

ETR: Outcome-Guided Elastic Trust Regions for Policy Optimization

Shijie Zhang^{1,2,†}, Kevin Zhang^{1,†}, Zheyuan Gu^{1,†}, Xiang Guo², Rujun Guo²,
Shaoyu Liu², Guanjun Jiang², Xiaozhao Wang²

¹Peking University ²Qwen Applications Business Group, Alibaba Group

Abstract

Reinforcement Learning with Verifiable Rewards (RLVR) has emerged as an important paradigm for unlocking reasoning capabilities in large language models, exemplified by the success of OpenAI o1 and DeepSeek-R1. Currently, Group Relative Policy Optimization (GRPO) stands as a prominent algorithm in this domain due to its stable training and critic-free efficiency. However, we argue that GRPO suffers from a structural limitation: it imposes a uniform, static trust region constraint across all samples. This design implicitly assumes signal homogeneity, a premise misaligned with the heterogeneous nature of outcome-driven learning, where advantage magnitudes and variances fluctuate significantly. Consequently, static constraints fail to fully exploit high-quality signals while insufficiently suppressing noise, often precipitating rapid entropy collapse. To address this, we propose **Elastic Trust Regions (ETR)**, a dynamic mechanism that aligns optimization constraints with signal quality. ETR constructs a signal-aware landscape through dual-level elasticity: at the micro level, it scales clipping boundaries based on advantage magnitude to accelerate learning from high-confidence paths; at the macro level, it leverages group variance to implicitly allocate larger update budgets to tasks in the optimal learning zone. Extensive experiments spanning mathematical reasoning, code generation, and open-ended generation across diverse model families and scales demonstrate that ETR consistently improves performance when integrated with GRPO, DAPO, and GSPO, while effectively mitigating policy entropy degradation to sustain exploration.

1 Introduction

Reinforcement Learning (RL) has become a central technique for improving the reasoning capabilities

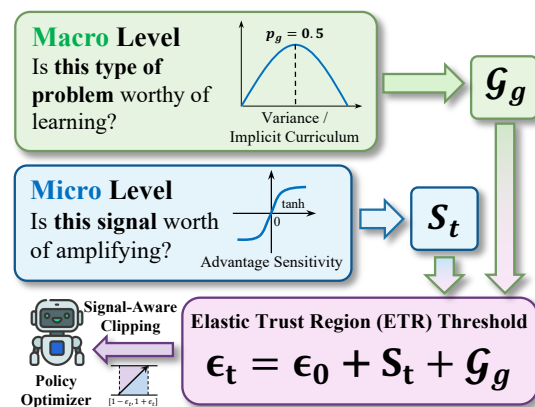


Figure 1: **Not all learning signals deserve equal trust.** ETR replaces GRPO’s fixed threshold with dynamic bounds that expand at the macro-level for valuable samples and contract at the micro-level for uncertain steps.

of Large Language Models (LLMs) (Achiam et al., 2023; Touvron et al., 2023; Grattafiori et al., 2024; Comanici et al., 2025), particularly in domains that require multi-step inference, such as mathematics, coding, and logical problem solving (Ouyang et al., 2022; Guo et al., 2025; Jaech et al., 2024; Shao et al., 2024; Zhang et al., 2025).

Early approaches primarily relied on Reinforcement Learning from Human Feedback (RLHF) (Schulman et al., 2017), where a learned reward model approximates human preferences and guides policy optimization. While effective, reward-model-based pipelines introduce additional sources of noise and biases, and often struggle to provide fine-grained supervision for tasks with objectively verifiable outcomes (Wen et al., 2026; Wang et al., 2025). To address these limitations, Reinforcement Learning with Verifiable Rewards (RLVR) has recently gained prominence (Wen et al., 2026; Zhang et al., 2026a) and is widely applied in existing reasoning models (Guo et al., 2025; Team et al., 2025). In RLVR settings, rewards are derived directly from ground-truth out-

[†]Core contributors.

comes, such as solution correctness or test case pass rates, providing precise and unambiguous optimization signals. This paradigm has proven particularly effective for mathematical reasoning and program synthesis, where correctness can be automatically verified. A popular approach, Group Relative Policy Optimization (GRPO) (Shao et al., 2024), has emerged due to its simplicity and efficiency. By computing relative advantages within sampled groups and eliminating the need for an explicit value function, GRPO significantly reduces computational overhead while retaining strong empirical performance.

Despite its empirical success, GRPO inherits PPO’s static clipping mechanism, which serves as a practical proxy for controlling the local policy update between the current policy π_θ and the roll-out policy $\pi_{\theta_{\text{old}}}$. This local clipping constraint is distinct from the KL regularization toward the reference policy π_{ref} . However, GRPO applies the same clipping threshold to all samples and groups, despite the heterogeneous nature of outcome-driven training signals. Advantage values vary substantially in magnitude and sign, while query prompts differ in difficulty and produce groups with different reward distributions. For binary outcomes, groups with pass rates near 50% attain the highest reward variance and contain a balanced mixture of successful and unsuccessful trajectories, making them particularly informative for group-relative optimization (Zhang et al., 2025). A uniform clipping threshold therefore misallocates optimization capacity: it over-constrains informative positive signals while providing insufficient protection against noisy negative updates. In practice, this mismatch can lead to inefficient learning and rapid policy entropy collapse, thereby limiting exploration and generalization.

To address this mismatch, we propose **Elastic Trust Regions (ETR)**, a dynamic constraint mechanism that adapts policy updates to observed outcome statistics. Instead of applying a uniform clipping threshold, ETR adjusts the effective trust region using both sample-level advantages and group-level reward distributions. At the micro level, it relaxes the clipping bound for positive-advantage trajectories while tightening it for negative-advantage trajectories. At the macro level, it allocates larger update budgets to groups with greater learning potential, thereby inducing an implicit curriculum over prompt difficulty.

ETR modifies only the clipping logic and re-

quires no additional networks or supervision. Because it operates directly on observed group-level rewards, it can be integrated with GRPO-style objectives and applied to binary rewards, fine-grained verifier rewards, and noisy reward-model signals.

Our contributions are summarized as follows:

1. We identify *The Static Mismatch*, a fundamental limitation of static trust regions in GRPO due to heterogeneous, outcome-driven training signals and systematically analyze their impact on policy diversity and learning efficiency.
2. We propose Elastic Trust Regions (ETR), a general dynamic constraint mechanism that adapts policy updates with both sample-level advantages and group-level variance, enabling an implicit curriculum learning over prompt difficulty.
3. Through extensive experiments spanning mathematical reasoning, code generation with verifiable rewards, and open-ended summarization with noisy rewards across model scales from 7B to 32B, we show that ETR consistently improves performance when integrated with GRPO, DAPO, and GSPO. On Qwen3-8B-Base, ETR improves four math benchmarks’ average Mean@32 and Best@32 by **4.4** and **5.7** points, respectively, while also improving OOD generalization and substantially slowing policy entropy decay.

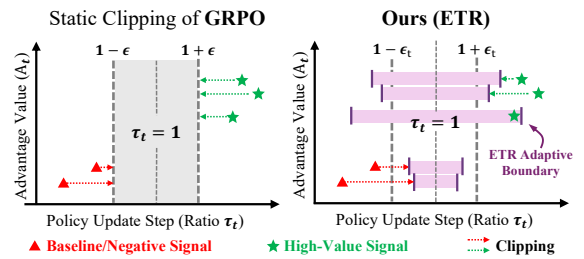


Figure 2: Comparison of optimization landscapes between GRPO and ETR. **Left (GRPO):** the static clipping range suffers from *static mismatch*, where high-value signals (green stars, large advantage) are aggressively truncated, limiting efficient learning. **Right (ETR):** our method creates a signal-aware adaptive boundary. By scaling the trust region proportional to signal strength, ETR successfully encapsulates both high-advantage samples while constraining negative signals (red triangles) to ensure stability.

2 Preliminaries

We consider the Reinforcement Learning with Verifiable Rewards (RLVR) setting for LLMs. The policy π_θ generates a response o given a query q .

