# Stabilizing Efficient Reasoning with Step-Level Advantage Selection

**Anonymous ACL submission**

## Abstract

Large language models (LLMs) achieve strong reasoning performance by allocating substantial computation at inference time, often generating long and verbose reasoning traces. While recent work on efficient reasoning seeks to reduce this overhead through length-based rewards or pruning mechanisms, many approaches rely on post-training with substantially shorter context windows than those used during base model training, a factor whose effect has not been systematically isolated. In this work, we first show that short-context post-training alone, using standard GRPO without any length-aware objectives, already induces substantial reasoning compression. However, this process leads to increasingly unstable training dynamics, characterized by accuracy degradation as training progresses. To address this issue, we propose Step-level Advantage Selection (SAS) that operates at the reasoning step level, selectively filtering noisy or redundant reasoning steps in correct rollouts, while preserving high-confidence intermediate reasoning steps even from verifier-failed rollouts, where failures may arise from truncation or verification issues rather than incorrect reasoning. Evaluating across diverse mathematical and general reasoning benchmarks, our approach reduces average reasoning length by more than 30% of tokens while consistently outperforming other baselines with average Pass@1 accuracy gains of 3.79 over a length-aware baseline. SAS achieves a consistently better accuracy–efficiency trade-off compared to strong length-aware baselines, demonstrating that stable and effective reasoning compression can be achieved with minimal modifications to standard reinforcement learning pipelines, highlighting the importance of disentangling context length effects from explicit length-control objectives in efficient reasoning.

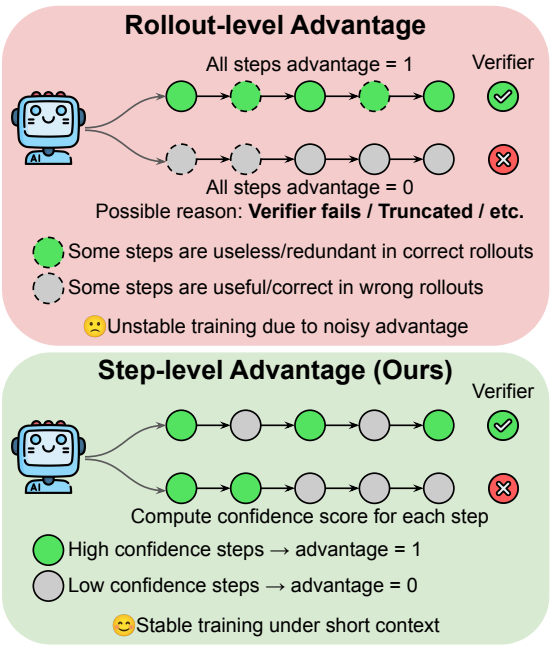



Figure 1: Rollout-level versus step-level advantage selection. Standard GRPO propagates rewards uniformly based on the final verifier outcome, suppressing all steps in verifier-failed rollouts and over-updating redundant steps in correct rollouts, which leads to unstable training under short context. In contrast, Step-level Advantage Selection (SAS) filters low-confidence redundant steps in correct rollouts and preserves high-confidence intermediate steps from verifier-failed rollouts, resulting in more stable training and efficient reasoning compression under short context.

## 1 Introduction

Large language models (LLMs) have demonstrated strong reasoning capabilities across a wide range of tasks, including mathematical problem solving, logical reasoning, and code generation. Recent progress has shown that allocating more computation at inference time—commonly referred to as test-time scaling—can substantially improve reasoning performance by encouraging models to generate longer chain-of-thought traces or explore mul-

tiple reasoning paths (Wei et al., 2022; Kojima et al., 2022; Wang et al., 2023; Yao et al., 2023). However, these gains often come at a significant computational cost, as models tend to produce excessively long and verbose reasoning even for relatively simple problems, leading to increased inference latency and reduced practical efficiency (Chen et al., 2025; Gema et al., 2025; Ghosal et al., 2025).

To address this issue, a growing body of work has focused on efficient reasoning, which aims to reduce reasoning length while preserving task performance. Many recent approaches adopt reinforcement learning–based post-training strategies that explicitly incorporate length-aware objectives, such as token-budget constraints, length-aware rewards, or pruning mechanisms (Aggarwal and Welleck, 2025; Wu et al., 2025a; Hou et al., 2025; Sui et al., 2025). While effective, these approaches share a critical, yet overlooked, training condition: they typically conduct post-training within a substantially restricted context window (e.g., 4K tokens), a sharp departure from the expansive windows (e.g., 16K–24K tokens) used during base model training. This discrepancy raises a fundamental question: to what extent does the observed reasoning compression stem from explicit length-aware objectives, as opposed to being a natural consequence of short-context post-training itself?

