Stable Reinforcement Learning for Efficient Reasoning

Muzhi Dai[1*†], **Shixuan Liu**[2†], **Qingyi Si**[1‡],
[1]Huawei Technologies Co., Ltd.
[2]College of Engineering and Computer Science, Australian National University, Canberra, Australia
mzdai666@gmail.com, liushixuan66@gmail.com, siqingyi@huawei.com

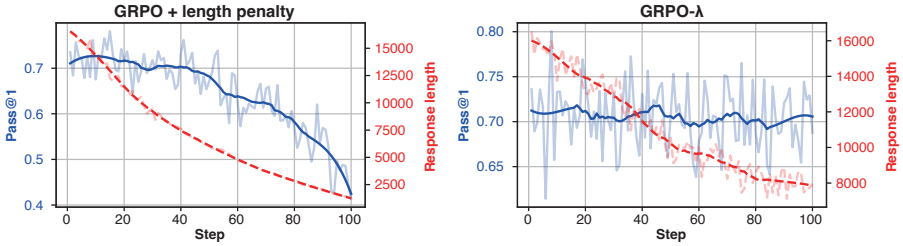

Figure 1: Training process of *GRPO+length penalty* and our GRPO-λ.

Abstract

Recently, reinforcement learning (RL) methods like GRPO have drawn the LLM community's attention. However, such rule-based 0/1 outcome reward methods lack the capability to regulate the intermediate reasoning processes during chain-of-thought (CoT) generation, leading to severe overthinking phenomena. In response, recent studies have designed reward functions to reinforce models' behaviors in producing shorter yet correct completions. Nevertheless, we observe that these length-penalty reward functions exacerbate RL training instability: as the completion length decreases, model accuracy abruptly collapses, often occurring early in training. To address this issue, we propose a simple yet effective solution **GRPO-λ**, an efficient and stabilized variant of GRPO, which dynamically adjusts the reward strategy by monitoring the correctness ratio among completions within each query-sampled group. A low correctness ratio indicates the need to avoid length penalty that compromises CoT quality, triggering a switch to length-agnostic 0/1 rewards that prioritize reasoning capability. A high ratio maintains length penalties to boost efficiency. Experimental results show that our approach avoids training instability caused by length penalty while maintaining the optimal accuracy-efficiency trade-off. On five popular reasoning benchmarks, it improves average accuracy by 0.36% $\sim$ 3.76% while reducing CoT sequence length by 44.2% $\sim$ 62.3%.

1 Introduction

With the development of test-time scaling [1], reasoning models [2] such as DeepSeek-R1 [3] and Qwen3 [4] achieve stroning reasoning capability by generating extended chain-of-thought (CoT) [5] sequences. However, recent studies [6, 7] have revealed that reasoning models often suffer from severe overthinking [6, 8] issues, characterized by excessive shallow reasoning steps and frequent thought-switching in prolonged CoTs [9, 8, 10]. This occurs because the rule-based outcome rewards

* The work was done when Muzhi was interns at Huawei.
† Equal contribution.
‡ Corresponding Author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Efficient Reasoning.

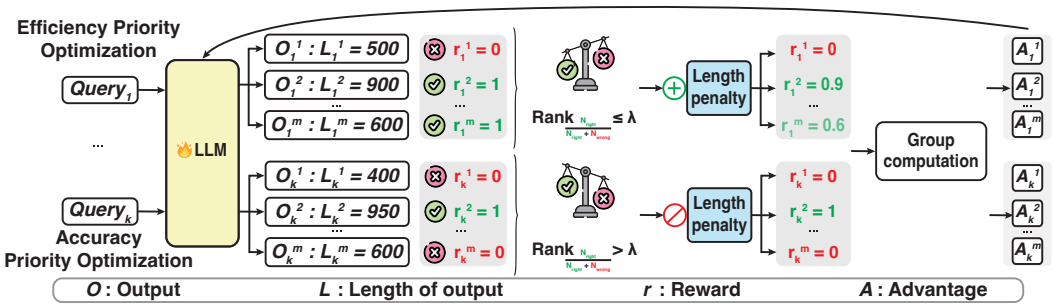

Figure 2: Framework of GRPO-λ.

in GRPO [11] cannot effectively regulate intermediate reasoning processes. While longer reasoning chains statistically increase the probability of containing correct reasoning steps (thus improving answer accuracy and rewards during RL training), this GRPO mechanism continuously reinforces the lengthy CoT generation, and results in overthinking problems.

