FastCuRL: Curriculum Reinforcement Learning with Stage-wise Context Scaling for Efficient Training R1-like Reasoning Models

Anonymous ACL submission

Abstract

Improving training efficiency continues to be one of the primary challenges in large-scale Reinforcement Learning (RL). In this paper, we investigate how ***context length*** and ***the complexity of training data*** influence *the scaling RL training process of R1-distilled small reasoning models, e.g., DeepSeek-R1-Distill-Qwen-1.5B. Our experimental results reveal that: (1) simply controlling the context length and curating the training data based on the input prompt length can effectively improve the training efficiency of scaling RL, achieving better performance with more concise CoT; (2) properly scaling the context length helps mitigate entropy collapse; and (3) choosing an optimal context length can improve the efficiency of model training and incentivize the model's chain-of-thought reasoning capabilities*. Inspired by these insights, we propose **FastCuRL**, a curriculum RL framework with stage-wise context scaling to achieve efficient training and concise CoT reasoning. Experiment results demonstrate that **FastCuRL-1.5B-V3** significantly outperforms state-of-the-art reasoning models on five competition-level benchmarks and achieves 49.6% accuracy on AIME 2024. Furthermore, **FastCuRL-1.5B-Preview** surpasses DeepScaleR-1.5B-Preview on five benchmarks while only using a single node with 8 GPUs and a total of 50% of training steps. The code, training data, and models will be publicly released.

1 Introduction

Large Language Models (LLMs) have emerged as immensely potent AI instruments, showcasing extraordinary proficiency in comprehending natural language and executing downstream tasks (Zhao et al., 2023; Minaee et al., 2024; Chen et al., 2025). Lately, test-time scaling (Snell et al., 2024; Muennighoff et al., 2025) has demonstrated a robust correlation between extending the generation length of Chain-of-Thought (CoT) (Wei et al., 2023) and improving the reasoning capabilities of LLMs.

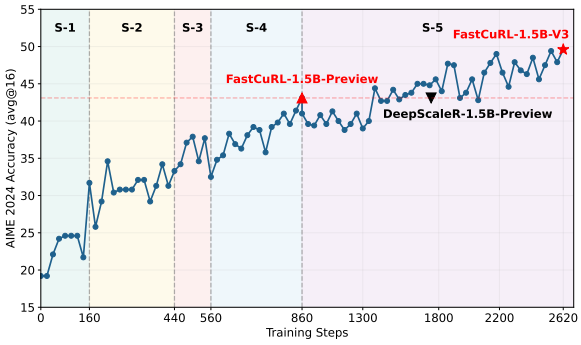

Figure 1: FastCuRL's accuracy on AIME 2024 as training progresses across five training stages. Specifically, S-5 indicates Stage 5 in the training process.

A primary finding from recent breakthroughs, exemplified by DeepSeek-R1 (DeepSeek-AI, 2025), reveals a scaling phenomenon in the training process of Reinforcement Learning (RL). Inspired by these findings, training LLMs through scaling RL has recently emerged as a promising paradigm for addressing complex reasoning tasks and many valuable research endeavours (Luo et al., 2025; Face, 2025; Hu et al., 2025; Zeng et al., 2025a; Liu et al., 2025) have emerged to explore and replicate reasoning models akin to DeepSeek-R1 (for example, starting from R1-distilled or pre-trained models) by extending the generation length of CoT.

However, generating excessively long CoT responses significantly increases computational overhead during model training and deployment. Moreover, recent studies (Yeo et al., 2025; Wu et al., 2025; Team, 2025a; Luo et al., 2025) have identified an inherent overthinking phenomenon in reasoning models, which includes irrelevant details and repetitive thinking patterns. This kind of information leads to inefficient use of computational resources and undermines reasoning accuracy, which causes models to stray from valid logical pathways, resulting in incorrect answers.

To this end, recent studies (Team, 2025a; Luo

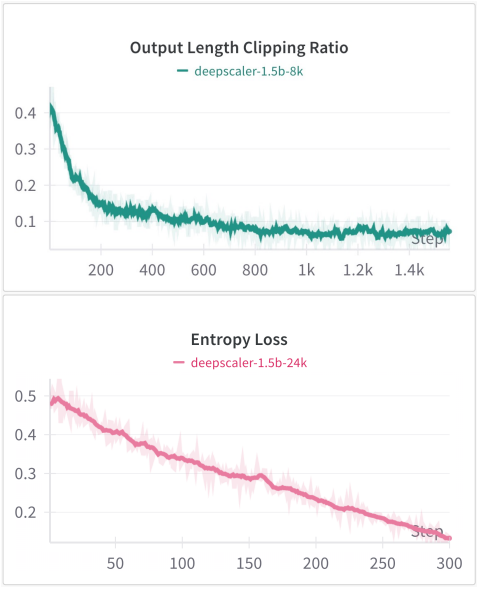

Figure 2: The training logs of DeepScaleR.

