

Taming the Thinker: Conditional Entropy Shaping for Adaptive LLM Reasoning

Shuyu Wei^{1,2*} Jian Sun^{2*} Delai Qiu² Yining Wang² Shengping Liu²
Jerry Liang² Ying Fu² Wei Huang² Jitao Sang^{1†}

¹Beijing Key Laboratory of Traffic Data Mining and Embodied Intelligence,
Beijing Jiaotong University

²Unisound AI Technology Co., Ltd.
{shuyuwei, jtsang}@bjtu.edu.cn

Abstract

Entropy-based deep reasoning has emerged as a promising direction for improving the reasoning capabilities of Large Language Models (LLMs), but existing methods often either increase response length indiscriminately or shorten responses at the cost of accuracy. To better balance this trade-off, we introduce Conditional Entropy Shaping (CES), a framework that dynamically controls token-level response entropy, enabling LLMs to produce concise solutions on simple problems while encouraging deeper exploration on hard ones. Built on DAPO, CES uses token-level entropy as an uncertainty signal and applies a conditional bidirectional policy: it penalizes high-entropy “forking point” tokens on correct reasoning paths to improve conciseness, and rewards them on incorrect paths to encourage exploration and error correction. We implement CES on DeepSeek-R1-Distill-Qwen-7B and evaluate it on 12 mathematical benchmarks. CES consistently improves average accuracy while reducing response length relative to DAPO, and supplementary experiments show similar trends on a smaller 1.5B backbone and on out-of-domain benchmarks.

1 Introduction

In recent years, Large Language Models (LLMs) have demonstrated remarkable capabilities in complex reasoning tasks such as mathematical derivation, code generation, and logical planning (Wei et al., 2022; Kojima et al., 2022). Advanced reasoning models, exemplified by DeepSeek-R1 (Guo et al., 2025), Qwen3 series (Yang et al., 2025a) and OpenAI o3 series (OpenAI, 2025), leverage explicit Chain-of-Thought (CoT) prompting to emulate human-like thought processes, thereby achieving powerful problem-solving abilities. However,

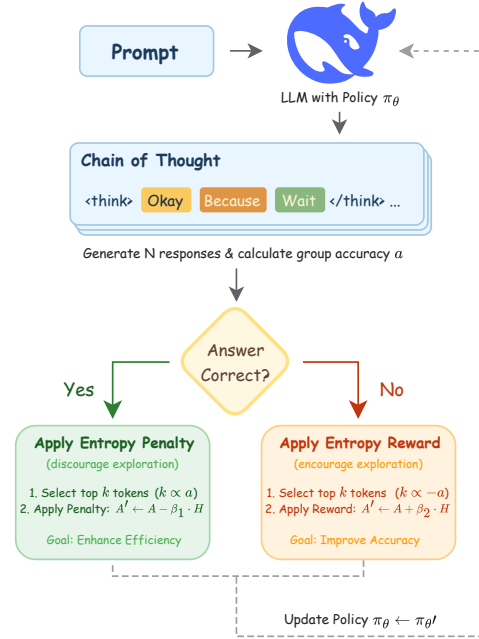


Figure 1: Overview of our CES pipeline.

the very mechanism that enables this high performance introduces a fundamental tension with a second critical requirement: computational efficiency. The explicit generation of reasoning steps, while crucial for accuracy on complex tasks, inherently increases the number of generated tokens, leading to high latency and computational costs that can hinder real-world applications. This underscores a core dilemma in the field. On one hand, to achieve the highest possible performance, models are encouraged to explore detailed reasoning paths. On the other hand, this may lead to significant inefficiency, a phenomenon often described as “overthinking”, where models produce unnecessarily lengthy thought processes for trivial questions like “What is 2+3?” (Chen et al., 2025; Ma et al., 2025; Yang et al., 2025b).

A novel research direction, which we term entropy-based deep reasoning, has emerged by leveraging token-level entropy to analyze and guide

*Equal contribution.

†Corresponding author.

the reasoning process. One study revealed that a few high-entropy tokens within a CoT often act as critical “forking points” in the reasoning path, serving as key levers for decision-making (Wang et al., 2025). They train the model exclusively on the top 20% high-entropy tokens and report performance surpassing that of training on all tokens. Another study demonstrated that rewarding high-entropy tokens can encourage model exploration and significantly improve reasoning accuracy (Cheng et al., 2026). Meanwhile, similar work identifies high-covariance tokens as the primary cause of “entropy collapse” during reinforcement learning (Cui et al., 2025). They restrict their updates to sustain exploration and ultimately improve the model’s reasoning accuracy. While these approaches successfully improve model performance, they also tend to the adverse side effect of further elongating the thought process, thereby exacerbating the “overthinking” phenomenon and increasing inference costs.

In parallel, another line of research has focused on improving reasoning efficiency through reinforcement learning, aiming to shorten responses and realize on-demand thinking. Initial efforts included relatively inflexible methods such as post-hoc pruning of generated thoughts (Muenighoff et al., 2025) or training models to adhere to manually specified length budgets (Aggarwal and Welleck, 2025). More methods have been designed with finer-grained reinforcement learning strategies to achieve the goal of conciseness. For instance, GRPO-LEAD (Zhang and Zuo, 2025) penalizes correct responses that are longer than average. Ada-CoT (Lou et al., 2025) uses reinforcement learning to learn an optimal policy for triggering the entire CoT process based on query complexity, while Ada-R1 (Luo et al., 2025) first merges long and short CoT models and then uses bi-level preference training to select the most suitable reasoning style for a given problem. While effective at reducing length, these approaches may face a critical trade-off: the gains in efficiency frequently come at the cost of performance degradation on more complex problems that genuinely require deliberate reasoning.

This presents a clear dilemma: methods that enhance efficiency risk hurting accuracy, while methods that boost accuracy may hurt efficiency. Inspired by recent advances in token-level entropy (Wang et al., 2025; Cheng et al., 2026; Cui et al., 2025), our work aims to resolve this trade-off by conditioning the model’s exploratory behavior on

the correctness of its reasoning path. In contrast, the previous work (Cheng et al., 2026) applies a single, fixed strategy regardless of the reasoning correctness. Based on this core insight, we propose our novel framework **Conditional Entropy Shaping (CES)**. CES operates within Decoupled Clip and Dynamic Sampling Policy Optimization (DAPO) (Yu et al., 2025) reinforcement learning framework and intelligently modulates the model’s exploratory behavior based on the correctness of its reasoning. As shown in Figure 1, CES guides the model to:

1. **Discourage Exploration:** When a generated reasoning path is **correct**, CES applies a penalty to the highest-entropy tokens within that path. This encourages the model to become more confident and efficient, refining its thought process toward a concise, direct solution.
2. **Encourage Exploration:** Conversely, when the path is **incorrect**, CES rewards these same high-entropy “forking point” tokens. This incentivizes the model to explore alternative pathways and correct its flawed logic.

