO** (DeepSeek-AI et al., 2025) is a reinforcement learning algorithm that improves LLM reasoning without training a separate value model. For each question, it samples multiple outputs from the current policy and optimizes the policy using a group-relative advantage, making it scalable for long chain-of-thought reasoning tasks.

- **DAPO** (Yu et al., 2025) is an enhanced RL method that introduces several improvements for LLM training. It prevents entropy collapse by using a higher clipping threshold to encourage exploration, applies dynamic sampling to filter prompts with zero variance, adopts token-level policy gradient loss to handle varying response lengths, and removes the KL divergence term for RL training.

- **W-REINFORCE** (Zhu et al., 2025) is a polarity-level advantage shaping method, which assigns higher weights to self-generated negative rollouts, enabling effective RLVR training.

- **DAPO w/Fork Tokens** (Wang et al., 2025b) is a token-level advantage shaping method that focuses policy gradient updates on high-entropy “forking tokens”. By masking gradients for the 80% lowest-entropy tokens and updating only the top 20% high-entropy tokens, it improves the reasoning performance of LLMs.

- **Lp-Reg** (Huang et al., 2025) is a token-level method designed to mitigate exploration collapse. It maintains useful low-probability tokens through regularization while filtering out noisy tokens, thereby sustaining exploration throughout RLVR training.

E Detailed Comparison of Related Work

In this section, we provide a detailed comparison with previous related work, highlighting the distinctions and contributions of our study from two perspectives: empirical analysis and methodology.

Different empirical analysis methods and findings. One related work (Mu et al., 2025) primarily studies distilled LLMs, and finds that combining positive and negative samples achieves the best results, and positive samples improve training fit while negative samples enhance generalization. Another closely related work (Zhu et al., 2025) similarly examines the distinct roles of positive and negative samples in reinforcement learning with verifiable reward. They formally decompose the RLVR objective into two separate components: positive sample reinforcement and negative sample reinforcement. Through separate training on one type of sample, they observe opposing effects between two sample types and report a surprising effectiveness of training solely with negative samples. Our work extends beyond this polarity-level analysis, as presented in Section 3. We analyze three categories of models: math-enhanced pretrained LLMs (*i.e.*, Qwen2.5-7B-Math), general pretrained LLMs

Method	Adaptive	Asymmetric	Advantage Shaping
GRPO	×	×	×
DAPO	×	✓	×
DAPO w/Fork Token	×	✓	Token-level
W-REINFORCE	×	✓	Polarity-level
Lp-Reg	×	✓	Token-level
ASPO	×	✓	Polarity-level
A3PO (Ours)	✓	✓	Token-level

Table 4: Comparison of our proposed method and other representative RLVR methods.

(*i.e.*, Qwen3-8B-Base), and distilled LLMs (*i.e.*, DeepSeek-R1-Distill-Qwen-7B). Through an examination of training dynamics and n-gram distributions in model rollouts, we demonstrate that positive samples encourage the sharpening of reasoning paths, while negative samples facilitate discovery and generalization. These findings further support and extend previous conclusions from multiple perspectives. In addition, our analysis reveals limitations in prior conclusions: we find that training with only one polarity impairs both reasoning capability and decision boundaries, which challenges the earlier claim that negative-only training improves reasoning performance. Furthermore, we are the first to conduct advantage shaping at both polarity and token levels to investigate how adjusting influence at finer granularities affects RLVR training, as described in Section 4.

Different methods. Regarding the methodology, we compare our approach with several common baselines, as summarized in Table 4. We propose an adaptive and asymmetric token-level advantage shaping method for policy optimization. This approach dynamically adjusts the advantages of high-probability tokens in negative samples and low-probability tokens in positive samples, enabling finer-grained advantage allocation and more effective and stable RLVR training.

F Different Training Dynamics of Base LLMs

In this part, we analyze the training dynamics of three RLVR training paradigms (*i.e.*, positive sample reinforcement (PSR), negative sample reinforcement (NSR), and DAPO) across three base LLMs (*i.e.*, Qwen2.5-7B-Math, Qwen3-8B-Base, and Deepseek-R1-Distilled-Qwen-7B). Specifically, we monitor accuracy changes on all validation samples from AIME24 and AIME25 during training and categorize them into five patterns:

- **Sharpen:** Accuracy improves by more than

$k\%$, indicating that training strengthens the model’s ability to solve the problem.

- **Degradation:** Accuracy drops by more than $k\%$, meaning training reduces reasoning ability

- **Fluctuation:** Accuracy fluctuates within $k\%$ of the original value, showing training has little effect.

- **Mastery:** Accuracy remains above $1 - k\%$, meaning the model consistently solves the problem correctly.

- **Struggle:** Accuracy stays below $k\%$, meaning the problem remains too difficult for the model.

We set k to 10 in our analyses. The results for Qwen2.5-7B-Math, Qwen3-8B-Base, and DeepSeek-R1-Distilled-Qwen-7B are shown in Figures 9, 10, and 11, respectively.

These results reveal distinct patterns of validation accuracy changes across different base LLMs during training. For Qwen2.5-7B-Math, which has been extensively exposed to reasoning data during pretraining, both PSR and NSR produce more sharpened samples than degraded ones. This shows that either polarity alone can improve performance, and combining them yields a further complementary boost. In contrast, Qwen3-8B-Base exhibits reward hacking when using only positive samples, causing degradation in most samples. Negative sample reinforcement leads to garbled outputs, leaving the majority of samples in the struggle phase. Only when both polarities are combined does RLVR training become effective and improve accuracy. For the distilled model DeepSeek-R1-Distilled-Qwen-7B, relying on a single sample polarity leads to significant degradation, and both polarities are needed together to achieve further performance improvement.

G Different Weighted Ratios of Token-level Advantage Shaping

In this part, we investigate how the proportion of tokens selected for token-level advantage shaping affects RLVR training dynamics. Following prior work (Wang et al., 2025b), our main experiments adopt a shaping ratio of 20%, meaning that advantages are reweighted for 20% of the tokens in each response. To assess the sensitivity of this choice, we conduct additional experiments with ratios of 5%, 10%, and 50%, while keeping the scaling factor for low-probability positive tokens fixed at $0.2\times$.

Figure 12 presents the results on Qwen2.5-7B-

Math. We observe that the shaping ratio mainly affects the magnitude and speed of training dynamics but does not change the overall learning trend. Specifically, in this setting, smaller ratios lead to faster entropy reduction and a smoother transition in response length (from an initial decrease to a later increase). Similarly, reward improvements occur more quickly in early training but slow down later when using smaller ratios. These results show that adjusting the proportion of advantage shaping tokens does not alter the fundamental training dynamics. Instead, it acts as a factor that modulates the rate of policy updates.

H Negative Samples Amplify the Training-Inference Mismatch

Training-inference mismatch is a critical issue in RLVR training, where token probabilities in training and inference engines exhibit significant discrepancies, potentially leading to training collapse. In this section, we investigate which sample types contribute to this mismatch.

Figure 13a shows the difference in token probabilities between training and inference engines for three training paradigms (*i.e.*, PSR, NSR, and DAPO). Our results reveal that utilizing negative samples widens the probability gap between training and inference engines. Furthermore, we study the effect of polarity-level advantage shaping on negative samples. As shown in Figure 13b, assigning higher advantages to negative samples further increases the training–inference probability difference. This phenomenon suggests that although weighting negative samples can raise model entropy and encourage exploration, consistently giving them higher weights enlarges the mismatch and may lead to instability.

Building on this observation, we adopt an adaptive advantage shaping strategy: we increase the weight of high-probability negative samples in early training to promote exploration, then gradually reduce it until it aligns with the weight of positive samples, thereby ensuring stable training.

I Detailed Analysis of A3PO

In this section, we present a detailed analysis of our proposed method, **A3PO**.

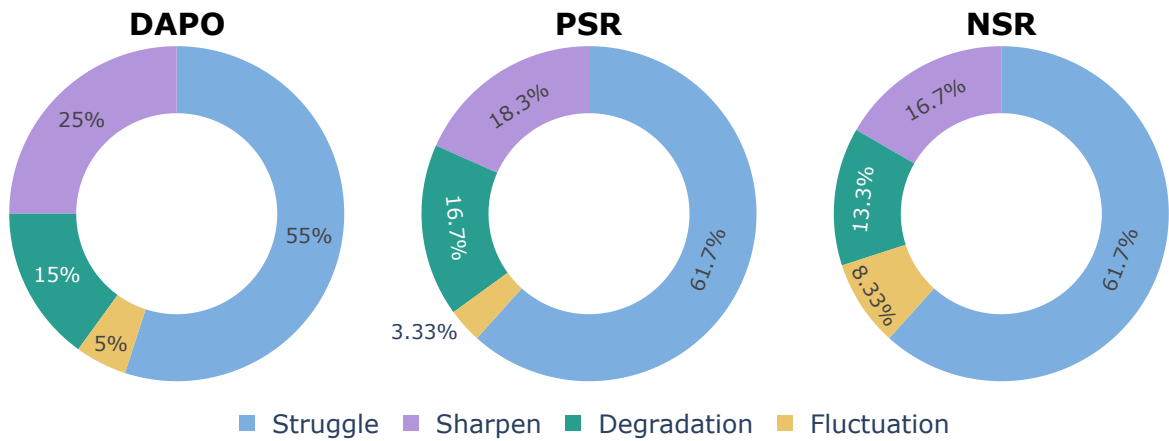


Figure 9: Training dynamics of sample accuracy changes in the validation set on Qwen2.5-7B-Math.

