ETR: Outcome-Guided Elastic Trust Regions for Policy Optimization

Anonymous ACL submission

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 the dominant 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 **E**lastic **T**rust **R**egions (**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 on AIME and MATH benchmarks demonstrate that ETR consistently outperforms GRPO, achieving superior accuracy while effectively mitigating policy entropy degradation to ensure sustained exploration.

1 Introduction

Reinforcement Learning (RL) has become a central technique for improving the reasoning capabilities of Large Language Models (LLMs) (Achiam et al., 2023; Touvron et al., 2023; Grattafiori et al., 2024; Team, 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., 2025b).

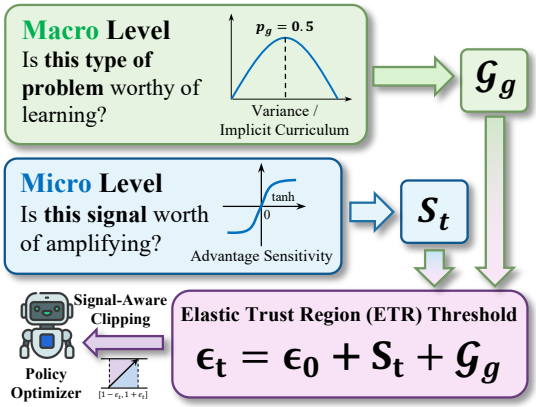

Figure 1: **Not all learning signals deserve equal trust.** ETR replaces GRPO's fixed threshold with dynamic bounds that expand at the macro-level for valuable samples and contract at the micro-level for uncertain steps.

Early approaches primarily relied on Reinforcement Learning from Human Feedback (RLHF) (Schulman et al., 2017b), 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., 2025b; Wang et al., 2025). To address these limitations, Reinforcement Learning with Verifiable Rewards (RLVR) has recently gained prominence (Wen et al., 2025a; Zhang et al., 2025a) 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 outcomes, 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 largely inherits from Proximal Policy Optimization's (PPO) (Schulman et al., 2017b) static trust region design, enforcing a fixed update constraint across all samples and groups. We argue that this assumption is poorly suited for RLVR, where training signals are inherently heterogeneous. Specifically, advantage values can vary widely in magnitude and sign, and query prompts differ substantially in difficulty, leading to groups with distinct statistical properties. Statistical principles suggest that groups with pass rates near 50% exhibit maximum variance and contain the richest gradient information(Zhang et al., 2025b). A uniform clipping threshold thus misallocates optimization capacity, over-constraining informative samples while insufficiently regulating low-quality and noisy ones. In practice, this mismatch leads to inefficient learning and rapid policy entropy collapse, thereby limiting exploration and generalization in reasoning tasks.

To address these issues, we **E**lastic **T**rust **R**egions (**ETR**), a dynamic constraint mechanism that adapts policy updates to outcome statistics. Instead of enforcing a fixed clipping threshold, ETR adjusts the effective trust region based on both sample-level advantages and group-level variance. This design allocates larger update budgets to more informative samples and groups, enabling difficulty-aware optimization and implicitly encouraging a curriculum over the given prompts, applying adequate constraints to groups with varying information density. ETR integrates seamlessly with GRPO-style objectives and improves training stability without introducing additional networks or supervision.

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 on mathematical reasoning benchmarks, we demonstrate that ETR consistently outperforms GRPO and its Clip-High variant across all base models. Specifically, on the Qwen3-8B baseline, ETR achieves a **+7.7%** gain on AMC23 (**+5.6%** over Clip-High) and a **+2.6%** gain on AIME25 (**+2.3%** over Clip-High) over GRPO, while also showing robust generalization on out-of-distribution tasks and substantially mitigating rapid policy entropy decay.

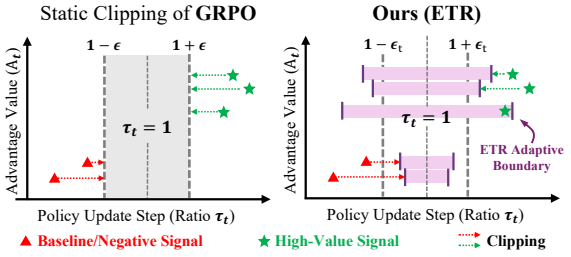

Figure 2: Comparison of optimization landscapes between GRPO and ETR. **Left (GRPO):** the static clipping range suffers from *static mismatch*, where high-value signals (green stars, large advantage) are aggressively truncated, limiting efficient learning. **Right (ETR):** our method creates a signal-aware 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., 2017b) 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:

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

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

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., 2015) 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 $\{o_1, \ldots, o_G\}$ from the old policy $\pi_{\theta_{old}}$. The advantage A_i is computed by normalizing the intra-group rewards in G.

