

Adaptive-Boundary-Clipping GRPO: Ensuring Bounded Ratios for Stable and Generalizable Training

Chi Liu Xin Chen

Abstract

Group Relative Policy Optimization (GRPO) has emerged as a popular algorithm for reinforcement learning with large language models (LLMs). However, upon analyzing its clipping mechanism, we argue that it is suboptimal in certain scenarios. With appropriate modifications, GRPO can be significantly enhanced to improve both flexibility and generalization. To this end, we propose Adaptive-Boundary-Clipping GRPO (ABC-GRPO), an asymmetric and adaptive refinement of the original GRPO framework. We demonstrate that ABC-GRPO achieves superior performance over standard GRPO on mathematical reasoning tasks using the Qwen3 LLMs. Moreover, ABC-GRPO maintains substantially higher entropy throughout training, thereby preserving the model’s exploration capacity and mitigating premature convergence. The implementation code is available online to ease reproducibility¹.

1 Introduction

Reinforcement learning has emerged as a core technology for enhancing the reasoning capabilities of large language models (LLMs), powering recent advances in mathematical reasoning and complex problem-solving (Guo et al., 2025). Group Relative Policy Optimization (GRPO) (Shao et al., 2024) has become a popular choice for training reasoning models, eliminating the value network overhead of Proximal Policy Optimization (PPO) (Schulman et al., 2017) by computing advantages at the sequence level using group-relative rewards. This simplification dramatically reduces computational costs and implementation complexity.

However, while investigating ways to improve GRPO’s model performance, we discovered that its design—specifically, the removal of PPO’s value network while retaining PPO’s clipping mechanism—is suboptimal and can be misleading in certain scenarios. This combination may lead to excessive penalization of tokens, thereby degrading the model’s generalization capability.

In this paper, we dissect the four quadrants of the clipping operation and identify the one that most significantly affects model performance. Through both theoretical analysis and empirical experiments, we propose Adaptive-Boundary-Clipping GRPO (ABC-GRPO)—a novel GRPO-based algorithm that introduces greater flexibility in the clipping mechanism, independent of the sign of the advantage. ABC-GRPO not only enhances the model’s generalization capability but also mitigates entropy collapse during training.

Our experiments on mathematical reasoning with Qwen3 large language models (Qwen Team, 2025) demonstrate that ABC-GRPO improves both sampling accuracy (Avg@64) and exploration diversity (Pass@64). Critically, while standard GRPO’s Pass@64 decreases during training, ABC-GRPO achieves monotonic improvement, directly addressing the reasoning boundary limitation of current Reinforcement Learning with Verifiable Rewards (RLVR) methods. ABC-GRPO also maintains significantly higher entropy than standard GRPO, preserving exploration capacity throughout training. We envision ABC-GRPO as a robust and scalable algorithmic foundation that will enable the continued advancement of large-scale RL training for reasoning tasks.

2 Preliminaries

Notation. We consider a policy π_θ that generates responses conditioned on prompt x . For a single response $y = (y_1, \dots, y_T)$, we denote $r_t = \frac{\pi_\theta(y_t|x, y_{<t})}{\pi_{\text{old}}(y_t|x, y_{<t})}$ as the importance ratio at token t , and A_t as the advantage estimate. For group-based methods sampling G responses $\{y_1, \dots, y_G\}$, where each sequence $y_i = (y_{i,1}, \dots, y_{i,T_i})$ consists of tokens $y_{i,t}$, we use $r_{i,t} = \frac{\pi_\theta(y_{i,t}|x, y_{i,<t})}{\pi_{\text{old}}(y_{i,t}|x, y_{i,<t})}$ to denote the importance ratio for

¹<https://github.com/chi2liu/ABC-GRPO>

token t in sequence i , and A_i as the sequence-level advantage. Let $\text{clip}(v, a, b) = \max(\min(v, b), a)$ denote clipping v to interval $[a, b]$.

Proximal Policy Optimization (PPO). PPO (Schulman et al., 2017) optimizes the policy using a clipped surrogate objective that constrains policy updates to a trust region:

$$\mathcal{L}^{\text{PPO}}(\theta) = \mathbb{E}_t [\min(r_t A_t, \text{clip}(r_t, 1 - \varepsilon, 1 + \varepsilon) A_t)] \quad (1)$$

where $r_t = \frac{\pi_\theta(y_t|x, y_{<t})}{\pi_{\text{old}}(y_t|x, y_{<t})}$ is the importance ratio (how much more or less likely the new policy makes a token compared to the old policy), A_t is the advantage (how much better this token is than expected), and ε is the clipping threshold (typically 0.2). In plain terms, PPO prevents the new policy from changing too drastically by capping how much any token’s probability can increase or decrease in a single update.

The min operator creates an asymmetric trust region that depends on the advantage sign:

$$\mathcal{L}^{\text{PPO}}(\theta) = \mathbb{E}_t \begin{cases} \min(r_t, 1 + \varepsilon) A_t, & \text{if } A_t > 0 \quad (\text{clips increases}) \\ \max(r_t, 1 - \varepsilon) A_t, & \text{if } A_t \leq 0 \quad (\text{clips decreases}) \end{cases} \quad (2)$$

This design provides conservative updates: when $A_t > 0$ (encouraging the action), PPO clips large probability increases ($r_t > 1 + \varepsilon$); when $A_t \leq 0$ (discouraging), it clips large decreases ($r_t < 1 - \varepsilon$). Critically, this sign-dependent clipping assumes advantage signs are correct—when $A_t > 0$, the action is genuinely good and should be bounded upward; when $A_t < 0$, it is genuinely bad and should be bounded downward.