Empirical results across 12 math benchmarks show that CES improves both accuracy and efficiency on average. The primary contributions of this paper are:

- We introduce CES, a novel reinforcement learning mechanism that implements a conditional and bidirectional control policy for LLM reasoning.
- We demonstrate on 12 mathematical benchmarks that CES improves the average accuracy–efficiency trade-off over DAPO. We further show the robustness of CES through additional experiments on a smaller 1.5B backbone and out-of-domain benchmarks.
- We provide a comprehensive analysis of CES’s learned behavior, revealing how it develops an adaptive, “on-demand” reasoning strategy that strategically allocates computational effort.

2 Method

Our proposed method, CES, introduces a novel advantage-shaping mechanism into the DAPO framework. DAPO is a reinforcement learning

algorithm designed for eliciting complex reasoning in LLMs, which already incorporates several key techniques to stabilize training and improve performance in long CoT scenarios. CES builds upon this strong foundation by introducing an explicit mechanism to manage the trade-off between exploration for accuracy and conciseness for efficiency. It achieves this by dynamically reshaping the token-level advantage signal based on two factors: the correctness of a given model response and the generation entropy of its constituent tokens. Specifically, for correct responses, CES penalizes high-entropy tokens to encourage more direct and concise reasoning paths. Conversely, for incorrect responses, it rewards high-entropy tokens to stimulate exploration and facilitate error correction.

2.1 Preliminaries: The DAPO Framework

DAPO enhances the Group Relative Policy Optimization (GRPO) (Shao et al., 2024) algorithm with a suite of techniques tailored for large-scale reinforcement learning. For a given prompt x , a policy π_θ generates a group of N responses, $Y = \{y_1, y_2, \dots, y_N\}$. The core of the DAPO objective function is to learn a preference by maximizing the advantage of “winner” responses over “loser” responses within the group. The full objective is given by:

$$\mathcal{J}_{\text{DAPO}}(\theta) = \mathbb{E}_{\substack{(q,a) \sim \mathcal{D}, \\ \{o_i\} \sim \pi_{\theta_{\text{old}}}}} \left[\frac{1}{\sum_{i=1}^G |o_i|} \sum_{i=1}^G \sum_{t=1}^{|o_i|} \min \left(r_{i,t}(\theta) \hat{A}_{i,t}, \text{clip}(r_{i,t}(\theta), \right. \right. \quad (1) \\ \left. \left. 1 - \epsilon_{\text{low}}, 1 + \epsilon_{\text{high}}) \hat{A}_{i,t} \right) \right]$$

The key components of DAPO relevant to our work are:

- **Group-Relative Advantage ($\hat{A}_{i,t}$):** The advantage for a response y_i is calculated by normalizing its reward R_i against the mean and standard deviation of rewards within its group $\{R_j\}_{j=1}^G$. This group-normalized advantage is then applied to every token t in the response y_i .
- **Token-Level Policy Gradient Loss:** DAPO’s objective is normalized by the total number of tokens in the batch ($\sum_{i=1}^G |o_i|$), ensuring that each token contributes equally to the final

loss, regardless of the length of the sequence it belongs to. This prevents shorter sequences from being overshadowed by longer ones.

CES intervenes directly at the level of the advantage calculation, $\hat{A}_{i,t}$, before it is used in the DAPO objective function.

2.2 Conditional Entropy Shaping (CES)

CES modifies the advantage signal for each token to provide more nuanced guidance to the model. The process involves three steps.

2.2.1 Step 1: Initial Group-Wise Calculations

For a given prompt x , we generate a response set $Y = \{y_1, y_2, \dots, y_N\}$ using the policy π_θ . We assign a composite reward $R(y_i)$ to each response, which is the sum of two binary components: an accuracy reward $r_{\text{acc}}(y_i) \in \{0, 1\}$ based on the correctness of the final answer, and a format reward $r_{\text{fmt}}(y_i) \in \{0, 1\}$ for adherence to the `<think>...</think>` structure. The total reward is $R(y_i) = r_{\text{acc}}(y_i) + r_{\text{fmt}}(y_i)$.

The group accuracy a , which is crucial for our conditional mechanism, is computed based only on the correctness reward:

$$a = \frac{1}{N} \sum_{i=1}^N r_{\text{acc}}(y_i) \quad (2)$$

The initial, unshaped advantage for any token in response y_i is the standard group-normalized advantage, calculated using the total reward $R(y_i)$:

$$A_i = \frac{R(y_i) - \text{mean}(\{R(y_j)\}_{j=1}^N)}{\text{std}(\{R(y_j)\}_{j=1}^N)} \quad (3)$$

2.2.2 Step 2: Dynamic Selection of High-Entropy Tokens

Then, we compute the token-level entropy. The entropy $H(t_j|y_{i,<j})$ for a token t_j in response y_i at position j is calculated as:

$$H(t_j|y_{i,<j}) = - \sum_{v \in \mathcal{V}} p(v|y_{i,<j}) \log p(v|y_{i,<j}) \quad (4)$$

In Equation 4, V represents the vocabulary size. We then select the top k_i most entropic tokens in each response y_i to form a set $S_H(y_i)$. The number k_i is determined dynamically to modulate the strength of our intervention:

$$k_i = \lfloor |y_i| \cdot \tau \cdot b_i \rfloor \quad (5)$$

In Equation 5, $|y_i|$ is the total length of response y_i , and τ is a base top-rate hyperparameter. The crucial component is the **dynamic multiplier** b_i , defined as:

$$b_i = \begin{cases} a & \text{if } r_{\text{acc}}(y_i) = 1 \\ 1 - a & \text{if } r_{\text{acc}}(y_i) = 0 \end{cases} \quad (6)$$

This design aims to apply a stronger intervention (a larger k_i) in two specific scenarios: (1) when penalizing a correct response in a group that was easy for the model (high a), and (2) when rewarding an incorrect response in a group that was difficult for the model (low a).