Proximal Policy Optimization (PPO). PPO (Schulman et al., 2017) stabilizes training by constraining policy updates within a trust region. It employs a clipped surrogate objective to prevent destructive large updates. The objective function is defined as:

$$\mathcal{J}_{PPO}(\theta) = \mathbb{E}_t [\min(r_t(\theta)A_t, L_t(\theta)A_t)] \quad (1)$$

$$L_t(\theta) = \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon) \quad (2)$$

where $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}$ is the probability ratio. To estimate the advantage A_t , PPO typically relies on a learned value function $V_\phi(s)$ and utilizes Generalized Advantage Estimation (GAE) (Schulman et al., 2015b) to balance bias and variance.

Group Relative Policy Optimization (GRPO). GRPO (Shao et al., 2024) eliminates the need for a value network. For each query q , it samples a group of G outputs $\{o_1, \dots, o_G\}$ from the old policy $\pi_{\theta_{old}}$. The advantage A_i is computed by normalizing the intra-group rewards in G .

The GRPO objective consists of two components: i) a clipped surrogate objective that maximizes the group-relative advantage while constraining the local update from $\pi_{\theta_{old}}$, and ii) a KL penalty that regularizes the current policy toward the reference model π_{ref} .

$$\mathcal{J}_{GRPO}(\theta) = \mathbb{E} \left[\frac{1}{G} \sum_{i=1}^G \frac{1}{|o_i|} \sum_{t=1}^{|o_i|} \left(\min \left(r_{i,t} A_t, \text{clip}(r_{i,t}, 1 - \epsilon, 1 + \epsilon) A_t \right) - \beta D_{KL}(\pi_\theta || \pi_{ref}) \right) \right] \quad (3)$$

where $r_{i,t} = \frac{\pi_\theta(o_t|q, o_{<t})}{\pi_{\theta_{old}}(o_t|q, o_{<t})}$ is the probability ratio between the current and old policies, and β is the coefficient for the KL regularization term towards the reference model.

3 Methodology

In this section, we formalize the optimization landscape of reasoning tasks. We first identify the theoretical limitations of static clipping (Sec 3.1) and derive the optimal dynamic boundary from a

weighted trust region framework (Sec 3.2). Finally, we instantiate this framework into the Elastic Trust Regions (ETR) algorithm (Sec 3.3).

3.1 Theoretical Diagnosis: The Static Mismatch

To understand the limitations of current RLVR methods, we revisit the local optimization objective of trust-region methods. Under the local quadratic approximation $D_{KL}(\pi_{\theta_{old}} || \pi_\theta) \approx \frac{1}{2}(r - 1)^2$, the penalized objective for a single sample becomes $\max_r rA - \frac{\eta}{2}(r - 1)^2$, where η is the local trust-region penalty coefficient. Its optimum satisfies $r^* = 1 + A/\eta$, showing that the preferred update magnitude depends on the advantage signal.

This relationship reveals that the optimal update magnitude is inherently dependent on the signal strength A . Ideally, tokens with strong and reliable advantages should receive a larger trust-region budget, while uncertain or noisy tokens should be updated more conservatively. However, existing algorithms such as PPO and GRPO impose the same fixed constraint $|r - 1| \leq \epsilon$ on all tokens, regardless of their signal quality. Because reasoning rewards are sparse and highly heterogeneous, this uniform constraint creates a "one-size-fits-all" mismatch: high-value reasoning paths can be prematurely clipped, whereas ambiguous samples may still receive unnecessarily large updates. We formalize this structural inefficiency as the *Static Mismatch*:

Definition 3.1: The Static Mismatch

The Static Mismatch denotes the misalignment between a fixed threshold ϵ and the heterogeneous nature of training signals. **Micro-level (Magnitude):** For high-advantage samples, the theoretical optimal step exceeds the static bound ($|r^* - 1| > \epsilon$), leading to aggressive gradient truncation and the under-utilization of informative signals. **Macro-level (Variance):** The fixed threshold ignores the varying difficulty across problem groups, applying identical constraints to distributions with vastly different variance profiles and information densities.

3.2 Theoretical Framework: Signal-Aware Trust Regions

To resolve this mismatch, we propose to replace the uniform trust region assumption with a *Signal-Aware Weighted Constraint*. Instead of enforcing a uniform budget $\mathbb{E}[D_{KL}] \leq \delta$ for all data points, we

Question: Find the sum of all integers x , where $x \geq 3$, such that 201020112012_x (interpreted as a base x number) is divisible by $x - 1$.

Task Difficulty (Macro):

Pass Rate (p_g) : 0.48

Macro Adjustment (G) : +0.10

Reasoning Trace:

Step 1 to Step 4: So the answer should be 16.

Step 3: *Wait*, noted that $x \equiv 1 \pmod{x-1}$.

Therefore, for any integer $k \geq 0$, we have $x^k \equiv 1^k \equiv 1 \pmod{x-1}$.

Step 4 to Step 6: The final answer should be 32.

Outcome Metrics (Micro):

Final Reward: $R = 1.0$

Estimated Advantage (A_t) : +1.85 (High Signal)

Micro Adjustment (S) : +0.10

Applying ETR:

Token Sequence: ".", "*Wait*", ",", "*Noted*"

Policy Ratio (r_t) : 1.10, **1.38 (Peak)**, 1.25, 1.15, 1.05

Clipping Bound: 0.30, **0.40**, 0.35, 0.32, 0.28

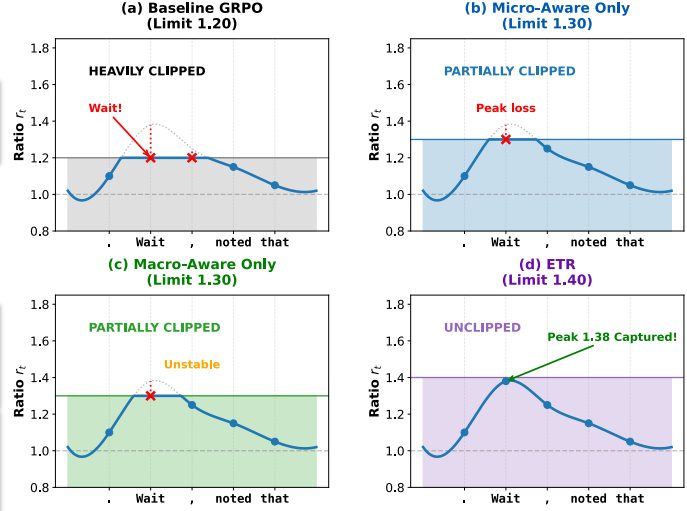


Figure 3: **A case study on a high-difficulty reasoning step.** We visualize the policy ratio trajectory for a critical step that requires a large update (≈ 1.38). (a) GRPO (limit 1.20) clips the gradient greatly, losing information. (b) & (c) Applying only Micro or Macro adjustments alleviates this issue, but still results in some clipping. (d) **Full ETR** combines both adjustments, raising the dynamic boundary to 1.40 to capture the peak learning signal. This demonstrates how ETR prevents the loss of crucial information in important sparse-reward reasoning tasks.

introduce a positive scaling factor ρ_t that controls the effective local trust-region radius of the t -th token:

$$\mathbb{E}_t \left[\frac{1}{\rho_t} D_{KL}(\pi_{\theta_{old}} || \pi_{\theta}) \right] \leq \delta \quad (4)$$

In this formulation, a larger ρ_t reduces the effective penalty weight of the KL divergence, while a smaller ρ_t tightens the local constraint.

Theorem 3.1: Optimal Dynamic Boundary

Under the signal-aware weighted constraint, the token-level clipping boundary satisfies

$$\epsilon_{\text{dynamic}}(t) = \epsilon_{\text{base}} \sqrt{\rho_t}. \quad (5)$$

The complete derivation is provided in Appendix A.

This adaptive boundary allows ETR to expand reliable positive updates and constrain noisy negative updates, balancing learning efficiency and optimization stability.