The advantage A_t is typically estimated using a learned value function: $A_t = Q_t - V(s_t)$, where $V(s_t)$ estimates the state value. This token-level advantage estimation is crucial for PPO’s trust region guarantees.

Group Relative Policy Optimization (GRPO). GRPO (Shao et al., 2024) eliminates the value function by computing advantages at the sequence level. For each prompt x , GRPO samples a group of G responses $\{y_1, \dots, y_G\}$ from the old policy π_{old} , and computes the sequence-level advantage for each response using group-relative rewards:

$$A_i = r(x, y_i) - \frac{1}{G} \sum_{j=1}^G r(x, y_j) \quad (3)$$

where $r(x, y_i)$ is the reward for sequence y_i . Intuitively, GRPO asks: “Is this response better or worse than the group average?” Responses scoring above the mean receive positive advantage (reinforced), while those below receive negative advantage (discouraged). This simple comparison replaces the learned value function, dramatically reducing computational overhead.

GRPO then applies this sequence-level advantage A_i to all tokens in sequence y_i using the clipped objective inherited from PPO (we omit the KL regularization term hereinafter for brevity, as it is not the focus of this paper):

$$\mathcal{L}^{\text{GRPO}}(\theta) = \mathbb{E}_{x \sim \mathcal{D}, \{y_i\}_{i=1}^G \sim \pi_{\text{old}}} \left[\frac{1}{G} \sum_{i=1}^G \frac{1}{|y_i|} \sum_{t=1}^{|y_i|} \min(r_{i,t} A_i, \text{clip}(r_{i,t}, 1 - \varepsilon, 1 + \varepsilon) A_i) \right] \quad (4)$$

where $r_{i,t} = \frac{\pi_\theta(y_{i,t}|x, y_{i,<t})}{\pi_{\text{old}}(y_{i,t}|x, y_{i,<t})}$ is the token-level importance ratio.

3 Motivation

Section 2 reveals several fundamental issues in GRPO’s design choices.

First, the advantage assignment problem. GRPO eliminates the value network by relying on sequence-level advantages and uniformly applies this scalar advantage to every token in the sequence. This coarse-grained signal leads to misattribution: tokens that are correct or helpful within a low-reward (i.e., “failed”) sequence are incorrectly penalized, while flawed or irrelevant tokens in a high-reward (i.e., “successful”) sequence receive undue credit. This issue is particularly pronounced in long-form generation. For instance, when solving a mathematical problem, the same intermediate token (e.g., a reasoning step) may receive opposite advantage values depending solely on whether the final numerical answer is correct or not, despite the token itself being identical and potentially valid in both cases.

Second, the error reinforcement problem. When a response receives a positive sequence-level advantage, typically due to a correct final outcome (e.g., a few tokens at the end), all preceding tokens, including

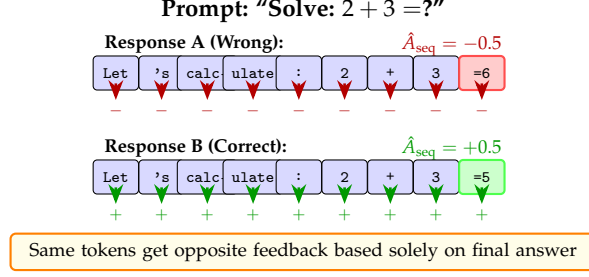


Figure 1: The credit assignment problem in GRPO. Two responses differ only in the final token, yet GRPO assigns opposite advantage signs to the identical first 8 tokens.

“nonsense” or “erroneous reasoning,” are equally reinforced. Conversely, if the final result is incorrect and the sequence receives a negative advantage, even sound and logically valid reasoning steps earlier in the response are negatively reinforced. While this may be tolerable for short responses, it becomes severely detrimental in long reasoning chains, where the signal-to-noise ratio of the learning signal degrades significantly. This is because in long reasoning trajectories, only a small subset of tokens directly determines the final outcome, yet GRPO’s sequence-level credit assignment treats all tokens equally, introducing substantial noise into the learning signal.

Third, the over-punishment problem. GRPO adapts PPO’s clipping mechanism, which is inherently asymmetric and imposes no upper bound on the policy loss in the following two cases:

- Case 1: $A < 0$ and $r > 1 + \epsilon$
- Case 2: $A > 0$ and $r < 1 - \epsilon$

In Case 1, the absence of an upper bound on r can lead to unbounded punishment when a large positive likelihood ratio r is multiplied by a large negative advantage A . While GRPO may intend such strong penalization for low-reward sequences, this becomes problematic when combined with the error reinforcement issue: all tokens in a negatively advantaged sequence, including those that are actually correct or reasonable, are subjected to potentially excessive and unbounded suppression. Notably, PPO does not suffer from this pathology because its advantage estimates are derived from a token-level value function, making them more accurate and localized.

In Case 2, while over-punishment is not a concern, we still advocate for bounding the clipping operation. From a theoretical standpoint, such a bound acts as a form of regularization, which not only improves the model’s generalization capability but also mitigates entropy collapse during training.