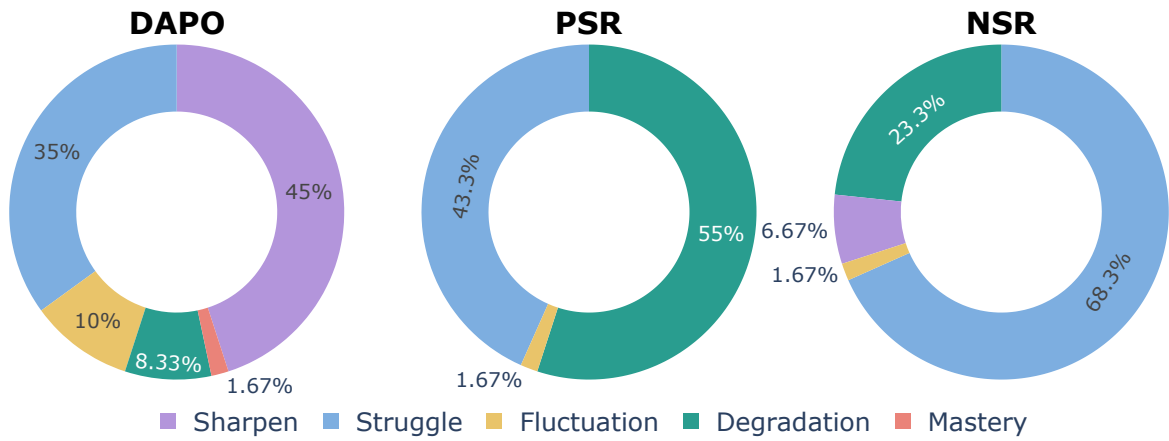


Figure 10: Training dynamics of sample accuracy changes in the validation set on Qwen3-8B-Base.

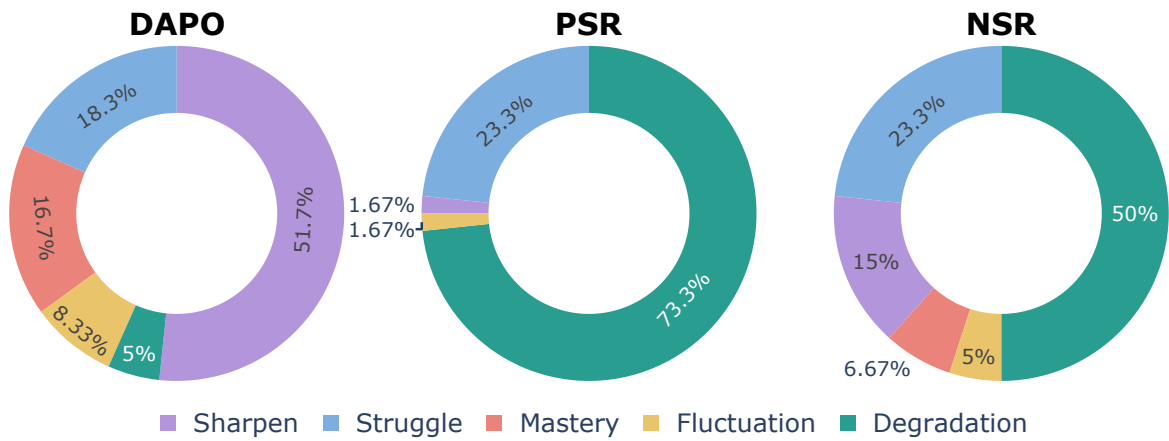


Figure 11: Training dynamics of sample accuracy changes in the validation set on DeepSeek-R1-Distilled-Qwen-7B.

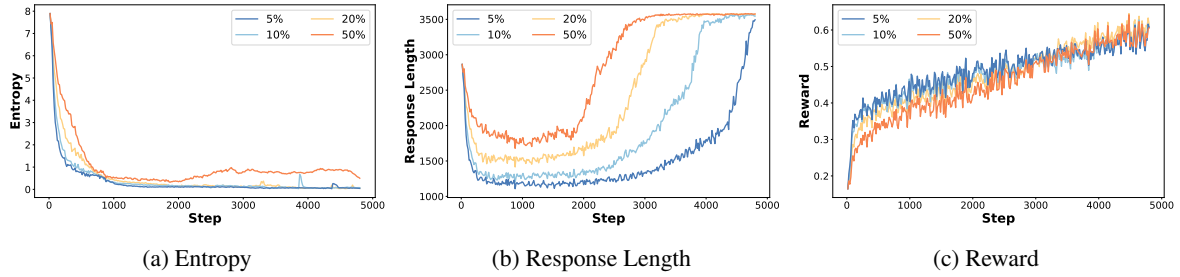


Figure 12: Impact of different ratios of advantage-shaped tokens when low-probability positive tokens are weighted at $0.2\times$ of their original values on Qwen2.5-7B-Math.

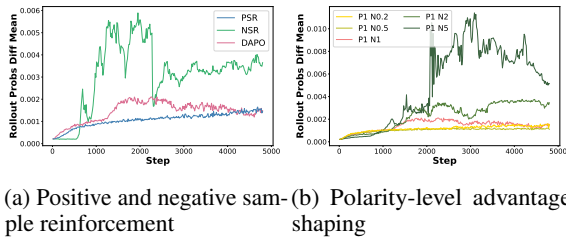


Figure 13: Difference in token probabilities between training and inference engines.

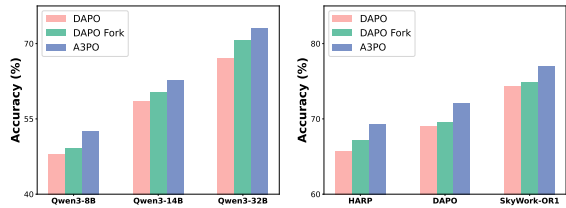


Figure 14: Different LLMs and training datasets

I.1 Different Scales of LLMs and Training Datasets

To assess the robustness and effectiveness of our proposed method, we conduct experiments using different scales of LLMs and datasets on Deepseek-R1-Distilled-Qwen-7B. The results are shown in Figure 14. We find that our proposed method consistently achieves the best performance across both varying model scales and different training datasets, demonstrating its effectiveness and generalizability.

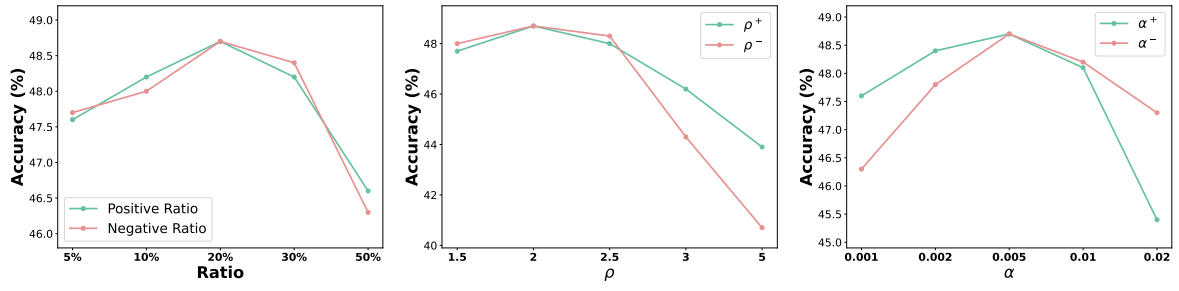
I.2 Hyperparameter Analysis

In this part, we analyze three key hyperparameters of **A3PO**: the token-shaped ratios, the initial scaling factors ρ , and the decay coefficients α . The results are shown in Figure 15. Our experimental results indicate that the model achieves optimal performance with the token-shaped ratios of 20%. If the ratio is too low, the model does not explore

the token space sufficiently. If it is too high, performance drops because many less relevant tokens receive advantage shaping. Next, we examine the initial scaling factors ρ . Setting appropriate values is important for stable training. If the hyperparameter is too high, the training–inference mismatch increases. If it is too low, the model does not explore the solution space effectively. Therefore, we set ρ to 2 in our main experiments. Finally, we find that the decay coefficients α of 0.005 yield the best performance. If α is too small, the gap between training and inference grows. If α is too large, exploration becomes insufficient and the model may converge to suboptimal solutions.