3.3 Elastic Trust Regions (ETR)

Guided by Theorem 3.1, we propose ETR. We parameterize the theoretical scaling factor through a bounded additive clipping radius to ensure numerical stability and decouple the hyperparameters. For a query group with rewards $\{R_i\}_{i=1}^G$, we define the pass rate as $p_g = \frac{1}{G} \sum_{i=1}^G \mathbb{I}(R_i > 0)$. Each token in response o_i uses the group-normalized advantage

A_i , denoted as A_t when discussing token-level clipping. This bounded additive parameterization lets ETR compute a token-specific trust-region budget using only statistics already available in standard group-based RLVR training.

Algorithm Formulation: Dynamic Threshold ϵ_t

We define the dynamic clipping threshold ϵ_t as the base threshold adjusted by micro-level and macro-level elasticity:

$$\epsilon_t = \epsilon_{\text{base}} + \underbrace{\mathcal{S}(A_t)}_{\text{Micro}} + \underbrace{\mathcal{G}(p_g)}_{\text{Macro}} \quad (6)$$

where $\mathcal{S}(A_t)$ accounts for sample-level signal strength, and $\mathcal{G}(p_g)$ accounts for group-level learnability.

This additive form is an exact instantiation of Theorem 3.1 under the definition

$$\rho_t = \left(\frac{\epsilon_{\text{base}} + \mathcal{S}(A_t) + \mathcal{G}(p_g)}{\epsilon_{\text{base}}} \right)^2. \quad (7)$$

Because $\mathcal{S}(A_t) \in [-\lambda_1, \lambda_1]$, $\mathcal{G}(p_g) \in [0, \lambda_2]$, and all our settings satisfy $\epsilon_{\text{base}} > \lambda_1$, the numerator is strictly positive. Substituting Eq. 7 into $\epsilon_{\text{dynamic}}(t) = \epsilon_{\text{base}} \sqrt{\rho_t}$ recovers Eq. 6. Thus, the implementation is not a heuristic departure from the multiplicative theory; it is a bounded parameterization of the same dynamic radius.

Micro-Level Adjustment (S). We define the sample-level term to adapt to both the magnitude and sign of the advantage. For positive-advantage trajectories, increasing their probability is a focused update, so ETR expands the clipping boundary to better preserve informative signals. For negative-advantage trajectories, decreasing the probability of sampled tokens redistributes probability mass across the vocabulary and can introduce diffuse gradient noise; ETR therefore tightens the boundary. We use a hyperbolic tangent to keep this asymmetric adjustment bounded:

$$S(A_t) = \lambda_1 \cdot \tanh(A_t) \quad (8)$$

Macro-Level Adjustment ($\mathcal{G}$). We define the group-level term using the variance of binarized success outcomes. For a group with $p_g = \frac{1}{G} \sum_{i=1}^G \mathbb{I}(R_i > 0)$, the corresponding Bernoulli variance is $p_g(1 - p_g)$.

$$\mathcal{G}(p_g) = \lambda_2 \cdot 4p_g(1 - p_g) \quad (9)$$

Insight: Implicit Curriculum Learning

The macro-level term $\mathcal{G}(p_g)$ is non-negative, reaches its maximum when $p_g \approx 0.5$ (maximum variance), and approaches zero when $p_g \approx 0$ or 1. This introduces an **Implicit Curriculum Learning** mechanism: the algorithm allocates extra update budget to tasks within the “optimal learning zone” (high variance), while keeping very hard or very easy groups close to the base trust-region radius rather than actively expanding them.

Final Objective. Integrating the dynamic boundary into the GRPO framework, ETR maximizes the following clipped objective:

$$\mathcal{J}_{ETR}(\theta) = \mathbb{E} \left[\frac{1}{G} \sum_{i=1}^G \frac{1}{|o_i|} \sum_{t=1}^{|o_i|} \left(\min \left(r_{i,t} A_i, \text{clip}(r_{i,t}, 1 - \epsilon_{i,t}, 1 + \epsilon_{i,t}) A_i \right) - \beta D_{KL}(\pi_\theta || \pi_{ref}) \right) \right] \quad (10)$$

Through this mechanism, ETR dynamically aligns the optimization constraints with the statistical properties of the data.

4 Experiments

In this section, we evaluate ETR on mathematical reasoning, code generation, and open-ended

summarization. We further examine its compatibility with modern RL objectives and larger model scales, analyze its training dynamics and policy exploration, and ablate each design component.

4.1 Experimental Setup

Benchmarks. We evaluate ETR across three domains with increasingly noisy reward signals. For mathematical reasoning, models are trained on DAPO-Math-17k (Yu et al., 2025) and evaluated with **MATH500** (Hendrycks et al., 2021b), **AMC23** (Li et al., 2024), **AIME 2024**, and **AIME 2025** as in-distribution benchmarks, reporting **Mean@32** and **Best@32**. MATH500 and AMC23 represent standard competition-level problems, while AIME24 and AIME25 contain more difficult problems with sparse correct reasoning paths. OOD generalization is measured with **Pass@1** on **GPQA-Main**, **GPQA-Diamond** (Rein et al., 2024), and **ACPbench** (Kokel et al., 2025).

For code generation, models are trained on the 18K-problem training subset derived from PRIME-RL (Cui et al., 2025) and evaluated on **TACO** (Li et al., 2023), **Codeforces** (MatrixStudio, 2024), **CodeContests** (Li et al., 2022), and **APPS** (Hendrycks et al., 2021a), using **Mean@4** and **Best@4**. A solution passing m of n unit tests receives reward m/n , providing a denser signal. Finally, we consider query-based summarization with keyphrase highlighting, using a reward combining Skywork-Reward-V2 (Liu et al., 2026) with rule-based length and faithfulness constraints. The adapted reward model achieves approximately 74% pairwise accuracy on a held-out human-preference set. We conduct blind human evaluation on 660 samples and report perfect responses (**Score 2**) and severe errors (**Score 0**).

Baselines and Implementation. For math and code, we report the **Base Model** and compare **GRPO** with a static clipping threshold of $\epsilon = 0.2$, **GRPO-Clip-High** with $\epsilon = 0.28$, and **GRPO + ETR**. Clip-High, inspired by DAPO (Yu et al., 2025), controls for whether ETR’s gains can be reproduced by simply relaxing the static trust region. The summarization experiment directly compares GRPO and GRPO + ETR.

We conduct experiments using the verl reinforcement learning framework, with FSDP for training and vLLM for inference. Math reasoning experiments use 8×H20 GPUs and cover **Qwen3-8B-Base** (Yang et al., 2025), **Llama-**

3.1-8B-Instruct (Grattafiori et al., 2024), and **Qwen2.5-7B-Math** (Yang et al., 2024). Code experiments use **Qwen2.5-7B-Instruct**, **Llama-3.1-8B-Instruct**, and **Qwen3-8B-Base** on $8 \times A100$ GPUs, with checkpoints at steps 250, 160, and 150, respectively. Summarization experiments use **Qwen3-8B**.

ETR modifies only the clipping logic in the loss function. Unless otherwise specified, we use a rollout group size $G = 8$ and a sampling temperature of 1.0. The math experiments use binary rewards, whereas the code and summarization experiments use fine-grained verifier rewards and reward-model signals. We additionally evaluate compatibility with DAPO and GSPO and scaling to 30B+ models with proportionally adjusted coefficients to match their tighter base trust regions. Further training details are provided in the appendix.

4.2 Main Results

Table 1 presents the performance comparison across all mathematical reasoning benchmarks. ETR achieves the highest average in-distribution Mean@32 and Best@32 within every model family, while also improving average OOD Pass@1 over GRPO.

Performance on Hard Tasks. ETR produces pronounced gains on difficult AIME settings, where correct reasoning paths are sparse. On Qwen3-8B-Base, ETR raises AIME24 Best@32 from 34.4 to 48.2, an absolute improvement of **13.8 points** over GRPO. On AIME25, it improves Best@32 by **8.1 points** for both Llama-3.1-8B-Instruct (12.2 to 20.3) and Qwen2.5-7B-Math (29.5 to 37.6). These results support our hypothesis that elastic clipping better preserves sparse, high-advantage trajectories that may otherwise be truncated by a uniform constraint.