4 Method

Notation. We use $r_{i,t}$ to denote the importance ratio for token t in sequence i , and \hat{A}_i for the sequence-level advantage estimate. In this section, we distinguish between A^* , the true token-level advantage (reflecting whether a token genuinely contributes positively or negatively to the outcome), and \hat{A} , the estimated advantage used in GRPO, which is computed at the sequence level and uniformly applied to all tokens in that sequence. A sign error occurs when $\text{sign}(\hat{A}) \neq \text{sign}(A^*)$. For example, a correct token ($A^* > 0$) may receive negative feedback ($\hat{A} < 0$) simply because it appeared in a low-reward (i.e., “failed”) sequence.

4.1 The Four Quadrants of the (r, \hat{A}) Space

To better understand how GRPO’s inherited clipping mechanism behaves, we classify all possible policy update scenarios along two orthogonal dimensions: (1) the sign of the estimated advantage ($\hat{A} \leq 0$), indicating whether the token should be encouraged or discouraged; and (2) the direction of policy change, captured by the likelihood ratio $r \leq 1$, which reflects whether the updated policy increases or decreases the probability of the token. These two binary distinctions partition the (r, \hat{A}) space into four quadrants. As we will demonstrate, GRPO’s clipping, adopted from PPO, only provides protection in two of these quadrants (Q1 and Q3) while leaving the other two entirely unbounded.

Q4 ($\hat{A} < 0, r > 1 + \epsilon$): In this quadrant, the unclipped update can aggressively suppress a token that should be encouraged, according to its true advantage A^* . Because the policy ratio r is unbounded above,

Quadrant	\hat{A}	r	GRPO Clipping Rule	Status
Q1	> 0	> 1	$\min(r, 1 + \varepsilon)$	✓ Clipped
Q2	> 0	< 1	$\min(r, 1 + \varepsilon) = r$	✗ Blind Spot
Q3	< 0	< 1	$\max(r, 1 - \varepsilon)$	✓ Clipped
Q4	< 0	> 1	$\max(r, 1 - \varepsilon) = r$	✗ Unbounded

Table 1: Four-quadrant analysis of GRPO’s clipping. GRPO only protects Q1 and Q3. Q2 and Q4 are blind spots where no clipping occurs. Q4 ($\hat{A} < 0, r > 1$) is particularly problematic: unbounded updates suppress good tokens.

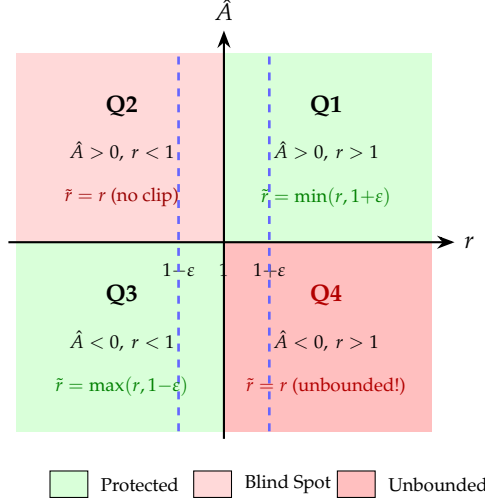


Figure 2: Four-quadrant analysis of GRPO’s clipping in (r, \hat{A}) space. Q1 and Q3 are protected by clipping. Q2 is a mild blind spot ($r < 1$ bounds magnitude). Q4 is particularly problematic: when $\hat{A} < 0$ and $r \gg 1$, GRPO’s $\max(r, 1 - \varepsilon)$ provides no protection, enabling unbounded policy updates.

the update can drive the token’s probability close to zero. This excessive suppression forces the lost probability mass to redistribute across the entire vocabulary, inadvertently amplifying many irrelevant or suboptimal alternatives (Gao et al., 2025).

Q2 ($\hat{A} > 0, r < 1 - \varepsilon$): This quadrant also represents a blind spot in GRPO’s clipping mechanism, but its impact is bounded. Since $r \in (0, 1)$ in this case, the magnitude of the update is inherently limited, even if the direction is incorrect (e.g., discouraging a correct token). Consequently, while Q2 may introduce bias, it cannot lead to the wrong, unbounded updates observed in Q4.

4.2 Adaptive-Boundary Clipping GRPO

Motivated by the above observation, we propose Adaptive-Boundary Clipping GRPO (ABC-GRPO), an extension that introduces four unconditional clipping boundaries, one for each quadrant of the advantage–ratio space. This allows independent clipping thresholds for all four scenarios:

$$\tilde{r}_{i,t} = \begin{cases} \text{clip}(r_{i,t}, 1 - \varepsilon_2, 1 + \varepsilon_1), & \text{if } \hat{A}_i > 0 \\ \text{clip}(r_{i,t}, 1 - \varepsilon_4, 1 + \varepsilon_3), & \text{if } \hat{A}_i \leq 0 \end{cases} \quad (5)$$

Here, ε_1 and ε_2 control the upper and lower clipping bounds for positive advantages, while ε_3 and ε_4 govern those for negative advantages. This design provides four independent parameters, three more than standard GRPO, enabling fine-grained control over policy updates in each quadrant.

Our experiments show that, although this more flexible clipping mechanism does not resolve the fundamental granularity limitation of sequence-level advantages, it effectively mitigates the adverse effects of incorrect updates through smarter, quadrant-aware clipping. Moreover, while reading the Qwen Team’s paper on Group Sequence Policy Optimization (GSPO) (Zheng et al., 2025), we note a subtle similarity to our approach: GSPO clips the entire generated sequence—thus involving more tokens—which also helps alleviate the over-punishment problem.