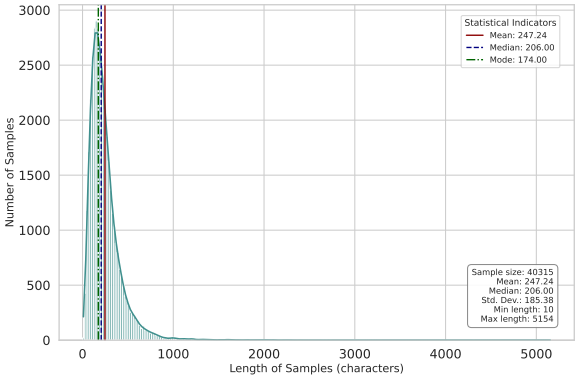

Figure 3: Prompt length distribution.

et al., 2025; Liu et al., 2025; Yu et al., 2025) focus on efficient reasoning for optimizing the model to generate more concise solutions. Among them, DeepScaleR (Luo et al., 2025) propose to iteratively increase the context length from 8K to 24K to train the DEEPSEEK-R1-DISTILL-QWEN-1.5B model toward more concise reasoning, outperforming OpenAI's o1-preview (OpenAI, 2024). By observing the training logs[1] of DeepScaleR in Figure 2, we find two issues:

- When the context length is 8K, about 42% of the model's outputs are clipped, which reduces the model's training efficiency.
- When the context length is 24K, the model's entropy collapses. Entropy reflects the exploration capability of an LLM during training. A rapid decrease in entropy might lead to premature convergence, preventing the model from achieving the expected performance.

The prior work and the aforementioned issues naturally motivate two research questions:

- ***Question 1***: *Does simultaneously controlling the model's context length and the complexity of the training dataset help the training process of R1-like reasoning models?*
- ***Question 2***: *What impact does setting different context lengths have on the RL training process of R1-like reasoning models?*

[1] https://github.com/agentica-project/rllm

To this end, in this paper, we investigate how the model's context length and the complexity of the training dataset influence the training process of R1-like reasoning models. Motivated by our observations, we propose **FASTCURL**, *a simple yet efficient* **Cu**rriculum **R**einforcement **L**earning framework with a stage-wise context scaling strategy to improve the RL training efficiency and achieve concise CoT reasoning for R1-like reasoning models. Experimental results demonstrate that our model **FASTCURL-1.5B-V3** outperforms recent state-of-the-art reasoning baselines across five competition-level benchmarks, AIME 2024, AMC 2023, MATH 500, Minerva Math, and OlympiadBench. Furthermore, our model **FASTCURL-1.5B-Preview** surpasses DeepScaleR-1.5B-Preview on five benchmarks and only uses 50% training steps on a single node with 8 GPUs. We hope the findings presented in this paper, the models we have released, and the open-sourced code will benefit future research.

2 Methodology

In this section, we introduce our investigation into how the model's context length and the complexity of training data influence the training process of R1-like reasoning models. Specifically, our method consists of two main components: (1) curating a complexity-aware, mathematics-focused dataset, and (2) implementing a resource-efficient reinforcement learning algorithm. These two components aim to balance a trade-off between achieving performance improvements and addressing practical limitations, such as reducing computational costs.

2.1 Complexity-Aware Data Curation

To ensure a fair comparison, we directly employ the dataset from DeepScaleR as the training data. The DeepScaleR dataset (Luo et al., 2025) consists

Example Problem *(Output Length=74706 characters)*: Ashley, Betty, Carlos, Dick, and Elgin went shopping. Each had a whole number of dollars to spend, and together they had 56 dollars. The absolute difference between the amounts Ashley and Betty had to spend was 19 dollars. The absolute difference between the amounts Betty and Carlos had was 7 dollars, between Carlos and Dick was 5 dollars, between Dick and Elgin was 4 dollars, and between Elgin and Ashley was 11 dollars. How many dollars did Elgin have?

Table 1: Example problem.

of 40,315 unique mathematics-specific problem-answer pairs collected from AIME (1984-2023), AMC (prior to 2023), Omni-MATH, and the Still dataset (Balunović et al., 2025; Gao et al., 2024; Min et al., 2024). The statistics of the DeepScaleR dataset are shown in Figure 3.

As illustrated in Figure 2, over 42% of training samples are clipped at the beginning of the training steps due to exceeding the maximum response length. By observing and analyzing the clipped responses, we find that they mainly correspond to two types of problems. The first type pertains to challenging problems requiring long CoT responses to solve. The second involves questions laden with numerous conditions, prompting the model to verify each condition repeatedly during problem-solving, e.g., the problem shown in Table 1. This repetitive verification may result in redundant thinking patterns, ultimately causing the reasoning responses to be unduly long. Both situations may impact the model's training efficiency during the 8K context.

After observing the above phenomenon, we utilize DEEPSEEK-R1-DISTILL-QWEN-1.5B to infer all the training data of DeepScaleR to obtain responses and analyze the response lengths, as shown in Figure 4. Specifically, the given figure examines the relationship between input length and output length. Interestingly, we find a correlation between the two-that is, the longer the input, the longer the corresponding output. Based on this observation, we assume a hypothesis that for complex reasoning tasks, there exists a relationship between the complexity of the problem prompt and the length of the output response generated by the model when solving it. Generally, the more complex the problem, the longer the output the model needs to produce to arrive at a solution. Based on this hypothesis, we directly divide the original training dataset (referred to as **L**2) into two training data subsets based on the average input prompt length: one representing a short CoT reasoning dataset (designated as **L**1) and the other constituting a long CoT reasoning dataset (labeled as **L**3). Finally, the average input length of each dataset as shown in Table 2.