Out-of-Distribution Generalization. Relative to GRPO, ETR improves average Pass@1 across GPQA-Main, GPQA-Diamond, and ACPBench by **4.3**, **2.5**, and **1.2 points** for Qwen3-8B-Base, Llama-3.1-8B-Instruct, and Qwen2.5-7B-Math, respectively. These results are consistent with the view that tightening constraints on negative updates while relaxing them for high-advantage trajectories helps preserve policy diversity.

4.3 Compatibility and Scaling

Because ETR only modifies the clipping boundary, it can be combined with stronger RL training

recipes rather than replacing them. To test this compatibility, we integrate ETR into a DAPO-style setup by applying the same dynamic clipping rule to DAPO’s weighted samples. Due to the substantially higher cost of DAPO training, we report results at step 100 on Qwen3-8B-Base.

We further evaluate ETR at larger model scales. The first setting applies ETR to GRPO on Qwen2.5-32B-Instruct. The second uses Qwen3-30B-A3B-Base with a tighter GSPO sequence-level trust region ($\epsilon_{base} = 4 \times 10^{-4}$), for which we proportionally set $\lambda_1 = \lambda_2 = 1 \times 10^{-4}$.

Table 2(a) shows that ETR improves DAPO-style training across all three math benchmarks, with the largest gain on AIME25 Best@32 (**+10.3 points**) over DAPO + Clip-High. Table 2(b) shows that the gains persist at larger scales: ETR improves AIME25 Best@32 by **4.3 points** on Qwen2.5-32B-Instruct and by **11.5 points** on Qwen3-30B-A3B-Base. These results indicate that ETR provides a compatible trust-region mechanism that transfers across GRPO, DAPO, and GSPO.

4.4 Generalization to Code Generation

To examine whether ETR generalizes beyond binary mathematical rewards, we evaluate it with fine-grained code-verification rewards. Results are reported in Table 3.

ETR consistently improves code generation performance over GRPO across all three model families. In average Best@4, ETR improves over GRPO by **5.0 points** on Qwen2.5-7B-Instruct, **2.2 points** on Llama-3.1-8B-Instruct, and **3.4 points** on Qwen3-8B-Base. ETR also outperforms Clip-High on every model, showing that the benefit does not simply arise from using a larger static trust region. These results indicate that signal-adaptive clipping transfers to structured non-mathematical tasks with fine-grained verifier rewards.

4.5 Generalization to Open-Ended Generation

We further test ETR under noisy reward-model supervision on open-ended summarization. Table 4 reports the blind human-evaluation results.

ETR increases the proportion of perfect responses from **76.0%** to **78.0%**, while reducing severe errors from **6.4%** to **5.8%**. Its benefits thus extend beyond objectively verifiable reasoning and coding tasks to open-ended generation with imperfect reward signals.

Table 1: Main results on in-distribution and out-of-distribution mathematical reasoning benchmarks. In-distribution metrics are reported as **Mean@32 / Best@32**. Highlighted rows indicate ETR. **Bold** denotes the best result within each model family.

Method	In-Distribution (Math Reasoning) (Mean@32 / Best@32)					Out-of-Distribution (Pass@1)		
	AMC23	AIME24	AIME25	MATH500	Avg.	GPQA-M	GPQA-D	ACP
<i>Model: Qwen3-8B-Base</i>								
Base Model	0.317 / 0.850	0.016 / 0.166	0.037 / 0.289	0.378 / 0.892	0.187 / 0.549	0.301	0.283	0.515
GRPO	0.713 / 0.949	0.160 / 0.344	0.228 / 0.497	0.782 / 0.890	0.471 / 0.670	0.353	0.384	0.562
GRPO + Clip-High	0.734 / 0.950	0.185 / 0.469	0.231 / 0.448	0.799 / 0.916	0.487 / 0.696	0.346	0.354	0.567
GRPO + ETR	0.790 / 0.967	0.195 / 0.482	0.254 / 0.532	0.819 / 0.926	0.515 / 0.727	0.393	0.444	0.590
<i>Model: Llama-3.1-8B-Instruct</i>								
Base Model	0.070 / 0.486	0.029 / 0.168	0.003 / 0.063	0.059 / 0.403	0.040 / 0.280	0.281	0.227	0.464
GRPO	0.161 / 0.538	0.051 / 0.165	0.007 / 0.122	0.073 / 0.336	0.073 / 0.290	0.324	0.333	0.478
GRPO + Clip-High	0.170 / 0.640	0.038 / 0.186	0.008 / 0.157	0.088 / 0.449	0.076 / 0.358	0.326	0.313	0.485
GRPO + ETR	0.178 / 0.693	0.049 / 0.202	0.016 / 0.203	0.203 / 0.604	0.112 / 0.426	0.339	0.354	0.516
<i>Model: Qwen2.5-7B-Math-Base</i>								
Base Model	0.233 / 0.797	0.002 / 0.042	0.000 / 0.000	0.349 / 0.870	0.146 / 0.427	0.250	0.253	0.428
GRPO	0.442 / 0.829	0.204 / 0.360	0.119 / 0.295	0.586 / 0.885	0.338 / 0.592	0.275	0.278	0.437
GRPO + Clip-High	0.605 / 0.919	0.199 / 0.316	0.168 / 0.370	0.677 / 0.914	0.412 / 0.630	0.266	0.268	0.443
GRPO + ETR	0.622 / 0.925	0.215 / 0.358	0.141 / 0.376	0.717 / 0.911	0.424 / 0.643	0.257	0.323	0.445

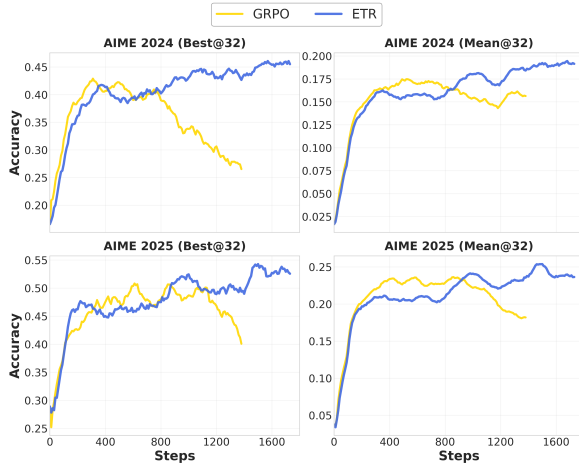


Figure 4: Validation accuracy on AIME 2024 and AIME 2025 using Qwen3-8B-Base. Standard GRPO (yellow) exhibits late-stage degradation, whereas ETR (blue) maintains stronger performance without the same sharp decline.

4.6 Analysis of Training Process

We analyze the evolution of validation accuracy and policy entropy to better understand the mechanism behind the performance gains.

Mitigating Late-Stage Collapse. Figure 4 presents validation performance on AIME 2024 and AIME 2025. Standard GRPO becomes unsta-

ble in the later stages of training: after reaching a peak, Best@32 begins to decline as policy entropy approaches zero. Under Best@32 evaluation, a near-deterministic policy produces less diverse responses; convergence to a suboptimal path therefore reduces the chance of sampling a correct alternative. In contrast, ETR avoids the same sharp late-stage degradation and maintains stronger performance. Together with the entropy curves, this pattern indicates that the dynamic boundary stabilizes policy updates while preserving exploration.

Entropy Evolution. Figure 5 depicts the evolution of policy entropy.

GRPO: Entropy drops rapidly to near zero, indicating premature convergence toward a deterministic policy.

GRPO-Clip-High: Increasing the static threshold to 0.28 prevents entropy collapse but produces severe oscillations and a continued rise in entropy during later stages. This suggests that an overly permissive static boundary introduces noisy updates that do not translate into performance gains.

ETR: ETR exhibits a dip-and-rebound pattern. Entropy initially decreases as the model learns and then stabilizes or recovers during later stages. This behavior is consistent with the elastic boundary allowing the model to expand its search space when

Table 2: Compatibility and scaling results, reported as Mean@32 / Best@32. **(a)** DAPO-style training on Qwen3-8B-Base at step 100. **(b)** Scaling to Qwen2.5-32B-Instruct with GRPO at step 200 and Qwen3-30B-A3B-Base with GSPO at step 150.