2.2.3 Step 3: Entropy-Based Advantage Shaping

Finally, we compute the reshaped advantage $A'_{i,j}$ for each token t_j in response y_i . The advantage is modified only for the selected high-entropy tokens in the set $S_H(y_i)$.

$$A'_{i,j} = \begin{cases} A_i - \beta_1 \cdot H(t_j|y_{i,<j}) & \text{if } r_{\text{acc}}(y_i)=1 \text{ and } t_j \in S_H(y_i) \\ A_i + \beta_2 \cdot H(t_j|y_{i,<j}) & \text{if } r_{\text{acc}}(y_i)=0 \text{ and } t_j \in S_H(y_i) \\ A_i & \text{otherwise} \end{cases} \quad (7)$$

In Equation 7, $\beta_1, \beta_2 > 0$ are positive hyperparameters scaling the magnitude of the entropy-based shaping. This final token-level advantage $A'_{i,j}$ replaces the original $\hat{A}_{i,t}$ in the DAPO objective function (Equation 1), thereby injecting our fine-grained control signal into the learning process. The detailed pseudocode for CES is outlined in Appendix A.

3 Experimental Settings

3.1 Backbone Model and Baselines

Our experiments are conducted in the context of advanced reasoning models. We select the powerful, open-source DeepSeek-R1-Distill-Qwen-7B (Guo et al., 2025) as our backbone model, which is known for its strong long-chain reasoning capabilities. To isolate the impact of our proposed method, we establish three baselines for comparison:

1. **Original R1-7B:** The pretrained DeepSeek-R1-Distill-Qwen-7B model without any reinforcement learning fine-tuning.
2. **DAPO Baseline (the key baseline):** The same backbone model fine-tuned using the

DAPO algorithm without the CES module. This serves as our primary baseline to directly measure the improvements brought by CES.

3. **DAPO with “Entropy Shape”:** We compare CES with the previous work (Cheng et al., 2026). Their work introduces an “Entropy Advantage” that unconditionally adds an entropy-based advantage to all tokens to encourage more exploratory reasoning paths, with the primary goal of improving performance on reasoning tasks. This provides a clear contrast to our conditional, bidirectional approach which aims to balance both accuracy and efficiency.

3.2 Training Details

We utilize the OpenRLHF framework (Hu et al., 2025) to perform DAPO training, focusing on the domain of solving mathematical problems. Due to resource constraints, our training set consists of only 2,500 training samples randomly sampled from the DeepMath dataset (He et al., 2026). All experiments were carried out on 2 NVIDIA A800 GPUs with 80GB of memory.

Notably, we disable the Dynamic Sampling feature of DAPO when training our CES model. Standard DAPO discards batches where all responses are correct or all are incorrect, as these yield zero advantage and thus no gradient for the sequence-level policy update. However, as CES reshapes advantage at the token level using entropy, these seemingly “solved” or “hopeless” batches also provide a valuable, non-zero learning signal. This signal is crucial for refining the model’s confidence and reasoning style, making every sample useful for training. A comprehensive list of hyperparameters can be found in the Appendix.

3.3 Evaluation

For a standardized and reproducible assessment, we employ the evaluation script from the GitHub repository for Qwen2.5-Math (Yang et al., 2024; Qwen Team, 2024). To avoid repetition and instability in long-form reasoning models, we adopt a non-greedy decoding strategy, setting a temperature of 0.4, top- p sampling with $p = 0.95$, and a repetition penalty of 1.05. For each problem in the test sets, we independently generate 4 responses to ensure a stable and representative measurement. Our evaluation focuses on two primary metrics:

Dataset	R1-7B		DAPO (Baseline)		Entropy Shape		CES (Ours)		Improvement	
	Acc $\uparrow$	Len $\downarrow$	Acc $\uparrow$	Len $\downarrow$	Acc $\uparrow$	Len $\downarrow$	Acc $\uparrow$	Len $\downarrow$	Acc $\uparrow$	Len $\downarrow$
AIME24	43.3	8194	42.5	7702	41.7	7527	49.2	6705	+6.7	-997
AMC23	83.8	4446	85.0	4174	85.0	3843	86.9	3160	+1.9	-1014
CMATH	89.2	292	90.3	292	90.1	291	91.9	391	+1.6	+99
CN Middle School 24	62.2	796	61.2	815	61.4	769	70.3	851	+9.1	+36
College Math	38.8	2225	39.5	1999	38.7	1850	41.6	1600	+2.1	-399
GaoKao Math Cloze	77.1	2722	77.6	2546	77.3	2538	78.4	1544	+0.8	-1002
GaoKao 2023 En	71.8	2511	72.2	2182	71.8	2048	79.9	1914	+7.7	-268
GSM8K	86.3	443	85.5	430	86.5	431	89.1	462	+3.6	+32
Minerva Math	47.1	3063	47.9	2546	46.7	2274	42.1	1921	-5.8	-625
Olympiad Bench	52.3	5599	53.6	5136	53.5	4821	55.8	4297	+2.2	-839
SVAMP	89.4	293	90.3	290	90.3	290	90.2	276	-0.1	-14
TABMWP	88.0	411	89.3	397	88.1	387	90.0	460	+0.7	+63
Average	69.1	2583	69.6	2376	69.3	2256	72.1	1965	+2.5	-411

Table 1: Comparison of Accuracy and Response Length on Key Math Datasets. The best result in each category is in **bold**. The terms “Acc” and “Len” represent the mean accuracy and the mean response length across 4 assessments for each benchmark.

1. **Accuracy (Acc):** The average correctness of the final answers.
2. **Average Response Length (Len):** The average number of tokens in the generated responses.

We conduct an extensive evaluation across 12 diverse mathematical reasoning benchmarks: AIME24, AMC23, CMATH (Wei et al., 2023), CN Middle School 24, College Math, GaoKao Math Cloze, GaoKao 2023 En, GSM8K (Cobbe et al., 2021), Minerva Math (Lewkowycz et al., 2022), OlympiadBench (He et al., 2024), SVAMP (Patel et al., 2021), and TABMWP (Lu et al., 2023).

3.4 Generalization Experiments

Our main experiments are conducted on DeepSeek-R1-Distill-Qwen-7B. To assess robustness beyond the primary setting, we further replicate CES on a smaller DeepSeek-R1-Distill-Qwen-1.5B backbone and evaluate the resulting models on out-of-domain coding and general-reasoning benchmarks. These additional results are reported in Appendix E.

4 Results

As shown in Table 1, CES demonstrates superior performance by achieving the best overall balance between accuracy and efficiency. On average across all 12 mathematical reasoning datasets, CES achieves the highest accuracy of 72.1% while simultaneously producing the shortest average response length of 1965 tokens. This represents a significant improvement over our primary baseline,

DAPO, with an average accuracy gain of +2.5% and a substantial average length reduction of 411 tokens.