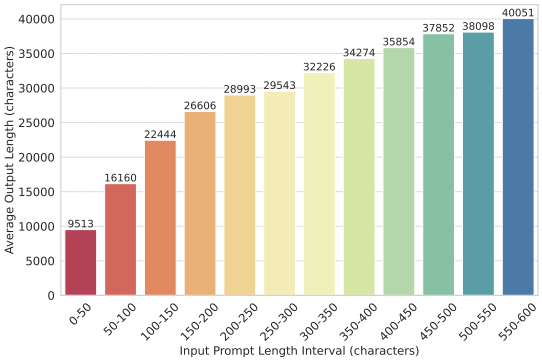

Figure 4: Relationship between input prompt length and output length of the training data. The output results are obtained from DEEPSEEK-R1-DISTILL-QWEN-1.5B.

Datasets	Average Input Prompt Length
L1	148.65
L2	247.24
L3	407.78

Table 2: Statistics of **L**1, **L**2, **L**3 datasets.

Next, we conduct experiments and analyses on these three datasets under different context lengths to observe and investigate the two questions raised in the prior section. It is important to note that this paper focuses on low-resource scenarios. Therefore, during training, when using different datasets at each stage, we train for only one epoch and utilize a single node with 8 GPUs.

2.2 Reinforcement Learning Algorithm

To train our model efficiently, we adopt the Group Relative Policy Optimization (GRPO) (Shao et al., 2024), which is utilized in DeepSeek-AI (2025). GRPO eliminates the necessity of maintaining a critic model, which is usually comparable in size to the policy model, by estimating baseline scores directly from group-level scores, significantly lowering the computational overhead. For each problem q, GRPO directly samples a group of G responses $\{o_1, o_2, ..., o_G\}$ from the old policy $\pi_{\theta_{\text{old}}}$ and optimizes the trained policy π_θ by maximizing the

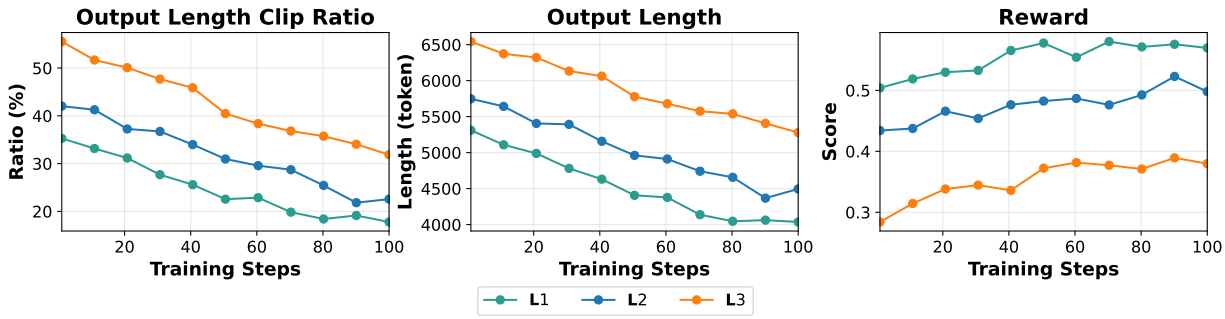

Figure 5: The average output length clip ratio, output length, and reward during Stage 1 in RL training on **L**1, **L**2, and **L**3 datasets. The curves shows the running average over a window size of 10.

following objective:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{[q\sim P(Q),\{o_i\}_{i=1}^{\text{G}}\sim\pi_{\theta_{\text{old}}}(\cdot|q)]} \frac{1}{\text{G}}\sum_{i=1}^{\text{G}}\left(\min\left(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{\text{old}}}(o_i|q)}\text{A}_i,\right.\right. \text{clip}\left(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{\text{old}}}(o_i|q)}, 1-\varepsilon, 1+\varepsilon\right)\text{A}_i\Bigg) \left.\left. -\beta\mathbb{D}_{\text{KL}}[\pi_\theta||\pi_{\text{ref}}] + \alpha\mathbb{H}(\pi_\theta(o_i|q))\right)\right, \quad (1)$$

where the advantage A_i is computed from a group of rewards $\{r_1, r_2, ..., r_{\text{G}}\}$:

$$\text{A}_i = \frac{r_i - \text{mean}(\{r_1, r_2, ..., r_G\})}{\text{std}(\{r_1, r_2, ..., r_{\text{G}}\})}. \quad (2)$$

Similar to the prior work (DeepSeek-AI, 2025; Luo et al., 2025), we leverage a rule-based reward model composed of two distinct criteria designed to balance answer correctness and clarity of structure without relying on an LLM-based reward model. To evaluate correctness objectively, we require the trained model to present its final answer enclosed within a **\boxed{}** format, assigning a binary score of 1 for correct answers and 0 for incorrect ones. To encourage structural clarity, the model must explicitly encapsulate its reasoning within tags, with compliance being rewarded positively.