The objective function of GRPO can be divided into two main parts: i) a clipped surrogate objective (to limit the update step from $\pi_{\theta_{old}}$) (that maximizes the reward / advantage of the sampled group, while the clipping limits the updated gradient step) and ii) a KL penalty term (to keep the policy close to the reference model π_{ref}).

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

where $r_t = \frac{\pi_\theta(o_t|q,o_{<t})}{\pi_{\theta_{old}}(o_t|q,o_{<t})}$ 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. For a single sample, the objective with a KL divergence penalty β is approximated via a second-order Taylor expansion as $\max_r r \cdot A - \frac{\beta}{2}(r-1)^2$. Solving for the optimal update step r^*, we obtain a linear relationship $|r^* - 1| \propto |A|/\beta$.

This relationship reveals that the optimal update magnitude is inherently dependent on the signal strength A. Crucially, this implies that the ideal trust region is not a constant, but a dynamic variable determined by the confidence of the policy's evaluation. However, existing algorithms like PPO and GRPO enforce a uniform, fixed constraint $|r-1| \leq \epsilon$, disregarding the specific utility of each token. This imposes a "one-size-fits-all" limitation: updates driven by high-quality reasoning paths are inadequately capped, while updates from ambiguous or noisy samples are allowed disproportionate freedom. 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 golden 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 introduce a scaling factor $\rho_t \geq 1$ representing the signal quality of the t-th token:

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

In this formulation, a larger ρ_t reduces the effective penalty weight of the KL divergence. This justifies allowing a larger deviation for samples with high signal quality. By solving the Lagrangian dual of this constrained optimization problem, we derive the optimal form of the clipping boundary. This theoretical result serves as the cornerstone of our method, explicitly dictating that the trust region radius should not be static, but must scale dynamically with the signal strength to balance stability and efficiency.

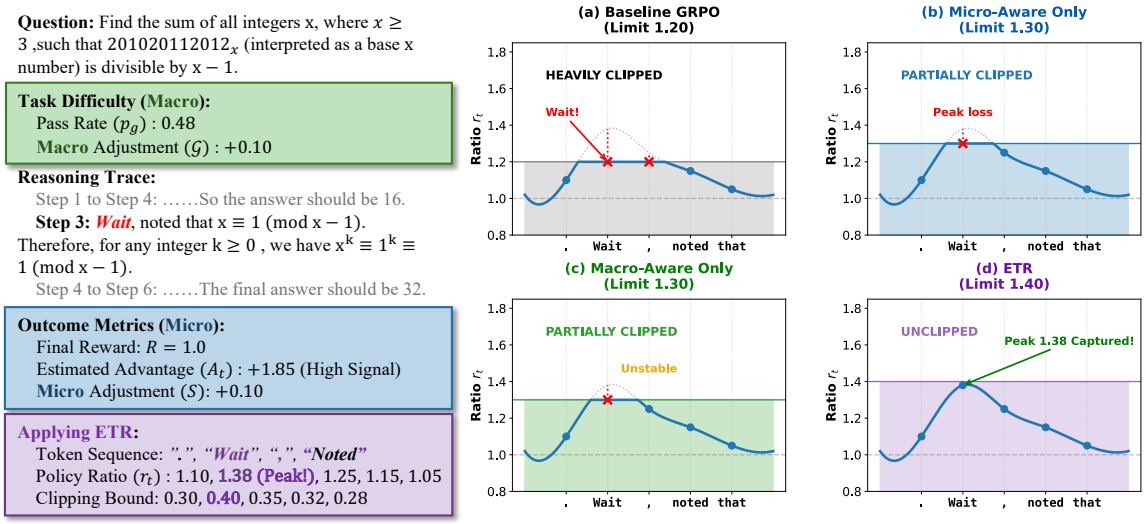

Figure 3: **A case study on a high-difficulty reasoning step.** We visualize the policy ratio trajectory for a critical step that requires a large update (≈ 1.38). **(a)** GRPO (limit 1.20) clips the gradient greatly, losing information. **(b) & (c)** Applying only Micro or Macro adjustments alleviates 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.