J Detailed Results

In this part, we provide detailed training dynamics in our experiments. Figure 16, 17, 18 present training dynamics of positive sample reinforcement, negative sample reinforcement, and DAPO across three different LLMs. Figure 19 and Figure 20 present training behaviors (*i.e.*, sharpen and discovery) on Qwen2.5-7B-Math and Deepseek-R1-Distilled-Qwen-7B. Figure 21 and Figure 22 show the results of polarity-level advantage shaping. Figure 23, Figure 24, Figure 25 and Figure 26 illustrate the entropy-based token-level advantage shaping. Figure 27, Figure 28, Figure 29 and Figure 30 illustrate the probability-based token-level advantage shaping. Figure 31 shows the results of different shaped token ratios in token-level advantage shaping.

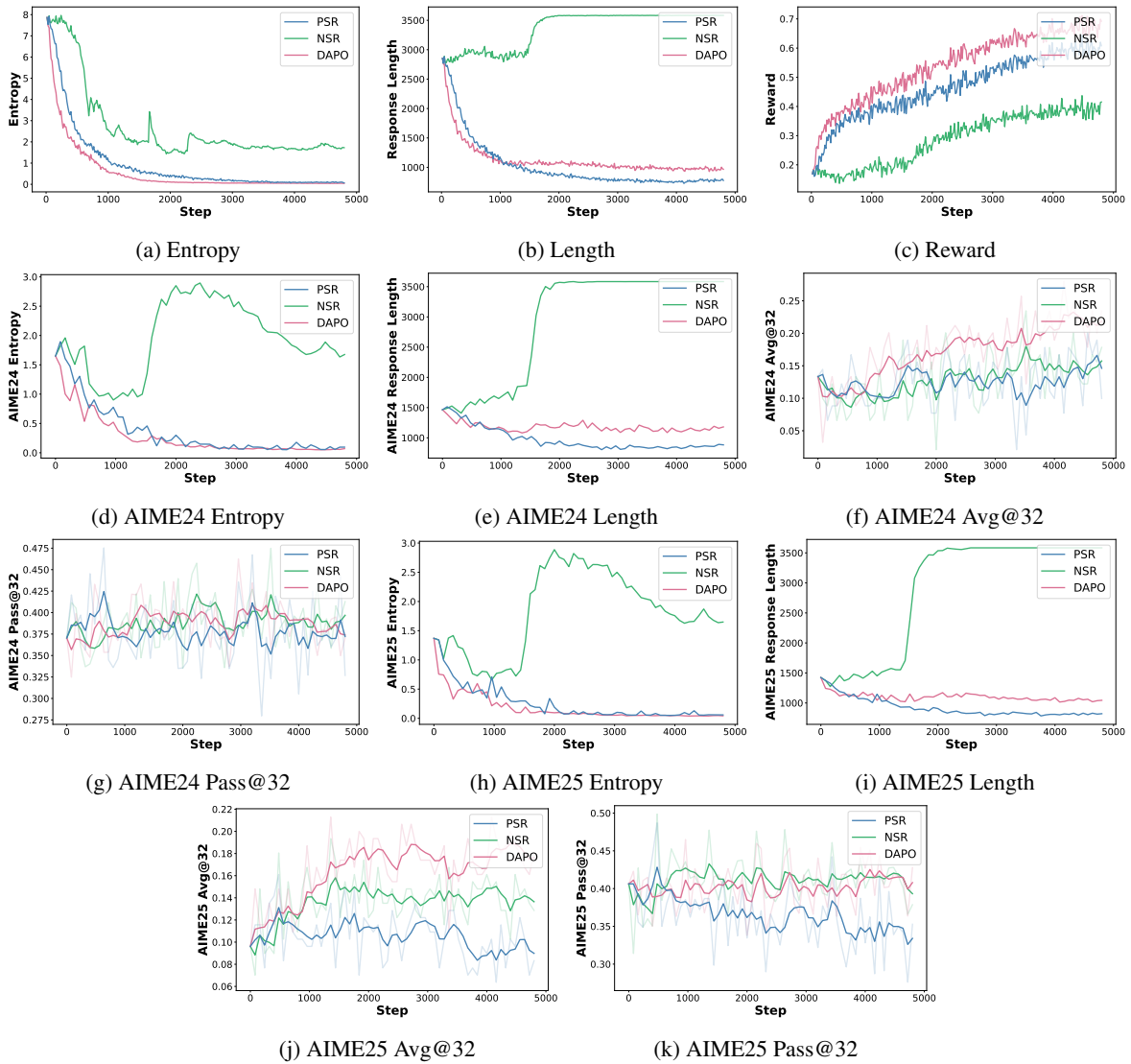


(a) Token-shaped ratios

(b) Decay coefficients

(c) Initial scaling factors

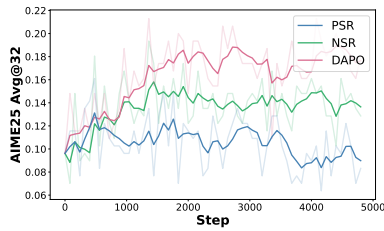
Figure 15: Hyperparameter Analysis.



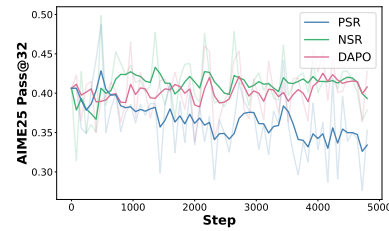
(g) AIME24 Pass@32

(h) AIME25 Entropy

(i) AIME25 Length



(j) AIME25 Avg@32



(k) AIME25 Pass@32

Figure 16: RLVR training dynamics on Qwen2.5-7B-Math.

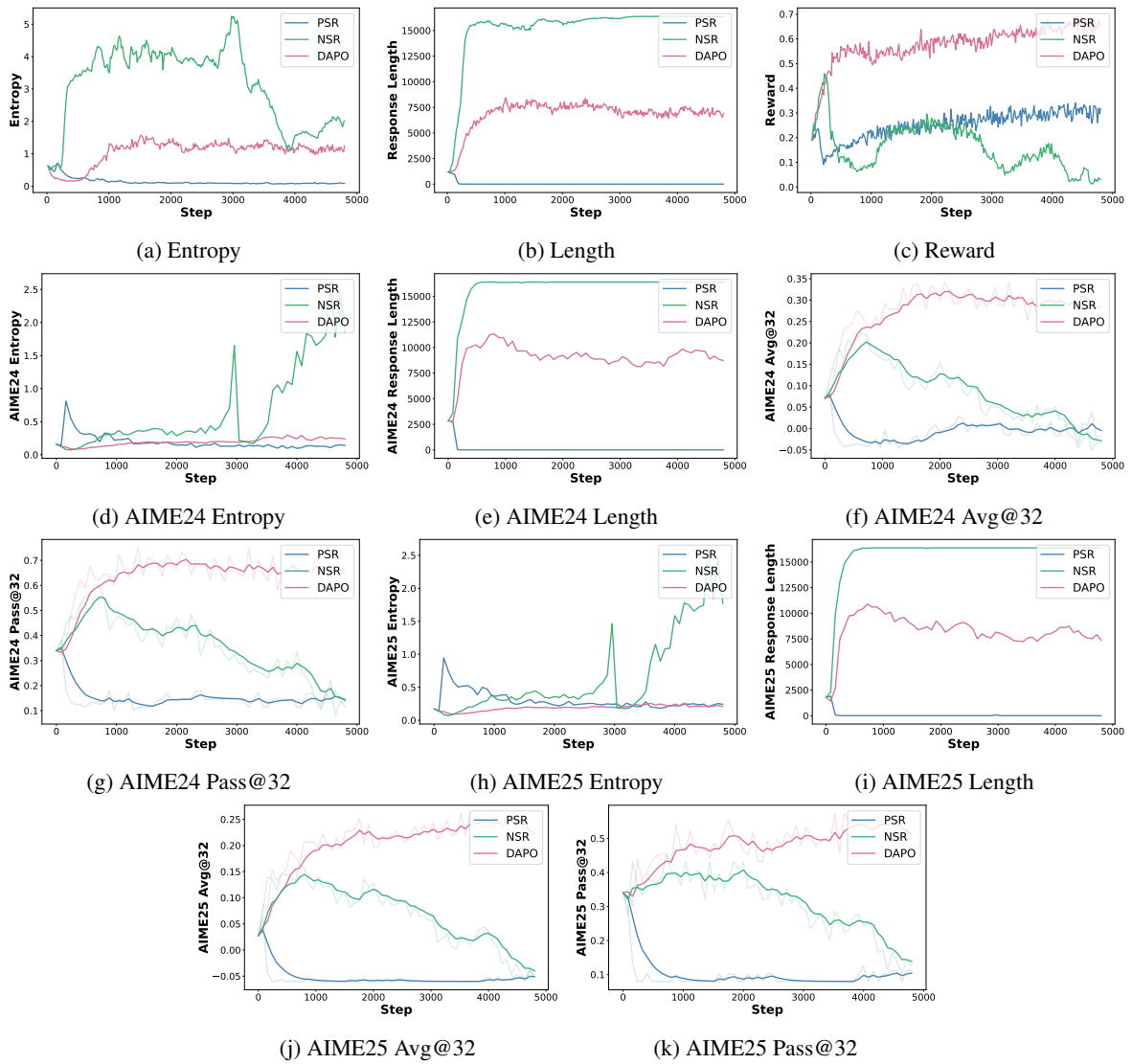


Figure 17: RLVR training dynamics on Qwen3-8B-Base.