(a) Compatibility with DAPO				
Method	AMC23	AIME24	AIME25	Avg.
Base Model	0.317 / 0.850	0.016 / 0.166	0.037 / 0.289	0.123 / 0.435
DAPO + Clip-High	0.642 / 0.941	0.163 / 0.383	0.204 / 0.387	0.336 / 0.570
DAPO + ETR	0.665 / 0.962	0.165 / 0.414	0.217 / 0.490	0.349 / 0.622

(b) Scaling to 30B+ Models			
Method	AMC23	AIME24	AIME25
<i>Qwen2.5-32B-Instruct (GRPO)</i>			
Base Model	0.650 / 0.921	0.081 / 0.212	0.101 / 0.344
GRPO	0.693 / 0.920	0.106 / 0.205	0.151 / 0.364
GRPO + ETR	0.695 / 0.930	0.122 / 0.243	0.180 / 0.407
<i>Qwen3-30B-A3B-Base (GSPO)</i>			
Base Model	0.450 / 0.924	0.011 / 0.115	0.009 / 0.112
GSPO	0.523 / 0.917	0.136 / 0.304	0.167 / 0.396
GSPO + ETR	0.578 / 0.919	0.152 / 0.348	0.190 / 0.511

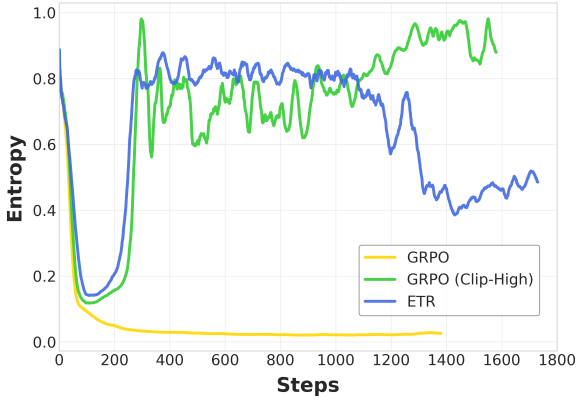


Figure 5: Policy entropy during training. GRPO approaches zero entropy, Clip-High produces high and unstable entropy, and ETR maintains a more stable intermediate level.

high-value signals are encountered, thereby balancing exploitation and exploration.

4.7 Ablation Study

We ablate the micro- and macro-level components of ETR and evaluate the direction of its asymmetric adjustment.

Component Effectiveness. Removing the micro-level adjustment ($\lambda_1 = 0$) causes the largest performance drop and slows convergence, confirming that advantage-aware clipping is the primary driver of ETR. Removing the macro-level adjustment ($\lambda_2 = 0$) also reduces training efficiency, supporting the benefit of allocating larger update bud-

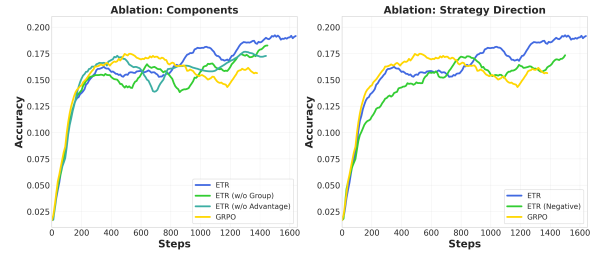


Figure 6: Ablation results. **Left:** contributions of the micro- and macro-level components. **Right:** comparison with the inverse adjustment direction.

gets to high-variance, moderately difficult groups.

Strategy Direction. The Inverse Variant, which relaxes the boundary for negative samples and tightens it for positive samples, performs substantially worse. This result supports our asymmetric design: positive updates benefit from a relaxed boundary, whereas negative updates require a tighter boundary to avoid indiscriminate probability redistribution over the vocabulary.

4.8 Computational Efficiency

ETR adds negligible computational overhead relative to GRPO. It applies only element-wise operations to advantages and group statistics that are already computed during rollout and advantage estimation, requiring no auxiliary model, additional rollout, or offline preprocessing.

Table 3: Results on code generation benchmarks, reported as Mean@4 / Best@4. **Bold** denotes the best result within each model family.

Method	TACO	Codeforces	CodeContests	APPS	Avg.
<i>Mean@4 / Best@4</i>					
<i>Model: Qwen2.5-7B-Instruct</i>					
Base Model	0.133 / 0.209	0.132 / 0.252	0.270 / 0.405	0.276 / 0.404	0.203 / 0.318
GRPO	0.135 / 0.164	0.302 / 0.354	0.266 / 0.312	0.293 / 0.341	0.249 / 0.293
GRPO + Clip-High	0.134 / 0.179	0.322 / 0.394	0.256 / 0.320	0.260 / 0.327	0.243 / 0.305
GRPO + ETR	0.168 / 0.215	0.345 / 0.436	0.280 / 0.349	0.313 / 0.373	0.277 / 0.343
<i>Model: Llama-3.1-8B-Instruct</i>					
Base Model	0.105 / 0.181	0.143 / 0.276	0.179 / 0.331	0.216 / 0.344	0.161 / 0.283
GRPO	0.198 / 0.243	0.378 / 0.469	0.334 / 0.401	0.332 / 0.394	0.311 / 0.377
GRPO + Clip-High	0.131 / 0.173	0.288 / 0.375	0.204 / 0.260	0.222 / 0.287	0.211 / 0.274
GRPO + ETR	0.209 / 0.264	0.403 / 0.499	0.336 / 0.422	0.356 / 0.412	0.326 / 0.399
<i>Model: Qwen3-8B-Base</i>					
Base Model	0.089 / 0.158	0.163 / 0.295	0.149 / 0.277	0.175 / 0.322	0.144 / 0.263
GRPO	0.155 / 0.192	0.293 / 0.372	0.246 / 0.306	0.255 / 0.316	0.237 / 0.297
GRPO + Clip-High	0.148 / 0.192	0.303 / 0.383	0.250 / 0.318	0.277 / 0.352	0.245 / 0.311
GRPO + ETR	0.177 / 0.229	0.309 / 0.403	0.263 / 0.334	0.282 / 0.358	0.258 / 0.331

Table 4: Blind human evaluation on the open-ended summarization task (660 samples). Higher is better for Score 2, while lower is better for Score 0.

Method	Score 2 (Perfect)	Score 0 (Severe Error)
GRPO	76.0%	6.4%
GRPO + ETR	78.0%	5.8%

4.9 Hyperparameter Robustness

Appendix Table 6 shows that halving either elasticity coefficient yields performance close to the default setting on Qwen3-8B-Base. The best scores are distributed across different configurations, indicating that ETR is not sensitive to a narrowly tuned coefficient pair. We therefore use the balanced setting $\lambda_1 = \lambda_2 = 0.10$ as the default in the main experiments.

5 Related Works

Reinforcement learning is widely used for LLM post-training, from RLHF with learned reward models (Ouyang et al., 2022; Lambert, 2025; Dong et al., 2024) to RLVR with verifiable rewards (Wen et al., 2026; Shao et al., 2024; Yu et al., 2025). Beyond mathematical reasoning, AFRL combines SFT, Stepwise-GRPO, and curriculum learning for search relevance (Zhang et al., 2026b), while ENVIS constructs verified supervision through environment-native search for long-

horizon GUI agents (Zhou et al., 2026).

Trust-region methods such as TRPO (Schulman et al., 2015a) and PPO (Schulman et al., 2017) stabilize policy updates, but standard clipping applies a uniform range to heterogeneous learning signals. Prior work has explored adaptive trust-region mechanisms (Wang et al., 2019; Su et al., 2025) and asymmetric clipping (Yu et al., 2025); ETR instead adapts the clipping radius using group-relative advantages and online outcome variance. ETR is also related to curriculum learning and adaptive optimization (Hanzely, 2023; Wang et al., 2024; Bengio et al., 2009; Lin et al., 2025; Okamoto et al., 2021; Song et al., 2025; Shi et al., 2025; Zhang et al., 2025), but derives its implicit curriculum directly from statistics already computed during GRPO training.