Theorem 3.1: Optimal Dynamic Boundary

To satisfy the signal-aware weighted constraint, the clipping threshold $\epsilon^*(t)$ for the t-th sample must scale with the square root of its signal strength ρ_t:

$$\epsilon_{dynamic}(t) = \epsilon_{base} \cdot \sqrt{\rho_t} \tag{5}$$

This theorem provides the mathematical justification for dynamic clipping: the trust region radius should inherently scale with the confidence of the learning signal.

3.3 Elastic Trust Regions (ETR)

Guided by Theorem 3.1, we propose ETR. We parameterize the scaling factor ρ_t using two additive terms to ensure numerical stability and decouple the hyperparameters.

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:

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

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

Micro-Level Adjustment ($\mathcal{S}$). We define the sample-level term to adapt to the magnitude and sign of the advantage. We employ the hyperbolic tangent function to bound the adjustment range.

- For positive samples ($A_t > 0$), we relax the upper bound to allow for larger updates on correct paths.
- For negative samples ($A_t < 0$), we adjust the lower bound to maintain stability against errors.

$$\mathcal{S}(A_t) = \lambda_1 \cdot \tanh(A_t) \tag{7}$$

Macro-Level Adjustment ($\mathcal{G}$). We define the group-level term based on the variance of the group outcomes. For a group with pass rate p_g, the variance is proportional to $p_g(1 - p_g)$.

$$\mathcal{G}(p_g) = \lambda_2 \cdot 4p_g(1 - p_g) \tag{8}$$

Insight: Implicit Curriculum Learning

The macro-level term $\mathcal{G}(p_g)$ 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 automatically allocates a larger update budget to tasks within the "optimal learning zone" (high variance), while maintaining conservative constraints on tasks that are currently too hard or too easy.

Final Objective. Integrating the dynamic boundary into the GRPO framework, the ETR loss function is computed by substituting ϵ_t into the standard clipped objective:

$$\mathcal{J}_{ETR}(\theta) = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\left(\min\left(r_{i,t}A_i,\right.\right.\right.$$
$$\left.\left.\left.\text{clip}(r_{i,t}, 1-\epsilon_{i,t}, 1+\epsilon_{i,t})A_i\right) - \beta D_{KL}(\pi_\theta||\pi_{ref})\right)\right] \tag{9}$$

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

4 Experiments

In this section, we empirically evaluate the proposed ETR method on multiple mathematical reasoning benchmarks. Our experiments aim to verify the performance of the dynamic trust region mechanism across different task difficulties, analyze its impact on training convergence and policy exploration, and validate the effectiveness of its individual design components.

4.1 Experimental Setup

Benchmarks. We assess the reasoning capabilities of the models using a diverse set of datasets. For in-distribution evaluation, we utilize **MATH500** (Hendrycks et al., 2021) and **AMC23**(Li et al., 2024) to represent standard competition-level problems, and **AIME 2024**(Lewkowycz et al., 2022) and **AIME 2025** to represent high-difficulty olympiad problems. The AIME datasets contain sparse correct paths, serving as a critical testbed for our signal-aware hypothesis. For these tasks, we report both **Mean@32** and **Best@32** accuracy to measure average performance and exploration ceilings, respectively. To evaluate generalization capabilities, we employ out-of-distribution (OOD) benchmarks including **GPQA-Main**, **GPQA-Diamond** (Rein et al., 2024), and **ACPBench** (Kokel et al., 2025), reporting **Pass@1** accuracy. The training data is derived from the DAPO-Math-17k (Yu et al., 2025) dataset.

Baselines and Implementation. We conduct experiments using the `verl` reinforcement learning framework, utilizing FSDP for training and vLLM for inference on 8×H20 GPUs. The experiments cover multiple model architectures: **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). Our primary baseline is **Group Relative Policy Optimization (GRPO)** with a static clipping threshold ($\epsilon = 0.2$). To investigate the effect of simply relaxing the constraint, we also compare against a **GRPO-Clip-High** variant inspired by (Yu et al., 2025) with $\epsilon = 0.28$. ETR is implemented within the same codebase, modifying only the clipping logic in the loss function. We set the group size $G = 8$ and the sampling temperature to 1.0. The reward function is strictly outcome-based (+1 for correct, -1 for incorrect). All models are trained using the AdamW optimizer. For ETR, the micro-level coefficient λ_1 and macro-level coefficient λ_2 are both set to default values of 0.1.

4.2 Main Results

Table 1 presents the performance comparison across all benchmarks. ETR consistently outperforms the static GRPO and the Clip-high baselines across different model families.