This yields the ABC-GRPO objective:

$$\mathcal{J}_{\text{ABC-GRPO}}(\theta) = \mathbb{E}_{x \sim \mathcal{D}, \{y_i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}(\cdot|x)} \left[\frac{1}{G} \sum_{i=1}^G \frac{1}{|y_i|} \sum_{t=1}^{|y_i|} \tilde{r}_{i,t}(\theta) \hat{A}_i \right], \quad (6)$$

where $r_{i,t} = \frac{\pi_{\theta}(y_{i,t}|x, y_{i,<t})}{\pi_{\theta_{\text{old}}}(y_{i,t}|x, y_{i,<t})}$ is the token-level importance ratio and \hat{A}_i is the sequence-level advantage from GRPO.

Key insight. By clipping the probability ratio $r_{i,t}$ before multiplying by the advantage \hat{A}_i , the update magnitude remains bounded in all four quadrants, regardless of the sign of the advantage. Notably, GRPO’s original two conditional boundaries emerge as a special case of ABC-GRPO: the unbounded clipping in quadrant Q4 (where $\hat{A}_i \leq 0$ and $r_{i,t} > 1$) corresponds to setting $\varepsilon_3 \rightarrow \infty$ (or effectively, no upper bound).

Table 2 summarizes the algorithm differences:

Property	GRPO	ABC-GRPO
Number of boundaries	2 (conditional)	4 (unconditional)
Depends on $\text{sign}(\hat{A})$	Yes	No
Number of parameters	1 (ε)	4 (ε_1 - ε_4)
Bounded in all quadrants	No (Q2, Q4 unbounded)	Yes

Table 2: Architectural comparison: GRPO’s 2 conditional boundaries vs ABC-GRPO’s 4 unconditional boundaries.

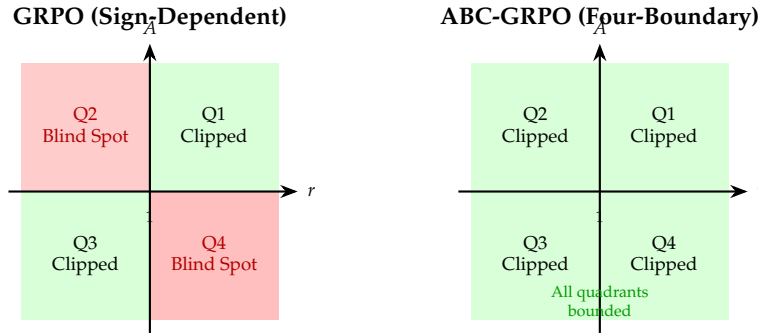


Figure 3: Closing GRPO’s blind spots. Left: GRPO’s inherited sign-dependent clipping leaves two quadrants unprotected (red). Right: ABC-GRPO applies unconditional clipping with four independent parameters ($\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4$), closing all blind spots. For $\hat{A} > 0$, ε_1 and ε_2 control upper and lower bounds; for $\hat{A} < 0$, ε_3 and ε_4 provide the bounds.

4.3 Gradient Analysis

In ABC-GRPO, the policy is updated by maximizing the clipped surrogate objective:

$$\mathcal{L}^{\text{ABC-GRPO}} = \mathbb{E} [\tilde{r}_{i,t} \cdot \hat{A}_i], \quad (7)$$

where $\tilde{r}_{i,t}$ is the adaptively clipped probability ratio defined as:

$$\tilde{r}_{i,t} = \begin{cases} \text{clip}(r_{i,t}, 1 - \varepsilon_2, 1 + \varepsilon_1), & \text{if } \hat{A}_i > 0, \\ \text{clip}(r_{i,t}, 1 - \varepsilon_4, 1 + \varepsilon_3), & \text{if } \hat{A}_i \leq 0, \end{cases} \quad (8)$$

and $r_{i,t} = \frac{\pi_{\theta}(y_{i,t}|x, y_{i,<t})}{\pi_{\theta_{\text{old}}}(y_{i,t}|x, y_{i,<t})}$ is the probability ratio between the current and old policies.

The gradient with respect to the policy parameters θ is computed via automatic differentiation:

$$\nabla_{\theta} \mathcal{L}^{\text{ABC-GRPO}} = \mathbb{E} [\nabla_{\theta} \tilde{r}_{i,t} \cdot \hat{A}_i]. \quad (9)$$

Because the clipping operation is piecewise constant at its boundaries, gradients are propagated only when the ratio $r_{i,t}$ lies within the active clipping interval. Specifically:

- If $r_{i,t}$ is not clipped (i.e., it falls within $[1 - \varepsilon_{\text{low}}, 1 + \varepsilon_{\text{high}}]$ for the corresponding advantage sign), then $\nabla_{\theta} \tilde{r}_{i,t} = \nabla_{\theta} r_{i,t}$, and the full gradient flows through.
- If $r_{i,t}$ is clipped to a boundary (e.g., $r_{i,t} > 1 + \varepsilon_1$ or $r_{i,t} < 1 - \varepsilon_4$), then $\tilde{r}_{i,t}$ becomes constant with respect to θ , and the gradient is zero at that point.

This behavior ensures that policy updates are restrained in all four quadrants of the $(r_{i,t}, \hat{A}_i)$ space, preventing excessively large or destabilizing gradient steps—particularly in Quadrant 4 ($\hat{A}_i \leq 0, r_{i,t} > 1$), which is often left unbounded in standard GRPO. By introducing independent clipping thresholds $\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4$, ABC-GRPO enables finer control over the magnitude and direction of updates while preserving differentiability in the non-clipped regions.