6 Conclusion

We identify the mismatch between static trust regions and heterogeneous outcome-driven signals and propose ETR, which adapts clipping boundaries using group-relative advantage magnitude and group outcome variance. Across mathematical reasoning, code generation, open-ended summarization, multiple policy optimization algorithms, and 30B+ models, ETR consistently improves performance and mitigates entropy collapse with negligible overhead, demonstrating the effectiveness of signal-aware trust regions for RLVR post-training.

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48.
- Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Naveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.
- Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Yuchen Zhang, Jiacheng Chen, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, et al. 2025. Process reinforcement through implicit rewards. *arXiv preprint arXiv:2502.01456*.
- Hanze Dong, Wei Xiong, Bo Pang, Haoxiang Wang, Han Zhao, Yingbo Zhou, Nan Jiang, Doyen Sahoo, Caiming Xiong, and Tong Zhang. 2024. Rlhf workflow: From reward modeling to online rlhf. *arXiv preprint arXiv:2405.07863*.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
- Slavomír Hanzely. 2023. Adaptive optimization algorithms for machine learning. *arXiv preprint arXiv:2311.10203*.
- Dan Hendrycks, Steven Basart, Saurav Kadavath, Mantas Mazeika, Akul Arora, Ethan Guo, Collin Burns, Samir Puranik, Horace He, Dawn Song, et al. 2021a. Measuring coding challenge competence with apps. *arXiv preprint arXiv:2105.09938*.
- Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021b. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.
- Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.
- Harsha Kokel, Michael Katz, Kavitha Srinivas, and Shirin Sohrabi. 2025. Acpbench: Reasoning about action, change, and planning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 26559–26568.
- Nathan Lambert. 2025. Reinforcement learning from human feedback. *arXiv preprint arXiv:2504.12501*.
- Jia Li, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Costa Huang, Kashif Rasul, Longhui Yu, Albert Jiang, Ziju Shen, Zihan Qin, Bin Dong, Li Zhou, Yann Fleureau, Guillaume Lample, and Stanislas Polu. 2024. [Numinamath](#). Hugging Face repository.
- Rongao Li, Jie Fu, Bo-Wen Zhang, Tao Huang, Zhihong Sun, Chen Lyu, Guang Liu, Zhi Jin, and Ge Li. 2023. [TACO: Topics in algorithmic COde generation dataset](#). *arXiv preprint arXiv:2312.14852*.
- Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, et al. 2022. Competition-level code generation with alphacode. *Science*, 378(6624):1092–1097.
- Siyu Lin, Qingwei Mi, and Tianhan Gao. 2025. [A survey of curriculum learning in deep reinforcement learning](#). In *2025 IEEE 15th Annual Computing and Communication Workshop and Conference (CCWC)*, pages 01141–01147.
- Yuhao Liu, Liang Zeng, Yuzhen Xiao, Jujie He, Jiacai Liu, Chaojie Wang, Rui Yan, Wei Shen, Fuxiang Zhang, Jiacheng Xu, et al. 2026. Skywork-reward-v2: Scaling preference data curation via human-ai synergy. In *International Conference on Learning Representations*, volume 2026, pages 133805–133838.
- MatrixStudio. 2024. [Codeforces python submissions](#). Hugging Face dataset.
- Wataru Okamoto, Hiroshi Kera, and Kazuhiko Kawamoto. 2021. Reinforcement learning with adaptive curriculum dynamics randomization for fault-tolerant robot control. *arXiv preprint arXiv:2111.10005*.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.
- David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. 2024. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.
- John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. 2015a. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897. Pmlr.

- John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. 2015b. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.
- 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 *Proceedings of the Twentieth European Conference on Computer Systems*, pages 1279–1297.
- Taiwei Shi, Yiyang Wu, Linxin Song, Tianyi Zhou, and Jieyu Zhao. 2025. Efficient reinforcement finetuning via adaptive curriculum learning. *arXiv preprint arXiv:2504.05520*.
- Mingyang Song, Mao Zheng, Zheng Li, Wenjie Yang, Xuan Luo, Yue Pan, and Feng Zhang. 2025. Fastcurl: Curriculum reinforcement learning with stage-wise context scaling for efficient training rl-like reasoning models. *arXiv preprint arXiv:2503.17287*.
- Mingyu Su, Jian Guan, Yuxian Gu, Minlie Huang, and Hongning Wang. 2025. [Trust-region adaptive policy optimization](#). *Preprint*, arXiv:2512.17636.
- Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. 2025. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Qi Wang, Feng Su, Shipeng Dai, Xiaojun Lu, and Yang Liu. 2024. Adagc: A novel adaptive optimization algorithm with gradient bias correction. *Expert Systems with Applications*, 256:124956.
- Yiping Wang, Qing Yang, Zhiyuan Zeng, Liliang Ren, Liyuan Liu, Baolin Peng, Hao Cheng, Xuehai He, Kuan Wang, Jianfeng Gao, et al. 2025. Reinforcement learning for reasoning in large language models with one training example. *arXiv preprint arXiv:2504.20571*.
- Yuhui Wang, Hao He, Xiaoyang Tan, and Yaozhong Gan. 2019. Trust region-guided proximal policy optimization. *Advances in Neural Information Processing Systems*, 32.
- Xumeng Wen, Zihan Liu, Shun Zheng, Shengyu Ye, Zhirong Wu, Yang Wang, Zhijian Xu, Xiao Liang, Junjie Li, Ziming Miao, Jiang Bian, and Mao Yang. 2026. Reinforcement learning with verifiable rewards implicitly incentivizes correct reasoning in base llms. In *International Conference on Learning Representations*.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.
- An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, et al. 2024. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*.
- Qiyang Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. 2025. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*.
- Shenao Zhang, Yaqing Wang, Yinxiao Liu, Tianqi Liu, Peter Grabowski, Eugene Ie, Zhaoran Wang, and Yunxuan Li. 2026a. Beyond markovian: Reflective exploration via bayes-adaptive rl for llm reasoning. In *International Conference on Learning Representations*, volume 2026, pages 120275–120294.
- Shijie Zhang, Xiang Guo, Rujun Guo, Shaoyu Liu, Xiaozhao Wang, Guanjin Jiang, and Kevin Zhang. 2026b. Answer first, reason later: Aligning search relevance via mode-balanced reinforcement learning. *arXiv preprint arXiv:2602.10006*.
- Shijie Zhang, Zheng Xiao, Shiyu Liu, Guohao Sun, Kevin Zhang, Xiang Guo, Rujun Guo, Shaoyu Liu, Wangxiao Zhao, and Guanjin Jiang. 2025. Clpo: Curriculum learning meets policy optimization for llm reasoning. *arXiv preprint arXiv:2509.25004*.
- Yincheng Zhou, Athena Zhuoming Zhong, Shijie Zhang, Kevin Zhang, Teresa Xiaotao Shang, and Shanghang Zhang. 2026. Envs: Environment-native verified search for long-horizon gui agents. *arXiv preprint arXiv:2606.22948*.

A Theoretical Derivation

In this section, we provide the complete mathematical derivation of the Elastic Trust Regions (ETR) framework. We detail the transition from the theoretical weighted constraint to the practical dynamic clipping algorithm.

A.1 Problem Formulation

We consider the standard reinforcement learning objective with a trust region constraint. Let π_θ denote the current policy and $\pi_{\theta_{old}}$ denote the behavior policy. The probability ratio is defined as $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}$.

Standard Trust Region Policy Optimization (TRPO) imposes a hard constraint on the KL divergence. To address the limitation of uniform risk, we introduce the **Signal-Aware Weighted Constraint**.

Given the space constraints of the column, we formulate the optimization problem as follows:

$$\begin{aligned} \max_{\theta} \quad & \mathbb{E}_t[r_t(\theta)A_t] \\ \text{s.t.} \quad & \mathbb{E}_t \left[\frac{1}{\rho(A_t, p_g)} D_{KL}(\pi_{\theta_{old}} || \pi_\theta) \right] \leq \delta \end{aligned} \quad (11)$$

where $\rho(\cdot) > 0$ is a scaling factor. A larger ρ reduces the effective cost of the KL divergence, allowing for larger policy updates when signal quality is high, while a smaller ρ tightens the constraint for low-confidence or noisy updates.