Performance on Hard Tasks. Notably, the performance gain provided by ETR increases with task difficulty. On the challenging AIME 2024 and AIME 2025 datasets, ETR achieves substantial improvements in the Best@32 metric. This confirms that in hard tasks, where correct reasoning paths are rare and yield high advantage values, the static clipping mechanism leads to under-fitting. ETR's elastic boundary effectively captures these sparse, high-value signals, allowing the model to capitalize on successful exploration.

Out-of-Distribution Generalization. ETR also demonstrates consistent improvements on OOD tasks (GPQA and ACPBench). We attribute this generalization capability to the suppression of overfitting. Static clipping constraints can force the model to overfit specific templates in the training set to maximize rewards within a restricted policy space. By tightening constraints on low-signal samples and relaxing them for high-confidence updates, ETR maintains higher policy diversity, encouraging the learning of robust reasoning logic rather than dataset-specific patterns.

4.3 Analysis of Training Process

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

Table 1: Main results on in-distribution and out-of-distribution benchmarks. In-distribution metrics are reported as **Mean@32 / Best@32**. Highlighted rows indicate our ETR method. **Bold** denotes the best performance 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
Model: Qwen3-8B-Base								
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**
Model: Llama-3.1-8B-Instruct								
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**
Model: Qwen2.5-7B-Math-Base								
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**

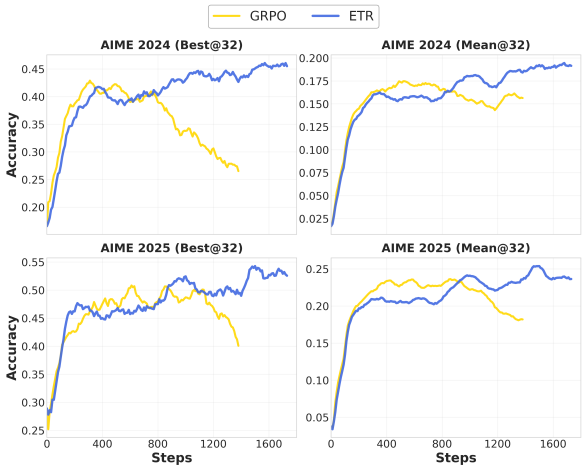

Figure 4: Val Accuracy on AIME 2024/2025 using Qwen3-8B-Base. Standard GRPO (yellow) suffers from performance collapse in later learning stages. ETR (blue) maintains a steady upwards trend in both Mean@32 and Best@32.

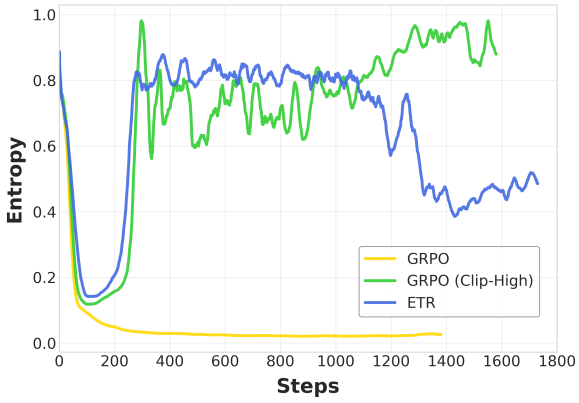

Figure 5: **Policy Entropy Evolution.** GRPO collapses to zero, limiting exploration. Clip-High leads to high, unstable entropy. ETR maintains healthy entropy levels, facilitating sustained learning.

Mitigating Late-Stage Collapse. Figure 4 illustrates the validation accuracy on AIME 2024. Standard GRPO exhibits significant instability in the later stages of training. After reaching a peak, the Best@32 metric begins to decline. We attribute this not to traditional overfitting, but to entropy collapse. In the multi-sampling setting ($G = 32$), if the policy becomes deterministic, the diversity of generated responses vanishes. Once the model converges to a sub-optimal path, it loses the capability to explore and sample correct answers, leading to the degradation of the Best@32 metric. In contrast, ETR (Blue line) maintains a robust upward trajectory without collapse. This proves that the dynamic boundary mechanism effectively stabilizes the policy update throughout the training process.

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

GRPO: The entropy drops rapidly to near-zero, indicating premature convergence to a deterministic distribution.
GRPO-Clip-High: Increasing the static threshold to 0.28 prevents entropy collapse but leads to severe oscillations and a continuous rise in entropy in later stages. This suggests that unconstrained exploration introduces harmful noise that does not translate into performance gains.
ETR: ETR exhibits a "dip-and-rebound" pattern. Entropy decreases initially as the model learns, but stabilizes or recovers in later stages. This resilience indicates that the elastic boundary allows the model to re-expand its search space when high-value signals are encountered, maintaining a balance between exploitation and exploration.