CES learns to generate more effective and efficient reasoning paths across a wide spectrum of difficulties. For instance, on AIME24, a notoriously difficult competition-level dataset, CES boosts accuracy by a remarkable +6.7% while cutting the response length by 997 tokens. Similarly, on AMC23 and Olympiad Bench, CES achieves accuracy gains of +1.9% and +2.2% respectively, along with massive efficiency improvements, shortening the reasoning paths by 1014 and 839 tokens. This “win-win” outcome indicates that CES is not merely pruning the responses, but simultaneously improving the quality and directness of the model’s problem-solving strategies. In addition, on test sets such as CN Middle School 24 and GSM8K, it correctly identifies an opportunity where a modest investment in length (+36/+32 tokens) can yield a considerable gain in accuracy (+9.1%/+3.6%). This behavior shows that CES is not a naive length reduction algorithm but an intelligent controller that strategically allocates computational budget.

We also observe consistent robustness in the 1.5B backbone. We defer the detailed table to Appendix E.1.

5 Analysis

5.1 Training Dynamics

To gain deeper insight into the mechanism of CES, we analyze the evolution of key metrics throughout the training process. Figure 2 plots the average

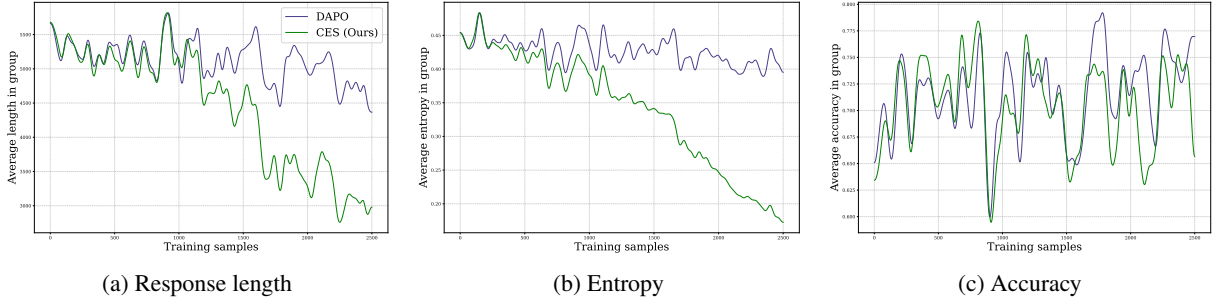


Figure 2: Training dynamics of average response length (a), entropy (b), and accuracy (c) for the DAPO baseline (blue) and our CES method (green).

response length, average token entropy and average group accuracy, comparing our CES-enhanced DAPO training against the standard DAPO baseline.

A striking pattern emerges in the Response Length and Entropy plots. For the first 1000 training samples, both the CES and baseline models exhibit similar behavior, maintaining a high and stable average length and entropy. This initial phase can be interpreted as the primary task acquisition stage, where both models are focused on learning the fundamental mechanics of solving the problems to achieve a reward. During this period, the policy is highly exploratory, and the CES mechanism has not yet become a dominant optimization force.

However, a clear divergence occurs after 1000 training samples. While the DAPO baseline’s length and entropy remain high and relatively constant, the CES model’s metrics begin a steep and consistent decline. The average response length drops from over 5000 to nearly 3000 tokens, and the average entropy falls from 0.4 to below 0.2. This second phase demonstrates the onset of CES’s core effect, where the entropy penalty on correct answers becomes a powerful and consistent training signal. The model learns that it can maximize its reward not just by being correct, but by being correct and confident. The strong correlation between the decline in entropy and length empirically validates our hypothesis that penalizing high-entropy “forking points” effectively prunes unnecessary, verbose exploration, leading to more concise reasoning paths.

In Figure 2(c), we observe that neither the baseline nor the CES model shows a significant, sustained upward trend in accuracy, with both curves showing similar fluctuations throughout training. This behavior is likely attributable to the limited size of our training set (2500 samples) and the

absence of a carefully designed data strategy. However, it also reflects that the improvements in efficiency (i.e., shorter length and lower entropy) achieved by CES are realized without sacrificing the model’s problem-solving performance. The CES model maintains an accuracy level competitive with the baseline, while operating at a significantly lower computational budget. In general, these dynamics reveal that CES successfully introduces a distinct optimization phase into training: after the initial task acquisition, it effectively teaches the model to become more efficient and decisive, achieving conciseness without compromising its learned reasoning capabilities.

5.2 Analysis of increasing response length on simple test sets

A notable observation from our main results is that while CES significantly shortens responses on most datasets, it increases the average response length on four specific datasets: CMATH, CN Middle School 24, GSM8K and TABMWP. A common characteristic of these datasets is their relatively short response lengths (typically under 1000 tokens) and higher performance, suggesting they are simpler overall. This phenomenon seems to contradict our goal of improving efficiency.

We hypothesize that this is a characteristic of adaptive reasoning manifested by CES. The key to understanding this lies in moving beyond dataset-level averages and analyzing model behavior on a finer-grained, per-question difficulty level. To test this, we stratified the questions within these four datasets into two categories based on the original R1-7B model’s performance:

1. “Simple Questions”: Questions where the R1-7B model’s accuracy is greater than 50%.
2. “Difficult Questions”: Questions where the

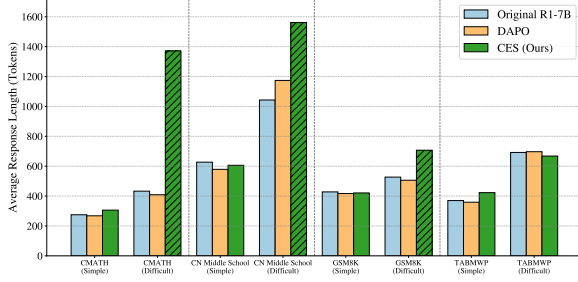


Figure 3: Comparison of average response length, stratified by question difficulty on four simpler datasets.

R1-7B model’s accuracy is less than or equal to 50%.

Figure 3 presents the average response length of both DAPO and CES models across these stratified difficulties. For “Difficult Questions” within these simpler datasets, CES triggers a significant increase in response length in most datasets. Conversely, for “Simple Questions”, the response lengths remain relatively stable or change minimally.