In this work, we systematically isolate this effect by training a long-context reasoning model (Luo et al., 2025b; Guo et al., 2025) using pure GRPO with short context, deliberately excluding any length-aware rewards or pruning techniques. Our findings reveal that short-context post-training alone serves as a strong and sufficient signal for reasoning compression, achieving reductions in output length comparable to state-of-the-art efficient reasoning methods. By identifying this "natural" compression effect, we highlight a critical variable that has been previously conflated with explicit length-control objectives, suggesting that the short-context training condition itself plays a more dominant role in model efficiency than previously recognized. However, further analysis reveals a critical limitation. As training progresses, standard GRPO under short-context post-training exhibits unstable behavior: task accuracy fluctuates and tends to degrade over time, while the KL divergence from the reference policy steadily increases—a phenomenon accompanied by a rapid collapse in exploration and increasingly brittle policy updates. We hypothesize that this instability stems from the truncation of reasoning traces induced by the restricted short context window. Although truncated rollouts are ultimately assigned zero reward by the verifier due to missing final answers, they often contain partially correct intermediate reasoning steps. In such cases, the resulting advantage signals become noisy and misaligned: correct intermediate steps are implicitly associated with failure, while incomplete reasoning fragments still contribute to policy updates. As these misaligned signals accumulate, the model is encouraged to optimize fragmented or shortcut reasoning patterns, leading to unstable behavior and degraded task performance. These observations indicate that simply reducing the training context length, while effective for compression, is insufficient for robust and stable efficient reasoning.

To resolve this tension between aggressive reasoning compression and stable performance preservation, we introduce Step-level Advantage Selection (SAS) (Fig. 1), which treats reasoning as a sequence of discrete, evaluable steps. As illustrated in Fig. 1, rather than applying a uniform reward to an entire trace, our method operates at the reasoning-step level to selectively filter out unreliable steps. Specifically, we mask the advantages of low-confidence reasoning steps in correct rollouts, while extending the mechanism to preserve high-quality intermediate steps even within verifier-failed rollouts, which may fail due to truncation under short context or verification issues rather than incorrect reasoning. By focusing the optimization signal on complete, high-confidence reasoning steps, SAS effectively mitigates noisy optimization signals and preserves performance during aggressive reasoning compression under short context. We evaluate SAS on a diverse set of mathematical and general reasoning benchmarks. Experimental results demonstrate that SAS consistently achieves a superior accuracy–efficiency trade-off compared to existing baselines. Specifically, relative to the base DeepScaleR-1.5B-Preview model, our method reduces the average output length by over 30% while simultaneously improving average Pass@1 accuracy by 1.51 points. Compared to length-aware baselines such as LAPO and ThinkPrune, our approach achieves higher accuracy while generating 15% fewer tokens on average. These improvements are reflected in the Accuracy–Efficiency Score, where our method exceeds all baselines by a clear margin, with an absolute gain of over 0.13 AES compared to the strongest competing approaches. Together, these results show that our method en-

ables aggressive reasoning compression while preserving, and in many cases improving, task performance relative to prior efficient reasoning methods.

## 2 Related Work

**Test-time Scaling in LLMs** Test-time scaling has been shown to improve the performance of large language models on various complex reasoning tasks, such as mathematical problem-solving and code generation (Wei et al., 2022; Kojima et al., 2022; Wu et al., 2025b). Such performance gains can often be achieved by allocating more inference-time computation, either through generating longer chain-of-thought reasoning traces (Madaan et al., 2023) or by exploring a larger number of reasoning paths (Wang et al., 2023; Yao et al., 2023; Wang et al., 2024). More recently, reinforcement learning has been used to directly induce extended reasoning behaviors, producing models that generate substantially longer and more elaborate reasoning traces at inference time, such as OpenAI-o1 (Jaech et al., 2024) and DeepSeek-R1 (Guo et al., 2025). While these approaches have demonstrated strong gains on challenging benchmarks, they often rely on generating significantly longer reasoning traces, which can be several times longer than those produced by short chain-of-thought models, leading to increased inference cost and latency. Subsequent analyses have found that such extended reasoning frequently contains redundant or unnecessary steps, including excessive verification or repetition, even on relatively simple problems, resulting in an "overthinking" phenomenon that degrades efficiency without proportional accuracy gains (Chen et al., 2025; Ghosal et al., 2025; Gema et al., 2025). As a result, despite their effectiveness, existing test-time scaling methods generally lack precise and dynamic control over the length of generated reasoning, motivating growing interest in approaches that retain the benefits of increased inference-time computation while enabling more efficient and controllable reasoning.

**Efficient Reasoning for LLMs** Motivated by the high computational cost and inefficiency introduced by test-time scaling, a growing body of work has focused on *efficient reasoning*, which aims to reduce reasoning overhead while preserving task performance. Rather than allocating additional computation at inference time, these approaches seek to shorten reasoning traces through training-time optimization, reasoning structure compression, or inference control (Sui et al., 2025). A prominent class of methods focuses on model-based efficient reasoning, particularly through leveraging traditional RL optimization techniques combined with explicit length-aware reward to control the length of CoT reasoning (Team et al., 2025; Arora and Zanette, 2025). Representative approaches include L1 (Aggarwal and Welleck, 2025), which introduces explicit length constraints during policy optimization, length-adaptive policy optimization methods such as LAPO (Wu et al., 2025a), and pruning-based strategies such as ThinkPrune (Hou et al., 2025), which iteratively remove redundant reasoning steps while maintaining correctness. Despite their effectiveness, most existing methods entangle explicit length-aware objectives with short-context post-training, making it difficult to isolate the role of training context and learning signal design. In contrast, our work revisits efficient reasoning from the perspective of credit assignment, showing that stable reasoning compression can be achieved without explicit length rewards by selectively assigning advantages at the level of reasoning steps.