4.4 Ablation Study

We verify the design choices of ETR through two sets of ablation studies.

Component Effectiveness. We compare the full ETR with variants where either the micro-level (sample-based) or macro-level (group-based) adjustment is disabled. Removing the micro-level adjustment ($\lambda_1 = 0$) results in the most significant performance drop and slower convergence. This confirms that adapting to the advantage magnitude is the primary driver of ETR, as static boundaries limit the utilization of high-value gradients. Removing the macro-level adjustment ($\lambda_2 = 0$) also degrades efficiency. This validates the effectiveness of implicit curriculum learning: allocating a larger update budget to groups with high variance (i.e., moderate difficulty) accelerates learning from the most informative batches.

Strategy Direction. We validate the necessity of the asymmetric design (expanding for positive samples, tightening for negative ones) by comparing ETR against an Inverse Variant (expanding for negative, tightening for positive). The results show that the Inverse Variant significantly underperforms. This is due to the asymmetry of gradient updates: increasing the probability of a correct token is a focused operation that benefits from a relaxed bound. In contrast, decreasing the probability of an error token redistributes probability mass to the rest of the vocabulary. If the constraint is too loose for negative samples, this redistribution indiscriminately boosts irrelevant tokens, introducing diffusion noise. Therefore, tightening the bound for negative samples is essential for stability.

4.5 Computational Efficiency

A significant advantage of ETR is its zero computational overhead relative to GRPO. Unlike methods that require auxiliary models or complex matrix computations, ETR relies solely on element-wise operations on tensors that are already computed during the standard rollout and advantage estimation phases. The calculation of the dynamic threshold adds negligible latency to the training step, making ETR highly scalable and easy to integrate into existing RLVR pipelines.

4.6 Hyperparameter Robustness

Finally, we examine the sensitivity of ETR to hyperparameters. Experimental results indicate that setting both the micro and macro coefficients to 0.1 is a robust choice. We applied this identical configuration across different model architectures (Qwen and Llama) and sizes, and consistently achieved performance gains over the baseline. This suggests that the effectiveness of ETR stems from its core signal-aware mechanism rather than overfitting to a specific set of hyperparameters.

5 Related Works

5.1 Reinforcement Learning for LLM Post-Training

Reinforcement learning has become a central paradigm for post-training large language models, most prominently through Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022; Lambert, 2026; Dong et al., 2024), where policies are optimized using learned reward models. While effective for alignment, reward-model-based approaches introduce additional approximation errors and often struggle to provide precise supervision for tasks with objectively verifiable outcomes. This has motivated the recent shift toward Reinforcement Learning with Verifiable Rewards (RLVR) (Wen et al., 2025a; Tian et al., 2025), particularly in domains such as mathematical reasoning and program synthesis, where correctness can be automatically evaluated (Cobbe et al., 2021; Lightman et al., 2023). We focus on improving the stability and efficiency of policy optimization under outcome-driven rewards in RLVR.

5.2 Trust-Region and PPO-Style Policy Optimization

Trust-region methods are widely used to stabilize policy optimization. TRPO (Schulman et al.,

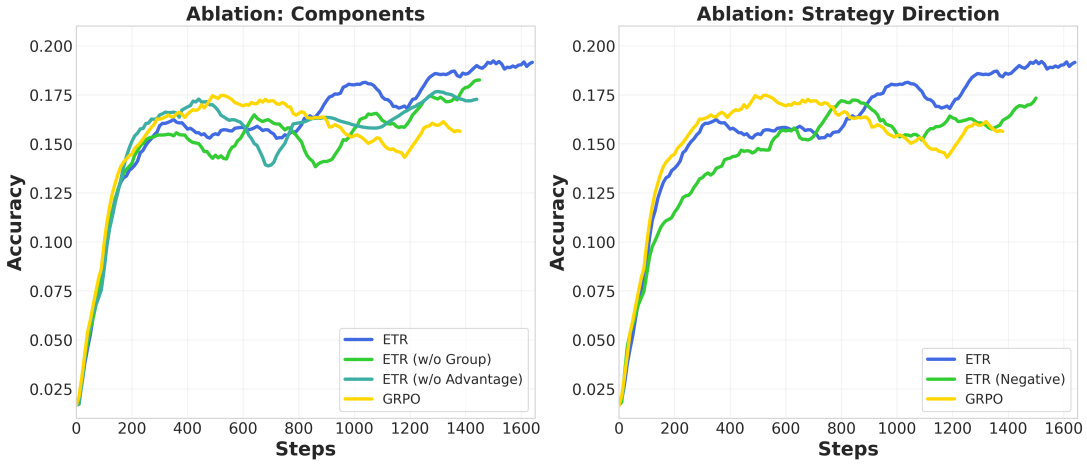

Figure 6: **Ablation Study. Left:** Contribution of Macro and Micro components. **Right:** Comparison of inverse directions. The Negative Variant (expanding boundaries for negative samples) fails due to noise, confirming the necessity of our asymmetric design.