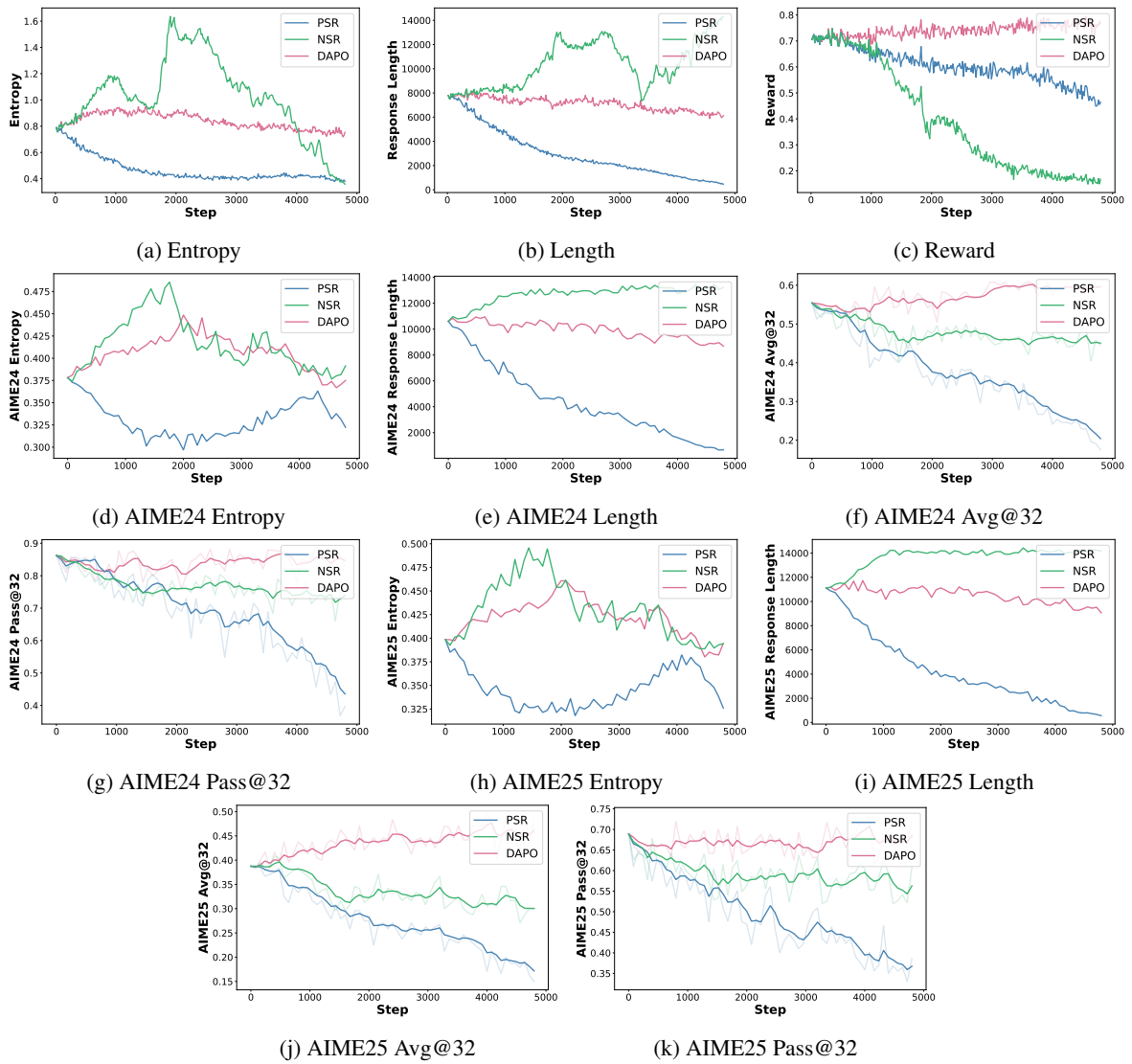


Figure 18: RLVR training dynamics on DeepSeek-R1-Distill-Qwen-7B.

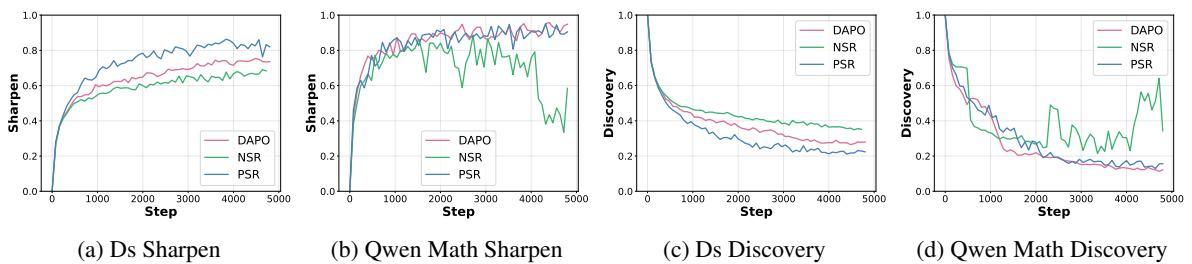


Figure 19: Training behaviors of different RLVR training when n_gram is 3.

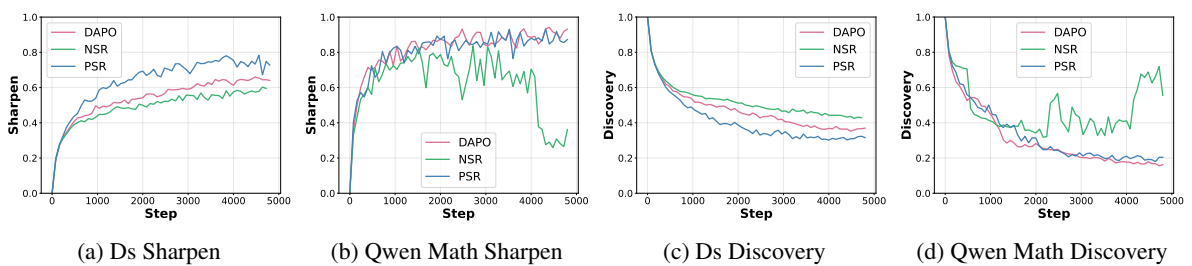


Figure 20: Training behaviors of different RLVR training when n_gram is 4.