## 3 Methodology

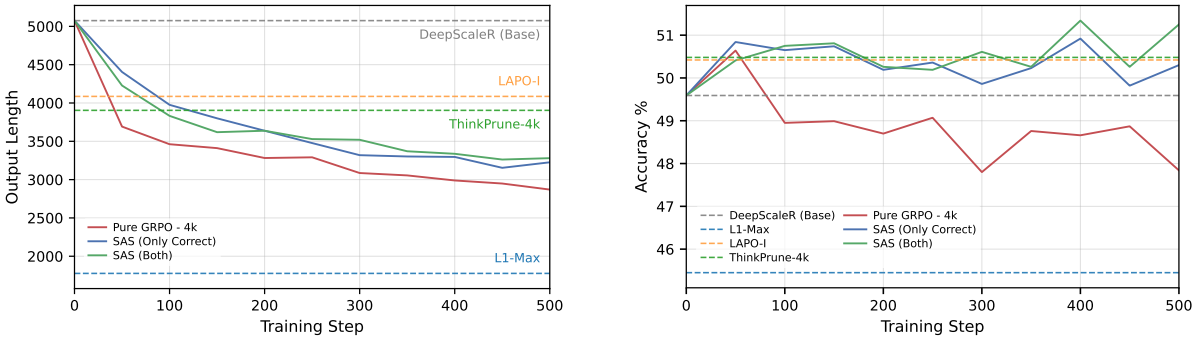
### 3.1 Preliminary: RL with Length-Aware Reward Design

Recent work on efficient reasoning commonly adopts reinforcement learning (RL) to explicitly control the length of model-generated reasoning. In this setting, a pretrained language model is further optimized using policy optimization, where the training objective combines task-level rewards with regularization terms that constrain deviations from a reference policy. A widely used framework is Group Relative Policy Optimization (GRPO), which optimizes the policy based on relative advantages computed from groups of sampled responses.

Formally, GRPO optimizes the following objective:

$$\mathcal{L}_{\text{GRPO}}(\theta) = E_{x\sim\mathcal{D},\, y\sim\pi_\theta(\cdot|x)}\left[\hat{A}(x,y)\right] - \beta\,\mathrm{KL}(\pi_\theta(\cdot \mid x)\,\|\,\pi_{\text{ref}}(\cdot \mid x)) \quad (1)$$

where $\pi_\theta$ denotes the current policy, $\pi_{\text{ref}}$ is a fixed reference policy, and $\hat{A}(x,y)$ represents a group-relative advantage estimated from sampled responses. The KL divergence term regularizes the policy update and limits distributional drift from the reference model.




(a) Output length vs. training step.

(b) Accuracy vs. training step.

Figure 2: Training dynamics of average output length and accuracy across six reasoning datasets under short-context (4K) post-training. Note that the dashed horizontal lines just indicate the performance of these baselines and do not represent their training trajectories.

To encourage concise reasoning, many prior methods incorporate length-aware reward design into the RL framework. A generic form of such rewards can be written as:

$$r(x, y) = r_{\text{task}}(x, y) - \lambda \cdot g(|y|) \qquad (2)$$

where $r_{\text{task}}(x, y)$ measures task correctness, $|y|$ denotes the output length, and $g(\cdot)$ is a monotonic function that penalizes longer reasoning traces. Different methods instantiate $g(\cdot)$ using target lengths, token budgets, or length-dependent penalties, and integrate this reward into the advantage estimation during policy optimization. Empirically, these approaches show that reasoning length can be directly influenced through RL-based post-training.

However, several empirical observations complicate this formulation. First, prior work reports that output length often exhibits non-monotonic behavior during RL training, where reasoning length initially increases before decreasing later. Second, most length-control methods perform post-training using a substantially shorter context window (e.g., 4k tokens) compared to the long-context pretraining setup of the base model. This context mismatch suggests that models trained under shorter contexts may naturally prefer shorter outputs, independent of the specific length-aware reward design. Together, these factors make it unclear to what extent reasoning compression should be attributed to explicit length rewards versus training dynamics and context length.

### 3.2 Short-Context Post-Training

To isolate the influence of training context length from explicit length-aware reward design, we conduct a controlled study using pure GRPO without any length-dependent reward or constraint. Starting from a reasoning-capable base model pretrained with a long context window, we perform GRPO-based post-training using a fixed 4k context window, which matches the post-training setup commonly adopted in prior efficient reasoning work. Importantly, our reward function is based solely on task correctness, introducing no explicit signal regarding output length.

Remarkably, despite the absence of length-aware objectives, short-context post-training alone induces a substantial reduction in reasoning length. As shown in Fig. 2a, the average output length decreases sharply during the early stages of training and continues to decline steadily thereafter, eventually reaching a level comparable to or even shorter than existing efficient reasoning baselines such as LAPO and ThinkPrune. At the same time, the trained policy achieves task accuracy that is initially comparable to these methods, as shown in Fig. 2b. These results indicate that short-context post-training itself provides a strong and previously underexamined compression signal, independent of explicit length-control mechanisms.

However, a closer inspection of the training dynamics reveals important limitations. While the reasoning length continues to decrease throughout training, task accuracy becomes increasingly volatile, characterized by noticeable fluctuations and a gradual performance decay in later training stages (Fig. 2b). This divergence suggests that naïvely applying pure GRPO under short-context constraints fails to reliably preserve performance over extended training.

Taken together, these observations highlight a fundamental tension between reasoning compres-

sion and performance stability. Although short-context GRPO post-training is effective at reducing output length, it does not guarantee stable accuracy as training proceeds. This motivates the need for a training strategy that can retain the strong compression effect induced by short-context post-training while explicitly stabilizing task performance.