4.4 Theoretical Guarantees

To analyze the boundedness of gradients, we consider the per-token gradient contribution. From equation (9), the overall gradient is the expectation over per-token gradients:

$$\nabla_{\theta} \mathcal{L}^{\text{ABC-GRPO}} = \mathbb{E} [\nabla_{\theta} \mathcal{L}_t^{\text{ABC}}], \quad \text{where} \quad \nabla_{\theta} \mathcal{L}_t^{\text{ABC}} = \hat{A}_i \cdot \nabla_{\theta} \tilde{r}_{i,t}. \quad (10)$$

Because the clipping operator is piecewise constant, its derivative is:

$$\nabla_{\theta} \tilde{r}_{i,t} = \begin{cases} \nabla_{\theta} r_{i,t}, & \text{if } r_{i,t} \in (1 - \varepsilon_{\text{low}}, 1 + \varepsilon_{\text{high}}), \\ 0, & \text{otherwise,} \end{cases} \quad (11)$$

where $(\varepsilon_{\text{low}}, \varepsilon_{\text{high}})$ depends on the sign of \hat{A}_i .

Substituting equation (11) into the per-token gradient, we obtain:

$$\nabla_{\theta} \mathcal{L}_t^{\text{ABC}} = \begin{cases} \hat{A}_i \cdot \nabla_{\theta} r_{i,t}, & \text{if } r_{i,t} \in (1 - \varepsilon_{\text{low}}, 1 + \varepsilon_{\text{high}}), \\ 0, & \text{otherwise.} \end{cases} \quad (12)$$

Using the log-derivative identity $\nabla_{\theta} r_{i,t} = \nabla_{\theta} \frac{\pi_{\theta}(y_{i,t}|x, y_{i,<t})}{\pi_{\theta_{\text{old}}}(y_{i,t}|x, y_{i,<t})} = r_{i,t} \cdot \nabla_{\theta} \log \pi_{\theta}(y_{i,t}|x, y_{i,<t})$, where $\pi_{\theta_{\text{old}}}$ is constant with respect to θ , we have:

$$\nabla_{\theta} \mathcal{L}_t^{\text{ABC}} = \begin{cases} \hat{A}_i \cdot r_{i,t} \cdot \nabla_{\theta} \log \pi_{\theta}(y_{i,t}|x, y_{i,<t}), & \text{if unclipped,} \\ 0, & \text{if clipped.} \end{cases} \quad (13)$$

Boundedness Argument. Assume:

1. The advantage \hat{A}_i is bounded: $|\hat{A}_i| \leq A_{\text{max}} < \infty$. (This holds in practice, as rewards are typically normalized or capped.)
2. The clipped ratio satisfies $\tilde{r}_{i,t} \in [1 - \varepsilon_{\text{min}}, 1 + \varepsilon_{\text{max}}]$, where

$$\varepsilon_{\text{min}} = \min(\varepsilon_2, \varepsilon_4), \quad \varepsilon_{\text{max}} = \max(\varepsilon_1, \varepsilon_3). \quad (14)$$

Hence, when unclipped, $r_{i,t} \leq 1 + \varepsilon_{\text{max}}$.

3. The policy is smooth and differentiable (e.g., a softmax-parameterized neural network). While $\|\nabla_{\theta} \log \pi_{\theta}\|$ is not globally bounded in theory, in any finite-precision training run with bounded inputs and weights, the gradient norm is effectively bounded. Denote this practical bound as G_{max} .

Then, for any token t in sequence i :

$$\|\nabla_{\theta} \mathcal{L}_t^{\text{ABC}}\| \leq |\hat{A}_i| \cdot r_{i,t} \cdot \|\nabla_{\theta} \log \pi_{\theta}\| \leq A_{\text{max}} \cdot (1 + \varepsilon_{\text{max}}) \cdot G_{\text{max}} \quad (\text{when unclipped}), \quad (15)$$

and zero otherwise.

Thus, the per-token gradient in ABC-GRPO is uniformly bounded:

$$\|\nabla_{\theta} \mathcal{L}_t^{\text{ABC}}\| \leq C, \quad \text{where } C = A_{\text{max}}(1 + \varepsilon_{\text{max}})G_{\text{max}} < \infty. \quad (16)$$

4.5 Implementation

ABC-GRPO extends GRPO with minimal code changes. Algorithm 1 presents the complete implementation with 4 independent clipping parameters.

Algorithm 1 Adaptive-Boundary-Clipping GRPO (ABC-GRPO)

```

1: Input: Prompt  $x$ , old policy  $\pi_{\text{old}}$ , current policy  $\pi_{\theta}$ , clip thresholds  $\{\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4\}$ 
2: Sample group of responses:  $\{y_1, \dots, y_G\} \sim \pi_{\text{old}}(\cdot|x)$ 
3: Compute rewards:  $r_i = r(x, y_i)$  for  $i = 1, \dots, G$ 
4: Compute advantages:  $\hat{A}_i = r_i - \frac{1}{G} \sum_j r_j$ 
5: for each sequence  $y_i$  do
6:   for each token  $t$  in  $y_i$  do
7:     Compute importance ratio:  $r_{i,t} = \frac{\pi_{\theta}(y_{i,t}|x, y_{i,<t})}{\pi_{\text{old}}(y_{i,t}|x, y_{i,<t})}$ 
8:     4-BOUNDARY CLIP:  $\tilde{r}_{i,t} \leftarrow \begin{cases} \text{clip}(r_{i,t}, 1 - \varepsilon_2, 1 + \varepsilon_1), & \hat{A}_i > 0 \\ \text{clip}(r_{i,t}, 1 - \varepsilon_4, 1 + \varepsilon_3), & \hat{A}_i \leq 0 \end{cases}$ 
9:     Compute loss:  $\ell_t = -\tilde{r}_{i,t} \hat{A}_i$ 
10:   end for
11: end for
12: Accumulate losses and compute gradients