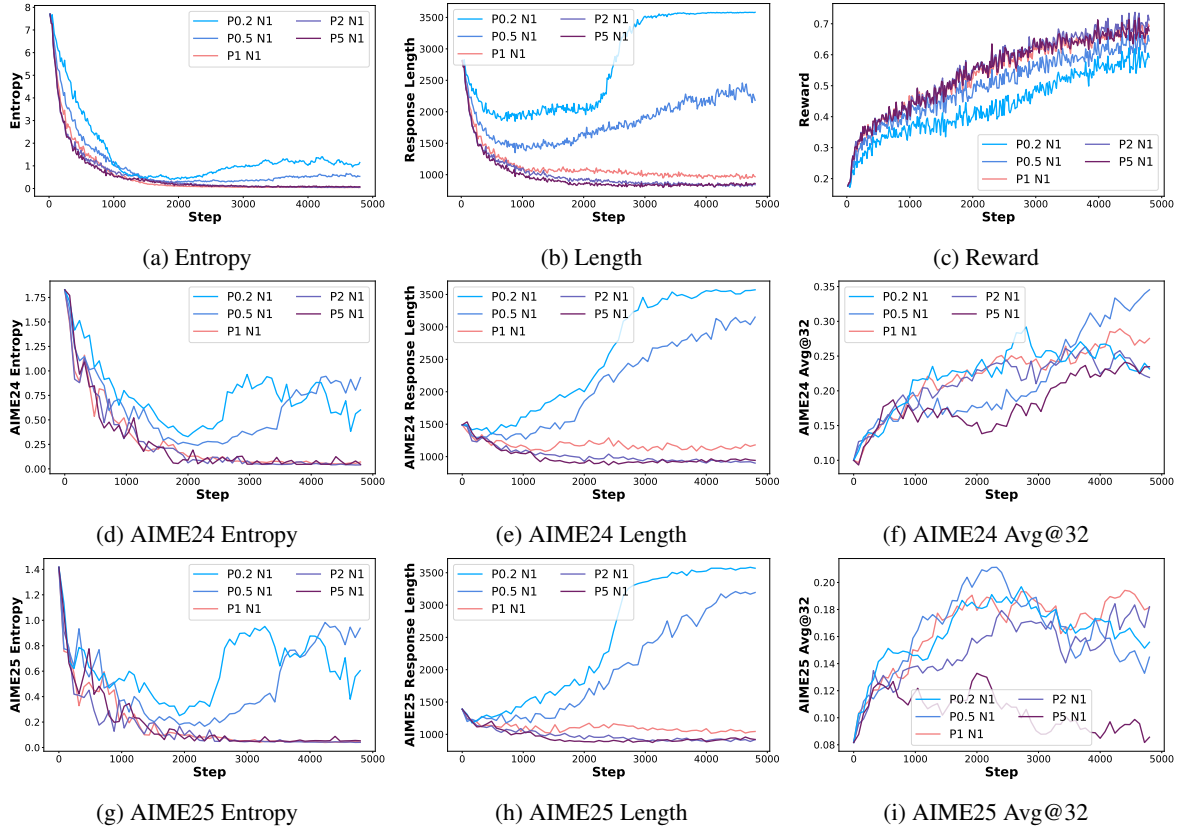


Figure 21: Different training dynamics of polarity-level positive sample advantage shaping on Qwen2.5-7B-Math.

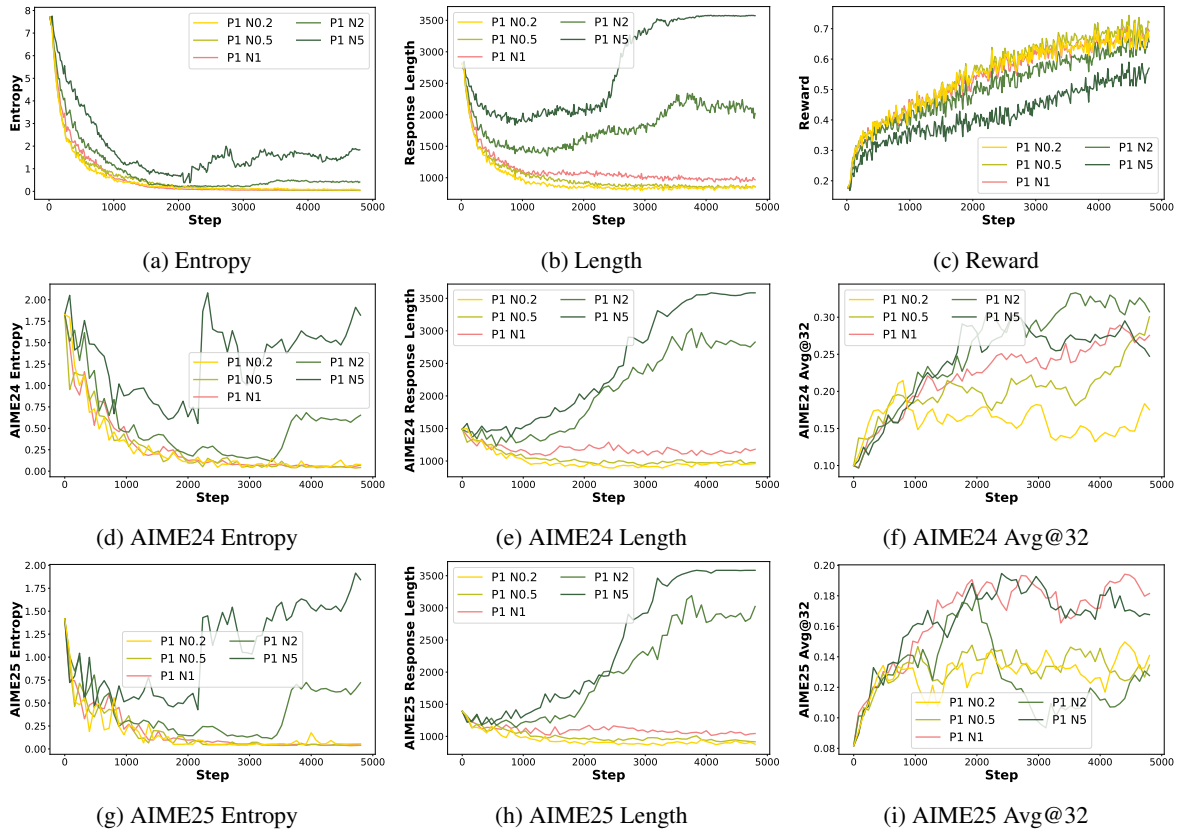


Figure 22: Different training dynamics of polarity-level negative sample advantage shaping on Qwen2.5-7B-Math.

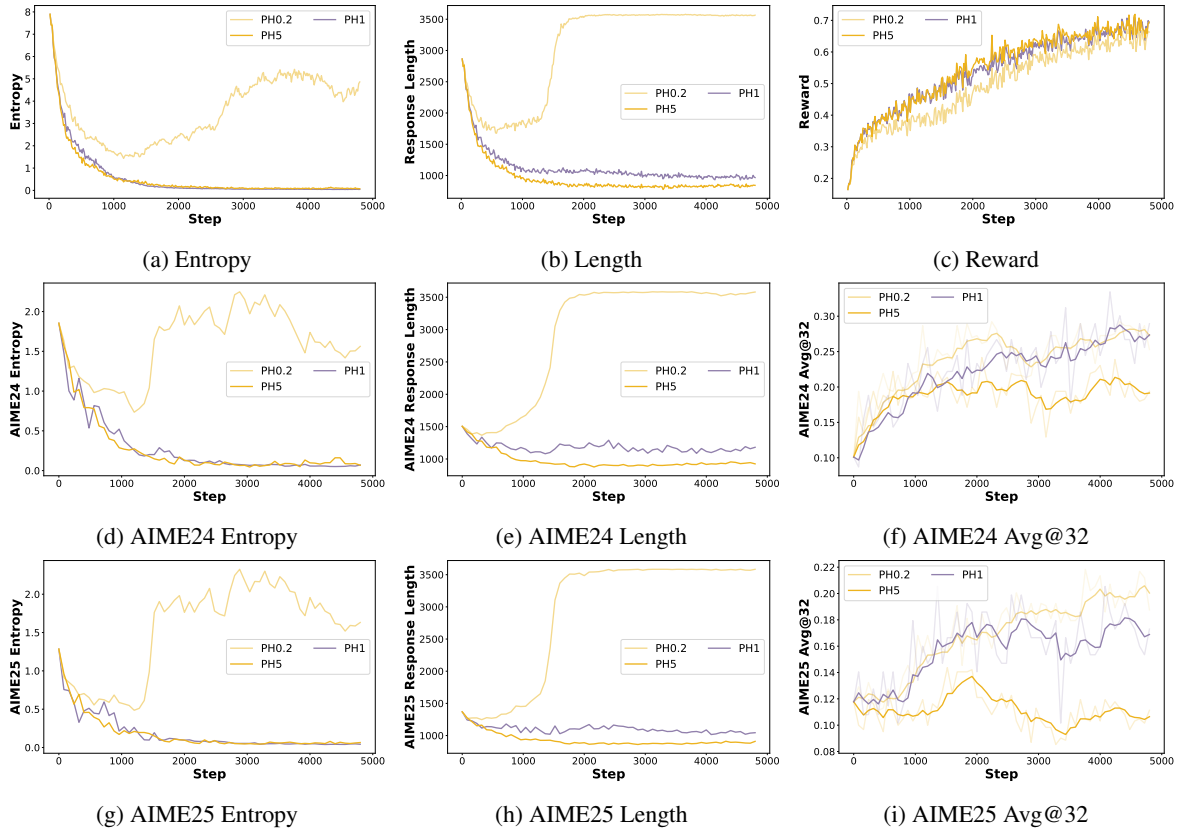


Figure 23: RLVR training dynamics on positive high entropy token advantage shaping.