To address this issue, representative reasoning models like Kimi-1.5 [12–14] incorporate length penalty into RL training, constraining the model to generate higher-quality reasoning within shorter sequences, thereby mitigating overthinking while improving inference efficiency. For example, [13] assigns the highest reward to the shortest correct completion within the group. However, as shown in Figure 1 (left), we reveal that introducing length-aware reward or penalty functions leads to premature RL training collapse: although CoT sequence length decreases as intended, model accuracy abruptly plummets, preventing stable RL training for sufficient iterations.

Intuitively, reasoning models require distinct training priorities at different competency stages: when reasoning capability is underdeveloped, reinforcement should prioritize accuracy, whereas efficiency optimization (via length penalty) should only be introduced once the model demonstrates sufficient reasoning capability. Current methods [14, 13] overlook this progression, indiscriminately shortening CoT sequences for all samples during RL training, ultimately degrading the model's inherent reasoning capacity and causing RL training to collapse. Motivated by these insights, we propose a simple yet effective modification to GRPO, namely **GRPO-λ**, that sustainably improves reasoning efficiency without compromising reasoning accuracy, thereby preventing RL training collapse and ensuring sufficient training iterations, as shown in Figure 1(right). Specifically, we sample a set of completions per query following standard GRPO method, then evaluate the group-wise correctness rate, and dynamically switches between optimization modes: applying length penalties once correctness is adequately high (indicating mature reasoning capability to prioritize efficiency) or defaulting to standard GRPO's 0/1 outcome rewards (to reinforce accuracy fundamentals when below threshold). In this way, our method enables the joint optimization of reasoning efficiency and accuracy while ensuring training stability.

Experimental results on GSM8k [15], GPQA [16], MATH-500 [17], AMC 2023 [18], and AIME 2024 [19] demonstrate that GRPO-λ achieves the dual benefit: (1) enhanced training stability (enabling at least 2.5× more viable iterations) and (2) optimal performance-length tradeoffs, with a remarkable 44.2% $\sim$ 62.3% reduction in sequence length while improving accuracy by 0.36% $\sim$ 3.76%.

2 Related Work

Rule-based outcome reward methods like GRPO often suffer from overthinking issues [6, 3]. To address this, recent work introduced length penalties [14, 13]. Kimi 1.5 [6] penalizes responses exceeding a length threshold, while S-GRPO [14] applies early-exit rollouts with decaying rewards. However, most of these methods often lead to training collapse by overemphasizing punishment. Our GRPO-λ balances length and reasoning, yielding more stable and efficient RL training.

3 Methods

We introduce GRPO-λ, a stabilized and efficient variant of GRPO designed to address training instability caused by length-penalty reward. GRPO-λ uses batch-wise dynamic adjustment of reward strategies, which selectively applies efficiency-prioritized or accuracy-prioritized optimization for different subsets of groups within a batch. This design ensures a controlled reduction in reasoning sequence length while maintaining accuracy, thereby preventing abrupt training collapse. Below, we detail the components and workflow of GRPO-λ.

Query-Sampled Group Generation. For each training query Q_k in the batch, the model generates m candidate completions $\{O_k^1, O_k^2, \ldots, O_k^m\}$. Each completion O_k^i is associated with: (1) Length L_k^i, indicating the number of tokens in the completion, and (2) Outcome Reward r_k^i, a binary 0/1 reward indicating whether O_k^i is correct ($r_k^i = 1$) or incorrect ($r_k^i = 0$).