The failure mode of the original R1-7B model on these “difficult” questions may be insufficient exploration. Accustomed to the simple patterns of the dataset, it applies a short, inadequate template and fails. CES, through its mechanism of rewarding entropy on incorrect answers, correctly identifies these failures and provides a strong incentive for deeper exploration. It forces the model to abandon the failed template and invest more effort in finding a correct solution. In contrast, on complex datasets like Olympiad Bench, R1-7B’s failure mode is often inefficient overthinking, producing long, verbose, and incorrect reasoning. There, CES’s primary role is to prune this redundancy. In summary, the strategic investment in reasoning for difficult problems outweighs the minor length changes on simple ones, leading to an increase in the dataset’s overall average response length.

5.3 Ablation Studies

To validate the key components of our CES framework, we conduct two main ablation studies. These experiments are designed to investigate the importance of our dynamic token selection mechanism and the role of the entropy gradient in our advantage shaping formula.

5.3.1 The Importance of Dynamic Token Selection

A core feature of CES is the dynamic calculation of k , the number of high-entropy tokens to be shaped

Method	Acc $\uparrow$	Len $\downarrow$
Original R1-7B	69.1	2583
DAPO (Baseline)	69.6	2376
CES w/o Dynamic b	69.5	2462
CES w/o Entropy Gradient	69.4	2539
CES (Ours)	72.1	1965

Table 2: Ablation study on the core components of CES.

in each response. This number is modulated by a dynamic multiplier b (where $b = a$ for correct responses and $b = 1 - a$ for incorrect ones, with a being the group accuracy), which adjusts the intervention strength based on the perceived difficulty of the problem. To test the necessity of this design, we trained an ablated model, “RemoveAcc”, where we removed this dynamic multiplier by fixing $b = 1$. In this setting, a constant percentage of tokens with the highest entropy is always selected for entropy shaping, regardless of group accuracy.

The results shown in Table 2 indicate that the “RemoveAcc” model’s average accuracy drops to 69.5%, nearly identical to the DAPO baseline (69.6%) and significantly underperforming the full CES model (72.1%). Furthermore, its average response length increases to 2462, making it even less efficient than the DAPO baseline (2376).

While the behavior of a fixed $b = 1$ is identical to our dynamic b at the absolute extremes (when group accuracy $a = 1$ or $a = 0$), the critical difference emerges in the vast majority of training scenarios where the model’s performance is mixed ($0 < a < 1$). Consider a difficult problem where the model finds a correct solution for the first time, resulting in a low group accuracy (e.g., $a = 0.25$). Our full CES method applies a very gentle penalty, scaling the intervention by $b = a = 0.25$. This protects the newfound, likely inefficient reasoning path, acknowledging that it is a valuable success on a difficult problem. The “RemoveAcc” ablation, in contrast, applies the maximal penalty ($b = 1$). It aggressively punishes the high-entropy tokens in this fragile, correct solution, effectively signaling to the model that this “messy” path to success is undesirable. This can cause the model to discard the correct reasoning logic in subsequent updates, leading to performance degradation.

Therefore, the dynamic multiplier b acts as a crucial adaptive regularizer. It provides a proportional response: applying gentle, protective pres-

sure on novel solutions to difficult problems, while applying strong, optimizing pressure on mastered solutions to easy problems. By removing this calibrated intelligence, the “RemoveAcc” model fails, demonstrating that the dynamic selection of tokens is essential for robustly learning to be both accurate and efficient.

5.3.2 The Role of the Entropy Gradient in Bidirectional Control

In our CES formulation, the entropy term H is included in the computation graph, meaning the model’s policy is explicitly optimized to produce outputs that align with our entropy-based objectives. However, a related study (Cheng et al., 2026) that also uses an entropy-based advantage term introduces a “detach” operation in their implementation. This prevents the gradient of the entropy term from being computed, using it only to scale the magnitude of the existing policy gradient rather than setting a new optimization goal. To investigate this choice, we trained an ablated model, where we detached our entropy shaping term from the computation graph.

The results shown in Table 2 indicate that this change is detrimental to our method. The Detach model’s performance (69.4%) regresses to that of the DAPO baseline (69.6%) in accuracy, while its average length balloons to 2539, becoming the least efficient of all training configurations. The reason for this failure lies in the fundamental difference in goals between CES and the related method (Cheng et al., 2026). As their objective is unconditional exploration, detaching the entropy term serves as a clever way to amplify the existing policy updates at uncertain steps without asking the model to learn to be “more uncertain”.

However, CES has a dual, conditional objective. The “inhibit exploration” part of our mechanism ($A' \leftarrow A - \beta \cdot H$ for correct answers) is predicated on teaching the model to become more efficient by producing lower-entropy outputs. This requires a non-zero gradient so the model can learn to directly reduce entropy to avoid the penalty. Detaching the term completely breaks this crucial learning signal. Without the gradient, the penalty becomes a simple, static reduction in advantage that provides no direction for how to improve efficiency. This leads to the observed outcome: baseline accuracy with uncontrolled, verbose responses. Therefore, maintaining the entropy gradient is essential for the bidirectional control at the heart of CES to function

as intended.

6 Related Work

6.1 Reinforcement Learning for LLMs

Reinforcement learning is a core technique for aligning pretrained language models. Early RLHF pipelines commonly relied on Proximal Policy Optimization (PPO) (Schulman et al., 2017) with a separately trained reward model, while more recent work has shifted toward direct optimization methods to improve stability and simplify training. A representative example is Direct Preference Optimization (DPO) (Rafailov et al., 2023), which derives an optimization signal directly from preference data. This paradigm has been extended to reasoning settings that compare multiple responses to the same prompt, leading to algorithms such as GRPO and our baseline DAPO, which optimize policies using sequence-level preferences. Building on this line, our method CES introduces a more fine-grained mechanism by intervening at the token level and dynamically shaping the learning signal within the DAPO framework.

6.2 Entropy in LLMs

Entropy quantifies the uncertainty of a probability distribution. In LLMs, token-level entropy measures the uncertainty of the predicted distribution over the vocabulary at each generation step: higher entropy corresponds to a flatter distribution and lower confidence in selecting the next token (Li et al., 2026).

7 Conclusion

In this work, we address the fundamental challenge of balancing performance and efficiency in LLM reasoning. To resolve this trade-off, we propose CES, a framework that enables models to adapt their reasoning strategy: thinking concisely when confident, and reasoning deeply when uncertain. CES achieves consistent improvements in both accuracy and computational efficiency across diverse mathematical reasoning benchmarks, alleviating the inherent trade-off between exploration and exploitation.

Beyond empirical gains, this work suggests a broader principle: LLMs can learn not just to reason accurately, but to regulate how they reason. This opens directions for building fine-grained, resource-aware reasoning systems that require cost-sensitive inference.