### 3.3 Step-level Advantage Selection (SAS)

Motivated by the training instability observed during the short-context post-training (Section 3.2), we propose Step-level Advantage Selection (SAS). The method stabilizes the learning process by selectively modulating the influence of individual reasoning steps on policy updates, without requiring explicit length-aware rewards or architectural changes. SAS operates directly at the advantage level, ensuring that only reliable reasoning steps contribute to the optimization signal.

**Mask Advantages in Correct Rollouts** Even when a rollout is verified as correct (reward = 1), the reasoning trace may contain redundant, erroneous or low-confidence steps that are weakly related to the final outcome. Reinforcing these steps can introduce noisy updates and exacerbate policy drift during training.

For each correct rollout, we partition the generated reasoning trace into a sequence of reasoning steps $\{s_i\}_{i=1}^{N}$, where steps are defined as contiguous text segments separated by double newline delimiters (`\n\n`). For each step $s_i$, we compute a step-level confidence score based on the model's token-level log probabilities. Let $\mathcal{T}_i$ denote the set of tokens belonging to step $s_i$. The confidence score is defined as:

$$c_i = \frac{1}{|\mathcal{T}_i|} \sum_{t \in \mathcal{T}_i} \log p_\theta(t \mid x, t_{<t}), \tag{3}$$

where $p_\theta$ denotes the current policy.

Given the confidence scores $\{c_i\}$, we sort the reasoning steps in ascending order of confidence and select a ratio $r \in (0, 1)$ of the lowest-confidence steps. For all tokens belonging to these steps, we mask their advantages by setting them to zero:

$$\tilde{A}_t = \begin{cases} 0, & \text{if } t \in s_i \text{ and } s_i \in \mathcal{S}^{+}_{\text{mask}}, \\ A_t, & \text{otherwise,} \end{cases} \tag{4}$$

where $A_t$ is the original GRPO advantage and $\mathcal{S}^{+}_{\text{mask}}$ denotes the selected subset of steps in a verifier-approved rollout. This operation suppresses the contribution of unreliable reasoning steps while preserving learning signals from more reliable ones.

**Preserving Signals in Incorrect Rollouts** We further extend this idea to rollouts that fail the rule-based verification (reward = 0). A key observation in short-context training is that many "incorrect" rollouts are not inherently flawed in logic; rather, they are naturally truncated "correct" traces where the model ran out of context before reaching the final answer. These traces often contain high-quality intermediate reasoning steps that are identical to those found in successful solutions. Rather than discarding these traces entirely—which would waste successful reasoning patterns—we preserve their "useful" segments to densify the learning signal.

For each verifier-rejected rollout, we again compute step-level confidence scores $\{c_i\}$ and sort the steps in descending order of confidence. We then select a ratio of $1 - r$ of the highest-confidence steps and explicitly preserve their learning signal by assigning them a positive advantage value 1:

$$\tilde{A}_t = \begin{cases} 1, & \text{if } t \in s_i \text{ and } s_i \in \mathcal{S}^{-}_{\text{mask}}, \\ A_t, & \text{otherwise,} \end{cases} \tag{5}$$

where $\mathcal{S}^{-}_{\text{mask}}$ denotes the selected subset of steps in a verifier-rejected rollout. By rewarding reliable intermediate steps even in failed rollouts, SAS provides a denser and more stable learning signal, which is critical for maintaining accuracy during aggressive reasoning compression

Together, these two components form a reward-conditioned, step-level advantage selection mechanism. Correct rollouts are prevented from reinforcing unreliable reasoning steps, while verifier-rejected rollouts are decomposed to preserve useful intermediate reasoning. As shown in our experiments, this symmetric treatment further stabilizes training and improves accuracy–efficiency trade-offs under short-context post-training.

## 4 Experiments

### 4.1 Experimental Setup

**Training Details** We use DeepScaleR-Preview-Dataset (Luo et al., 2025b) for training, which includes approximately 40K mathematics problem–answer pairs collected from AIME (1984–2023), AMC (prior to 2023), Omni-MATH (Gao et al., 2025), and Still (Min et al., 2024). Our base model is DeepScaleR-1.5B-Preview (Luo et al., 2025b), a 1.5B-parameter model

| | AIME24 | | AIME25 | | MATH | | AMC | | OlympiadBench | | Average | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Pass@1 | #Tok | Pass@1 | #Tok | Pass@1 | #Tok | Pass@1 | #Tok | Pass@1 | #Tok | Pass@1 | #Tok | AES |
| DeepScaleR | 33.75 | 6755 | 26.88 | 6444 | 86.36 | 2809 | 67.62 | 4761 | 47.23 | 4824 | 52.37 | 5118 | 0.00 |
| GRPO-4K | 38.75 | 5282 | 25.42 | 4812 | 85.09 | 1976 | **71.69** | 3395 | 47.09 | 3411 | 53.61 | 3775 | 0.33 |
| L1-Max | 25.63 | 2158 | 21.67 | 2032 | 83.60 | 1480 | 64.61 | 1768 | 44.69 | 1703 | 48.04 | 1828 | 0.23 |
| LAPO-I | 34.58 | 5627 | **27.92** | 5290 | 86.04 | 2220 | 69.73 | 3765 | 48.18 | 3735 | 53.29 | 4127 | 0.25 |
| ThinkPrune-4k | 36.04 | 5468 | 26.67 | 5177 | 86.23 | 2090 | 70.18 | 3636 | 47.62 | 3651 | 53.35 | 4004 | 0.27 |
| SAS (ours) | **39.79** | 4876 | 26.67 | 4295 | **86.58** | 1768 | 71.46 | 3090 | **48.19** | 3008 | **54.54** | 3407 | **0.46** |