3 Experiments

To investigate the research question described in Section 1—namely, how the model's context length and the complexity of the training data influence the RL training process of R1-like reasoning models—we designed a set of experiments under computational resource constraints. We aim to analyze the training behavior of small LLMs and find practical insights. These experiments are intended not only to provide empirical evidence of performance gains but also to offer clear and actionable guidance for both future academic research and practical industry implementations.

3.1 Experimental Setup

In this work, we choose a 1.5B parameter model DEEPSEEK-R1-DISTILL-QWEN-1.5B (DeepSeek-AI, 2025) as the base model. We utilize the AdamW optimizer with a constant learning rate of 1×10^{-6} for optimization. For rollout, we set the temperature to 0.6 and sample 16 responses per prompt. We do not utilize a system prompt; instead, we add "Let's think step by step and output the final answer within **\boxed{}**." at the end of each problem. Detailed parameters are shown in Figure 3.

3.2 Benchmarks

To comprehensively evaluate the performance, we select five competition-level benchmarks: MATH 500 (Hendrycks et al., 2021), AIME 2024[2], AMC 2023[3], Minerva Math (Lewkowycz et al., 2022), and OlympiadBench (He et al., 2024).

3.3 Baselines

In this paper, we conduct evaluations against 1.5B and 7B parameter language models, which includes DEEPSEEK-R1-DISTILL-QWEN-1.5B (DeepSeek-AI, 2025), QWEN2.5-MATH-7B-Instruct (Yang et al., 2024), DeepScaleR-1.5B-Preview (Luo et al., 2025), QWEN2.5-7B-SimpleRL (Zeng et al., 2025a), RSTAR-MATH-7B (Guan et al., 2025), STILL-3-1.5B-Preview (Team, 2025b), and EURUS-2-7B-PRIME (Cui et al., 2025).

[2] https://huggingface.co/datasets/AI-MO/aimo-validation-aime

[3] https://huggingface.co/datasets/AI-MO/aimo-validation-amc

Experiments	Stages	Context Length		Training Data	Batch Size	Rollout	α (for Entropy)	β (for KL)	Avg.
		Input	Output (K)						
Exp-1	3	1K	8, 16, 24	**L**1, **L**2, **L**3	128, 64, 64				0.550
Exp-2	3	1K	8, 16, 24	**L**1, **L**3, **L**2	128, 64, 64	8, 8, 8	0.001	0.001	0.540
Exp-3	3	1K	8, 16, 24	**L**1, **L**2, **L**2	128, 64, 64				0.552
Exp-4	4	1K	8, 16, 24, **32**	**L**1, **L**2, **L**3, **L**2	128, 64, 64, 64				0.566
Exp-5	4	1K	8, 16, 24, **24**	**L**1, **L**2, **L**3, **L**2	128, 64, 64, 64	8, 8, 8, 16	0.001	0.001	0.565
Exp-6	4	1K	8, 16, 24, **16**	**L**1, **L**2, **L**3, **L**2	128, 64, 64, 64				0.575
Exp-7	5	1K	8, 16, 24, 16, **24**	**L**1, **L**2, **L**3, **L**2, **L**2	128, 64, 64, 64, 64				0.556
Exp-8	5	1K	8, 16, 24, 16, **16**	**L**1, **L**2, **L**3, **L**2, **L**2	128, 64, 64, 64, 64	8, 8, 8, 16, 16	0.001	0.001	0.567
Exp-9	5	1K	8, 16, 24, 16, **8**	**L**1, **L**2, **L**3, **L**2, **L**2	128, 64, 64, 64, 64				0.535
Exp-10	5	1K	8, 16, 24, 16, **16**	**L**1, **L**2, **L**3, **L**2, **L**2	128, 64, 64, 64, 64	8, 8, 8, 16, 16	**0.000001**	**0.000**	0.600
Exp-11	5	1K	8, 16, 24, 16, **16**	**L**1, **L**2, **L**3, **L**2, **L**2	128, 64, 64, 64, 64	8, 8, 8, 16, 16	**0.000**	**0.000**	0.616

Table 3: Experimental setups combining different context lengths and data complexities.

3.4 Evaluation Metric

Following the prior work (DeepSeek-AI, 2025), we set the maximum context length to 32,768 tokens and use Pass@1 as the evaluation metric. Specifically, we adopt **a sampling temperature of 0.6** and **a top-p value of 1.0** to generate k responses for each question, typically $k = 16$. Specifically, Pass@1 is then calculated as:

$$\text{Pass@1} = \frac{1}{k}\sum_{i=1}^{k} p_i, \tag{3}$$

where p_i is the correctness of the i-th response.

3.5 Main Processes and Results

In this section, we first validate the effectiveness of the complexity-aware data curation strategy. Then, we design a series of progressive experiments with varying context lengths and data complexities and analyze the experimental results.