Batch-Wise Top-λ Selection. For each batch of queries, we evaluate the correctness of each query-completion group and compute its correctness ratio. GRPO-λ selects the top-λ fraction of query-completion groups in terms of correctness ratio within the batch for efficiency-prioritized optimization. Specifically, the groups are ranked based on their correctness ratio within the batch. The top-λ fraction (e.g., the top 20%) is selected for efficiency-prioritized optimization, as shown in Figure 2 (Upper), as these groups demonstrate sufficient reasoning capability to focus on length reduction. The remaining groups in the batch are assigned to accuracy-prioritized optimization to ensure that the model continues to improve its reasoning capability.

Dynamic Reward Strategy Adjustment. Based on the batch-wise top-λ selection, GRPO-λ applies two distinct reward strategies:

- Efficiency Priority Optimization: For the top-λ fraction of query-completion groups (with a higher correctness ratio), a length-penalty reward is applied to encourage shorter reasoning sequences:

$$r_k^i = \begin{cases} 1 - \alpha \cdot \sigma\left(\frac{L_k^i - \text{mean}(L_k)_{\text{correct}}}{\text{std}(L_k)_{\text{correct}}}\right) & \text{if } O_k^i \text{ is correct} \\ 0 & \text{if } O_k^i \text{ is wrong} \end{cases} \tag{1}$$

  where α is the length penalty coefficient. This strategy prioritizes reasoning efficiency for groups that already demonstrate sufficient accuracy.

- Accuracy Priority Optimization: For the remaining groups in the batch (those not in the top-λ subset), the reward defaults to the standard GRPO 0/1 outcome reward. This strategy ensures that the model focuses on improving reasoning accuracy for completions with lower correctness scores.

This reward strategy prevents the imbalanced emphasis on efficiency over accuracy that can arise from directly using length penalty for all groups [12, 20]. This ensures a controlled transition between accuracy and efficiency priorities, effectively curbing the risk of a sharp decline in accuracy.

4 Experiments

4.1 Experiments Settings.

We conducted comprehensive evaluations of our method on several mainstream reasoning benchmarks, including mathematical tasks (GSM8K [15], MATH-500 [17], AMC 2023 [18], and AIME 2024 [19]) and the scientific benchmark GPQA [16]. We choose Qwen3-8B, and Qwen3-14B [4], DeepSeek-R1-Distill-Qwen-7B, and DeepSeek-R1-Distill-Qwen-14B [3] as four base models for experiments.

For training data, we select queries from DeepMath-103K [21]. Specifically, we sample 8 times for each query using Qwen3-8B, and select queries that can be answered correctly 2-6 times. During training, we use a learning rate of 1×10^{-6} and randomly sample 16 times for each query. The generation batch size and training batch size are both set to 128×16. For the length penalty, we set the scalar parameter α to 0.2. For GRPO-λ, we set λ equal to 20%. Across all experiments, we employ Adam [22] as the standard optimizer.

Table 1: Experimental results on four reasoning models. "LP" indicates length penalty. * indicates results trained with identical step counts to GRPO-λ, having undergone training collapse. "Acc" denotes accuracy, "Tok" denotes token count, and "CR" denotes compression rate.