Table 1: Comparison of methods across five math reasoning benchmarks. SAS achieves a superior accuracy–efficiency trade-off, reducing reasoning length while maintaining or improving accuracy over strong baselines.

| | GPQA-Diamond | | LSAT | | MMLU | | Average | | |
|---|---|---|---|---|---|---|---|---|---|
| | Pass@1 | #Tok | Pass@1 | #Tok | Pass@1 | #Tok | Pass@1 | #Tok | AES |
| DeepScaleR | 35.70 | 4860 | **28.64** | 6934 | 47.99 | 1454 | 37.44 | 4416 | 0.00 |
| GRPO-4K | 32.48 | 1515 | 28.45 | 5025 | 48.73 | 950 | 36.55 | 2496 | 0.32 |
| L1-Max | 36.05 | 3878 | 26.66 | 1949 | 48.94 | 900 | 37.22 | 2242 | **0.46** |
| LAPO-I | 36.17 | 3404 | 28.42 | 5382 | 48.71 | 1207 | 37.77 | 3331 | 0.27 |
| ThinkPrune-4k | 35.83 | 3278 | 28.13 | 5131 | 48.91 | 973 | 37.62 | 3127 | 0.31 |
| SAS (ours) | **37.18** | 2998 | 28.32 | 4312 | **49.39** | 876 | **38.30** | 2729 | 0.45 |

Table 2: Performance comparison of methods on three out-of-domain general reasoning benchmarks. SAS consistently improves Pass@1 accuracy while substantially reducing reasoning length, resulting in a strong and competitive accuracy–efficiency trade-off under short-context post-training.

originally RL fine-tuned from DeepSeek-R1-Distill-Qwen-1.5B (Guo et al., 2025) on the same dataset using three training stages with increasing context lengths (8K $\rightarrow$ 16K $\rightarrow$ 24K). For GRPO training, we adopt the same hyperparameters as in DeepScaleR-1.5B-Preview. In particular, we use a learning rate of 1e-6 and a training batch size of 128. For our method, we set the hyperparameter ratio $r$ to 0.3. For a fair comparison, we set the maximum context length as 4K tokens during training. All experiments are conducted for 500 training steps using the VeRL framework (Sheng et al., 2025), with 8 rollouts sampled per prompt. We use AIME24 as validation set to select the checkpoint with the highest Accuracy–Efficiency Score (AES). All experiments are conducted using 8 AMD MI250 with 64 GB memory.

**Evaluation Details** We evaluate on five different math reasoning datasets: AIME2024, AIME2025, AMC (MAA, 2024), MATH (Hendrycks et al., 2021b), OlympiadBench (He et al., 2024). In addition, we include GPQA-Diamond (Rein et al., 2024), LSAT (Zhong et al., 2024), MMLU (Hendrycks et al., 2021a), three general reasoning benchmarks to test the ability to generalize to out-of-domain data. For each problem, we sample 16 responses with a temperature of 0.6, a top-p of 0.95, and a maximum generation length of 8K tokens. Following standard practices Guo et al. (2025), we report Pass@1 accuracy, computed by averaging over 16 sampled responses per problem. We also report the average number of output tokens to measure reasoning length. Finally, we compute the Accuracy-Efficiency Score (AES; Luo et al., 2025a) to evaluate the trade-off between improving accuracy and reducing computational cost.

**Baselines** We evaluate our method against the following baselines, all of which are initialized from DeepScaleR-1.5B-Preview post-trained using a maximum context length of 4K tokens:

**GRPO-4K**: trained with standard GRPO post-training under a 4K training context window, without any additional RL techniques.

**L1-Max** (Aggarwal and Welleck, 2025): trained with Length Controlled Policy Optimization (LCPO) to generate outputs of varying lengths while respecting a specified maximum length constraint.

**ThinkPrune-4k** (Hou et al., 2025): An RL–based pruning method that iteratively enforces progressively stricter token limits to remove redundant reasoning steps.

**LAPO-I** (Wu et al., 2025a): A two-stage RL approach that adaptively controls reasoning length by modeling and leveraging successful solution length distributions.

### 4.2 Experimental Results

Table 1 presents the performance of SAS and baselines across five mathematical reasoning datasets. Our method achieves the best accuracy–efficiency trade-off, outperforming all baselines. Notably, short-context post-training alone already induces substantial reasoning compression: compared to the base DeepScaleR-1.5B-Preview, GRPO-4K substantially reduces the average output length while slightly improving average Pass@1 accuracy, confirming that context length itself provides a strong compression signal. However, this compression comes with limited stability (as shown in Fig. 2). Among length-aware baselines, L1-Max achieves the most aggressive compression but at a significant cost to task accuracy. Conversely, ThinkPrune and LAPO prioritize accuracy preservation but achieve only moderate length reduction, resulting in lower overall efficiency gains. In contrast, SAS consistently outperforms all the baselines, improving average Pass@1 accuracy by more than 2 points over the base model while reducing output length by approximately 1,700 tokens. This dual improvement results in the highest AES of 0.46. Notably, our method maintains or exceeds the accuracy of GRPO-4K across all math benchmarks while generating shorter traces, demonstrating a better accuracy–efficiency trade-off.