3.5.1 Dataset Complexity Verification

To validate the effectiveness of complexity-aware data curation, we train three models with the same setting on **L**1, **L**2, and **L**3 under the 8K context length as seen from Figure 5, whether the experiment results meet expectations in clipping ratio, response length, and reward scores. These experimental results support our hypothesis that the more complex the problem, the longer the output the model needs to produce to arrive at a solution.

3.5.2 Multi-Stage Experimental Results

We conduct three sets of multi-stage experiments, with specific parameter settings shown in Table 3. These experiments include ones with 3, 4, and 5 training stages, respectively. The experimental results are presented in Table 3.

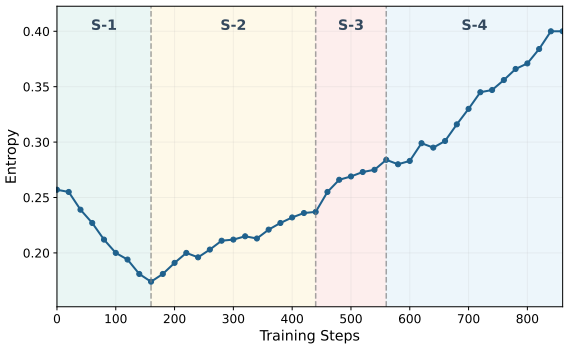

Figure 6: Entropy curves of FastCuRL-1.5B-Preview as training progresses across four training stages. Specifically, S-4 indicates Stage 4 in the training. The curve shows the running average over a window size of 10.

For the first set of experiments, Exp-3 achieves better performance compared to Exp-1, but it required more training steps (Exp-3 is trained based on the **L**2 dataset twice). Therefore, by comparing the differences in effectiveness and the computational cost in terms of training steps, we select Exp-1 as the output of the first stage and adopt it as the base model for the second stage.

In the first set of experiments, we observe that the average response length in its final stage is between 6,000 and 7,000 tokens. Therefore, we test context lengths that are longer, shorter, and equal to the 24K context length. As shown in Table 3, setting the context length to 16K yielded the best performance, rather than longer contexts of 24K or 32K tokens. Therefore, we select Exp-6 as the base model for the third stage.

Inspired by the second set of experiments, we conduct a third set in which we set the context lengths to 24K, 16K, and 8 K. As shown in Table 3, the 16K context still achieves the best performance,

Model	MATH 500	AIME 2024	AMC 2023	Minerva Math	OlympiadBench	Avg.
Qwen2.5-Math-7B-Instruct	79.8	13.3	50.6	34.6	40.7	43.8
rStar-Math-7B	78.4	26.7	47.5	-	47.1	-
Eurus-2-7B-PRIME	79.2	26.7	57.8	38.6	42.1	48.9
Qwen2.5-7B-SimpleRL	82.4	26.7	62.5	39.7	43.3	50.9
DeepSeek-R1-Distill-Qwen-1.5B	82.8	28.8	62.9	26.5	43.3	48.9
Still-3-1.5B-Preview	84.4	32.5	66.7	29.0	45.4	51.6
DeepScaleR-1.5B-Preview	87.8	43.1	73.6	30.2	50.0	57.0
FastCuRL-1.5B-Preview	**88.0**	**43.1**	**74.2**	**31.6**	**50.4**	**57.5**
FastCuRL-1.5B-V2	**89.3**	**47.5**	**77.0**	**32.8**	**53.3**	**60.0**
FastCuRL-1.5B-V3	**90.5**	**49.6**	**78.5**	**34.7**	**54.5**	**61.6**

Table 4: Pass@1 accuracy is reported, averaged over 16 samples for each problem. † indicates results obtained by re-evaluating using the checkpoints provided by the corresponding work.

Model	Training Steps	Training Stages	Number of GPUs Used in Each Stage
DeepScaleR-1.5B-Preview	$\sim 1,750$	3	8, 16, 32
FastCuRL-1.5B-Preview (Exp-6)	~ 860	4	8, 8, 8, 8
FastCuRL-1.5B-V2 (Exp-10)	$\sim 1,710$	5	8, 8, 8, 8, 8
FastCuRL-1.5B-V3 (Exp-11)	$\sim 2,620$	5	8, 8, 8, 8, 8

Table 5: Training Details. To ensure consistency in counting training steps, we standardized the batch size to 128. This means that two steps with a batch size of 64 are considered equivalent to one step with a batch size of 128.

but there is virtually no difference compared to the fourth stage. Analyzing this phenomenon, we find that during progressive context extension training, the model's output length is initially constrained by the short context in the first stage. This constraint compresses the length of the thoughts but improves their quality. As the context increases in the second and third stages, the model begins to explore problems that require longer thought. However, this extension also introduces repetitive thought patterns. These repetitive patterns do not enhance the model's reasoning capabilities; on the contrary, they may decrease the model's exploratory efficiency, especially when the context length becomes excessively long. Therefore, further compressing the context length (as in the fourth stage) is necessary to improve the quality of the chain-of-thought and enhance the model's exploratory efficiency.