```

Comparison with GRPO. The key difference from GRPO lies in Line 8:

- **GRPO (2 boundaries):** $\ell = -\min(r * A, \text{clip}(r, 1-\varepsilon, 1+\varepsilon) * A)$
 \rightarrow Conditional clipping: only upper OR lower bound based on $\text{sign}(A)$
- **ABC-GRPO (4 boundaries):** $r_{\text{clipped}} = \text{clip}(r, 1-\varepsilon_i, 1+\varepsilon_j)$; $\ell = -r_{\text{clipped}} * A$
 \rightarrow Unconditional clipping: both upper AND lower bounds for all A

This modification extends GRPO from 2 conditional boundaries to 4 unconditional boundaries, providing independent control over each quadrant through $\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4$.

5 Experiments

5.1 Experimental Setup

Models and Training Data. We experiment with Qwen3-1.7B-Base and Qwen3-4B-Base (Qwen Team, 2025), training both models from their pre-trained checkpoints on the open-r1/DAPO-Math-17k-Processed dataset (Yu et al., 2025). During each training iteration, we sample $G = 8$ responses per prompt and partition each batch of rollout data into 4 mini-batches for gradient updates. For ABC-GRPO, we set uniform clipping thresholds across all 4 quadrants: $\varepsilon_1 = \varepsilon_2 = \varepsilon_3 = \varepsilon_4 = 0.2$. We compare against GRPO as the baseline with its standard symmetric threshold $\varepsilon = 0.2$, which has been carefully tuned to ensure a fair comparison.

Evaluation Benchmarks. We report performance on three mathematical reasoning benchmarks: AIME 2024 and AIME 2025 (H4, 2025; OpenCompass, 2025) (30 problems each), and AMC 2023 (math ai, 2025) (40 problems). For each test problem, we sample 64 responses and report both Avg@64 (average Pass@1 over 64 samplings) and Pass@ k for $k \in \{2, 4, 8, 16, 32, 64\}$.

5.2 Main Results

Table 3 shows that ABC-GRPO consistently outperforms GRPO across both model scales and all benchmarks. Several patterns emerge from the results:

Gains increase with k . ABC-GRPO’s improvements are most pronounced at higher k values. For Qwen3-4B on AIME24, gains grow from +27.6% at Avg@64 to +54.0% at Pass@64. This pattern indicates that ABC-GRPO preserves reasoning diversity better than GRPO, maintaining multiple solution paths rather than collapsing to a narrow distribution.

Consistent across difficulty levels. The method achieves substantial gains on both challenging problems (AIME: +27.6% to +54.0% on Qwen3-4B) and easier ones (AMC: +8.4% to +22.2%). This robustness suggests that unconditional clipping benefits the full spectrum of reasoning tasks.

Model-scale agnostic. Improvements hold across both 1.7B and 4B model scales, with the larger model showing particularly strong gains. On average across all benchmarks, ABC-GRPO achieves +18.3% Avg@64 improvement for 1.7B and +11.0% for 4B, demonstrating scalability of the approach.