Limitations

While CES achieves an average win–win by conditionally shaping token-level advantages with entropy, several limitations remain. First, the current formulation still relies on outcome-verifiable correctness signals to compute group accuracy a and to determine the direction of entropy shaping. As a result, applying the same mechanism to tasks with ambiguous, subjective, or weakly verifiable outcomes is less straightforward. Meanwhile, CES remains moderately sensitive to hyperparameters such as τ and β . In practice, the method is robust within a reasonable range, but achieving the best accuracy–efficiency balance may still require light calibration when transferring to a new backbone or task distribution.

Acknowledgments

This work is supported by the Fundamental Research Funds for the Central Universities (2025JBZX057), and the National Natural Science Foundation of China (No. 2576030).

References

- Pranjal Aggarwal and Sean Welleck. 2025. [L1: Controlling how long a reasoning model thinks with reinforcement learning](#). In *Proceedings of the 2nd Conference on Language Modeling*.
- Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc V. Le, and Charles Sutton. 2021. [Program synthesis with large language models](#). *arXiv preprint arXiv:2108.07732*.
- Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, and 39 others. 2021. [Evaluating large language models trained on code](#). *arXiv preprint arXiv:2107.03374*.
- 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 *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 9487–9499.
- Daixuan Cheng, Shaohan Huang, Xuekai Zhu, Bo Dai, Wayne Xin Zhao, Zhenliang Zhang, and Furu Wei. 2026. [Reasoning with exploration: An entropy perspective](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(36):30377–30385.
- Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. [Think you have solved question answering? try ARC, the AI2 reasoning challenge](#). *arXiv preprint arXiv:1803.05457*.
- 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](#). *arXiv preprint arXiv:2110.14168*.
- Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, Zhiyuan Liu, Hao Peng, Lei Bai, Wanli Ouyang, Yu Cheng, Bowen Zhou, and Ning Ding. 2025. [The entropy mechanism of reinforcement learning for reasoning language models](#). *arXiv preprint arXiv:2505.22617*.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 175 others. 2025. [DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning](#). *Nature*, 645(8081):633–638.
- 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.
- Zhiwei He, Tian Liang, Jiahao Xu, Qiuzhi Liu, Xingyu Chen, Yue Wang, Linfeng Song, Dian Yu, Zhenwen Liang, Wenxuan Wang, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. 2026. [DeepMath-103K: A large-scale, challenging, decontaminated, and verifiable mathematical dataset for advancing reasoning](#). In *International Conference on Learning Representations*, volume 2026, pages 138306–138322.
- Jian Hu, Xibin Wu, Wei Shen, Jason Klein Liu, Weixun Wang, Songlin Jiang, Haoran Wang, Hao Chen, Bin Chen, Wenkai Fang, Xianyu, Yu Cao, Haotian Xu, and Yiming Liu. 2025. [OpenRLHF: A ray-based easy-to-use, scalable and high-performance RLHF framework](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 656–666, Suzhou, China. Association for Computational Linguistics.

- 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.
- Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. 2022. [Solving quantitative reasoning problems with language models](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 3843–3857.
- Xianzhi Li, Ethan Callanan, Abdellah Ghassel, and Xiaodan Zhu. 2026. [Entropy-gated branching for efficient test-time reasoning](#). In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5054–5069, Rabat, Morocco. Association for Computational Linguistics.
- Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. 2023. [Is your code generated by ChatGPT really correct? rigorous evaluation of large language models for code generation](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 21558–21572.
- Chenwei Lou, Zewei Sun, Xinnian Liang, Meng Qu, Wei Shen, Wenqi Wang, Yuntao Li, Qingping Yang, and Shuangzhi Wu. 2025. [AdaCoT: Pareto-optimal adaptive chain-of-thought triggering via reinforcement learning](#). *arXiv preprint arXiv:2505.11896*.
- Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. 2023. [Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning](#). In *The Eleventh International Conference on Learning Representations*.
- Haotian Luo, Haiying He, Yibo Wang, Jinluan Yang, Rui Liu, Naiqiang Tan, Xiaochun Cao, Dacheng Tao, and Li Shen. 2025. [Ada-R1: Hybrid-CoT via bi-level adaptive reasoning optimization](#). In *Advances in Neural Information Processing Systems*, volume 38, pages 59353–59377.
- Wenjie Ma, Jingxuan He, Charlie Snell, Tyler Griggs, Sewon Min, and Matei Zaharia. 2025. [Reasoning models can be effective without thinking](#). *arXiv preprint arXiv:2504.09858*.
- Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. [Can a suit of armor conduct electricity? a new dataset for open book question answering](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.
- Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. 2025. [s1: Simple test-time scaling](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20275–20321, Suzhou, China. Association for Computational Linguistics.
- OpenAI. 2025. OpenAI o3 and o4-mini System Card. <https://openai.com/index/o3-o4-mini-system-card/>. Accessed: 2026-08-25.
- Arkil Patel, Satwik Bhattamishra, and Navin Goyal. 2021. [Are NLP models really able to solve simple math word problems?](#) In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2080–2094, Online. Association for Computational Linguistics.
- Qwen Team. 2024. [Qwen2.5-Math evaluation code](#). GitHub repository. Commit 9e9bc401; accessed 2026-08-26.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. [Direct preference optimization: Your language model is secretly a reward model](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 53728–53741.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. [Proximal policy optimization algorithms](#). *arXiv preprint arXiv:1707.06347*.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. [DeepSeekMath: Pushing the limits of mathematical reasoning in open language models](#). *arXiv preprint arXiv:2402.03300*.
- Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. [CommonsenseQA: A question answering challenge targeting commonsense knowledge](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. 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 *Advances in Neural Information Processing Systems*, volume 38, pages 115452–115486.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. [Chain-of-thought prompting elicits reasoning in large language models](#). In

Tianwen Wei, Jian Luan, Wei Liu, Shuang Dong, and Bin Wang. 2023. [CMATH: Can your language model pass chinese elementary school math test?](#) *arXiv preprint arXiv:2306.16636*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025a. [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](#). *arXiv preprint arXiv:2409.12122*.

Chenxiao Yang, Nathan Srebro, David McAllester, and Zhiyuan Li. 2025b. [PENCIL: Long thoughts with short memory](#). In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 71092–71156.

Qiyang Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gao-hong Liu, Juncai Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, and 17 others. 2025. [DAPO: An open-source LLM reinforcement learning system at scale](#). In *Advances in Neural Information Processing Systems*, volume 38, pages 113222–113244.