In the third set of experiments, we find that neither increasing nor decreasing the context length is as effective as maintaining the context length at 16K. Does this phenomenon suggest that there is a "sweet spot" for context length in R1-like models, and that for the DeepSeek-R1-Distill-Qwen-1.5B, 16K is the optimal sweet spot? Or is it that 16K is closer to the sweet spot compared to 24K and 8K? Based on this question, we conduct a series of experiments where we train the model with different context lengths and set the entropy coefficient equal to 1×10^{-6} to observe the changes in the entropy. As shown in Figure 7, we find that when the context lengths are 4K, 8K, and 12K, the entropy rapidly decreases to a small value, indicating that the model has lost its exploratory capability. Interestingly, when the context lengths are 16K, 20K, and 24K, the entropy stabilizes at a fixed value and does not decrease rapidly.

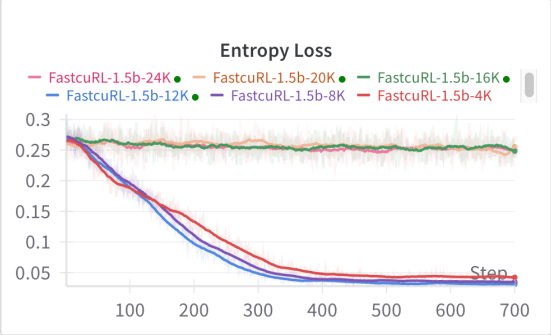

Figure 7: Entropy curves of different context lengths.

Inspired by the above findings, we continue to train **FastCuRL-1.5B-Preview** under a 16K context and adjust the coefficients of KL and Entropy (Table 3). Results in Table 4 show that after being incentivized in the prior stages, the performance of **FastCuRL-1.5B-V3** gradually increases in Stage 5 and achieves an accuracy of 49.6% on AIME 2024, supporting the above raised question.

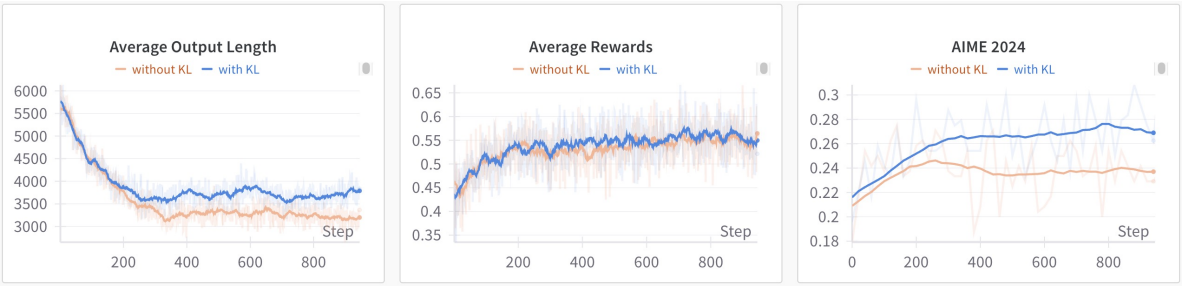

Figure 8: Performance comparison of training with and without KL penalty at 8k context length.

Figure 9: Performance comparison of training with and without Entropy loss at 8k context length.

3.5.3 Overall Comparison Results

Table 4 present the overall PASS@1 performance of QWEN2.5-MATH-7B-Instruct, DEEPSEEK-R1-DISTILL-QWEN-1.5B, STILL-1.5B, QWEN2.5-7B-SimpleRL, RSTAR-MATH-7B, EURUS-2-7B-PRIME, and DEEPSCALER-1.5B-Preview. Specifically, our models achieve the best overall performance on five competition-level benchmarks.

Meanwhile, FASTCURL-1.5B-Preview has better generalization on the AMC 2023 and Minerva Math test sets than the baseline DEEPSCALER-1.5B-Preview. Furthermore, as shown in Table 5, compared with the baseline DEEPSCALER-1.5B-Preview, we only use 50% of the training steps during training and only one node with 8 GPUs, saving more than half of the training resources. Moreover, our model can achieve better results when using the same or more training steps.

3.5.4 The Effectiveness of KL and Entropy

The KL penalty and entropy loss are very important in RL training. Therefore, we conduct simple ablation experiments on the KL penalty and entropy loss. As presented in Figure 8, we find that when training the DEEPSEEK-R1-DISTILL-QWEN-1.5B model without the KL penalty, even when the average output length was compressed to between 3500-4000 tokens, the model's output length does not show a significant increasing trend. From the results in Figure 9, we can see that removing the entropy loss caused the model's output length to decrease significantly around step 800. In Figure 8 and Figure 9, the blue lines represent the original experimental setup, but these are results from two different experiments. This paper primarily focuses on exploring the impact of context length and data complexity on the training process. Therefore, we do not provide an extensive analysis of the effects of the KL penalty and entropy loss.

3.5.5 Analyzing Generated Responses

Table 6 presents comparative statistics on the response characteristics of DEEPSEEK-R1-DISTILL-QWEN-1.5B and FASTCURL-1.5B-Preview. The results focus on two key metrics: average output length and frequency of the term "wait"/"Wait" in responses. The DEEPSEEK-R1-DISTILL-QWEN-1.5B produces significantly longer responses overall (50.5% longer than FASTCURL-1.5B-Preview. Interestingly, both models show a pattern where incorrect responses tend to be substantially longer than correct ones. The frequency of "wait"/"Wait" terms is indicative of reflection behaviors in the R1-like reasoning models. DEEPSEEK-R1-DISTILL-QWEN-1.5B uses these terms approximately 36% more frequently than FASTCURL-1.5B-Preview overall. Similarly, both models show significantly higher usage of these terms in incorrect responses compared to correct ones.