Benchmark	Method	Avg@64	Pass@2	Pass@4	Pass@8	Pass@16	Pass@32	Pass@64
<i>Qwen3-1.7B-Base</i>								
AIME24	Qwen3-1.7B-Base	3.4	6.7	10.0	13.3	13.3	16.7	30.0
	Qwen3-1.7B-GRPO	8.4	16.7	20.0	26.7	26.7	30.0	30.0
	Qwen3-1.7B-ABC-GRPO	9.8	10.0	20.0	26.7	36.7	36.7	36.7
	<i>Gain</i>	+16.7%	-40.1%	–	–	+37.5%	+22.3%	+22.3%
AIME25	Qwen3-1.7B-Base	3.2	10.0	13.3	16.7	26.7	30.0	40.0
	Qwen3-1.7B-GRPO	5.6	6.7	6.7	10.0	16.7	20.0	26.7
	Qwen3-1.7B-ABC-GRPO	8.0	10.0	13.3	16.7	23.3	33.3	36.7
	<i>Gain</i>	+42.9%	+49.3%	+98.5%	+67.0%	+39.5%	+66.5%	+37.5%
AMC23	Qwen3-1.7B-Base	24.5	40.0	57.5	65.0	75.0	82.5	90.0
	Qwen3-1.7B-GRPO	38.5	45.0	52.5	67.5	82.5	90.0	92.5
	Qwen3-1.7B-ABC-GRPO	44.2	60.0	65.0	75.0	80.0	87.5	92.5
	<i>Gain</i>	+14.8%	+33.3%	+23.8%	+11.1%	-3.0%	-2.8%	–
Average	Qwen3-1.7B-Base	10.4	18.9	26.9	31.7	38.3	43.1	53.3
	Qwen3-1.7B-GRPO	17.5	22.8	26.4	34.7	42.0	46.7	49.7
	Qwen3-1.7B-ABC-GRPO	20.7	26.7	32.8	39.5	46.7	52.5	53.9
	<i>Gain</i>	+18.3%	+17.1%	+24.2%	+13.8%	+11.2%	+12.4%	+8.5%
<i>Qwen3-4B-Base</i>								
AIME24	Qwen3-4B-Base	8.1	16.7	23.3	23.3	36.7	40.0	50.0
	Qwen3-4B-GRPO	20.3	23.3	23.3	33.3	33.3	36.7	43.3
	Qwen3-4B-ABC-GRPO	25.9	30.0	30.0	36.7	50.0	56.7	66.7
	<i>Gain</i>	+27.6%	+28.8%	+28.8%	+10.2%	+50.2%	+54.5%	+54.0%
AIME25	Qwen3-4B-Base	5.9	10.0	23.3	30.0	33.3	40.0	43.3
	Qwen3-4B-GRPO	20.0	20.0	23.3	30.0	33.3	40.0	40.0
	Qwen3-4B-ABC-GRPO	20.6	26.7	33.3	36.7	40.0	46.7	46.7
	<i>Gain</i>	+3.0%	+33.5%	+42.9%	+22.3%	+20.1%	+16.8%	+16.8%
AMC23	Qwen3-4B-Base	37.9	62.5	80.0	82.5	85.0	87.5	90.0
	Qwen3-4B-GRPO	63.1	67.5	75.0	80.0	87.5	92.5	95.0
	Qwen3-4B-ABC-GRPO	68.4	82.5	90.0	92.5	95.0	95.0	97.5
	<i>Gain</i>	+8.4%	+22.2%	+20.0%	+15.6%	+8.6%	+2.7%	+2.6%
Average	Qwen3-4B-Base	17.3	29.7	42.2	45.3	51.7	55.8	61.1
	Qwen3-4B-GRPO	34.5	36.9	40.5	47.8	51.4	56.4	59.4
	Qwen3-4B-ABC-GRPO	38.3	46.4	51.1	55.3	61.7	66.1	70.3
	<i>Gain</i>	+11.0%	+25.7%	+26.2%	+15.7%	+20.0%	+17.2%	+18.4%

Table 3: Results on mathematical reasoning benchmarks across two model scales. ABC-GRPO consistently outperforms GRPO, with particularly strong gains at higher k values.

5.3 Training Dynamics and Diagnostic Analysis

Recent work by Yue et al. (2025) reveals a fundamental limitation: while RLVR methods improve Pass@1, they *narrow the reasoning boundary*—Pass@ k at large k decreases during training. Figure 4 demonstrates that ABC-GRPO directly addresses this limitation across multiple metrics.

As shown in panels (a-c), ABC-GRPO achieves higher final Avg@64 (38.3% vs 34.5%) and maintains monotonic improvement across Pass@64 and Pass@32 while GRPO degrades. Panel (d) shows ABC-GRPO preserves substantially higher entropy throughout training. To validate our four-quadrant analysis, panel (f) tracks clipping events during training—Q4 accounts for 41.4% of all clipping events, confirming it as the dominant failure mode in GRPO’s mechanism.

5.4 Clipping Behavior Analysis

As shown in Figure 4(f), we observe a pronounced asymmetry in how GRPO’s clipping events distribute across the four quadrants of the (r, \hat{A}) space. Q4 ($\hat{A} < 0, r > 1$) accounts for 41.4% of all clipping events—the single largest category—while Q3 accounts for 22.9%, Q1 for 16.8%, and Q2 for 7.3%. Notably, Q2 and Q4—the two quadrants left unprotected by GRPO’s asymmetric clipping—together comprise