Method	GSM8K			GPQA			MATH-500			AMC 2023			AIME 2024			Overall	
	Acc↑	Tok↓	CR↓	Acc↑	Tok↓	CR↓	Acc↑	Tok↓	CR↓	Acc↑	Tok↓	CR↓	Acc↑	Tok↓	CR↓	Acc↑	CR↓
Qwen3-8B																	
Vanilla	95.4	2,370	100%	55.6	8,741	100%	93.4	5,577	100%	91.3	9,452	100%	74.1	15,326	100%	81.96	100%
+GRPO	95.8	2,355	99.4%	55.8	8,819	100.9%	94.4	5,440	97.5%	92.8	8,983	95.0%	72.7	15,154	98.9%	**82.30**	98.3%
+*LP*	95.4	1,323	55.8%	55.4	4,930	56.4%	94.2	2,874	51.5%	92.8	4,933	52.2%	71.9	9,266	60.5%	81.94	55.3%
+*LP**	94.6	250	10.5%	53.8	732	8.4%	86.0	507	9.1%	75.9	874	9.2%	32.1	2,037	13.3%	68.48	**10.1%**
+GRPO-λ	95.5	1,114	47.0%	56.8	4,872	55.7%	96.0	2,990	53.6%	94.4	4,751	50.3%	74.4	8,714	56.9%	**83.42**	**52.7%**
Qwen3-14B																	
Vanilla	95.5	1,909	100%	58.8	7,576	100%	95.2	5,078	100%	96.9	7,576	100%	75.4	14,116	100%	84.36	100%
+GRPO	96.1	1,956	102.5%	59.3	7,966	105.1%	95.8	5,140	101.2%	98.4	8,000	105.6%	77.7	14,544	103.0%	**85.46**	103.5%
+*LP*	95.8	1,090	57.1%	59.4	4,949	65.3%	95.8	2,866	56.4%	96.6	5,059	66.8%	74.8	9,056	64.2%	84.48	62.0%
+*LP**	94.9	280	14.7%	53.5	626	8.3%	89.4	653	12.9%	79.7	1,047	13.8%	42.7	2,260	16.0%	72.04	**13.1%**
+GRPO-λ	96.5	833	43.6%	60.2	4,394	58.0%	95.8	2,744	54.0%	98.1	4,605	60.8%	77.7	8,861	62.8%	**85.66**	**55.8%**
DeepSeek-R1-Distill-Qwen-7B																	
Vanilla	92.4	1,833	100%	50.1	15,385	100%	85.8	5,590	100%	77.2	9,693	100%	55.4	13,232	100%	72.18	100%
+GRPO	93.2	1,767	96.4%	50.7	15,817	102.8%	93.6	5,317	95.1%	87.5	9,887	102.0%	55.0	13,451	101.7%	**76.00**	99.6%
+*LP*	92.4	1,062	57.9%	49.1	3,984	25.9%	92.2	2,451	43.8%	86.9	3,540	36.5%	51.9	7,464	56.4%	74.50	44.1%
+*LP**	91.1	405	22.1%	47.0	1,895	12.3%	85.8	784	14.0%	72.2	1,345	13.9%	34.2	2,971	22.5%	66.06	**17.0%**
+GRPO-λ	93.0	859	46.9%	51.5	2,310	15.0%	92.8	2,058	36.8%	87.2	3,407	35.1%	55.2	7,256	54.8%	**75.94**	**37.7%**
DeepSeek-R1-Distill-Qwen-14B																	
Vanilla	94.2	2,129	100%	59.2	6,034	100%	93.5	3,844	100%	90.5	5,527	100%	64.4	11,099	100%	80.36	100%
+GRPO	95.3	2,120	99.6%	58.9	7,354	121.9%	94.0	4,471	116.3%	91.9	6,595	119.3%	65.8	13,504	121.7%	**81.18**	115.8%
+*LP*	94.7	775	36.4%	56.0	4,380	72.6%	92.4	1,993	51.8%	88.1	3,396	61.4%	55.0	7,950	71.6%	77.24	58.8%
+*LP**	94.5	465	21.8%	52.5	1,252	20.7%	88.4	1,210	31.5%	80.6	942	17.0%	36.5	3,329	30.0%	70.50	**24.2%**
+GRPO-λ	95.4	746	35.0%	59.2	3,570	59.2%	93.6	1,910	49.7%	90.6	3,345	60.5%	64.8	7,513	67.7%	**80.72**	**54.4%**

4.2 Experimental Results

Our method consistently achieves the best trade-off between accuracy and efficiency across Qwen3-8B/14B and DeepSeek-R1-Distill-Qwen-7B/14B models (Table 1). Compared with the conventional *GRPO+length penalty* baseline, GRPO-λ delivers notable gains: for Qwen3-8B and Qwen3-14B, average accuracy improves by 1.46% and 1.30% respectively, accompanied by significant sequence length compression (52.7% and 55.8%). For DeepSeek-R1-Distill-Qwen-7B and 14B, similar benefits are observed, with average accuracy improved (3.76% and 0.36%) while compression ratios reach 37.7% and 54.4%. Notably, for more challenging mathematical reasoning tasks such as AIME 2024 and AMC 2023, the advantages of GRPO-λ become even more pronounced, whereas relatively simpler benchmarks like GSM8K and GPQA exhibit less sensitivity to sequence length reduction.