Model	# Average Output Length			# Average Frequency of "Wait" and "wait"		
	TOTAL	CORRECT	INCORRECT	TOTAL	CORRECT	INCORRECT
DEEPSEEK-R1-DISTILL-QWEN-1.5B	43176	21859	52629	109	49	138
FASTCURL-1.5B-Preview	28681	18970	36044	80	48	104

Table 6: Statistics of the responses of DEEPSEEK-R1-DISTILL-QWEN-1.5B and **FASTCURL-1.5B-Preview**.

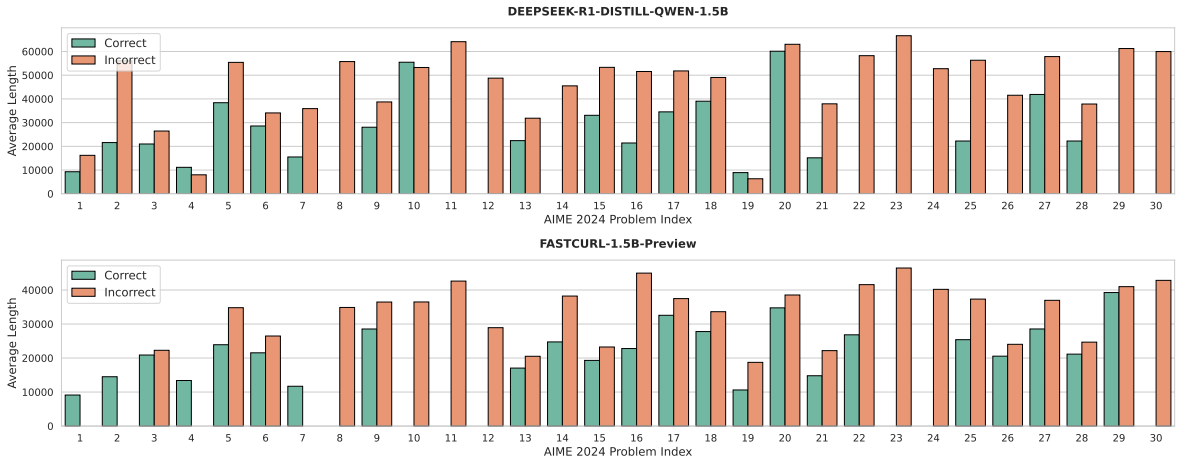

Figure 10: Comparison of average response length (character-level) between correct and incorrect answers. Green bars represent correct answers, while red bars represent incorrect answers. Each problem's analysis is based on 16 samples. A few problems have no green bars, indicating no correct answers are provided for those problems.

Figure 10 compares DEEPSEEK-R1-DISTILL-QWEN-1.5B and FASTCURL-1.5B-Preview on the AIME 2024, measuring the average response length between correct and incorrect answers at the problem level to observe and analyse whether the long incorrect response is related to the difficulty of the problem. Across both models, incorrect answers (red bars) almost universally have greater average response lengths than correct answers (green bars). This suggests that models tend to generate more verbose content when producing incorrect answers, potentially reflecting "over-explanation" or "verbose reasoning" when the model is uncertain.

4 Related Work

Advancements in RL methodologies have considerably enhanced the reasoning capabilities of LLMs. A pivotal development in this domain is OpenAI's o1 (OpenAI, 2024), which employs RL training to promote the development of long CoT reasoning in LLMs. This approach has significantly enhanced performance on complex mathematical and programming benchmarks. Building upon this foundation, DeepSeek-R1 (DeepSeek-AI, 2025) demonstrates that pure RL post-training via Group Reinforcement Policy Optimization (GRPO), without needing supervised pre-training, can directly perform robust CoT reasoning capabilities. Notably, this method not only achieves performance competitive with o1 but also exhibits emergent behaviors such as self-verification and multi-step planning. Building on these advancements, the research community has been collectively working to study and apply DeepSeek-R1's methodology to enhance the reasoning capabilities of various sizes of language models, yielding remarkable progress, such as (Face, 2025; Luo et al., 2025; Zeng et al., 2025b; Liu et al., 2025; Yu et al., 2025).

5 Conclusion

We investigate how the model's context length and the complexity of the training dataset influence the training process of R1-like reasoning models. Motivated by our findings, we propose **FASTCURL**, a simple yet effective curriculum reinforcement learning framework incorporating a stage-wise context scaling strategy. This framework is designed to accelerate the training efficiency and improve the model's long CoT reasoning capabilities. Experimental results demonstrate that FASTCURL-1.5B-Preview achieves better performance and reduces computational resource consumption by more than 50%, with all training phases efficiently executed using a single node with 8 GPUs.