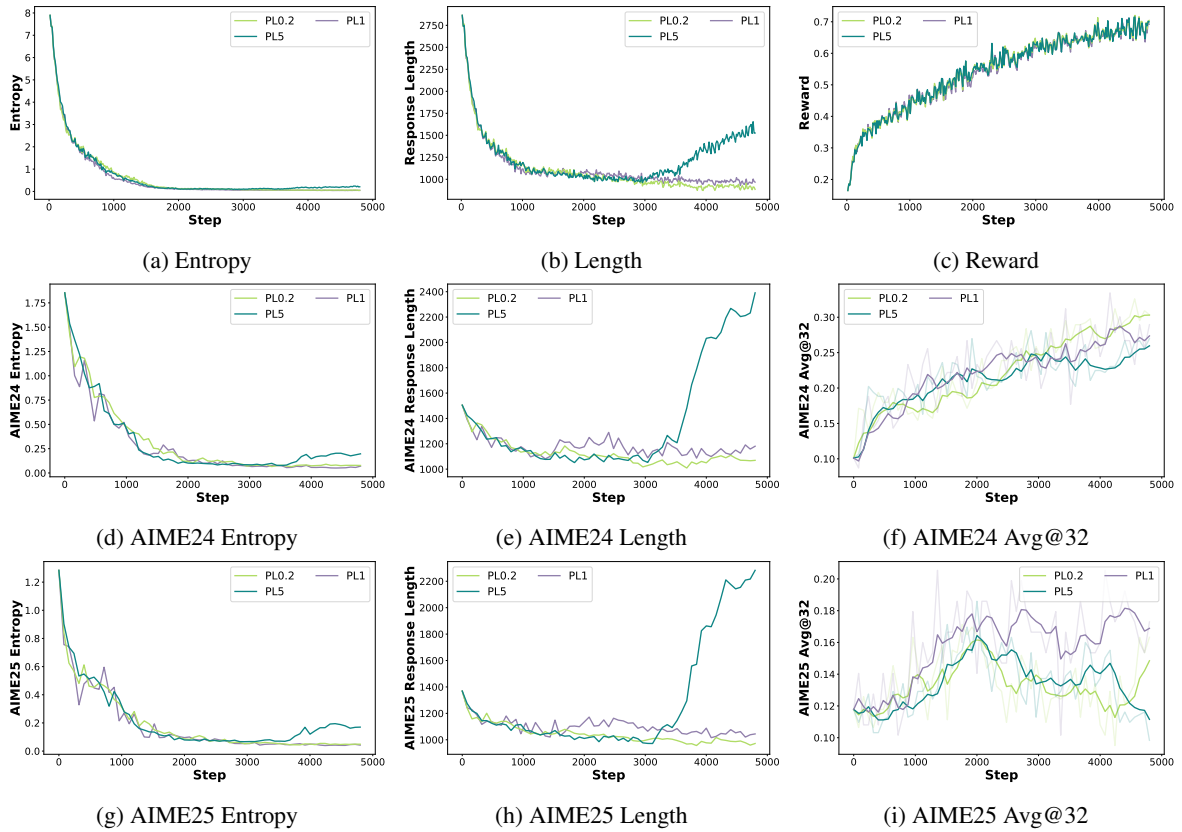


Figure 24: RLVR training dynamics on positive low entropy token advantage shaping.

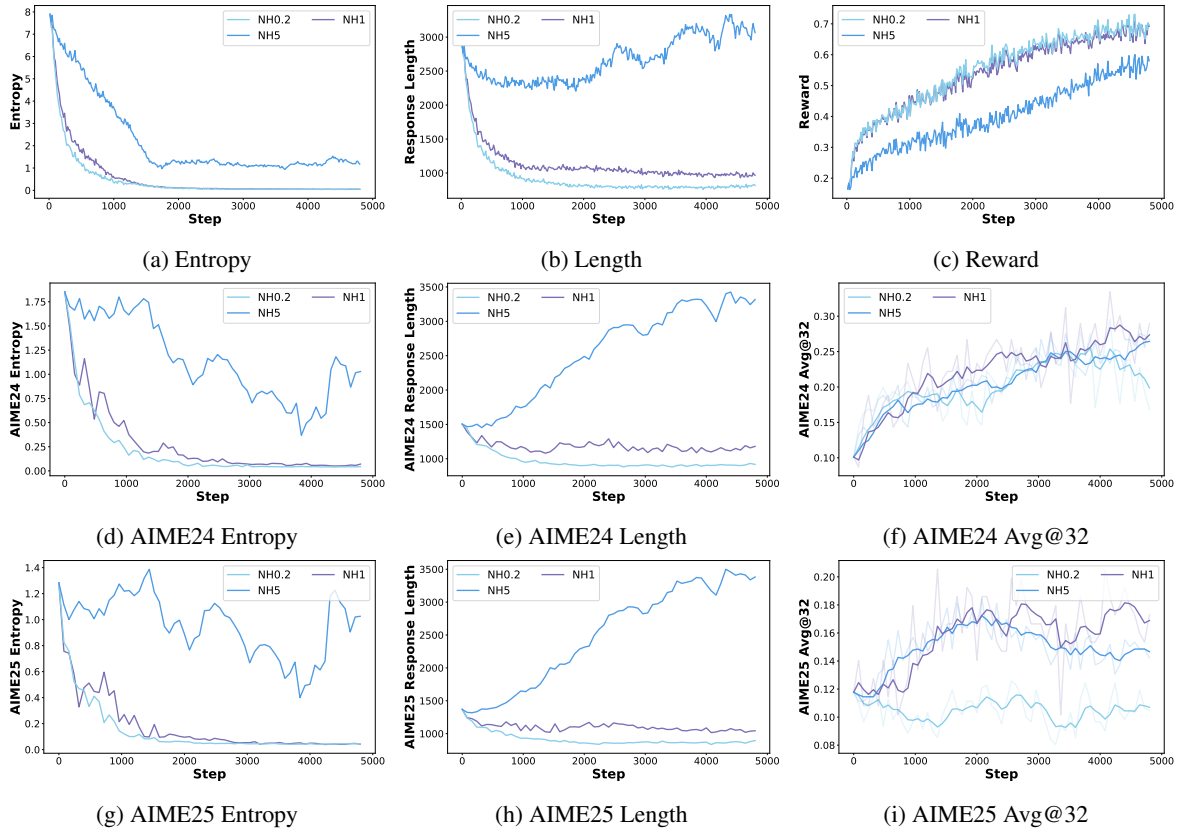


Figure 25: RLVR training dynamics on negative high entropy token advantage shaping.

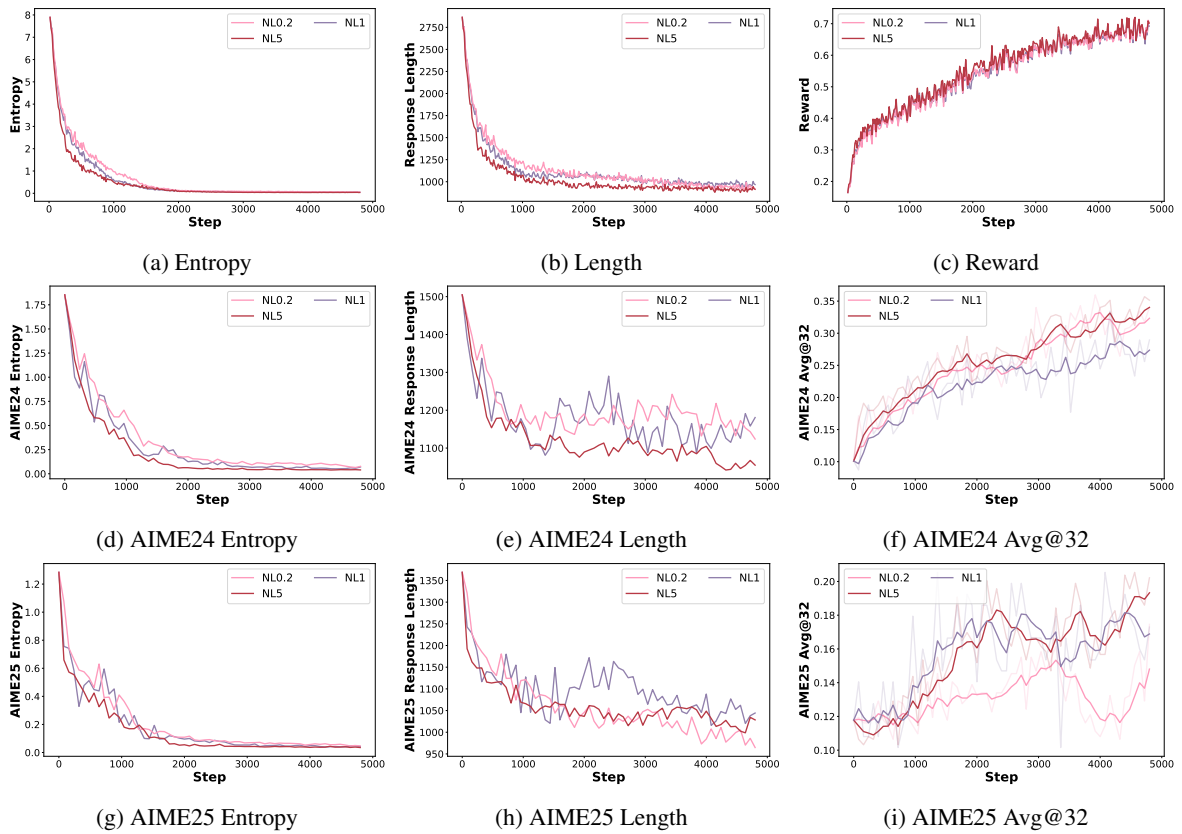


Figure 26: RLVR training dynamics on negative low entropy token advantage shaping.