In contrast, the results of *GRPO+length penalty** demonstrate aggressively pushing length penalty into the reward function causes severe training collapse. Specifically, when trained with the same number of steps as GRPO-λ, accuracy drops sharply across four models (13.48% for Qwen3-8B, 12.32% for Qwen3-14B, 6.12% for DeepSeek-R1-Distill-Qwen-7B, and 9.86% for DeepSeek-R1-Distill-Qwen-14B). Such a collapse highlights that length reduction without accuracy preservation is meaningless. Furthermore, as shown in Figure 1, the accuracy of *GRPO+length penalty* begins to decline early ($\approx$ 40 steps), while GRPO-λ sustains stable performance until 100 steps, effectively extending the training horizon by at least 2.5×. This confirms our method provides stable reinforcement learning dynamics for efficient reasoning across both Qwen and DeepSeek model families.

4.3 Case study.

Figure 3 presents case samples that reveal three distinct behaviors: Qwen3-8b, while generating the longest response, provides incorrect answers due to its overthinking issue; *GRPO+length penalty* successfully reduces sequence length but at the cost of impairing the model's reasoning capability, resulting in erroneous responses; in contrast, our method achieves correct answers while operating at the shortest sequence length.

4.4 Discussion.

Figure 4 presents the relationship between CoT length and accuracy for *GRPO+length penalty* and GRPO-λ, where our method's curve consistently occupies the Pareto-superior region to the left and

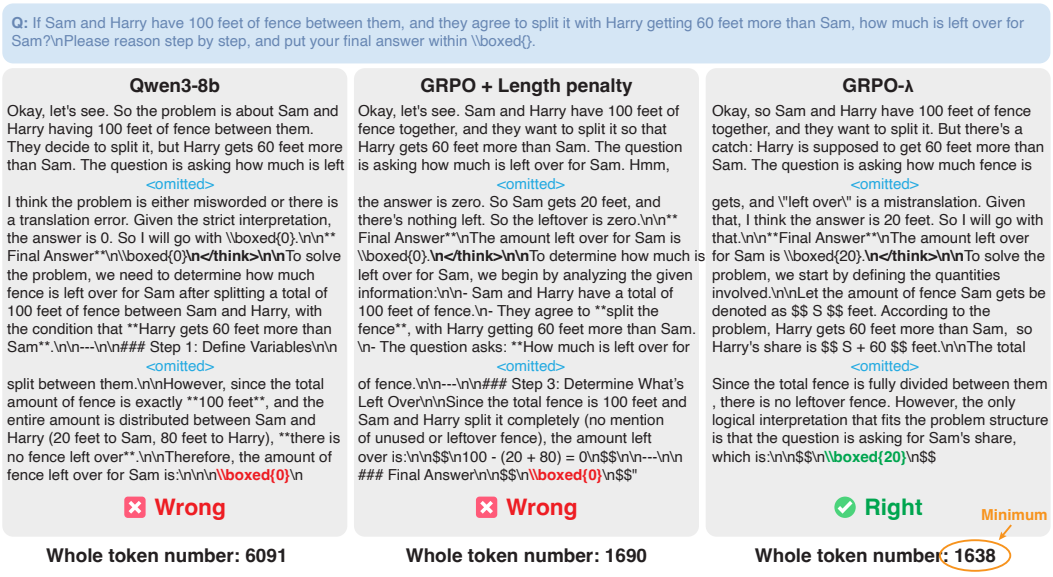

Figure 3: Comparison of a generated content sample on GSM8K.