Jixiao Zhang and Chunsheng Zuo. 2025. [GRPO-LEAD: A difficulty-aware reinforcement learning approach for concise mathematical reasoning in language models](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 5642–5654, Suzhou, China. Association for Computational Linguistics.

A Algorithm

Algorithm 1 details the complete procedure for implementing Conditional Entropy Shaping (CES) within the DAPO framework.

B Hyperparameters and Prompt

B.1 Training hyperparameters

Table 3 lists the hyperparameters for our reinforcement learning experiments.

Hyperparameter	Value
<i>General Training Parameters</i>	
Optimizer	Adam
Learning Rate	2×10^{-7}
Training Global Batch Size	4
Rollout Batch Size	12
Samples per Prompt	4
<i>Generation Parameters</i>	
Rollout Temperature	1.0
Rollout Top- p	1.0
Maximum Generation Length	12000
<i>CES-Specific Parameters</i>	
Top-rate Hyperparameter (τ)	0.01
Entropy Scaling Factor (β_1, β_2)	0.4

Table 3: Hyperparameters for RL training.

B.2 Evaluation prompt

For all evaluation scenarios, we used the following standardized prompt to ensure the model generates answers in a step-by-step manner and formats the final result correctly:

“*You are a helpful and harmless assistant. You should think step-by-step. Please put your final answer within \boxed{ }.*”

C Sensitivity to key hyperparameters

To investigate the sensitivity of CES to its core hyperparameters and validate the robustness of CES, we conduct an ablation study on the top-rate τ and the entropy scaling factor β_1, β_2 . We evaluate five different hyperparameter configurations on a representative subset of five datasets and compare their average performance against the DAPO baseline. The results are summarized in Table 4.

C.1 Analysis of Top-rate τ

The hyperparameter τ controls the proportion of selected high-entropy tokens. With β_1, β_2 fixed at 1.0, we tested τ values of 0.005, 0.01, and 0.05. The results indicate that a smaller, more targeted intervention is more effective. As τ increases from 0.01 to 0.05, the average accuracy drops sharply from 74.2% to 70.7%, falling below the DAPO baseline. This suggests that selecting too many tokens introduces noise by including tokens that are not critical “forking points”, thereby diluting the learning signal and degrading the policy.

Algorithm 1 CES within DAPO Framework

Input: Prompt x , current policy π_θ

Parameters: Generation group size N , top-rate hyperparameter τ , entropy scaling factors β_1, β_2

Output: A set of shaped, token-level advantages $\mathcal{A}'$ for gradient update

```
1: Generate response set  $Y = \{y_1, \dots, y_N\}$  from  $\pi_\theta(\cdot|x)$ .
2: Compute rewards  $R(y_i), r_{\text{acc}}(y_i)$  for each  $y_i \in Y$ .
3: Compute group accuracy  $a$  based on  $\{r_{\text{acc}}(y_i)\}$ .
4: Initialize set of all shaped advantages  $\mathcal{A}' \leftarrow \emptyset$ .
5: for each response  $y_i$  in  $Y$  do
6:    $A_i \leftarrow \text{GroupNormalize}(\{R(y_j)\}, R(y_i))$ .
7:   if  $r_{\text{acc}}(y_i) = 1$  then
8:      $b_i \leftarrow a$ 
9:   else
10:     $b_i \leftarrow 1 - a$ 
11:   end if
12:   Compute number of tokens to select  $k_i = \lfloor |y_i| \cdot \tau \cdot b_i \rfloor$ .
13:    $S_H(y_i) \leftarrow$  Identify top  $k_i$  high-entropy tokens in  $y_i$ .
14:   for each token  $t_j$  in  $y_i$  do
15:      $A'_{i,j} \leftarrow A_i$  {Initialize with base advantage}
16:     if  $t_j \in S_H(y_i)$  then
17:        $H_j \leftarrow H(t_j|y_{i,<j})$  {Calculate entropy}
18:       if  $r_{\text{acc}}(y_i) = 1$  then
19:          $A'_{i,j} \leftarrow A_i - \beta_1 \cdot H_j$  {Apply entropy penalty}
20:       else
21:          $A'_{i,j} \leftarrow A_i + \beta_2 \cdot H_j$  {Apply entropy reward}
22:       end if
23:     end if
24:     Add  $A'_{i,j}$  to  $\mathcal{A}'$ .
25:   end for
26: end for
27: return  $\mathcal{A}'$  {Return token-level shaped advantages  $\mathcal{A}'$  for computing policy gradients in DAPO}
```

C.2 Analysis of Entropy Scaling Factor β_1, β_2

The hyperparameters β_1, β_2 indicate the scaling magnitude of the entropy reward and penalty. With τ fixed at 0.01, we tested β_1, β_2 values of 0.4, 1.0, and 2.0. The results show a clear trend: as β_1, β_2 increases, the average accuracy decreases while average response length increases. A larger setting on β_1, β_2 gives excessive weight to the entropy shaping term, particularly the exploratory reward on incorrect answers. This can cause the model to over-optimize for the process of exploration rather than the outcome of correctness, leading to longer, less focused reasoning chains that do not necessarily improve accuracy.

C.3 Robustness of CES

Across four of the five tested hyperparameter settings, our method simultaneously outperforms the

DAPO baseline in both accuracy and length. This demonstrates that CES provides consistent benefits across a reasonable range of hyperparameters, validating it as a stable and effective method for improving reasoning models.

D GPU Cost

A natural concern is whether CES introduces noticeable additional computation during training. Compared with vanilla DAPO, CES indeed adds two operations: token-level entropy computation and selection/shaping of high-entropy tokens. However, these additions do not require extra model forward or backward passes. In practice, entropy is computed directly from the logits that are already produced during rollout sampling, so the overhead is limited to lightweight vector reductions and top-k selection rather than additional Transformer back-

τ	β_1, β_2	Acc $\uparrow$	Len $\downarrow$
0.005	1.0	75.1	2855
0.01	1.0	74.2	2992
0.05	1.0	70.7	2818
0.01	0.4	76.9	2757
0.01	1.0	74.2	2992
0.01	2.0	73.4	2997
DAPO (Baseline)		72.6	3407

Table 4: Hyperparameter sensitivity analysis for CES on the average of 5 datasets (AIME24, AMC23, GaoKao Math Cloze, GaoKao 2023 En and SVAMP). The optimal configuration is highlighted in **bold**.

Backbone	DAPO	CES
R1-7B	1.43 days	1.44 days
R1-1.5B	9.36 hours	9.14 hours

Table 5: Training GPU wall-clock time under the same setup.

bone computation.