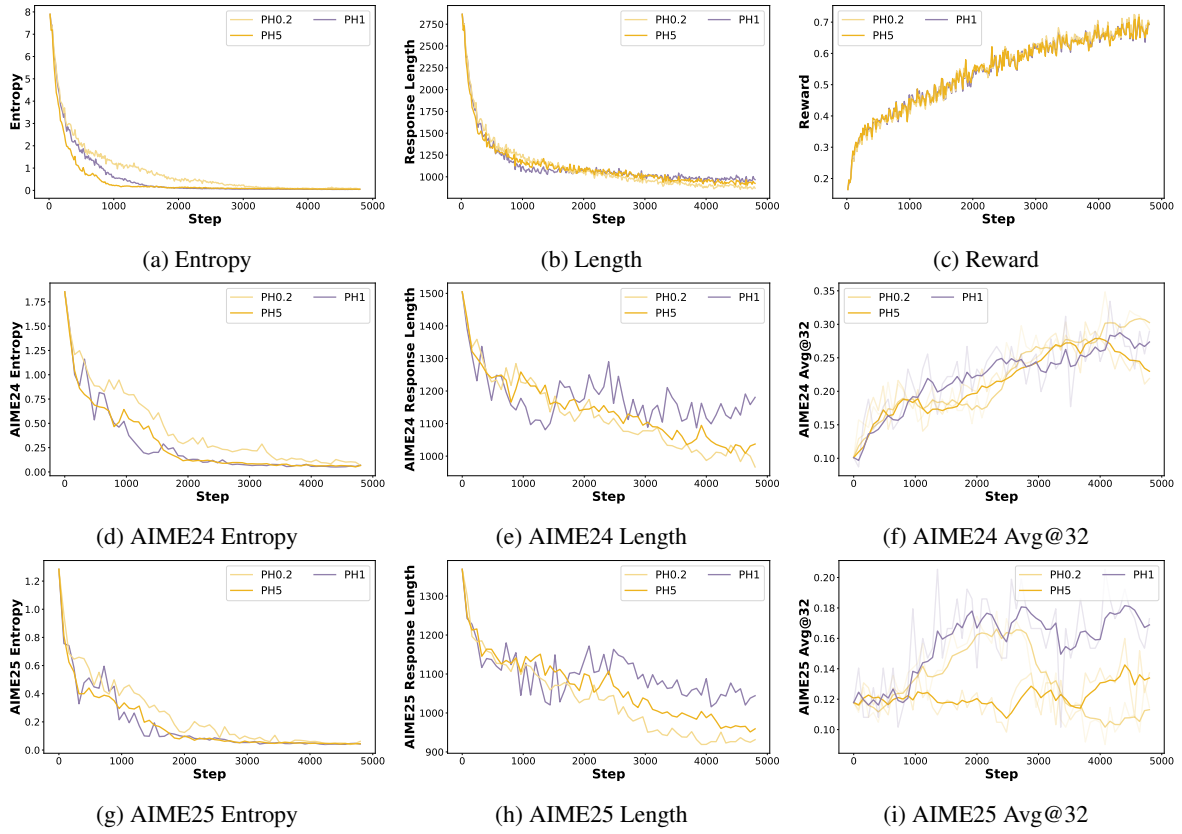


Figure 27: RLVR training dynamics on positive high probability token advantage shaping.

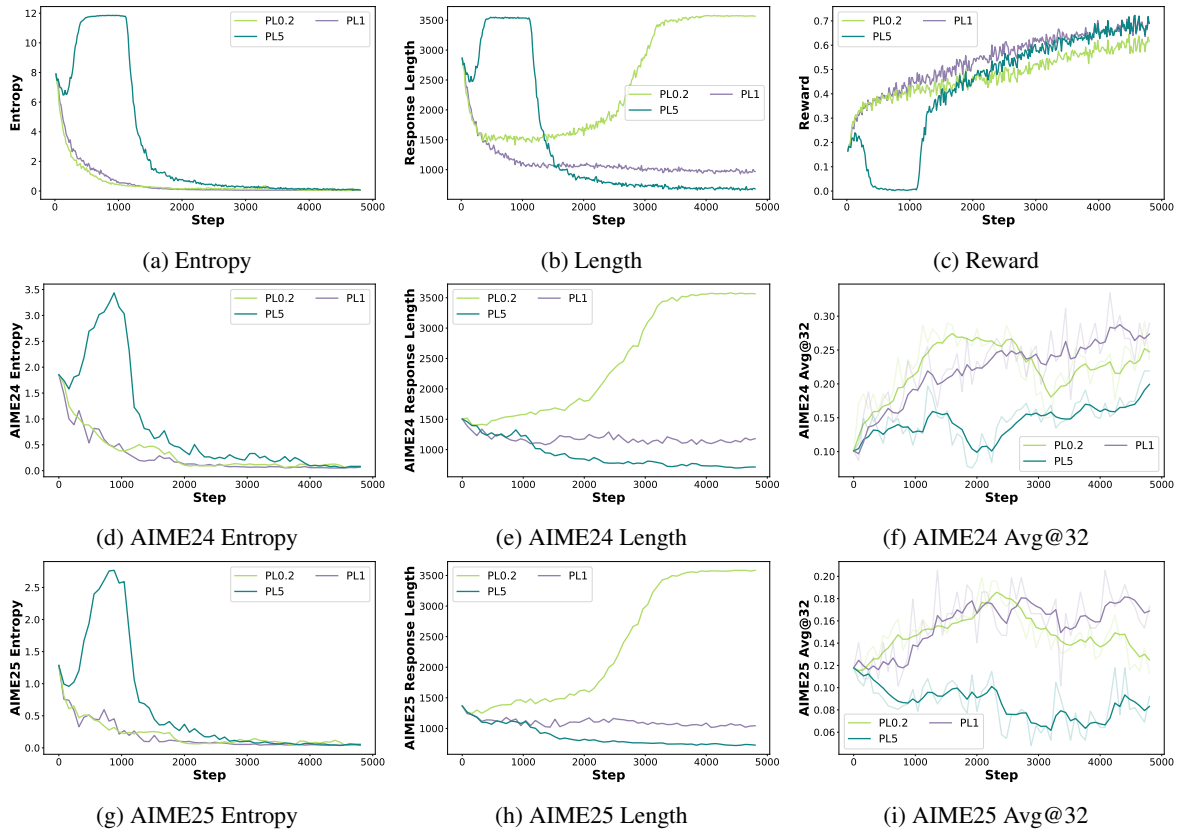


Figure 28: RLVR training dynamics on positive low probability token advantage shaping.