We extend our evaluation to three general reasoning benchmarks in Table 2. Although these tasks typically require more concise reasoning traces, we observe similar trends. Pure short-context post-training (GRPO-4K) again leads to reduced output length but degrades accuracy relative to the base model, indicating over-compression. In contrast, our method preserves or improves accuracy across all general reasoning datasets with substantially shorter outputs, achieving the best overall efficiency score among stable methods. These results demonstrate that the benefits of our approach are not limited to mathematical reasoning, but extend to broader reasoning settings where controlling training stability is critical.

## 5 Analysis

### 5.1 Ablation Study

We evaluate the impact of our design choices on five math reasoning datasets, as shown in Table 3.

**Value of Verifier-Failed Rollouts.** We first examine the necessity of leveraging verifier-failed rollouts. Restricting SAS to only correct rollouts

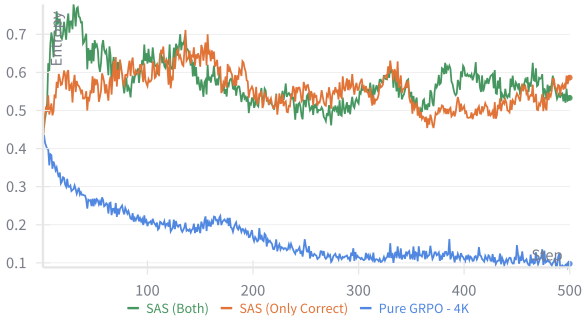



Figure 3: Policy entropy throughout training. SAS maintains higher entropy compared to the rapid entropy collapse observed in pure GRPO-4K, indicating more robust exploration.

("Only Correct") leads to a consistent drop in overall efficiency, reducing the average AES from 0.46 to 0.43 and increasing the average output length. This suggests that preserving high-confidence intermediate steps from verifier-failed rollouts provides a critical learning signal that densifies the training data and stabilizes the model during aggressive compression. We further analyze the training stability by monitoring the policy entropy, as shown in Fig. 3. While the entropy of pure GRPO-4K (blue line) drops rapidly, signaling a collapse in exploration and the adoption of brittle, repetitive reasoning patterns, both SAS (Both) and SAS (Only Correct) maintain a significantly higher and more stable entropy level throughout training. This suggests that by selectively filtering noisy advantage signals and preserving high-confidence steps from various rollouts, SAS prevents the model from converging prematurely to suboptimal "shortcut" reasoning, thereby preserving the diversity and robustness of the reasoning process.

**Confidence-Aware Step Selection.** Replacing our confidence-based selection with random step selection ("Random Steps") results in a clear degradation in efficiency (AES 0.38) and longer reasoning traces. This indicates that the gains of our method do not arise merely from sparsifying the advantage signal, but critically depend on selectively filtering low-confidence reasoning steps.

**Step-Level vs. Token-Level Granularity.** The token-level variant ("Token Level") underperforms our step-level design, achieving lower average accuracy and AES (0.39 vs. 0.46), while producing longer outputs. This confirms that aggregating tokens into semantically meaningful reasoning steps yields more stable and effective advantage selection than fine-grained token-level selection.

| | AIME24 | | AIME25 | | MATH | | AMC | | OlympiadBench | | Average | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Pass@1 | #Tok | Pass@1 | #Tok | Pass@1 | #Tok | Pass@1 | #Tok | Pass@1 | #Tok | Pass@1 | #Tok | AES |
| DeepScaleR | 33.75 | 6755 | **26.88** | 6444 | 86.36 | 2809 | 67.62 | 4761 | 47.23 | 4824 | 52.37 | 5118 | 0.00 |
| GRPO-4K | 38.75 | 5282 | 25.42 | 4812 | 85.09 | 1976 | 71.69 | 3395 | 47.09 | 3411 | 53.61 | 3775 | 0.33 |
| SAS (ours) | **39.79** | 4876 | 26.67 | 4295 | **86.58** | 1768 | 71.46 | 3090 | **48.19** | 3008 | **54.54** | 3407 | **0.46** |
| - Only Correct | 36.46 | 4686 | **26.88** | 4294 | 86.18 | 1801 | 71.99 | 3082 | 47.99 | 2965 | 53.90 | 3366 | 0.43 |
| - Random Steps | 37.29 | 4892 | 24.79 | 4461 | 86.10 | 1780 | 71.16 | 3086 | 47.51 | 3055 | 53.37 | 3455 | 0.38 |
| - Token Level | 36.25 | 4771 | 25.00 | 4295 | 86.11 | 1912 | **72.36** | 3148 | 47.56 | 3071 | 53.46 | 3439 | 0.39 |

Table 3: Experimental results for the ablation study.

Overall, these ablations demonstrate that all components of our method—leveraging both correct and incorrect rollouts, step-level granularity, and confidence-based selection—jointly contribute to the strongest accuracy–efficiency trade-off under short-context post-training.