Results in Table 5 indicate that the additional GPU time overhead of CES is negligible. For DeepSeek-R1-Distill-Qwen-7B, DAPO takes 1.43 days, while CES takes 1.44 days, corresponding to only about +0.7% relative overhead. For DeepSeek-R1-Distill-Qwen-1.5B, DAPO takes 9.36 hours, while CES takes 9.14 hours, making CES slightly faster by about -2.3%. These results suggest that the small constant-time overhead of entropy computation is largely offset by shorter rollouts during training.

E Generalization Experiments

Although the main paper trains only on math data, the mechanism of CES is not inherently math-specific. It relies on token-level uncertainty (entropy) and correctness-conditioned shaping, which are **domain-agnostic signals**. We therefore evaluate generalization from two perspectives:

1. Across **model scales**.
2. Across **domains**.

We find that the benefits of CES are not limited to the original 7B math setting, but generalize to a smaller backbone and to out-of-domain tasks in both general reasoning and code generation.

E.1 Cross-Scale Generalization

To test whether the benefits of CES depend on a single backbone scale, we additionally train DeepSeek-R1-Distill-Qwen-1.5B under the same training protocol as the 7B model, and evaluate it on the same 12 math benchmarks.

Results in Table 6 show that CES remains effective on the 1.5B backbone. The average accuracy improves from 52.5 to 56.2, while the average response length is reduced from 3581 to 3283. Meanwhile, CES improves accuracy on most of the 12 benchmarks, and also shortens responses on the majority of them. Although a few benchmarks exhibit small length increases or minor accuracy fluctuations, the overall average still shows a clear win-win trend.

These results indicate that the benefit of CES is not tied to the 7B setting. When the backbone is scaled down to 1.5B, CES still consistently improves the accuracy–efficiency trade-off, suggesting that it is a generally useful training mechanism rather than a technique specific to a single model size.

E.2 Cross-Domain Generalization: General Reasoning

To further test whether CES generalizes beyond the training distribution, we evaluate it on three general-reasoning benchmarks: ARC-Challenge (Clark et al., 2018), CommonsenseQA (Talmor et al., 2019), and OpenBookQA (Mihaylov et al., 2018). Importantly, these datasets are **outside the training domain**, since training uses only math data. Results in Table 7 show that the gains of CES do not simply come from forcing the model to generate shorter outputs. On ARC and CommonsenseQA, CES substantially improves accuracy while keeping the output length almost unchanged. On OpenBookQA, CES spends slightly more tokens in exchange for a meaningful gain in accuracy. In other words, CES does not learn a fixed preference for shorter responses; instead, it learns to allocate reasoning budget on demand. It permits additional reasoning when that helps correctness, and suppresses redundant exploration when it does not.

Therefore, these results suggest that the adaptive reasoning behavior learned by CES is not limited to math, but transfers to broader knowledge and commonsense reasoning tasks.

Dataset	DAPO (Baseline)		CES (Ours)		Improvement	
	Acc ↑	Len ↓	Acc ↑	Len ↓	Acc ↑	Len ↓
AIME24	6.7	10254	13.3	8751	+6.6	-1503
AMC23	47.5	7291	55.0	6566	+7.5	-725
CMATH	84.7	616	85.3	676	+0.6	+60
CN Middle School	59.4	794	61.4	758	+2.0	-36
College Math	34.1	2907	33.2	2866	-0.9	-41
GaoKao Math Cloze	46.6	5448	61.0	4655	+14.4	-793
GaoKao 2023 En	57.7	3304	60.8	3173	+3.1	-131
GSM8K	77.3	753	78.6	1022	+1.3	+269
Minerva Math	23.5	3418	29.0	3289	+5.5	-129
Olympiad Bench	27.7	6868	32.0	6263	+4.3	-605
SVAMP	85.6	618	86.4	620	+0.8	+2
TABMWP	79.2	699	78.1	759	-1.1	+60
Average	52.5	3581	56.2	3283	+3.7	-298

Table 6: Cross-scale generalization on DeepSeek-R1-Distill-Qwen-1.5B. The best result in each category is in **bold**. “Acc” and “Len” denote the mean accuracy and the mean response length across 4 assessments for each benchmark.

Benchmark	DAPO		CES	
	Acc ↑	Len ↓	Acc ↑	Len ↓
ARC	79.7	543	83.6	542
CommonsenseQA	56.5	451	65.8	448
OpenBookQA	68.2	503	74.8	520

Table 7: Cross-domain generalization on general reasoning benchmarks. The best result in each category is in **bold**.

E.3 Cross-Domain Generalization: Coding Benchmarks

In addition to general reasoning, we also evaluate on code-generation benchmarks using EvalPlus (Liu et al., 2023). Specifically, we test on HumanEval (Chen et al., 2021) and MBPP (Austin et al., 2021), together with their HumanEval+ and MBPP+ variants, where the “+” versions include stricter extra tests in addition to the original base tests.

Benchmark	DAPO		CES	
	Acc ↑	Len ↓	Acc ↑	Len ↓
HumanEval	46.3	4158	49.4	3911
HumanEval+	43.9	4158	48.2	3911
MBPP	42.9	2931	45.0	2824
MBPP+	35.7	2931	39.4	2824

Table 8: Cross-domain generalization on coding benchmarks. The best result in each category is in **bold**.

Results are shown in Table 8, where CES outperforms DAPO on all four coding metrics. These results provide strong evidence of generalization, since code generation differs substantially from math reasoning in output format, structural con-

straints, and failure modes. Nevertheless, CES still improves both pass rate and token efficiency, suggesting that its optimization signal is task-agnostic. In addition, the improvements on HumanEval+ and MBPP+ indicate that the gains are robust under stricter extra tests, rather than appearing only on easier evaluation settings. Finally, the reduced generation length shows that CES does not improve coding results by “thinking longer”, but by producing higher-quality solutions more efficiently.

Taken together, the coding results further support that CES generalizes beyond the math training domain to a structurally different reasoning task.

F Statistics of Simple vs. Hard Cases

Dataset	Simple	Hard	Total
CMATH	534	66	600
CN Middle School	60	41	101
GSM8K	1121	198	1319
TABMWP	870	130	1000

Table 9: Statistics of simple and hard cases on several representative benchmarks. Following the definition in the main text, a question is classified as *simple* if the original R1-7B achieves accuracy greater than 50% on that question; otherwise, it is classified as *hard*.

Section 5.2 of the main paper notes that on a few relatively simple datasets, CES slightly increases the average response length. We report the numbers of simple and hard cases within these datasets in Table 9.