A.2 Derivation of the Dynamic Boundary

Since optimizing the expectation constraint directly is computationally expensive, we convert the global constraint into a local constraint using a second-order approximation.

Taylor Expansion of KL Divergence. The KL divergence between the old and new policy can be expressed as the expectation of the negative log-ratio:

$$\begin{aligned} D_{KL}(\pi_{\theta_{old}} || \pi_\theta) &= \mathbb{E}_{x \sim \pi_{\theta_{old}}} \left[-\log \frac{\pi_\theta(x)}{\pi_{\theta_{old}}(x)} \right] \\ &= \mathbb{E}_{x \sim \pi_{\theta_{old}}} [-\log r_t] \end{aligned} \quad (12)$$

We perform a second-order Taylor expansion of the function $f(r) = -\log r$ around the point $r = 1$ (since $\pi_\theta \approx \pi_{\theta_{old}}$ locally). The derivatives are $f'(1) = -1$ and $f''(1) = 1$. The expansion is:

$$-\log r \approx -(r - 1) + \frac{1}{2}(r - 1)^2 \quad (13)$$

Substituting this back into the expectation:

$$\begin{aligned} D_{KL} &\approx \mathbb{E}_{x \sim \pi_{\theta_{old}}} [-(r_t - 1)] \\ &\quad + \mathbb{E}_{x \sim \pi_{\theta_{old}}} \left[\frac{1}{2}(r_t - 1)^2 \right] \end{aligned} \quad (14)$$

Crucially, the first-order term vanishes because the expectation of the probability ratio is 1:

$$\mathbb{E}[r_t - 1] = \mathbb{E}[r_t] - 1 = \int \pi_{old} \frac{\pi}{\pi_{old}} - 1 = 1 - 1 = 0 \quad (15)$$

Thus, the KL divergence is dominated by the second-order term:

$$D_{KL}(\pi_{\theta_{old}} || \pi_\theta) \approx \frac{1}{2}(r_t - 1)^2 \quad (16)$$

Solving for the Boundary. We now substitute Eq. 16 into the local version of the weighted constraint (Eq. 11). For a specific sample t , let $\rho_t = \rho(A_t, p_g)$. We require:

$$\frac{1}{\rho_t} \cdot \frac{1}{2}(r_t - 1)^2 \leq \delta_{local} \quad (17)$$

Rearranging to solve for the maximum allowable deviation $|r_t - 1|$:

$$(r_t - 1)^2 \leq 2\delta_{local} \cdot \rho_t \implies |r_t - 1| \leq \sqrt{2\delta_{local}} \sqrt{\rho_t} \quad (18)$$

Defining the baseline clipping threshold as $\epsilon_{base} = \sqrt{2\delta_{local}}$, we obtain the dynamic threshold ϵ_t :

$$\epsilon_t = \epsilon_{base} \cdot \sqrt{\rho(A_t, p_g)} \quad (19)$$

This derivation theoretically justifies why the clipping range should scale with the square root of the signal-dependent factor.

A.3 Connection to the Additive Implementation

The implementation in Eq. 6 uses an additive dynamic threshold for numerical stability:

$$\epsilon_t = \epsilon_{base} + \mathcal{S}(A_t) + \mathcal{G}(p_g). \quad (20)$$

This is not a heuristic approximation to the multiplicative result. It is obtained by defining the signal-dependent scaling factor as

$$\rho(A_t, p_g) = \left(\frac{\epsilon_{base} + \mathcal{S}(A_t) + \mathcal{G}(p_g)}{\epsilon_{base}} \right)^2. \quad (21)$$

Substituting this definition into $\epsilon_t = \epsilon_{base} \sqrt{\rho(A_t, p_g)}$ recovers the exact additive

threshold above. The practical benefit of this parameterization is that $\mathcal{S}(A_t) = \lambda_1 \tanh(A_t)$ is bounded, preventing outlier advantages from causing trust-region explosion, while $\mathcal{G}(p_g)$ can be tuned independently to capture group-level learnability.

B Limitations

Despite the consistent improvements across the evaluated settings, two limitations remain. First, ETR has not yet been evaluated on models larger than 70B, and its scaling behavior at such scales requires further study. Second, our experiments do not cover a broader range of tasks, particularly agentic tasks, where rewards may be more heterogeneous. We leave these extensions for future work. ETR may improve the general reasoning capabilities of language models and could therefore amplify both beneficial and harmful downstream uses. Moreover, optimization against imperfect reward signals may reinforce reward-model biases or unintended reward-hacking behavior. Domain-specific safety evaluation and human oversight are therefore necessary before deployment.

C Usage of Large Language Models

During the writing of this manuscript, we use Large Language Models (LLMs) as a writing assistant. The usage of LLMs was for improving the fluency, clarity, and grammatical correctness of the language, such as rephrasing sentences or correcting grammatical errors. LLMs were not involved in the core research ideation or the formulation of key conclusions presented in this paper.

D Algorithm Pseudocode

Algorithm 1 outlines the complete training process of GRPO with Elastic Trust Regions. The core innovation lies in lines 8-11, where the clipping threshold is dynamically adjusted based on signal quality.

E Implementation Details

E.1 Experimental Setup

Our experiments are conducted using the ver1 framework (Sheng et al., 2025) on a computational node equipped with $8 \times$ NVIDIA H20 GPUs. We utilize the DAPO-Math-17k dataset for the math reasoning experiments and the 18K code subset of PRIME-RL for code generation. The prompt

Algorithm 1 Group Relative Policy Optimization with ETR

```

1: Input: Dataset  $\mathcal{D}$ , policy  $\pi_\theta$ , ref policy  $\pi_{\text{ref}}$ , group size  $G$ .
2: Hyperparams: Base clip  $\epsilon_{\text{base}}$ , weights  $(\lambda_1, \lambda_2)$ , KL coef  $\beta$ .
3: for each training step do
4:   Sample query  $q \sim \mathcal{D}$ .
5:   Generate outputs  $O = \{o_1, \dots, o_G\}$  via  $\pi_\theta(\cdot|q)$ .
6:   Compute rewards  $\mathbf{r} = \{r_1, \dots, r_G\}$ .
7:   // 1. Compute Group Statistics (Macro)
8:    $p_g \leftarrow \frac{1}{G} \sum_{i=1}^G \mathbb{I}(r_i > 0)$ .
9:    $\mathcal{G}_{\text{macro}} \leftarrow \lambda_2 \cdot 4p_g(1 - p_g)$ .
10:  // 2. Compute Advantages (Standard GRPO)
11:   $A_i \leftarrow (r_i - \text{mean}(\mathbf{r})) / (\text{std}(\mathbf{r}) + \xi)$  for all  $i$ .
12:  // 3. Compute Dynamic Thresholds (Micro)
13:  Initialize effective clips  $\mathcal{E} \leftarrow \emptyset$ .
14:  for  $i = 1$  to  $G$  do
15:     $\mathcal{S}_{\text{micro}} \leftarrow \lambda_1 \cdot \tanh(A_i)$ .
16:     $\epsilon_i \leftarrow \epsilon_{\text{base}} + \mathcal{S}_{\text{micro}} + \mathcal{G}_{\text{macro}}$ .
17:     $\mathcal{E} \leftarrow \mathcal{E} \cup \{\epsilon_i\}$ .
18:  end for
19:  // 4. Policy Update
20:  Compute loss  $\mathcal{J}_{\text{ETR}}$  using dynamic bounds  $\mathcal{E}$  (Eq. 6).
21:  Update  $\pi_\theta$  via gradient descent  $\nabla_\theta \mathcal{J}_{\text{ETR}}$ .
22: end for
```

templates follow the standard chat formats corresponding to each base model (e.g., Qwen-Chat or Llama-Instruct).

E.2 Hyperparameters

Table 5 lists the detailed hyperparameters used in our experiments. To ensure fair comparison, we maintain consistent settings between GRPO and ETR, differing only in the clipping mechanism.

For the model-specific configurations, we adjust the `max_response_length` to accommodate the capacity of different architectures:

- Qwen3-8B-Base and Llama-3.1-8B-Instruct: Set to 8192 tokens.
- Qwen2.5-Math-7B-Base: Set to 3584 tokens due to its embedding positions limits.
- Code generation experiments: Set to 4096 tokens.