### 5.2 Effect of the Selection Ratio

We study the effect of the selection ratio $r$ in Step-level Advantage Selection (SAS), which controls the fraction of reasoning steps selected for advantage assignment during training. We vary $r$ from 0.1 to 0.9 while keeping all other training settings fixed, and compare against both the base DeepScaleR-1.5B-Preview model and pure GRPO-4K. As shown in Table 4, SAS consistently outperforms the base DeepScaleR-1.5B-Preview model across all tested ratios, yielding higher accuracy and substantially shorter reasoning traces. The best overall performance is achieved at $r = 0.3$, which attains the highest Pass@1 accuracy (54.54), the strongest accuracy–efficiency trade-off (AES = 0.46), and a large reduction in output length compared to the base model. However, performance differences across ratios are relatively small: even extreme values $r = 0.9$ remain competitive, with AES scores above 0.36 and stable accuracy. One possible explanation is that, even when most steps are retained, SAS still induces a highly non-uniform learning signal in which only a small subset of informative reasoning steps meaningfully influences policy updates. This aligns with recent findings that reinforcement learning for LLM reasoning is primarily driven by a minority of high-entropy or decision-critical steps, while most tokens in long reasoning traces are effectively redundant from an optimization perspective (Wang et al., 2025). Overall, these results show that SAS is practically robust and easy to deploy. While moderate ratios offer the best accuracy–efficiency trade-off, the method performs reliably across a wide range of settings. This reinforces the central conclusion that *which* reasoning steps are selected is more important than *how many*, highlighting step-level advantage selection as the key driver of stable and effective reasoning compression.

| Selection Ratio ($r$) | Pass@1 | #Tok | AES |
|---|---|---|---|
| DeepScaleR | 52.73 | 5118 | 0.00 |
| GRPO-4K | 53.61 | 3775 | 0.33 |
| 0.1 | 53.52 | 3259 | 0.43 |
| 0.3 | **54.54** | 3407 | **0.46** |
| 0.5 | 53.39 | 3407 | 0.39 |
| 0.7 | 53.22 | 3412 | 0.38 |
| 0.9 | 53.06 | 3482 | 0.36 |

Table 4: Average performance of SAS under different selection ratio $r$ on five math reasoning datasets.

## 6 Discussion and Conclusion

We revisit efficient reasoning through the lens of post-training context length and advantage selection. We show that short-context post-training alone is sufficient to induce substantial reasoning compression in long-context reasoning models, but also reveals a previously overlooked fundamental limitation of rollout-level reinforcement learning: truncated or verifier-failed rollouts introduce noisy and misaligned learning signals that undermine training stability. These findings highlight that efficiency gains in reasoning models are not solely determined by reward design, but are deeply intertwined with how learning signals are assigned under constrained training environments. To address this challenge, we propose Step-level Advantage Selection (SAS), a simple yet effective strategy that refines advantage assignment within a rollout. By selectively filtering low-confidence steps and preserving high-confidence intermediate reasoning—even in verifier-failed rollouts—our approach stabilizes training and achieves a stronger accuracy–efficiency trade-off than existing length-aware methods.

## Limitations

Our proposed method Step-level Advantage Selection(SAS) demonstrates strong empirical results on both in-domain and out-of-domain tasks. However, our experiments focus on a single base model, and it remains to be seen how well SAS generalizes to models with different sizes, pretraining or post-training paradigms. In addition, all experiments are conducted under a fixed short-context post-training setting, and the behavior of SAS across varying training context length has not been systematically studied. Further, while our ablation results suggest that the selective advantage strategy is broadly effective, we leave a deeper theoretical understanding of advantage selection in reasoning language models to future work. We do not foresee any particular risks associated with the application of our method.

## References


Pranjal Aggarwal and Sean Welleck. 2025. L1: Controlling how long a reasoning model thinks with reinforcement learning. In *Second Conference on Language Modeling*.

Daman Arora and Andrea Zanette. 2025. Training language models to reason efficiently. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuzhi Liu, Mengfei Zhou, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. 2025. Do NOT think that much for 2+3=? on the overthinking of long reasoning models. In *Forty-second International Conference on Machine Learning*.

Bofei Gao, Feifan Song, Zhe Yang, Zefan Cai, Yibo Miao, Qingxiu Dong, Lei Li, Chenghao Ma, Liang Chen, Runxin Xu, Zhengyang Tang, Benyou Wang, Daoguang Zan, Shanghaoran Quan, Ge Zhang, Lei Sha, Yichang Zhang, Xuancheng Ren, Tianyu Liu, and Baobao Chang. 2025. Omni-MATH: A universal olympiad level mathematic benchmark for large language models. In *The Thirteenth International Conference on Learning Representations*.

Aryo Pradipta Gema, Alexander Hägele, Runjin Chen, Andy Arditi, Jacob Goldman-Wetzler, Kit Fraser-Taliente, Henry Sleight, Linda Petrini, Julian Michael, Beatrice Alex, Pasquale Minervini, Yanda Chen, Joe Benton, and Ethan Perez. 2025. Inverse scaling in test-time compute. *Transactions on Machine Learning Research*. Featured Certification, J2C Certification.

Soumya Suvra Ghosal, Souradip Chakraborty, Avinash Reddy, Yifu Lu, Mengdi Wang, Dinesh Manocha, Furong Huang, Mohammad Ghavamzadeh, and Amrit Singh Bedi. 2025. Does thinking more always help? mirage of test-time scaling in reasoning models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

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*.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. 2024. OlympiadBench: A challenging benchmark for promoting AGI with olympiad-level bilingual multimodal scientific problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3828–3850, Bangkok, Thailand. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021a. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

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. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

Bairu Hou, Yang Zhang, Jiabao Ji, Yujian Liu, Kaizhi Qian, Jacob Andreas, and Shiyu Chang. 2025. Thinkprune: Pruning long chain-of-thought of llms via reinforcement learning. *arXiv preprint arXiv:2504.01296*.