2017a) enforces an explicit constraint on policy updates via KL divergence, while PPO (Schulman et al., 2017b) introduces a clipped surrogate objective as a computationally efficient approximation. PPO-style clipping has become the de facto standard in RLHF and RLVR pipelines due to its simplicity and robustness (Team, 2022, 2024). However, this design relies on a static trust region that applies uniformly across samples, implicitly assuming homogeneous training signals. While prior work has explored alternative constraints and adaptive penalties (Wang et al., 2019; Su et al., 2025; Yu et al., 2025), the suitability of static trust regions for heterogeneous, outcome-driven learning remains underexplored. Our work revisits this assumption and proposes a data-adaptive alternative.

5.3 Adaptive Optimization and Curriculum Learning

Adapting optimization behavior to data difficulty and information content has been widely studied in reinforcement learning and optimization. Prior works (Schulman et al., 2017a; Hanzely, 2023; Wang et al., 2024) explore adaptive learning rates, trust-region adjustments, and variance-aware update rules to improve stability and sample efficiency. In parallel, curriculum learning (Bengio et al., 2009; Lin et al., 2025; Okamoto et al., 2021) methods aim to prioritize informative or appropriately challenging data, either through manually designed schedules or learned difficulty estimators (Song et al., 2025; Shi et al., 2025; Zhang et al., 2025b). While effective, many curriculum strategies rely on task-specific heuristics or additional supervision that may not be readily available in large-scale RL settings. In the context of LLM post-training, GRPO can be viewed as an implicit step toward difficulty-aware learning by computing relative advantages within groups, reducing reliance on absolute reward scales. However, GRPO still applies a uniform trust region across samples and groups, limiting its ability to adapt optimization dynamics to heterogeneous outcome statistics. Our work builds on this line of research by introducing Elastic Trust Regions, which directly leverage advantage magnitude and group-level variance to dynamically adjust optimization constraints, enabling implicit curriculum learning without additional supervision.

6 Conclusion

We identify static trust regions as a key limitation of GRPO when applied to heterogeneous, outcome-driven reasoning tasks. To address this, we propose Elastic Trust Regions (ETR), a data-adaptive constraint mechanism that adjusts policy updates based on advantage magnitude and group-level variance. Experiments on mathematical reasoning benchmarks show that ETR improves accuracy and stabilizes training by mitigating policy entropy collapse. These results highlight the importance of outcome-adaptive optimization for effective post-training of reasoning-oriented language models. Given its simplicity and negligible computational overhead, ETR can serve as a plug-and-play enhancement for existing RLVR pipelines. Ultimately, bridging signal quality and constraints enables robust, sample-efficient training for reasoning models and their future large-scale applications.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 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.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *CoRR*, abs/2110.14168.

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. *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, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 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. *Preprint*, arXiv:2311.10203.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *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, and 1 others. 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 *AAAI*. AAAI Press.

Nathan Lambert. 2026. Reinforcement learning from human feedback. *Preprint*, arXiv:2504.12501.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, and 1 others. 2022. Solving quantitative reasoning problems with language models. *Advances in neural information processing systems*, 35:3843–3857.

Jia Li, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Huang, Kashif Rasul, Longhui Yu, Albert Q Jiang, Ziju Shen, and 1 others. 2024. Numinamath: The largest public dataset in ai4maths with 860k pairs of competition math problems and solutions. *Hugging Face repository*, 13(9):9.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. *Preprint*, arXiv:2305.20050.

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.

Wataru Okamoto, Hiroshi Kera, and K. Kawamoto. 2021. Reinforcement learning with adaptive curriculum dynamics randomization for fault-tolerant robot control. *ArXiv*, abs/2111.10005.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 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, Philipp Moritz, Michael I. Jordan, and Pieter Abbeel. 2017a. Trust region policy optimization. *Preprint*, arXiv:1502.05477.

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

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017b. 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, and 1 others. 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 r1-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.