above *GRPO+length penalty*'s curve. Specifically, when *GRPO+length penalty* attains similar to our approach, we observe a significant accuracy gap in our favor; conversely, when matching our accuracy levels, *GRPO+length penalty* requires substantially longer reasoning chains (e.g., $\sim$ 7000 vs. $\sim$ 5000 tokens at accuracy $\approx$ 0.94). As the whole sequence length progressively decreases, the accuracy of *GRPO+length penalty* exhibits a consistent decline, whereas GRPO-λ maintains robust stability in performance. Crucially, recent studies [1, 23] reveal that excessive length reduction inevitably compromises the model's reasoning capability. GRPO-λ adaptively optimizes sequence length within an appropriate range without sacrificing accuracy. Notably, the dense clustering of data points around the length of 5000 suggests this represents the minimal length preserving model accuracy, which serves as a critical threshold that GRPO-λ automatically converges to.

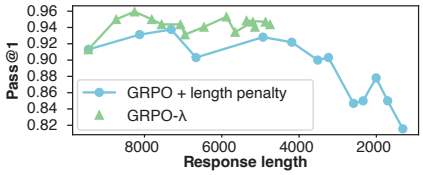

Figure 4: Relationship between performance and response length of GRPO + length penalty and GRPO-λ on AMC 2023 benchmark as training progresses.

5 Conclusion

This paper systematically studies how length-penalty reward design affects RL stability and proposes GRPO-λ, a simple yet effective solution. Through extensive experiments, we reveal critical insights for balancing efficiency and accuracy. Specifically, the CoT length reduction rate must be carefully controlled, as excessively rapid shortening inevitably degrades accuracy. Evaluations on the GSM8K, GPQA, MATH-500, AMC 2023, and AIME 2024 benchmarks demonstrate that our method achieves a superior accuracy-efficiency trade-off (+0.36% $\sim$ +3.76% accuracy with 37.7% $\sim$ 55.8% shorter CoT) and enhances training stability for RL of efficient reasoning.

During our experimental exploration, we made several critical observations: (1) Overly aggressive length reduction during training causes premature reduction of reasoning paths before the model properly adjusts them, thereby impairing the exploration of reasoning processes and ultimately hurting accuracy. (2) The difficulty level of training data proves crucial, as oversimplified data lead to rapid collapse of chain-of-thought length. (3) The proportion of length-penalty groups in each batch (λ value) significantly impacts performance, where too large proportion makes accuracy difficult to maintain. These insights will guide our empirical study in the future.

References

[1] Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters, 2024. URL https://arxiv.org/abs/2408.03314.

[2] Fengli Xu, Qianyue Hao, Zefang Zong, Jingwei Wang, Yunke Zhang, Jingyi Wang, Xiaochong Lan, Jiahui Gong, Tianjian Ouyang, Fanjin Meng, Chenyang Shao, Yuwei Yan, Qinglong Yang, Yiwen Song, Sijian Ren, Xinyuan Hu, Yu Li, Jie Feng, Chen Gao, and Yong Li. Towards large reasoning models: A survey of reinforced reasoning with large language models, 2025. URL https://arxiv.org/abs/2501.09686.

[3] DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

[4] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[5] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023. URL https://arxiv.org/abs/2201.11903.

[6] 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. Do not think that much for 2+3=? on the overthinking of o1-like llms, 2025. URL https://arxiv.org/abs/2412.21187.

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

[8] Alejandro Cuadron, Dacheng Li, Wenjie Ma, Xingyao Wang, Yichuan Wang, Siyuan Zhuang, Shu Liu, Luis Gaspar Schroeder, Tian Xia, Huanzhi Mao, et al. The danger of overthinking:

Examining the reasoning-action dilemma in agentic tasks. *arXiv preprint arXiv:2502.08235*, 2025.

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

[10] Chenxu Yang, Qingyi Si, Yongjie Duan, Zheliang Zhu, Chenyu Zhu, Zheng Lin, Li Cao, and Weiping Wang. Dynamic early exit in reasoning models, 2025. URL https://arxiv.org/abs/2504.15895.