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*.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.

Haotian Luo, Li Shen, Haiying He, Yibo Wang, Shiwei Liu, Wei Li, Naiqiang Tan, Xiaochun Cao, and Dacheng Tao. 2025a. O1-pruner: Length-harmonizing fine-tuning for o1-like reasoning pruning. In *2nd AI for Math Workshop @ ICML 2025*.

Michael Luo, Sijun Tan, Justin Wong, Xiaoxiang Shi, William Y. Tang, Manan Roongta, Colin Cai, Jeffrey Luo, Li Erran Li, Raluca Ada Popa, and Ion Stoica. 2025b. DeepscaleR: Effective RL scaling of reasoning models via iterative context lengthening. Notion Blog.

Mathematical Association of America MAA. 2024. American mathematics competitions (amc). https://maa.org/student-programs/amc/.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, pages 46534–46594. Curran Associates, Inc.

Yingqian Min, Zhipeng Chen, Jinhao Jiang, Jie Chen, Jia Deng, Yiwen Hu, Yiru Tang, Jiapeng Wang, Xiaoxue Cheng, Huatong Song, and 1 others. 2024. Imitate, explore, and self-improve: A reproduction report on slow-thinking reasoning systems. *arXiv preprint arXiv:2412.09413*.

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*.

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*, EuroSys '25, page 1279–1297, New York, NY, USA. Association for Computing Machinery.

Yang Sui, Yu-Neng Chuang, Guanchu Wang, Jiamu Zhang, Tianyi Zhang, Jiayi Yuan, Hongyi Liu, Andrew Wen, Shaochen Zhong, Na Zou, Hanjie Chen, and Xia Hu. 2025. Stop overthinking: A survey on efficient reasoning for large language models. *Transactions on Machine Learning Research*.

Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, and 1 others. 2025. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*.

Han Wang, Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. 2024. Soft self-consistency improves language models agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 287–301, Bangkok, Thailand. Association for Computational Linguistics.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xiong-Hui Chen, Jianxin Yang, Zhenru Zhang, Yuqiong Liu, An Yang, Andrew Zhao, Yang Yue, Shiji Song, Bowen Yu, Gao Huang, and Junyang Lin. 2025. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for LLM reasoning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837.

Xingyu Wu, Yuchen Yan, Shangke Lyu, Linjuan Wu, Yiwen Qiu, Yongliang Shen, Weiming Lu, Jian Shao, Jun Xiao, and Yueting Zhuang. 2025a. Lapo: Internalizing reasoning efficiency via length-adaptive policy optimization. *arXiv preprint arXiv:2507.15758*.

Yangzhen Wu, Zhiqing Sun, Shanda Li, Sean Welleck, and Yiming Yang. 2025b. Inference scaling laws: An empirical analysis of compute-optimal inference for LLM problem-solving. In *The Thirteenth International Conference on Learning Representations*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 11809–11822. Curran Associates, Inc.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2024. AGIEval: A human-centric benchmark for evaluating foundation models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2299–2314, Mexico City, Mexico. Association for Computational Linguistics.

## A Accuracy-Efficiency Score (AES)

The AES score (Luo et al., 2025a) compares a tuned model against its corresponding base model to measure the trade-off between accuracy and computational efficiency. Let $\mathrm{L}_{\text{base}}$ and $\mathrm{L}_{\text{model}}$ denote the output lengths, and $\mathrm{Acc}_{\text{base}}$ and $\mathrm{Acc}_{\text{model}}$ denote the accuracies of the base and tuned models, respectively. Defining $\Delta\mathrm{L} = \frac{\mathrm{L}_{\text{base}} - \mathrm{L}_{\text{model}}}{\mathrm{L}_{\text{base}}}$ and $\Delta\mathrm{Acc} = \frac{\mathrm{Acc}_{\text{model}} - \mathrm{Acc}_{\text{base}}}{\mathrm{Acc}_{\text{base}}}$, the AES is calculated as:

$$\mathrm{AES} = \begin{cases} \alpha \cdot \Delta\mathrm{L} + \beta \cdot \Delta\mathrm{Acc}, & \text{if } \Delta\mathrm{Acc} \geq 0 \\ \alpha \cdot \Delta\mathrm{L} - \gamma \cdot |\Delta\mathrm{Acc}|, & \text{if } \Delta\mathrm{Acc} < 0 \end{cases}$$

where $\alpha > 0$, $\beta > 0$, and $\gamma > 0$. Following the original work, we set $\alpha = 1$, $\beta = 3$, and $\gamma = 5$, with $\gamma > \beta$ to emphasize the penalization of accuracy degradation.

## B Licenses

Datasets are released under the following licenses:

- DeepScaleR-Preview-Dataset: MIT license
- AIME24: Apache-2.0 license
- AIME25: Apache-2.0 license
- MATH: MIT license
- AMC: Apache-2.0 license
- OlympiadBench: MIT license
- GPQA: MIT license
- LSAT: MIT license
- MMLU: MIT license

The models we use have the following licenses:

- DeepScaleR-1.5B-Preview: MIT license
- L1-Max: MIT license
- LAPO-I: Apache-2.0 license
- ThinkPrune-4k:Apache-2.0 license