Constitutional AI Team. 2022. Constitutional ai: Harmlessness from ai feedback. *Preprint*, arXiv:2212.08073.

Gemini Team. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, Chuning Tang, Congcong Wang, Dehao Zhang, Enming Yuan, Enzhe Lu, Fengxiang Tang, Flood Sung, Guangda Wei, Guokun Lai, and 77 others. 2025. Kimi k1.5: Scaling reinforcement learning with llms. *Preprint*, arXiv:2501.12599.

OpenAI Team. 2024. Openai o1 system card. *Preprint*, arXiv:2412.16720.

Wanxin Tian, Shijie Zhang, Kevin Zhang, Xiaowei Chi, Yulin Luo, Junyu Lu, Chunkai Fan, Qiang Zhou, Yiming Zhao, Ning Liu Siyu Lin, and 1 others. 2025. Seea-r1: Tree-structured reinforcement fine-tuning for self-evolving embodied agents. *arXiv preprint arXiv:2506.21669*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 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, and 1 others. 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*. Curran Associates Inc., Red Hook, NY, USA.

Xumeng Wen, Zihan Liu, Shun Zheng, Zhijian Xu, Shengyu Ye, Zhirong Wu, Xiao Liang, Yang Wang, Junjie Li, Ziming Miao, and 1 others. 2025a. Reinforcement learning with verifiable rewards implicitly incentivizes correct reasoning in base llms. *arXiv preprint arXiv:2506.14245*.

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. 2025b. Reinforcement learning with verifiable rewards implicitly incentivizes correct reasoning in base llms. *Preprint*, arXiv:2506.14245.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 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, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. 2024. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement. *Preprint*, arXiv:2409.12122.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, and 1 others. 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. 2025a. Beyond markovian: Reflective exploration via bayes-adaptive rl for llm reasoning. *Preprint*, arXiv:2505.20561.

Shijie Zhang, Guohao Sun, Kevin Zhang, Xiang Guo, and Rujun Guo. 2025b. Clpo: Curriculum learning meets policy optimization for llm reasoning. *Preprint*, arXiv:2509.25004.

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:

$$\begin{aligned} &\max_\theta \ \mathbb{E}_t[r_t(\theta)A_t] \\ &\text{s.t.} \quad \mathbb{E}_t\left[\frac{1}{\rho(A_t, p_g)} D_{KL}(\pi_{\theta_{old}}||\pi_\theta)\right] \leq \delta \end{aligned} \tag{10}$$

where $\rho(\cdot) \geq 1$ 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.

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:

$$\begin{aligned} D_{KL}(\pi_{\theta_{old}}||\pi_\theta) &= \mathbb{E}_{x\sim\pi_{\theta_{old}}}\left[-\log\frac{\pi_\theta(x)}{\pi_{\theta_{old}}(x)}\right] \\ &= \mathbb{E}_{x\sim\pi_{\theta_{old}}}[-\log r_t] \end{aligned} \tag{11}$$

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:

$$-\log r \approx -(r-1) + \frac{1}{2}(r-1)^2 \tag{12}$$

Substituting this back into the expectation:

$$\begin{aligned} D_{KL} \approx\ &\mathbb{E}_{x\sim\pi_{\theta_{old}}}[-(r_t-1)] \\ &+ \mathbb{E}_{x\sim\pi_{\theta_{old}}}\left[\frac{1}{2}(r_t-1)^2\right] \end{aligned} \tag{13}$$

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

$$\mathbb{E}[r_t - 1] = \mathbb{E}[r_t] - 1 = \int \pi_{old}\frac{\pi}{\pi_{old}} - 1 = 1 - 1 = 0 \tag{14}$$

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

$$D_{KL}(\pi_{\theta_{old}}||\pi_\theta) \approx \frac{1}{2}(r_t-1)^2 \tag{15}$$

Solving for the Boundary. We now substitute Eq. 15 into the local version of the weighted constraint (Eq. 10). For a specific sample t, we require:

$$\frac{1}{\rho_t} \cdot \frac{1}{2}(r_t-1)^2 \leq \delta_{local} \tag{16}$$

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

$$(r_t-1)^2 \leq 2\delta_{local}\cdot\rho_t \implies |r_t-1| \leq \sqrt{2\delta_{local}}\sqrt{\rho_t} \tag{17}$$

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

$$\epsilon_t = \epsilon_{base} \cdot \sqrt{\rho(A_t, p_g)} \tag{18}$$

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

B Limitations

The main limitations of this work are the scope of evaluation. First, experiments are limited to medium-scale models, with no validation on larger-scale LLMs (e.g., 10B+ parameters). The computational efficiency, convergence stability, and performance of ETR's adaptive thresholding when scaled to such sizes remain unexamined. Second, the study focuses solely on structured tasks (e.g., math reasoning) and is not extended to open-ended tasks (e.g., open-domain dialogue, creative writing). The applicability of ETR's assessment proxies to scenarios with high output diversity and ambiguous evaluation criteria remain unexplored.