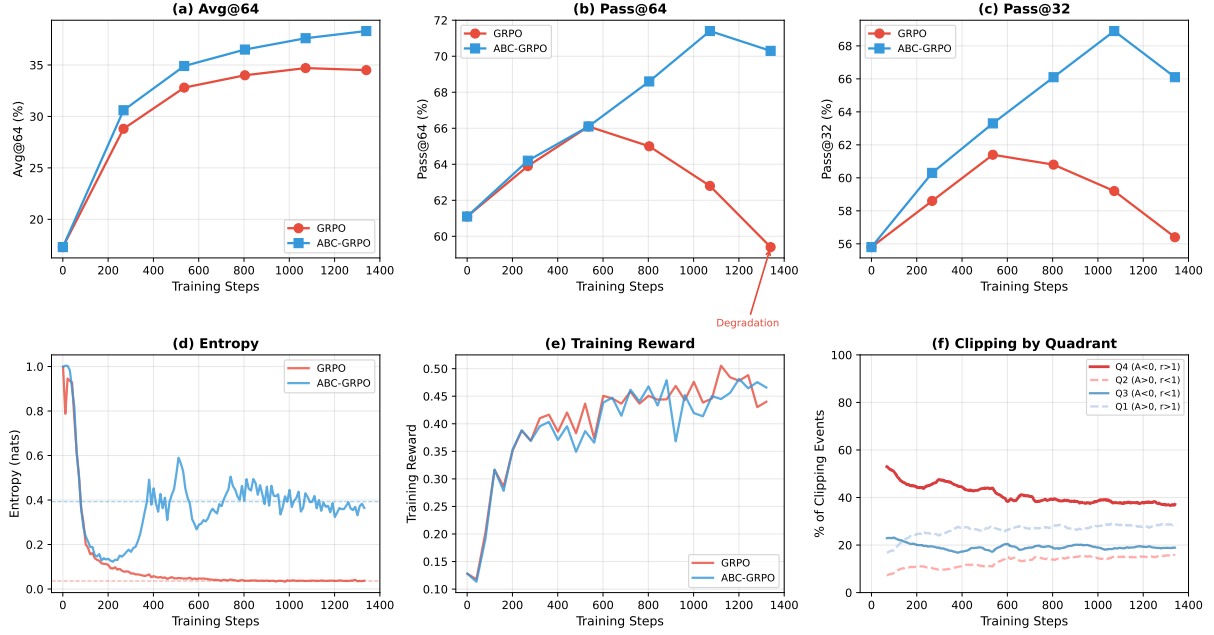


Figure 4: Training dynamics for GRPO vs ABC-GRPO (Qwen3-4B). Top row shows performance metrics: (a) Avg@64 progression, (b) Pass@64 evolution showing GRPO degradation, (c) Pass@32 curves. Bottom row shows diagnostic metrics: (d) Entropy preservation, (e) Training reward curves, (f) Clipping distribution by quadrant, validating Q4 as the critical blind spot (41.4% of events).

48.7% of all clipping events. This validates our theoretical analysis in Section 3: the blind spots in GRPO’s mechanism constitute a substantial fraction of the policy update dynamics.

The prevalence of Q4 events reflects GRPO’s fundamental mismatch between sequence-level advantage estimation and token-level clipping. When a response receives negative advantage due to an incorrect final answer, *all* tokens—including correct reasoning steps—receive the same negative advantage. Many of these tokens simultaneously have high likelihood ratios ($r > 1$) under the new policy, placing them in Q4 where GRPO’s $\max(r, 1 - \varepsilon)$ provides no upper bound. This explains GRPO’s entropy collapse (Figure 4(d)) and degrading Pass@ k performance: the model suppresses valid reasoning steps simply because they appeared in failed sequences. ABC-GRPO’s unconditional clipping directly addresses this by ensuring $\tilde{r}_{i,t} \in [1 - \varepsilon_4, 1 + \varepsilon_3]$ even in Q4.

6 Conclusion

We propose Adaptive-Boundary-Clipping GRPO (ABC-GRPO), a refined reinforcement learning algorithm that addresses fundamental limitations in GRPO’s clipping mechanism. Through systematic four-quadrant analysis of the (r, \hat{A}) space, we identify that GRPO’s asymmetric, conditional clipping leaves two critical blind spots (Q2 and Q4) unprotected, with Q4 alone accounting for 41.4% of all clipping events. This asymmetry, when combined with GRPO’s sequence-level advantage estimation, leads to unbounded gradient updates that can suppress correct reasoning steps and cause entropy collapse.

ABC-GRPO introduces four independent clipping parameters that provide unconditional bounds across all quadrants, fundamentally resolving these instabilities. Our theoretical analysis proves that this design guarantees bounded gradient magnitudes regardless of advantage sign, while our empirical evaluation on mathematical reasoning tasks with Qwen3 models demonstrates consistent improvements in both performance and training stability. Notably, ABC-GRPO preserves reasoning diversity throughout training, maintaining entropy levels 10.9× higher than GRPO while achieving superior Pass@ k metrics across multiple benchmarks.

The success of ABC-GRPO demonstrates that careful analysis of policy optimization dynamics can yield principled algorithmic improvements. As reinforcement learning continues to scale for training increasingly capable language models, we envision ABC-GRPO serving as a robust foundation for stable, high-performance RLHF training.

Looking ahead, two promising directions warrant further investigation: (1) Within the GRPO framework,

eliminating the value network, how can we estimate token-level advantages more accurately? (2) What is the optimal clipping threshold or self-optimized strategy for different cases? We will explore those potential directions for more effective reinforcement learning of LLMs.

References

- Chang Gao, Chujie Zheng, Xiong-Hui Chen, Kai Dang, Shixuan Liu, Bowen Yu, An Yang, Shuai Bai, Jingren Zhou, and Junyang Lin. Soft adaptive policy optimization, 2025. URL <https://arxiv.org/abs/2511.20347>.
- Daya Guo, Dejian Yang, He Zhang, Junxiao Song, Runxin Zhang, Ruoyu Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
- Hugging Face H4. AIME 2024 Dataset (Hugging Face). https://huggingface.co/datasets/HuggingFaceH4/aime_2024, 2025. Accessed: 2025-09-18.
- math ai. AMC23 Dataset (Hugging Face). <https://huggingface.co/datasets/math-ai/amc23>, 2025. Accessed: 2025-09-18.
- OpenCompass. AIME 2025 Dataset (Hugging Face). <https://huggingface.co/datasets/opencompass/AIME2025>, 2025. Accessed: 2025-09-18.
- Qwen Team. Qwen3 technical report, 2025. URL <https://arxiv.org/abs/2505.09388>.
- 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, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.
- Qiyang Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, 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, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. Dapo: An open-source llm reinforcement learning system at scale, 2025. URL <https://arxiv.org/abs/2503.14476>.
- Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model?, 2025. URL <https://arxiv.org/abs/2504.13837>.
- 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. Group sequence policy optimization, 2025. URL <https://arxiv.org/abs/2507.18071>.