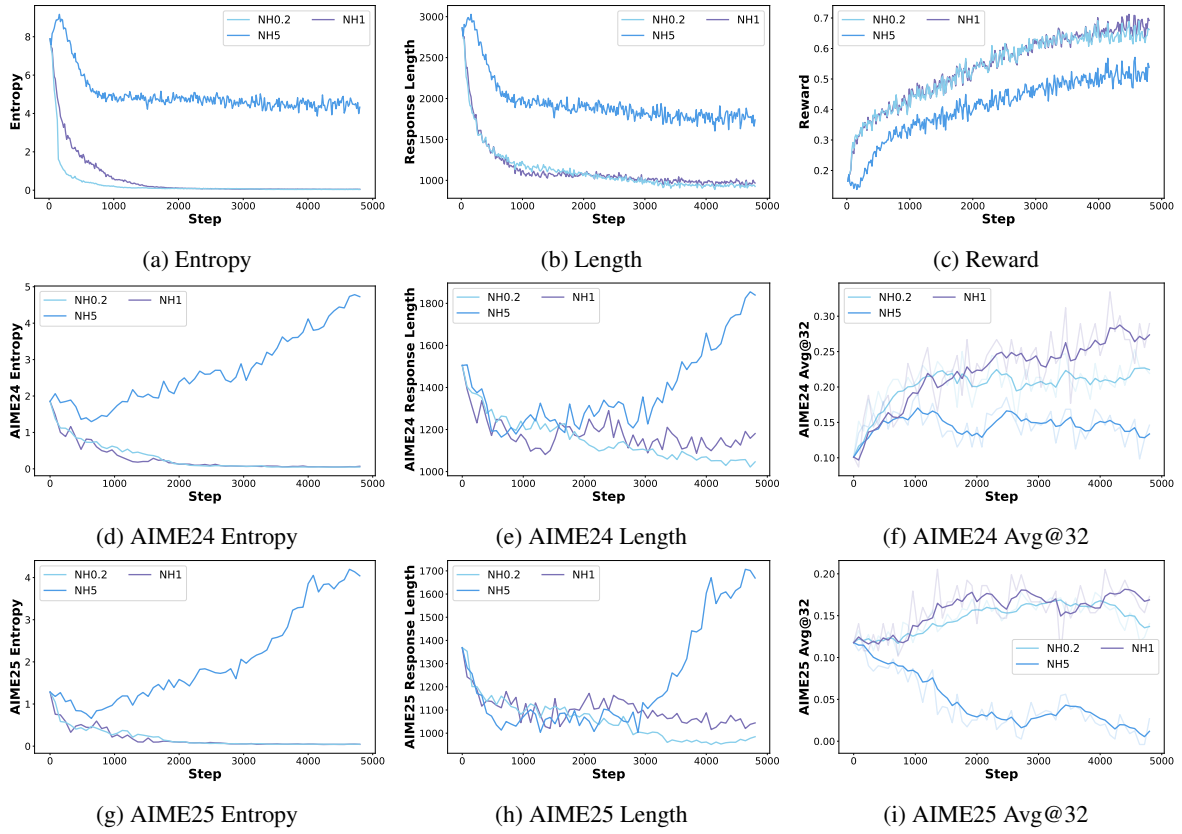


Figure 29: RLVR training dynamics on negative high probability token advantage shaping.

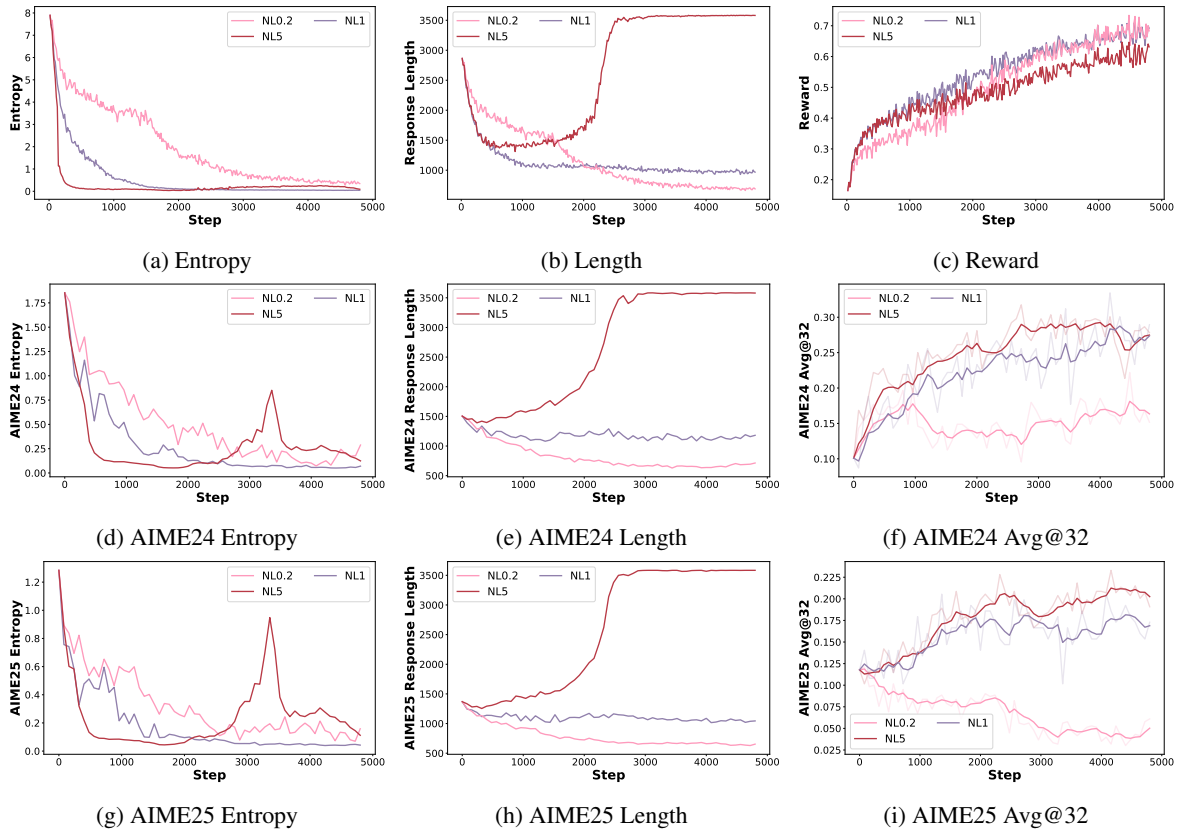


Figure 30: RLVR training dynamics on negative low probability token advantage shaping.

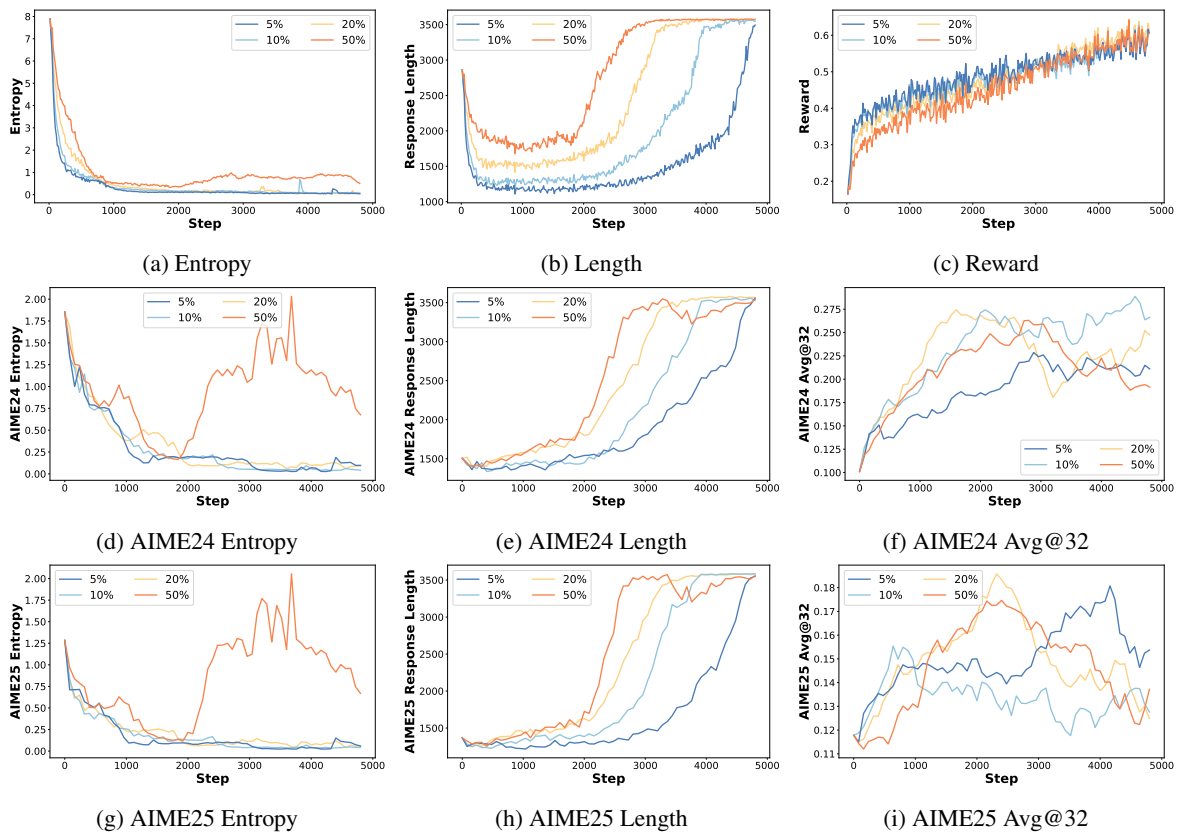


Figure 31: RLVR training dynamics under different token-shaped ratios for low-probability positive tokens (scaled by 0.2).