For our proposed ETR method, we set the elasticity coefficients $\lambda_1 = 0.1$ (micro-level) and $\lambda_2 = 0.1$ (macro-level) for the main GRPO experiments without further tuning. For the Qwen3-30B-A3B scaling experiment, which uses a much tighter base trust region, we proportionally set $\lambda_1 = \lambda_2 = 1 \times 10^{-4}$.

Table 5: Hyperparameters for training and evaluation. ETR introduces λ_1 and λ_2 , while other settings remain identical to the GRPO baseline.

Hyperparameter	Value
<i>General Training Config</i>	
Optimizer	AdamW
Learning Rate	1e-6 (Constant)
Global Batch Size	64
Rollouts per Prompt (G)	8
Gradient Clipping	1.0
KL Coefficient (β)	0.001
Reward Function	$\{0, 1\}$ (Math), m/n (Code)
<i>Generation Config</i>	
Training Temperature	1.0
Training Top-p	1.0
Evaluation Temperature	1.0
Max Response Length	{8192, 4096, 3584}
<i>ETR Specifics</i>	
Base Clip (ϵ_{base})	0.2
Micro-Elasticity (λ_1)	0.1 (main), 1×10^{-4} (30B)
Macro-Elasticity (λ_2)	0.1 (main), 1×10^{-4} (30B)

Table 6: Hyperparameter sensitivity of ETR on Qwen3-8B-Base at step 200. Results are reported as Mean@32 / Best@32.

Config.	AMC23	AIME24	AIME25
$\lambda_1=0.05, \lambda_2=0.10$	0.627 / 0.941	0.144 / 0.352	0.211 / 0.479
$\lambda_1=0.10, \lambda_2=0.05$	0.623 / 0.945	0.148 / 0.347	0.206 / 0.481
$\lambda_1=0.10, \lambda_2=0.10$	0.628 / 0.938	0.145 / 0.343	0.207 / 0.482

F Analysis

In this section, we analyze the training dynamics to understand the source of ETR’s performance gains. We focus on the clipping behavior and response length evolution, while referring back to the accuracy curves presented in Figure 4.

F.1 Asymmetric Noise Suppression

Figure 7 illustrates the fraction of samples where the policy update is clipped. A counter-intuitive observation is that ETR triggers clipping much more frequently (spikes $> 1.0\%$) than the static GRPO baseline (stable at $\approx 0.1\%$), despite ETR’s ability to expand the trust region.

This phenomenon validates the **asymmetric design** of our dynamic boundary. For negative samples ($A_t < 0$), the threshold tightens ($\epsilon_t < \epsilon_{base}$). Since incorrect reasoning paths typically constitute the majority of generated data in difficult reasoning tasks, these negative samples frequently hit the tightened boundary. Therefore, the high clipping fraction indicates that ETR is actively **suppressing noise**. By strictly limiting the gradient

updates from erroneous paths, ETR prevents the model from "unlearning" useful logic due to low-quality negative signals, while implicitly reserving the trust region budget for the sparser, positive signals where the boundary is expanded.

F.2 Regularization against Reward Hacking

Figure 8 reveals a common failure mode in RLVR: *Reward Hacking via Length*. The GRPO baseline (Yellow) rapidly drifts towards generating excessively long responses (> 5000 tokens), likely attempting to maximize the probability of hitting a correct answer through verbose exploration. This bloating increases training latency without proportional accuracy gains.

In contrast, ETR (Blue) acts as a geometric regularizer. When the model generates long, incorrect chains, the advantage is negative, causing the trust region to shrink. This penalizes the model heavily if it drifts too far from the reference policy on wrong paths. As a result, ETR stabilizes the response length at a more efficient level (≈ 3500 tokens), encouraging concise and precise reasoning.

F.3 Prevention of Policy Collapse

As shown previously in Figure 4, standard GRPO suffers from a "performance collapse" in the later stages of training, where the *Best@32* accuracy degrades significantly. This collapse is often linked to the unchecked accumulation of policy shifts induced by noisy updates.

By implementing the asymmetric constraints described above—tightening on errors and relaxing on successes—ETR maintains policy stability. The dynamic boundary ensures that the policy does not drift excessively on ambiguous or incorrect samples. Consequently, ETR sustains a robust upward trajectory in accuracy throughout the training process, avoiding the degradation seen in the baseline.

F.4 Comparison with Adaptive KL

We also compare ETR with an adaptive KL baseline, which adjusts the global KL penalty coefficient during training. Although adaptive KL is a natural way to regulate policy drift, it operates at the global regularization level rather than the local clipping level. This distinction is important in sparse reasoning tasks: when the model discovers a rare correct trajectory, the resulting policy shift may temporarily increase KL, and a global controller can mistakenly suppress the useful breakthrough.

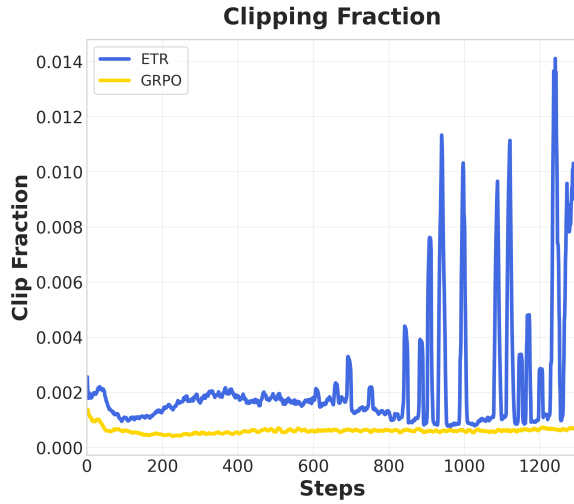


Figure 7: **Clipping Fraction.** ETR triggers clipping significantly more often than GRPO. This reflects the *tightening* of constraints on negative samples, effectively filtering out diffusion noise from incorrect paths.

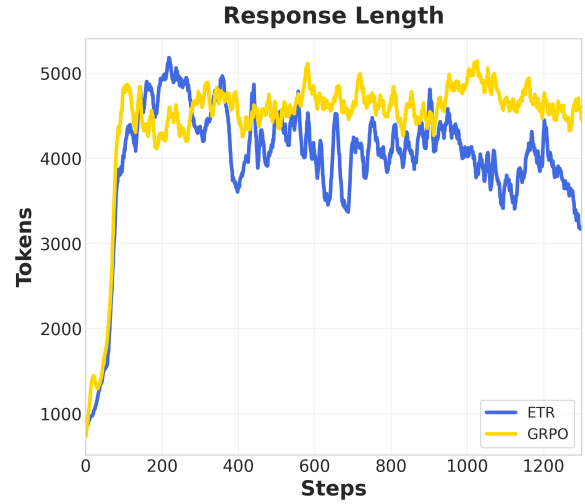


Figure 8: **Response Length.** ETR (Blue) effectively curbs the "length hacking" phenomenon observed in GRPO (Yellow), promoting more efficient reasoning chains.

Table 7: Comparison with adaptive KL on Qwen3-8B-Base at step 500. Results are reported as Mean@32 / Best@32.

Method	AMC23	AIME24	AIME25	Avg.
Base Model	0.317 / 0.850	0.016 / 0.166	0.037 / 0.289	0.123 / 0.435
GRPO + Fixed KL	0.680 / 0.948	0.169 / 0.315	0.215 / 0.460	0.355 / 0.574
GRPO + Adaptive KL	0.482 / 0.905	0.094 / 0.248	0.169 / 0.433	0.248 / 0.529
GRPO + ETR	0.710 / 0.953	0.176 / 0.446	0.226 / 0.509	0.371 / 0.636

Table 7 shows that adaptive KL underperforms the fixed-KL GRPO baseline in this setting. We attribute this to *breakthrough suppression*: global KL adaptation penalizes the model precisely when sparse correct paths induce a useful policy movement. ETR instead keeps the reference-model KL term unchanged and adjusts only the local trust region, allowing high-confidence samples to receive larger updates while preserving global regularization.