Algorithm 1 Group Relative Policy Optimization with ETR

1: **Input:** Dataset $\mathcal{D}$, policy π_θ, ref policy π_{ref}, group size G.
2: **Hyperparams:** Base clip ϵ_{base}, weights (λ_1, λ_2), KL coef β.
3: **for** each training step **do**
4: Sample query $q \sim \mathcal{D}$.
5: Generate outputs $O = \{o_1, \ldots, o_G\}$ via $\pi_\theta(\cdot|q)$.
6: Compute rewards $\mathbf{r} = \{r_1, \ldots, r_G\}$.
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 π_θ via gradient descent $\nabla_\theta \mathcal{J}_{\text{ETR}}$.
22: **end for**

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 `verl` framework (Sheng et al., 2025) on a computational node equipped with 8× NVIDIA H20 GPUs. We utilize the DAPO-Math-17k dataset for training across all models. The prompt templates follow the standard chat formats corresponding to each base model (e.g., Qwen-Chat or Llama-Instruct).

E.2 Hyperparameters

Table 2 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.

For our proposed ETR method, we set the elasticity coefficients $\lambda_1 = 0.1$ (micro-level) and $\lambda_2 = 0.1$ (macro-level) across all experiments without further tuning.

Table 2: Hyperparameters for training and evaluation. ETR introduces λ_1 and λ_2, while other settings remain identical to the GRPO baseline.

Hyperparameter	**Value**
General Training Config	
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	± 1 (Binary)
Generation Config	
Training Temperature	1.0
Training Top-p	1.0
Evaluation Temperature	1.0
Max Response Length	{8192, 3584}
ETR Specifics	
Base Clip (ϵ_{base})	0.2
Micro-Elasticity (λ_1)	**0.1**
Macro-Elasticity (λ_2)	**0.1**

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

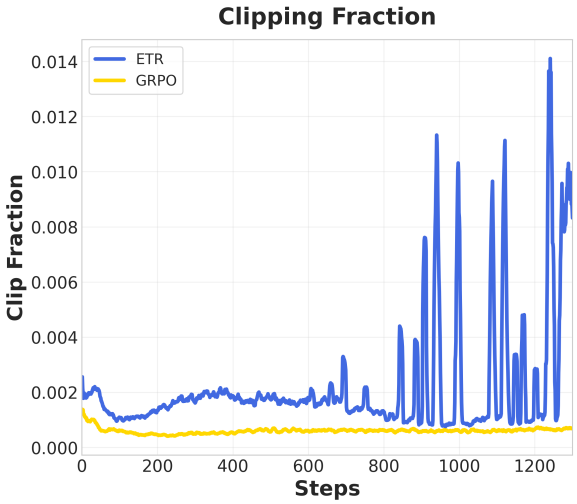

Figure 7: **Clipping Fraction.** ETR triggers clipping significantly more often than GRPO. This reflects the *tightening* of constraints on negative samples, effectively filtering out diffusion noise from incorrect paths.

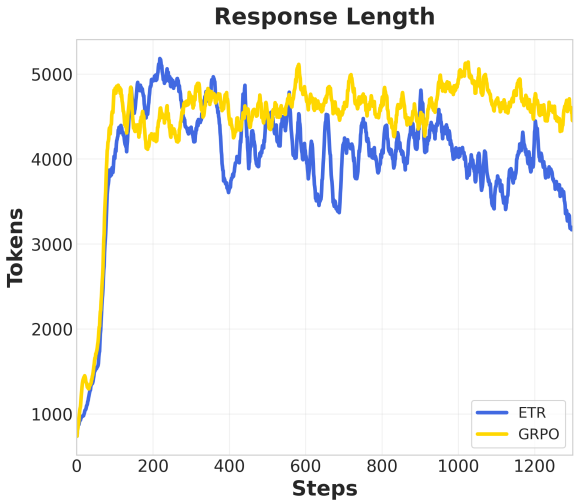

Figure 8: **Response Length.** ETR (Blue) effectively curbs the "length hacking" phenomenon observed in GRPO (Yellow), promoting more efficient reasoning chains.

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.