6 Limitations

Due to limited resources, this paper verifies the effectiveness of the proposed method, FastCuRL, only on a 1.5B language model. Generally, validating its effectiveness on models of varying sizes is a worthwhile direction for future research. Furthermore, in this paper, we investigate the influence of using complexity-aware training data by employing the simplest separation method to validate the efficacy of separating the training data by complexity, and achieves significant results. If more sophisticated separation methods were adopted, achieving even more promising results might be possible.

Training over multiple stages, rather than in a single training stage, involves more than changes in parameters like context length; it also fundamentally alters the reference policy. In a multi-stage training strategy, the KL penalty imposed by the reference policy on the model is gradually relaxed, which allows the trained model to explore a broader range of solutions. Delving into dynamic control of context lengths or implementing a dynamic KL penalty may be valuable directions.

References

Mislav Balunović, Jasper Dekoninck, Ivo Petrov, Nikola Jovanović, and Martin Vechev. 2025. Matharena: Evaluating llms on uncontaminated math competitions.

Qiguang Chen, Libo Qin, Jinhao Liu, Dengyun Peng, Jiannan Guan, Peng Wang, Mengkang Hu, Yuhang Zhou, Te Gao, and Wanxiang Che. 2025. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. *Preprint*, arXiv:2503.09567.

Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, Qixin Xu, Weize Chen, Jiarui Yuan, Huayu Chen, Kaiyan Zhang, Xingtai Lv, Shuo Wang, Yuan Yao, Xu Han, Hao Peng, Yu Cheng, and 4 others. 2025. Process reinforcement through implicit rewards. *Preprint*, arXiv:2502.01456.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Hugging Face. 2025. Open r1: A fully open reproduction of deepseek-r1.

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. 2024. Omni-math: A universal olympiad level mathematic benchmark for large language models. *CoRR*, abs/2410.07985.

Xinyu Guan, Li Lyna Zhang, Yifei Liu, Ning Shang, Youran Sun, Yi Zhu, Fan Yang, and Mao Yang. 2025. rstar-math: Small llms can master math reasoning with self-evolved deep thinking. *Preprint*, arXiv:2501.04519.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng 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 *ACL (1)*, pages 3828–3850. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the MATH dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, and Heung-Yeung Shum Xiangyu Zhang. 2025. Open-reasoner-zero: An open source approach to scaling reinforcement learning on the base model. `https://github.com/Open-Reasoner-Zero/Open-Reasoner-Zero`.

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. *Preprint*, arXiv:2206.14858.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. 2025. Understanding r1-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*.

Michael Luo, Sijun Tan, Justin Wong, Xiaoxiang Shi, William Y. Tang, Manan Roongta, Colin Cai, Jeffrey Luo, Tianjun Zhang, Li Erran Li, Raluca Ada Popa, and Ion Stoica. 2025. Deepscaler: Surpassing o1-preview with a 1.5b model by scaling rl. `https://github.com/agentica-project/deepscaler`. Notion Blog.

Yingqian Min, Zhipeng Chen, Jinhao Jiang, Jie Chen, Jia Deng, Yiwen Hu, Yiru Tang, Jiapeng Wang, Xiaoxue Cheng, Huatong Song, Wayne Xin Zhao, Zheng Liu, Zhongyuan Wang, and Ji-Rong Wen. 2024. Imitate, explore, and self-improve: A reproduction report on slow-thinking reasoning systems. *Preprint*, arXiv:2412.09413.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large language models: A survey. *Preprint*, arXiv:2402.06196.

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. *arXiv preprint arXiv:2501.19393*.

OpenAI. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

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. *Preprint*, arXiv:2402.03300.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *Preprint*, arXiv:2408.03314.

Kimi Team. 2025a. Kimi k1.5: Scaling reinforcement learning with llms. *Preprint*, arXiv:2501.12599.

RUCAIBox STILL Team. 2025b. Still-3-1.5b-preview: Enhancing slow thinking abilities of small models through reinforcement learning.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Yuyang Wu, Yifei Wang, Tianqi Du, Stefanie Jegelka, and Yisen Wang. 2025. When more is less: Understanding chain-of-thought length in llms. *Preprint*, arXiv:2502.07266.

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. *CoRR*, abs/2409.12122.

Edward Yeo, Yuxuan Tong, Morry Niu, Graham Neubig, and Xiang Yue. 2025. Demystifying long chain-of-thought reasoning in llms. *Preprint*, arXiv:2502.03373.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, and 16 others. 2025. Dapo: An open-source llm reinforcement learning system at scale. *Preprint*, arXiv:2503.14476.

Weihao Zeng, Yuzhen Huang, Qian Liu, Wei Liu, Keqing He, Zejun Ma, and Junxian He. 2025a. Simplerl-zoo: Investigating and taming zero reinforcement learning for open base models in the wild. *Preprint*, arXiv:2503.18892.

Weihao Zeng, Yuzhen Huang, Qian Liu, Wei Liu, Keqing He, Zejun Ma, and Junxian He. 2025b. Simplerl-zoo: Investigating and taming zero reinforcement learning for open base models in the wild. *arXiv preprint arXiv:2503.18892*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, and 3 others. 2023. A survey of large language models. *Preprint*, arXiv:2303.18223.