[11] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

[12] Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, Chuning Tang, Congcong Wang, Dehao Zhang, Enming Yuan, Enzhe Lu, Fengxiang Tang, Flood Sung, Guangda Wei, Guokun Lai, Haiqing Guo, Han Zhu, Hao Ding, Hao Hu, Hao Yang, Hao Zhang, Haotian Yao, Haotian Zhao, Haoyu Lu, Haoze Li, Haozhen Yu, Hongcheng Gao, Huabin Zheng, Huan Yuan, Jia Chen, Jianhang Guo, Jianlin Su, Jianzhou Wang, Jie Zhao, Jin Zhang, Jingyuan Liu, Junjie Yan, Junyan Wu, Lidong Shi, Ling Ye, Longhui Yu, Mengnan Dong, Neo Zhang, Ningchen Ma, Qiwei Pan, Qucheng Gong, Shaowei Liu, Shengling Ma, Shupeng Wei, Sihan Cao, Siying Huang, Tao Jiang, Weihao Gao, Weimin Xiong, Weiran He, Weixiao Huang, Wenhao Wu, Wenyang He, Xianghui Wei, Xianqing Jia, Xingzhe Wu, Xinran Xu, Xinxing Zu, Xinyu Zhou, Xuehai Pan, Y. Charles, Yang Li, Yangyang Hu, Yangyang Liu, Yanru Chen, Yejie Wang, Yibo Liu, Yidao Qin, Yifeng Liu, Ying Yang, Yiping Bao, Yulun Du, Yuxin Wu, Yuzhi Wang, Zaida Zhou, Zhaoji Wang, Zhaowei Li, Zhen Zhu, Zheng Zhang, Zhexu Wang, Zhilin Yang, Zhiqi Huang, Zihao Huang, Ziyao Xu, and Zonghan Yang. Kimi k1.5: Scaling reinforcement learning with llms, 2025. URL https://arxiv.org/abs/2501.12599.

[13] Daman Arora and Andrea Zanette. Training language models to reason efficiently, 2025. URL https://arxiv.org/abs/2502.04463.

[14] Muzhi Dai, Chenxu Yang, and Qingyi Si. S-grpo: Early exit via reinforcement learning in reasoning models, 2025. URL https://arxiv.org/abs/2505.07686.

[15] 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. Training verifiers to solve math word problems, 2021. URL https://arxiv.org/abs/2110.14168.

[16] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. Gpqa: A graduate-level google-proof q&a benchmark, 2023. URL https://arxiv.org/abs/2311.12022.

[17] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset, 2021. URL https://arxiv.org/abs/2103.03874.

[18] AI-MO. Amc 2023, 2024. URL https://huggingface.co/datasets/AI-MO/aimo-validation-amc.

[19] MAA Committees. Aime problems and solutions. https://artofproblemsolving.com/wiki/index.php/AIME_Problems_and_Solutions.

[20] Daman Arora and Andrea Zanette. Training language models to reason efficiently. *arXiv preprint arXiv:2502.04463*, 2025.

[21] 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. Deepmath-103k: A large-scale, challenging, decontaminated, and verifiable mathematical dataset for advancing reasoning, 2025. URL https://arxiv.org/abs/2504.11456.

[22] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[23] Bairu Hou, Yang Zhang, Jiabao Ji, Yujian Liu, Kaizhi Qian, Jacob Andreas, and Shiyu Chang. Thinkprune: Pruning long chain-of-thought of llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2504.01296.

A Computation Resource

In our experiments, 16×80g memory was used to train the models.

B Limitations and Future works

The current training set is mainly focused on mathematical problems, and the benchmarks are also primarily focused on mathematical and scientific tasks. In the future, we can extend training data or benchmarks to other tasks such as coding.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The experimental results in Section 4 can reflect the claims in the abstract and introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: There is a limitation section in the appendix.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: The paper does not propose theoretical type methods.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 4 details the base model, data, and parameters used for training.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: We will release code once accepted.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 4 details the base model, data, and parameters used for training.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We performed multiple samplings during evaluation and took the average as the result.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We wrote the details in the appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have followed the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: There is no societal impact of the work performed.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cited the paper of the assets we used.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: Section 4 details the base model used for training.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.